Pontifícia Universidade Católica
do Rio de Janeiro

Leila Figueiredo Dantas

Predicting the Acquisition of Resistant Pathogens in ICUs using Machine Learning Techniques

Tese de Doutorado

Thesis presented to the Programa de Pós-graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção.

Advisor: Prof. Silvio Hamacher
Co-advisor: Prof. Fernando Augusto Bozza

Rio de Janeiro
December 2020

Leila Figueiredo Dantas

Predicting the Acquisition of Resistant Pathogens in ICUs using Machine Learning Techniques

Thesis presented to the Programa de Pós-graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção. Approved by the Examination Committee:

Prof. Silvio Hamacher
Advisor
Departamento de Engenharia Industrial – PUC-Rio

Prof. Fernando Augusto Bozza
Co-advisor
Fundação Oswaldo Cruz

Prof. Fernanda Araújo Baião Amorim
Departamento de Engenharia Industrial – PUC-Rio

Prof. Fernando Luiz Cyrino Oliveira
Departamento de Engenharia Industrial – PUC-Rio

Prof. Thiago Costa Lisboa
Hospital de Clínicas de Porto Alegre

Prof. Benjamin Dalmas
École des Mines de Saint-Etienne

Rio de Janeiro, December 21st, 2020

Leila Figueiredo Dantas

Graduated in Production Engineering at the Federal University of Sergipe in 2014 and obtained her M.Sc. Degree in Production Engineering from the Pontifical Catholic University of Rio de Janeiro in 2016.

Bibliographic data

Dantas, Leila Figueiredo

Predicting the acquisition of resistant pathogens in ICUs using machine learning techniques / Leila Figueiredo Dantas ; advisor: Silvio Hamacher ; co-advisor: Fernando Augusto Bozza. – 2020.
253 f. : il. color. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2020.
Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Aprendizado de máquina. 3. Estratégias de balanceamento. 4. Modelo preditivo. 5. Resistência aos carbapenêmicos. 6. Bactérias gram-negativas. I. Hamacher, Silvio. II. Bozza, Fernando Augusto. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

To my parents and brothers, for their support
and encouragement.

Acknowledgments

To my advisors, Professor Silvio Hamacher and Fernando Bozza, for the encouragement, help, and shared knowledge to carry out this work.

To CAPES, CNPq, and PUC-Rio, for the aid granted, without which this work could not have been performed.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and PDSE 88881.190042/2018-01.

To my friends at Tecgraf/PUC-Rio for all their friendship, studies, and comprehension.

To my parents and brothers, for their dedication, support, and confidence always.

To the teachers who participated in the examining committee for their availability and suggestions for improvement.

To all professors and Department staff for their teachings and help.

Abstract

Dantas, Leila Figueiredo; Hamacher, Silvio (Advisor); Bozza, Fernando (Co-Advisor). **Predicting the acquisition of resistant pathogens in ICUs using machine learning techniques**. Rio de Janeiro, 2020. 253p. Tese de Doutorado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Infections by Carbapenem-Resistant Gram-negative bacteria (CR-GNB) are among the most significant contemporary health concerns, especially in intensive care units (ICUs), and may be associated with increased hospitalization time, morbidity, costs, and mortality. This thesis aims to develop a comprehensive and systematic approach applying machine-learning techniques to build models to predict the CR-GNB acquisition in ICUs from Brazilian hospitals. We proposed screening models to detect ICU patients who do not need to be tested and a risk model that estimates ICU patients' probability of acquiring CR-GNB. We applied feature selection methods, machine-learning techniques, and balancing strategies to build and compare the models. The performance criteria chosen to evaluate the models were Negative Predictive Value (NPV) and Matthews Correlation Coefficient (MCC) for the screening model and Brier score and calibration curves for the CR-GNB acquisition risk model. Friedman's statistic and Nemenyi post hoc tests are used to test the significance of differences among techniques. Information gain method and association rules mining assess the importance and strength among features. Our database gathers the patients, antibiotic, and microbiology data from five Brazilian hospitals from May 8th, 2017 to August 31st, 2019, involving hospitalized patients in 24 adult ICUs. Information from the laboratory was used to identify all patients with a positive or negative test for carbapenem-resistant GNB, *A. baumannii, P. aeruginosa*, or Enterobacteriaceae. We have a total of 539 positive and 7,462 negative tests, resulting in 3,604 patients with at least one exam after 48 hours hospitalized. We proposed to the hospital's decision-maker two screening models. The random forest's model would reduce approximately 39% of the

unnecessary tests and correctly predict 92% of positives. The Neural Network model avoids unnecessary tests in 64% of the cases, but 24% of positive tests are misclassified as negatives. Our results show that the sampling, SMOTEBagging, and UnderBagging approaches obtain better results. The linear techniques such as Logistic Regression with regularization give a relatively good performance and are more interpretable; they are not significantly different from the more complex classifiers. For the acquisition risk model, the Nearest Shrunken Centroids is the best model with a Brier score of 0.152 and a calibration belt acceptable. We developed an external validation of 624 patients from two other hospitals in the same network, finding good Brier score (0.128 and 0.079) values in both. The antibiotic and invasive procedures used, especially mechanical ventilation, are the most important attributes for the colonization or infection of CR-GNB. The predictive models can help avoid screening tests and inappropriate treatment in patients at low risk. Infection control policies can be established to control these bacteria's spread. Identifying patients who do not need to be tested decreases hospital costs and laboratory waiting times. We concluded that our models present good performance and seem sufficiently reliable to predict a patient with these pathogens. These predictive models can be included in the hospital system. The proposed methodology can be replicated in different healthcare settings.

Keywords

Machine learning; Balancing strategies; Predictive model; Carbapenem-Resistant; Gram-negative bacteria.

Resumo

Dantas, Leila Figueiredo; Hamacher, Silvio; Bozza, Fernando. **Prevendo a aquisição de patógenos resistentes em UTIs utilizando técnicas de aprendizado de máquina.** Rio de Janeiro, 2020. 253p. Tese de Doutorado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

As infecções por bactérias Gram-negativas Resistentes aos Carbapenêmicos (CR-GNB) estão entre as maiores preocupações atuais da área da, especialmente em Unidades de Terapia Intensiva (UTI), e podem estar associadas ao aumento do tempo de hospitalização, morbidade, custos e mortalidade. Esta tese tem como objetivo desenvolver uma abordagem abrangente e sistemática aplicando técnicas de aprendizado de máquina para construir modelos para prever a aquisição de CR-GNB em UTIs de hospitais brasileiros. Propusemos modelos de triagem para detectar pacientes que não precisam ser testados e um modelo de risco que estima a probabilidade de pacientes de UTI adquirirem CR-GNB. Aplicamos métodos de seleção de características, técnicas de aprendizado de máquina e estratégias de balanceamento para construir e comparar os modelos. Os critérios de desempenho escolhidos para avaliação foram *Negative Predictive Value* (NPV) and *Matthews Correlation Coefficient* (MCC) para o modelo de triagem e *Brier score* e curvas de calibração para o modelo de risco de aquisição de CR-GNB. A estatística de Friedman e os testes post hoc de Nemenyi foram usados para testar a significância das diferenças entre as técnicas. O método de ganho de informações e a mineração de regras de associação avaliam a importância e a força entre os recursos. Nosso banco de dados reúne dados de pacientes, antibióticos e microbiologia de cinco hospitais brasileiros de 8 de maio de 2017 a 31 de agosto de 2019, envolvendo pacientes hospitalizados em 24 UTIs adultas. As informações do laboratório foram usadas para identificar todos os pacientes com teste positivo ou negativo para CR-GNB, *A. baumannii, P. aeruginosa* ou Enterobacteriaceae. Há um total de 539 testes positivos e 7.462 negativos, resultando em 3.604 pacientes com pelo menos um exame após 48 horas de hospitalização. Dois modelos de triagem foram

propostos ao tomador de decisão do hospital. O modelo da floresta aleatória reduz aproximadamente 39% dos testes desnecessários e prevê corretamente 92% dos positivos. A rede neural evita testes desnecessários em 64% dos casos, mas 24% dos testes positivos são classificados incorretamente. Os resultados mostram que as estratégias de amostragem tradicional, SMOTEBagging e UnderBagging obtiveram melhores resultados. As técnicas lineares como Regressão Logística com regularização apresentam bom desempenho e são mais interpretáveis; elas não são significativamente diferentes dos classificadores mais complexos. Para o modelo de risco de aquisição, o Centroides Encolhidos Mais Próximos é o melhor modelo com um Brier score de 0,152 e um cinto de calibração aceitável. Desenvolvemos uma validação externa a partir de 624 pacientes de dois outros hospitais da mesma rede, encontrando bons valores de Brier score (0,128 and 0,079) em ambos. O uso de antibióticos e procedimentos invasivos, principalmente ventilação mecânica, são os atributos mais importantes e significativos para a colonização ou infecção de CR-GNB. Os modelos preditivos podem ajudar a evitar testes de rastreamento e tratamento inadequado em pacientes de baixo risco. Políticas de controle de infecção podem ser estabelecidas para controlar a propagação dessas bactérias. A identificação de pacientes que não precisam ser testados diminui os custos hospitalares e o tempo de espera do laboratório. Concluímos que nossos modelos apresentam bom desempenho e parecem suficientemente confiáveis para prever um paciente com esses patógenos. Esses modelos preditivos podem ser incluídos no sistema hospitalar. A metodologia proposta pode ser replicada em diferentes ambientes de saúde.

Palavras-chave

Aprendizado de máquina; Estratégias de balanceamento; Modelo preditivo; Resistência aos Carbapenêmicos; Bactérias Gram-negativas.

Summary

1. Introduction 17
1.1. Contributions 20
1.2. How to follow this document? 21

2. Related Work 22
2.1. Learning methods in the healthcare context 22
2.2. Multi-resistant Bacteria Acquisition 27
2.2.1. Hospital-Acquired Infections (HAI) 27
2.2.1.1. Hospital surveillance 29
2.2.2. Literature review on multidrug-resistant 30
2.3. Data Mining and Machine Learning in Health Care 44
2.4. Overview of classification techniques 47
2.4.1. Linear Classification Models 48
2.4.1.1. Logistic Regression 48
2.4.1.2. Linear Discriminant Analysis 50
2.4.1.3. Nearest Shrunken Centroids 51
2.4.2. Nonlinear Classification Models 51
2.4.2.1. Support Vector Machine 51
2.4.2.2. k-Nearest Neighbors 52
2.4.2.3. Naive Bayes 53
2.4.2.4. Neural Network 53
2.4.3. Classification Trees 54
2.4.3.1. Decision Tree 54
2.4.3.2. Ensemble learning 57
2.4.3.2.1. Bagging and Boosting 57
2.4.3.2.2. Random Forests 58
2.5. Methods for Imbalanced Learning 59
2.5.1. Sampling 62
2.5.2. Data Cleaning Techniques 64
2.5.2.1. Cleaning under-sampling techniques 64
2.5.2.2. Cleaning and sampling techniques both classes 65
2.5.3. Algorithm level 65
2.5.4. Ensemble-Based Methods for Class Imbalance Problem 66

2.6. Feature Selection 67

3. Case Study Setting and Methodology 69
3.1. Study Overview 69
3.2. Database settings 71
3.3. A framework to machine learning analysis 72
3.3.1. Data Preparation 73
3.3.1.1. Import database 73
3.3.1.2. Feature Engineering 73
3.3.2. Visualization and Data Cleaning 77
3.3.2.1. Descriptive statistical analysis 77
3.3.2.2. Missing values 77
3.3.2.3. Outlier Detection and Treatment 78
3.3.3. Data Splitting 79
3.3.4. Data Preprocessing 79
3.3.4.1. Dimension Reduction 81
3.3.4.1.1. Zero- and Near Zero-Variance Predictors 81
3.3.4.1.2. Between-Predictor Correlations Analysis 81
3.3.4.2. Imputation 82
3.3.4.3. Feature Selection 83
3.3.4.4. Normalization 85
3.3.4.5. Dummy variables 85
3.3.5. Balancing data 86
3.3.6. Building Models - Training 87
3.3.6.1. Hyperparameter tuning and input selection 87
3.3.6.2. Cross-validation 90
3.3.7. Testing 90
3.3.8. Model Evaluation 91
3.3.8.1. Classification 91
3.3.8.2. Prediction 94
3.3.8.2.1. Calibration 95
3.3.9. Statistical comparison of classifiers 95
3.3.10. Running 97
3.4. Important Factors 98
3.5. Association Rules Mining 98

3.6. Differences between the models 99

4. Screening Model 100
4.1. Setting and study population 100
4.2. Conducting a machine learning analysis 102
4.2.1. Visualization and Data Cleaning 103
4.2.2. Data Splitting 107
4.2.3. Data Preprocessing 108
4.2.4. Building models - Training 112
4.2.5. Model Evaluation and Comparison 119
4.2.5.1. Computational Time 129
4.2.6. Model Analysis 131
4.2.6.1. Interpretability 133

5. Risk Model for the acquisition of CR-GNB 137
5.1. Setting and study population 137
5.2. Database Preparation 139
5.3. Model building and evaluation 145
5.3.1. General model 146
5.3.2. Model by hospital 151
5.4. External Validation 154
5.5. Importance of variables 156
5.6. Association Rules 158

6. Discussion 162
6.1. Main Findings 163
6.2. Comparing Related Works 167
6.3. Limitations 171
6.4. Future Researches 172
6.5. Final Consideration 172
6.6. Publications 173

7. References 176

Appendices 192
Appendix A 192
Appendix B 193
Appendix C 194
Appendix D 196
Appendix E 197
Appendix F 203
Appendix G 206
Appendix H 209
Appendix I 216
Appendix J 219
Appendix K 230
Appendix L 231
Appendix M 233
Appendix N 234
Appendix O 239
Appendix P 241
Appendix Q 243
Appendix R 248
Appendix S 250

List of tables

Table 1 - Healthcare techniques and their applications. 23
Table 2 - Classification techniques by authors. 31
Table 3 - Overview of empirical studies selected by data, balancing strategies, variables, and bacteria type. 33
Table 4 - Overview of empirical studies selected by missing value analysis, feature selection, best results, and risk factors. 38
Table 5 - Structure of each hospital 71
Table 6 - The description of each feature included in this work by category and type. 75
Table 7 - Summary of the classification methods. 80
Table 8 - Chronological overview of the balancing strategies used in this work. 86
Table 9 - Hyperparameter ranges. 89
Table 10 - Confusion matrix for a two-class problem (adapted by (KUHN; JOHNSON, 2013)). 91
Table 11 - Difference between the screening and acquisition risk models. 99
Table 12 - The number of patients and culture tests in each hospital. 102
Table 13 - Descriptive statistical analysis comparing the negative and positive culture tests. 103
Table 14 - Number and percentage of missing values for each category and variable. 107
Table 15 - Deviance rate between the fitted model and the perfect model from each imputed dataset and original dataset using a generalized linear model via the lasso penalty. 109
Table 16 - Comparison of all methods' performance on each classifier by AUC values, Average Ranked (AR), and the number of variables. 109
Table 17 - Average of the metric estimates using 10-fold cross-validation for the best hyperparameters based on AUC values. 114
Table 18 - Mean of the best model to AUC, PPV, NPV, sensitivity, and specificity by cross-validation for each new combination. 118
Table 19 - NPV of all 16 classifiers on 11 different balancing strategies, the Average Ranked (AR) among sampling approaches, and the descriptive analysis for each strategy and method. The highest NPV for each strategy is highlighted. 121
Table 20 - MCC of all 16 classifiers on 11 different balancing and descriptive analysis strategies for each strategy and method. The highest MCC for each strategy is highlighted. 125
Table 21 - The best strategy for each metric and method. 128
Table 22 - The best performance for each metric and methods. 128
Table 23 - Summary of computational time for each strategy in ascending order by the median. 130
Table 24 - Summary of computational time for each method in ascending order by the median. 130

Table 25 - Confusion matrix and its metrics for Naïve Bayes, Random Forest, and Logistic Regression regularized methods. The values predicted as false negatives are highlighted in red and true negatives in green. 131
Table 26 - Confusion matrix and its metrics for the methods Neural Network and SVM Radial. The values predicted as false negatives are highlighted in red and true negatives in green. 132
Table 27 - The output from regularized logistic regression, including coefficients (β) and odds ratio (OR). 134
Table 28 - The number of patients and tests in each hospital. 138
Table 29 - The number of patients considered by the hospital after the matching process. 139
Table 30 - Descriptive statistical analysis comparing the Positive and Negative patients. 140
Table 31 - Brier score and MCC of all 16 methods. The best values of the Brier score are highlighted. 146
Table 32 - P-values and the ranges of the predicted probabilities where the belt significantly deviates from the bisector (under and over) to the 80% and 95% confidence level using the testing set, in decreasing order of p-value. 149
Table 33 - Brier score result for each model using test data. 151
Table 34 - Average Ranked for each model using test data. 152
Table 35 - Comparison of models using t-test. 153
Table 36 – Information and results of external validation. 154
Table 37 - List of the 20 rules with higher lift generated from the association rule mining. 158

List of figures

Figure 1 - Knowledge Discovery in Databases (adapted by (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)) 45
Figure 2 - Illustration of the supervised learning techniques used in this work. 48
Figure 3 - Illustration of the methods for imbalanced learning used in this work. 61
Figure 4 - The essential process to conduct a machine learning analysis (created by the author). 72
Figure 5 - Model building and evaluating process. 98
Figure 6 - Illustration of how the screening tests were selected. We considered only the first episode of carbapenem-resistant Gram-negative bacterial isolation for each patient. 101
Figure 7 - Exclusion of variables during the process. Of the 112 initials, only 35 remain on the final base. 111
Figure 8 - Boxplots representing the ROC values from the cross-validation process for each strategy and method. 116
Figure 9 - Boxplots representing the NPV of each method (points) for all strategies. 122
Figure 10 - Boxplots representing the NPV of each strategy (points) for all methods. 122
Figure 11 - Boxplots representing the MCCs of each method (points) for all strategies. 126
Figure 12 - Boxplots representing the MCCs of each strategy (points) for all methods. 126
Figure 13 - Flowchart for a screening culture strategy. 136
Figure 14 - Summary of this Chapter steps. 137
Figure 15 - Illustration of how the tests for each patient were selected. 139
Figure 16 - Cleaning, Splitting, and Preprocessing data. 144
Figure 17 - Boxplots representing the Brier score from the cross-validation process for each method. 146
Figure 18 - Calibration belts for the Nearest Shrunken Centroids at two confidence levels. CI:0-80% (light shaded area) and CI:0-95% (dark shaded area). 148
Figure 19 - Predicted values by NB and NSC. 151
Figure 20 - Classifiers´ mean rank across datasets. The point corresponds to the AR. 153
Figure 21 - Calibration belts for the Nearest Shrunken Centroids at two confidence levels. CI:0-80% (light shaded area) and CI:0-95% (dark shaded area) using external validation data. 155
Figure 22 - Top 20 attributes ranked by their Information Gain for all hospitals. 156
Figure 23 - Top 10 attributes ranked by their Information Gain. 157
Figure 24 - Dendrogram: similarity between items. 161

1
Introduction

Infections by antibiotic-resistant bacteria are among the most significant current threats to global health. Although these bacteria have become a cause of community-acquired infections, in general, they are associated with hospital-acquired infections (HAIs) (CARDOSO et al., 2015; VAN DUIN; PATERSON, 2016). These infections are frequently related to increased mortality, hospitalization time, and economic costs, mainly in the context of Intensive Care Units (ICUs), where severely ill patients have a higher risk of developing a hospital infection and frequently require antibiotics and invasive procedures (CHANG et al., 2011; ESCOLANO et al., 2000; JARRELL et al., 2018; MACVANE, 2017).

Although both Gram-positive and Gram-negative bacteria have demonstrated increasingly resistant patterns, the recent appearance of Gram-negative strains resistant to almost all antibiotics is an additional concern (FALAGAS et al., 2008; VARDAKAS et al., 2013). According to the global priority list of antibiotic-resistant bacteria developed by the World Health Organization (WHO) (TACCONELLI et al., 2017), the Carbapenem-Resistant Gram-negative pathogens (*Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and Enterobacteriaceae) are a critical priority (number 1) for research and development. In Brazilian hospitals, these Gram-negative bacteria cause a significant proportion of infections, and they are a substantial concern on infection control initiatives in ICUs (BRAGA et al., 2018).

First introduced during the 1980s, carbapenems play a critical role as some of the last-line agents for treating antibiotic-resistant Gram-negative pathogens (PAPP-WALLACE et al., 2011). Due to the global importance and the epidemiological relevance in Brazil and other LMICs, we decided to focus our research on Carbapenem-Resistant Gram-Negative Bacteria (CR-GNB).

Studies evaluating populations with Multidrug-Resistant Gram-Negative Bacteria (MDR-GNB) have shown high mortality rates ranging from 26% to 80% (JARRELL et al., 2018). A meta-analysis reported 1.78 times higher mortality in patients with MDR-GNB infections than patients with non-MDR-GNB infections. Still, the actual death rate attributable to resistant infections is unknown (ESCOLANO et al., 2000; FALAGAS et al., 2008; VARDAKAS et al., 2013).

Although some studies have addressed the influence of risk factors on infection, the prediction of CR-GNB acquisition, including colonization and infection, is also relevant. Infection usually occurs after colonization, and the timing of colonization is essential to determine the origin of the multidrug-resistant bacteria (VAN DUIN; PATERSON, 2016). Efforts directed at identifying colonization can help avoid transmission risks and decrease future infections (KOLLEF; FRASER, 2001).

Colonization refers to all patients who had any positive test for Carbapenem-Resistant Gram-negative bacteria that did not require antimicrobial treatment. Infection refers to patients with an infection documented by the clinicians with therapy initiation (EHRENTRAUT et al., 2018; LYE et al., 2012). Hospital-acquired infections were defined according to the Brazilian National Healthcare Safety Network criteria (NHSN, 2020).

Several predisposing factors have been associated with increased risks of infection or colonization, such as the patient's demographic characteristics, comorbidities, prior antibiotic use, and use of invasive procedures. Statistical methods have been used to evaluate the relationship between these factors and MDR-GNB acquisition, such as logistic regression, multi-state Markov models, decision tree analysis, and artificial neural networks (CHANG et al., 2011; ESCOLANO et al., 2000; GOODMAN et al., 2016). Predictive models can monitor and forecast resistant bacteria's possibility to be acquired in the hospital before it occurs, thereby reducing deaths, complications, and hospital costs (FERREIRA et al., 2017).

Although previous studies have analyzed the factors associated with MDR, we have not found research well-structured in this area using different techniques or evaluation methods to predict the risk of acquiring these pathogens or developing

screening models, notably in low- and middle-income countries' (LMIC) hospitals. Thus, the thesis's main objective is to develop a comprehensive and systematic approach applying machine-learning techniques to build models to predict CR-GNB acquisition in ICUs from Brazilian hospitals.

We divide this work into four specific objectives. The first is to predict Carbapenem-Resistant Gram-Negative Bacteria acquisition in ICUs, to assess the impact of this acquisition on mortality rate, and determine its risk factors using the logistic regression technique. In this first objective, we considered only positive results. This work was published in the "Journal of Hospital Infection" and is presented in Appendix A. Models developed displayed good results with an accuracy of ~90%, and patients who acquired CR-GNB were 2.72 times more likely to die than non-CR-GNB acquisition patients.

Some hospitals perform weekly culture tests in all inpatients, known as a screening process, to detect the existence of CR-GNB, independently of the risk of colonization. Since patients colonized are prone to spread these bacteria by contact without any symptoms, the screening allows isolating them and, if necessary, to treat them before compromising other patients and workers. However, despite the benefit of screening, they increase hospital costs and laboratory waiting times since hospitals and practitioners dedicate a significant amount of time and resources to the surveillance of these infections. That said, our second objective is to build a screening model that reliably detects ICU patients who do need to be tested since the high cost of surveillance testing can be avoided for some specific patients. The model aims to investigate the amount of non-colonization patients detected if surveillance activities follow a predictive model. The model is applied in hospital databases, and we explore additional data science techniques.

Since the dataset is imbalanced, containing a smaller number of observations in positive antibiotic-resistant tests, we combined machine learning techniques with balancing strategies. Thus, our third objective was to compare these combinations' performance regarding their discrimination power using different evaluation metrics. Moreover, we evaluate the trade-off between model performance and computational time.

The final objective identifies risk factors and develops a risk model that estimates ICU patients' probability of acquiring CR-GNB. We develop a matched case-control study and assess the acquisition probability by measuring the predictions' calibration. The model establishes the patient likelihood of acquiring the pathogens, including other clinical exams in addition to the screening tests. We also build an individual model for each hospital, discuss the factors' importance, and use association rule mining to identify the features that often occur together.

In short, we aim to reduce the number of necessary screening tests, know the probability of CR-GNB acquisition, identify the risk factors, and understand the techniques' behavior, extracting from Electronic Health Record (EHR) the administrative, pharmacy, and clinical data from ICU patients. Our methodology followed a framework developed by us about "how to conduct a machine learning analysis" that can be replicated in different healthcare settings.

1.1. Contributions

We show the potential of data mining and machine learning to complement the existing medical and engineering research. The main contributions of this thesis can be divided into literature, methodological, and applied in the following aspects:

Literature

• A literature review on prediction in the healthcare context, focusing on multi-resistant bacteria acquisition using a keyword-driven search strategy;

• Evaluation of how the different machine learning techniques and balancing strategies behave between the different metrics;

Methodological

• A framework about "how to conduct a machine learning analysis" that can be replicated in different healthcare settings;

Applied

• Combination of feature selection and cluster techniques to be applied to unbalanced problems comparing them to other methods;

• Rules of strongly associated features that indicate that a patient is at risk of acquired CR-GNB;

• An approach to screening modeling considering weekly tests and variables that consider actions that happened between one test and another;

• Two screening models to the hospital's decision-maker, one more conservative and the other moderate, and a CR-GNB acquisition risk model. These predictive models can be included in the hospital system and applied to each patient during hospitalization;

1.2. How to follow this document?

This work is divided into Chapters that describe the stages of the thesis. Chapter 1 presented the introduction and contributions of the thesis. Chapter 2 will provide a literature review of works conducted in the healthcare context, focusing on multi-resistant bacteria acquisition, classification techniques, and imbalanced learning methods. Chapter 3 will present the study problem and how we will conduct our machine learning analysis. Chapter 4 explains and develops the screening model, evaluating and discussing the different machine learning techniques and balancing strategies. Chapter 5 identifies risk factors, their associations, and develops a CR-GNB acquisition risk model. Finally, Chapter 6 will present the discussions, conclusions, and limitations of the thesis. References and appendices are given at the end.

2
Related Work

Chapter 2 provides some works on prediction conducted in the healthcare context, focusing on multi-resistant bacteria acquisition. We also present an overview of data mining and machine learning in healthcare, giving a detailed explanation of the classification techniques and imbalanced learning strategies used in our analysis.

2.1.
Learning methods in the healthcare context

A wide range of classification techniques has already been proposed in the healthcare literature, including statistical methods, such as Logistic Regression (LR) and Linear Discriminant Analysis (LDA), and non-parametric models, such as decision trees and Support Vector Machine (SVM).

Table 1 provides a selection of recent papers that have employed multiple comparisons of healthcare algorithms for predicting, along with the references and applications. They were selected through a non-systematic literature search for healthcare works that use machine learning techniques. Since our goal is to compare supervised machine learning techniques, we do not include the list of works that evaluate a simple algorithm or unsupervised machine learning. All these works examine and evaluate different predictive models. At the end of Table 1, we include information about our current thesis.

Table 1 - Healthcare techniques and their applications.

Authors [reference]	Application in a healthcare context	LR	SVM	NN	KNN	NB	Decision Tree	RF	Boosting /Bagging	Others (such as...)
(KANG et al., 2020)	Continuous renal replacement therapy		x	x	x			x	x	Multivariate adaptive regression splines
(LORETO; LISBOA; MOREIRA, 2020)	ICU admissions					x	x	x	x	Jrip, SMO, Logit Boost (LB), Iterative classifier (ICO)
(GANGGAYAH et al., 2019)	Breast cancer	x	x	x			x	x	x	
(GOODMAN et al., 2019)	ESBL infection	x					x			
(KUO et al., 2019)	Pneumonia	x	x		x	x	x	x		
(LIN; HU; KONG, 2019)	Acute Kidney Injury (AKI)		x	x				x		
(SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019)	Hospital Associated Disability	x						x		
(HARTVIGSEN et al., 2018)	MRSA Infections	x	x					x		
(KAUR; KUMARI, 2018)	Diabetes		x	x	x					Multifactor Dimensionality Reduction (MDR)
(TAN et al., 2017)	Multidrug-Resistant Tuberculosis	x					x			
(BACH et al., 2017)	Osteoporosis				x	x	x	x	x	
(LI; TANG; HE, 2016)	Multidrug-Resistant Tuberculosis						x		x	
(KELTCH; LIN; BAYRAK, 2014)	Liver fibrosis	x		x		x	x			
(PARK et al., 2013)	Breast câncer		x	x						semisupervised learning models
(KIM; KIM; PARK, 2011)	Mortality	x	x	x			x			
(PERIWAL et al., 2011)	Tuberculosis screening programs					x	x	x		Sequential Minimal Optimization (SMO)
Dantas's thesis (2020)	Carbapenem-resistant Gram-negative Bacteria	x	x	x	x	x	x	x	x	Linear Discriminant Analysis (LDA); Nearest Shrunken Centroids (NSC)

Legend: LR – Logistic Regression; SVM – Support Vector Machine; NN – Neural Network; RF – Random Forest

We can see that the most common technique is the decision tree, followed by SVM, NN, and LR, respectively. Of those cited, NB and kNN are still the least used for predicting. Complex algorithms are often very flexible and can learn many tasks, but they are often uninterpretable and function mostly as "black boxes," such as SVM and NN (BEAM; KOHANE, 2018).

In the literature, we can find many studies that use a single machine-learning algorithm to predict diseases, such as liver disease diagnosis by SVM (HASHEM; MABROUK, 2014), bronchitis symptoms among school-aged children using the gradient Boosting approach in a longitudinal framework (DENG et al., 2019), early identification of patients at risk for sepsis (DELAHANTY et al., 2019), among others.

Recently, Shillan et al. (2019) systematically reviewed 169 papers from 1991 to 2018 that applied machine learning to predict complications, mortality, length of stay, or health improvement using collected ICU data, following some inclusion

criteria. The predictions were evaluated in 161 of these studies. They found that the most common machine learning techniques were neural networks, support vector machines, and classification trees. However, since 2015, the random forest method and SVM methods use has increased.

In addition to Table 1, other techniques have already been proposed in the general literature, such as Linear Discriminant Analysis (LDA) (BROWN; MUES, 2012) and Nearest Shrunken Centroids (NSC) (DUALE et al., 2020). It is currently unclear which technique is the most appropriate for predicting disease. Hence, it is important to conduct studies applying various classification techniques based on a real-life healthcare data set.

The works presented in Table 1 compare the performance of some techniques for a given dataset sample, but most of them not consider the imbalanced data problem; that is, the studies have ignored the sample distribution.

Of 16 papers cited, only eight included data balancing strategies (BACH et al., 2017; HARTVIGSEN et al., 2018; KUO et al., 2019; LI; TANG; HE, 2016; LORETO; LISBOA; MOREIRA, 2020; PARK et al., 2013; PERIWAL et al., 2011; SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019). To deal with the imbalanced class problem, these papers applied the strategies of cost-sensitive learning (LORETO; LISBOA; MOREIRA, 2020; PERIWAL et al., 2011; SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019), random sampling (HARTVIGSEN et al., 2018; PARK et al., 2013; SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019), and SMOTE (KUO et al., 2019; LORETO; LISBOA; MOREIRA, 2020). Li et al. (2016) compared three strategies based on Bagging, Undersampling+Bagging, and EasyEnsemble; and Bach et al. (2017) analyzed different solutions of random undersampling, Edited Nearest Neighbors (ENN), and SMOTE. Our thesis compares 13 balancing strategies.

Several technique types have been compared in the literature to ascertain the most effective way of overcoming the class imbalance problem. Still, most current healthcare research on imbalanced learning focuses on data sampling methods and algorithm improvement.

Batista et al. (2014) identified some alternative techniques in dealing with class imbalances and tested them on different data sets (including healthcare) using

the k-NN algorithm. The techniques were Tomek Link, CNN, OSS, CNN + Tomek links, NCL, SMOTE, SMOTE + Tomek links, and SMOTE + ENN. They also analyzed under-sampling and over-sampling methods, and the findings suggested that, generally, over-sampling provides more accurate results than under-sampling methods.

Another essential issue to be considered is the reduction of dimensionality. Some methods have automatic feature selection, but others are sensitive to irrelevant predictors. Thus, feature selection methods can be used to improve results. Saeys et al. (2007) reviewed feature selection techniques for classification in bioinformatics. According to them, feature selection aims to avoid overfitting and improve model performance, providing more cost-effective and faster models, besides understanding better the generated data. They summarize the advantages and disadvantages of the three different feature selection types: filter, wrapper, and embedded.

In short, the filter technique is the fastest but ignores the interaction with the classifier, such as t-test, ANOVA, and regression. The wrapper interacts with the classifier but has the risk of overfittings, such as sequential search and genetic algorithms. The embedded techniques have better computational complexity than wrapper methods, using random forest or weight vector of SVM. The dependence on classifiers can be advantageous (SAEYS; INZA; LARRANAGA, 2007).

The most common methods were wrapper algorithm (KAUR; KUMARI, 2018; LORETO; LISBOA; MOREIRA, 2020) and stepwise variable selection (GOODMAN et al., 2019; KIM; KIM; PARK, 2011; TAN et al., 2017). Li et al. (2016) used only the automatic feature selection. Loreto et al. (2020) compared four different sets of attributes, including the PCA and wrapper method with Naïve Bayes. The remaining works did not present any technique for feature selection (BACH et al., 2017; GANGGAYAH et al., 2019; HARTVIGSEN et al., 2018; KANG et al., 2020; KELTCH; LIN; BAYRAK, 2014; KUO et al., 2019; LIN; HU; KONG, 2019; PARK et al., 2013; PERIWAL et al., 2011; SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019). This thesis analyzed four different feature selection methods explained in Chapter 3, section 3.3.4.3.

Most studies worry only about predictive accuracy, failing to analyze other objectives such as interpretability and the problem's objective (BAESENS et al., 2003). Some methods, such as NN and SVM, can report results slightly better than linear approaches but are more complex, increasing resources and reducing the interpretability. The choice of the model also depends on the decision-maker.

We develop and compare the following techniques in this thesis: Logistic Regression (LR), LR with regularization, Linear Discriminant Analysis (LDA), Nearest Shrunken Centroids (NSC), linear and radial Support Vector Machines (SVM), Neural Networks (NN), k-Nearest Neighbors (kNN), Naive Bayes (NB), decision trees (C4.5, CART, and C50), Random Forest (RF), a Gradient Boosting Machine (GBM), Bagging, and AdaBoost.

Besides that, our real-world data set is imbalanced, containing a much smaller number of observations in the positive class than the negative category. Therefore, we also use different imbalanced data techniques to solve the data imbalance and overlap problem. Some methods have already been used in healthcare applications (BACH et al., 2017; KUO et al., 2019; LI; TANG; HE, 2016; LORETO; LISBOA; MOREIRA, 2020; PARK et al., 2013; PERIWAL et al., 2011; SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019). Others can be seen in Batista et al. in different application areas (BATISTA; PRATI; MONARD, 2004). For us, misclassifying a negative class observation (false negative) is more critical than misclassifying a positive class observation. Our screening predictive model´s objective is to find not infected people (true negative) to avoid screening.

The classification techniques and balancing strategies applied in this work will be explained in sections 2.4 and 2.5, respectively.

2.2 .
Multi-resistant Bacteria Acquisition

2.2.1.
Hospital-Acquired Infections (HAI)

Hospital-Acquired Infections (HAI) represents a public health priority in most countries worldwide (RABHI; JAKUBOWICZ; METZGER, 2019). According to the World Health Organization (WHO) (OECD, 2018), 7% of hospitalized patients in high-income countries acquire some infection during hospitalization, raising this ratio to 10% in low-income countries. In Brazil, the last national survey conducted by the Ministry of Health in 2019 estimated that the rate of hospital infections reaches 14% of admissions (AGÊNCIA BRASIL, 2019).

HAI is determined by patient factors, such as the degree of immunocompromise, the excessive use of antibiotics, or natural broking barriers by interventions that increase risk (for example, surgeries or catheter implantation). It might also be developed in wounds after surgery or occur when microorganisms spread from person to person (EHRENTRAUT et al., 2018; GIRARD et al., 2002).

According to the Ministry of Health, the most common types of HAI are urinary and bloodstream infections associated with catheter use and pneumonia associated with mechanical ventilation (AGÊNCIA BRASIL, 2019). These rates vary related to the ICU's nature and the population studied (BOUZBID et al., 2011). In addition to the ICUs, the highest prevalence occurs in the surgical and orthopedic units (GIRARD et al., 2002).

Since the ICU activities are indispensable, complex, and expensive, it is exciting to assess, compare, and improve their quality of care and resource use. The ICU is the last line of defense for the critically ill (HALPERN; PASTORES, 2015; LI et al., 2019).

The Brazilian Health Regulatory Agency (ANVISA) is responsible for coordinating hospital infection control actions, aiming to reduce the national incidence of HAI in health services (ANVISA, 2016). According to ANVISA, the main action of prevention and control is hand hygiene to avoid infection through contact between patients or health professionals. Besides, it is also essential to

sanitize environments and beds, isolate patients who are already contaminated, and apply prevention protocols (AGÊNCIA BRASIL, 2019). Girard et al. (2002) affirm that the prevention of nosocomial infections includes such as: limiting transmission of organisms between patients through adequate handwashing, glove use, and isolation strategies; controlling environmental risks; minimize invasive procedures; optimal antimicrobial use; identifying and controlling outbreaks; prevention of infection in staff members.

Regulation 930, created in 1992, affirms that all hospitals must maintain the Hospital Infection Control Committee (CCIH), regardless of the entity responsible. In low- and middle-income countries such as Brazil, only a minority of hospitals have active infection control committees (PRADE et al., 1995). Surveillance of HAIs usually requires trained staff and a systematic approach, which is difficult to achieve when restricted. For this reason, monitoring is frequently overlooked in these countries.

The problems related to infections can be aggravated when these are caused by multiple drugs resistant pathogens. Thus, preventing hospital infections is becoming more and more critical in the current context of multidrug-resistant bacteria since some usual antibiotics are no longer sufficient for the treatment. The drugs to treat MDR are usually more toxic, less efficient, and frequently more expensive (AGÊNCIA BRASIL, 2019). The induction of antibiotic resistance is due to excessive use of these products in health care, communities, or animal breeding worldwide (GIRARD et al., 2002).

Infections caused by resistant bacteria to commonly utilized antibiotics are rapidly increasing, mainly in ICUs. At the beginning of the century, the Gram-positive bacteria were considered major healthcare threats. Nowadays, however, the attention has shifted to the multi-resistant Gram-negative bacteria due to outbreaks and increasing infection rates, especially the Klebsiella pneumoniae, which produces both extended-spectrum β-lactamases and carbapenemases. Only a few antibiotics remain active against these bacteria (BONTEN, 2012).

Recent data from the U.S. National Healthcare Safety Network indicate that Gram-negative bacteria are responsible for more than 30% of HAI, predominant in cases of ventilator-associated pneumonia and urinary tract infections (HIDRON et

al., 2008). According to previous studies of Gram-negative infections in Brazil, *A. baumannii* is the most prevalent carbapenem-resistant MDRGN bacterial (LUNA et al., 2014; RUBIO et al., 2013).

In addition to HAI, there are the community-associated infections, defined as infections manifested and diagnosed within 48 hours after patients´ admission without any prior medical assistance (TAPLITZ; RITTER; TORRIANI, 2017) - not analyzed in this work.

2.2.1.1.
Hospital surveillance

Hospital surveillance systems are becoming crucial to control and prevent infections and colonization acquired in the hospital. The study of nosocomial infection control (HALEY et al., 1985) demonstrated that surveillance should be included in infection control activities. In Brazil, hospitals have used standardized HAI surveillance methods adapted from the National Nosocomial Infections Surveillance (NNIS) system and developed by the Centers for Disease Control (CDC). They are based on the manual collection of clinical data from medical records, clinical laboratories, and antibiotic prescriptions. However, these active surveillance methods are both time-consuming and costly and do not focus on preventing infections (BOUZBID et al., 2011). In many hospitals, the consequence is probably underestimating the true HAIs incidence in acute care hospitals (RABHI; JAKUBOWICZ; METZGER, 2019).

Approaches focused on automated surveillance systems are emerging, consisting of cross-analyzing electronic data in different medical information systems. The disease automated or electronic surveillance is the process of obtaining information from interrelated electronic. It is possible due to the increase in the amount of data generated within different health institutions, such as administrative data (e.g., admission and discharge date, hospitalization characteristics), laboratory data (e.g., microbiology results), and clinical information system data (e.g., electronic health records, antibiotic prescriptions, use of invasive procedures) (RABHI; JAKUBOWICZ; METZGER, 2019).

A recent development in automated HAI surveillance is the adoption of machine learning techniques. Instead of using static rule-based algorithms, computers can learn by identifying data patterns (SIPS; BONTEN; VAN MOURIK, 2017). Some automated methods of infections have already been proposed in the literature, exemplified by Rabhi et al. (2019). The most frequently employed algorithms to identify HAI patients are classification algorithms or simple rule-based decision trees (SIPS; BONTEN; VAN MOURIK, 2017). However, these studies did not systematically analyze ML models' performances to approach multi-resistant bacteria detection; neither consider actions between one test and another.

Moreover, they were not developed in low- and middle-income countries' (LMIC) hospitals, where antibiotics use, invasive devices, and hospital settings are different. Other works use natural language processing methods for identifying infections, but we will not discuss them here. These models are task-specific and not easily generalizable (RABHI; JAKUBOWICZ; METZGER, 2019).

Few hospitals are conducting surveillance cultures to identify colonization by resistant Gram-negative organisms because of high test costs, lack of staff, limited laboratory resources, and the long wait time to get results (SONG; JEONG, 2018). Thus, a surveillance model can help select those who have a low probability of acquiring CR-GNB, avoiding the culture test.

2.2.2. Literature review on multidrug-resistant

There are many applications in the healthcare context. However, since our goal is to predict the acquisition of Carbapenem-Resistant Gram-negative pathogens, we focused on studies about multidrug-resistant (MDR) bacteria, Gram-negative (GN) bacteria, carbapenem resistance, and infection/colonization.

For data collection, we used the Scopus database and performed a keyword-driven search strategy. Our search's unified query was as follows: (multiresistant or MDR or multidrug-resistant or Enterobacteriaceae or Acinetobacter or Pseudomonas) and (predictive model or predicting or machine learning). Our search spanned 1,243 publications from 2005 until 2020 and comprised the fields “title,”;

"abstracts,"; and "keywords" with no limitations with regards to the field "journals."

The including criteria were papers with more than 100 samples that applied at least a multivariate method and analyzed hospital-acquired infection or colonization. Univariate statistical analysis was not considered because they do not adjust all variables to the model. Documents with only microbiology records, strictly related to clinical treatment, emergency departments, or pediatric patients, were excluded. We also manually screened the references of selected papers after eligibility criteria.

In the end, we selected 35 papers. Our goal is to provide an overview of each one of these works. Table 2 shows the classification methods used by each author; Table 3 summarizes information about the number of data sets, amount of cases and control, balancing strategies, dependent and independent variables, and bacteria type; and Table 4 includes some issues related to missing values, feature selection, results, and risk factors. This thesis results and our published work from the first objective also were added in this section.

Table 2 - Classification techniques by authors.

Classification techniques	Reference
Logistic Regression (LR)	(ALEXIOU et al., 2012; AN et al., 2017; CHAISATHAPHOL; CHAYAKULKEEREE, 2014; CHANG et al., 2011; DANTAS et al., 2019; DEBBY et al., 2012; FALCONE et al., 2018; FERREIRA et al., 2017; GOMILA et al., 2018; GOODMAN et al., 2019; HU et al., 2016; HUANG et al., 2012; JUNG et al., 2010; KENGKLA et al., 2016; KIDDEE et al., 2018; LEE et al., 2017; MARCHENAY et al., 2015; PARK et al., 2011; PATEL et al., 2014; PLAYFORD; CRAIG; IREDELL, 2007; ROMANELLI et al., 2009; ROUTSI et al., 2013; SCHWABER et al., 2008; SONG; JEONG, 2018; SURASARANG et al., 2007; TACCONELLI et al., 2008; TAN et al., 2017; TSENG et al., 2017; TUMBARELLO et al., 2011a, 2011b; VARDAKAS et al., 2015; VASUDEVAN et al., 2014; WILLMANN et al., 2014; YANG et al., 2016)
Neural Network (NN)	(CHANG et al., 2011)
Decision Tree	(GOODMAN et al., 2019; LI; TANG; HE, 2016; SONG; JEONG, 2018; TAN et al., 2017)
Bagging/Boosting	(LI; TANG; HE, 2016)

As shown in Table 2, most of the selected studies used the traditional application originating from logistic regression. Of the 35 papers, only five developed any other learning technique, such as Neural Network (CHANG et al., 2011), Decision Tree (GOODMAN et al., 2019; LI; TANG; HE, 2016; SONG; JEONG, 2018; TAN et al., 2017), and Bagging/Boosting (LI; TANG; HE, 2016). Besides, only four works used more than one method (CHANG et al., 2011; LI; TANG; HE, 2016; SONG; JEONG, 2018; TAN et al., 2017). It is possible to

conclude a lack of studies that use efficient alternatives to traditional methods to predict multi-resistant acquisitions. As shown in Table 1, the NN, k-NN, SVM, and NB are becoming increasingly popular in disease analysis and prediction in recent years. However, it is unclear from the literature which techniques are the most appropriate for each application.

Table 3 - Overview of empirical studies selected by data, balancing strategies, variables, and bacteria type.

Authors	Data sets	Good/negative cases	Bad/positive cases	Sample size ratio	Balancing strategies	# Independent Variables	Dependent Variable	Infection/ Colonization/ bacteria
(DANTAS et al., 2019)	1	1029	343	3.00	cut-off points - A matched case-control study according to the admission period (3:1)	23	carbapenem-resistant Gram-negative acquisition X non-acquisition	Gram-negative bacilli
(GOODMAN et al., 2019)	1	1094	194	5.64	cut-off points	14	Extended-spectrum beta-lactamases (ESBL) infection X ESBL non-infection	Enterobacteriaceae (Escherichia coli or Klebsiella spp bacteremia)
(FALCONE et al., 2018)	1	131	122	1.07	cut-off points	24	Bloodstream infections (BSI) caused by MDR-GNB X BSI due to susceptible GNB	Gram-negative bacilli
(GOMILA et al., 2018)	multicenter (together)	691	257	2.69	none	37	MDR in Gram-negative bacteria infections X Susceptible	Gram-negative bacilli
(KIDDEE et al., 2018)	2 (together)	243	32	7.59	none	19	Carbapenem-Resistant Gram-Negative Bacteria (CR-GNB) X Non-CR-GNB	Gram-negative bacilli
(SONG; JEONG, 2018)	1	355	89	3.99	cut-off points	37	Carbapenem-Resistant Enterobacteriaceae (CRE) colonized X non-colonized	Enterobacteriaceae
(AN et al., 2017)	1	947	168	5.64	none	21	Carbapenem-resistant A. baumannii (CRAB) acquisition X non-acquisition	*A. baumannii*
(FERREIRA et al., 2017)	1	198	66	3.00	none - Controls were selected in a ratio of 3:1 by the admission date	31	Healthcare-associated infections (HCAIs) x non-HCAIs	infection
(LEE et al., 2017)	2 (together)	1076	65	16.55	cut-off points	21	extended-spectrum b-lactamase (ESBL) producers X non-ESBL producers	Enterobacteriaceae
(TAN et al., 2017)	1	95	74	1.28	none	7	multidrug-resistant tuberculosis (MDR-TB) X without tuberculosis	tuberculosis
(TSENG et al., 2017)	1	873	122	7.16	cut-off points	16	Colonized by MDR-GNB X Non-Colonized	Gram-negative bacilli
(HU et al., 2016)	1	65	65	1.00	none - Matched for the year of ICU admission and site of infection.	27	Carbapenem-resistant Klebsiella pneumoniae (CRKP) X Carbapenem-susceptible Klebsiella pneumoniae (CSKP)	Enterobacteriaceae (Klebsiella pneumoniae)
(KENGKLA et al., 2016)	1	367	443	0.83	cut-off points	10	extended-spectrum beta-lactamase-producing Escherichia coli (ESBL-EC) X non-ESBL-EC	Enterobacteriaceae (Escherichia coli)
(LI; TANG; HE, 2016)	multicenter (together)	8709	86	101.27	Bagging; Undersampling+Bagging;EasyEnsemble	no information	multidrug-resistant tuberculosis (MDR-TB) X tuberculosis (TB) patients	tuberculosis
(YANG et al., 2016)	1	740	370	2.00	cut-off points - Matched case-control study according to the month of admission, ward, and interval days.	27	Carbapenem-resistant Klebsiella pneumoniae (CR-KP) infection X no CR-KP infection	Enterobacteriaceae (Klebsiella pneumoniae)
(MARCHENAY et al., 2015)	1	324	23	14.09	none	41	Carbapenem-resistant Gram-negative bacilli (CR-GNB) X non-CR-GNB	Gram-negative bacilli

Authors	Data sets	Good/negative cases	Bad/positive cases	Sample size ratio	Balancing strategies	# Independent Variables	Dependent Variable	Infection/ Colonization/ bacteria
(VARDAKAS et al., 2015)	1	18	73	0.25	none	48	Carbapenem-resistant Klebsiella pneumoniae (CRKp) infections X carbapenem-susceptible (CSKp)	Enterobacteriaceae (Klebsiella pneumoniae)
(CHAISATHAPHOL; CHAYAKULKEEREE, 2014)	1	110	105	1.05	none	14	Multidrug-Resistant Gram-Negative Bacteria infection X non-infection	Gram-negative bacilli
(PATEL et al., 2014)	1	195	103	1.89	none - A matched case-control (1:2)	31	Healthcare-associated Infections Caused by Extremely Drug-resistant Gram-Negative Bacilli (XDR-GNB HAIs) X non-XDR-GNB HAIs	Gram-negative bacilli
(VASUDEVAN et al., 2014)	1	1398	76	18.39	cut-off points	35	Resistant gram negative bacteria (RGNB) Infection X No GNB Infection/colonization	Gram-negative bacilli
(WILLMANN et al., 2014)	1	93	31	3.00	cut-off points - A matched case-control (1:3)	29	extensively drug-resistant P. aeruginosa (XDR-PA) colonization X non-colonization – screening model	*P. aeruginosa*
(ROUTSI et al., 2013)	1	630	85	7.41	none	26	Carbapenem-resistant (CR) GNB X without Gram-negative	Gram-negative bacilli
(DEBBY et al., 2012)	1	132	48	2.75	none	31	Carbapenem resistant Klebsiella pneumoniae (CRKP) colonization X non-colonization	Enterobacteriaceae (Klebsiella pneumoniae)
(HUANG et al., 2012)	1	164	62	2.65	none	39	Carbapenem-resistant A. baumannii (CRAB) X carbapenem-susceptible A. baumannii (CSAB)	*A. baumannii*
(ALEXIOU et al., 2012)	1	52	48	1.08	none	57	Patients with infections caused by MDR-GNB X Patients without infection caused by MDRGNB	Gram-negative bacilli
(CHANG et al., 2011)	1	1376	476	2.89	cut-off points - Matched case-control study according to time	16	HAI X non-HAI	infection
(PARK et al., 2011)	8 (together)	66	33	2.00	none - Matched by the hospital (2:1)	26	acquisition of extensively drug-resistant P. aeruginosa X non-acquisition	*P. aeruginosa*
(TUMBARELLO et al., 2011a)	1	226	113	2.00	cut-off points - Two control subjects were enrolled for each case	38	ESBL-producing Escherichia coli X non-ESBL	Enterobacteriaceae (Escherichia coli)
(TUMBARELLO et al., 2011b)	2 (together)	66	40	1.65	none - Matched case-control study according to the patients admitted to the same ward during the same period.	26	Multidrug-resistant (MDR) P. aeruginosa bloodstream infections X non-MDR	*P. aeruginosa*
(JUNG et al., 2010)	1	108	92	1.17	none	93	Multi-drug resistant A. baumannii bacteremia X nonBacteremic	*A. baumannii*
(ROMANELLI et al., 2009)	1	102	51	2.00	none - Matched with a 2:1 proportion	13	Carbapenem-resistant A. baumannii X non-resistant	*A. baumannii*
(SCHWABER et al., 2008)	1	59	48	1.23	none - Matched case-control study according to time	23	Carbapenem-resistant Klebsiella pneumoniae (CRKP) X no Klebsiella spp	Enterobacteriaceae (Klebsiella pneumoniae)
(TACCONELLI et al., 2008)	1	120/137	120/137	1.00	cut-off points - Matched case-control study according to number of days from admission and duration of hospitalization	16	Multidrug-resistant A. baumannii calcoaceticus (MDR-Abc)	*A. baumannii*

Authors	Data sets	Good/negative cases	Bad/positive cases	Sample size ratio	Balancing strategies	# Independent Variables	Dependent Variable	Infection/ Colonization/ bacteria
							colonization X non-colonization; MDR-Abc infection X non-infection	
(SURASARANG et al., 2007)	1	310	155	2.00	none - The cases were matched with controls by age and ward of admission with a ratio of 1:2.	22	Multi-Drug Resistant A. baumannii Infection X non-infection	*A. baumannii*
(PLAYFORD; CRAIG; IREDELL, 2007)	1	128	64	2.00	none - Each case was matched with two controls	16	Carbapenem-resistant A. baumannii (CR-AB) acquisition X non-acquisition;	*A. baumannii*
Dantas's thesis (2020) – screening model	**4 (together)**	**3,517**	**394**	**11.3**	**undersampling; oversampling; SMOTE; Tomek Links; NCL; OSS; SMOTE+Tomek; SMOTE+NCL; SMOTE+OSS; SMOTEBoost;RUSBoost; SMOTEBagging;UnderBagging.**	**112**	**Positive X negative culture test for CR-GNB;**	**Gram-negative bacilli**
Dantas's thesis (2020) – CR-GNB acquisition risk model	**5**	**3,604**	**527**	**14.6**	**Matched case-control by the hospital and admission date**	**98**	**Carbapenem-resistant Gram-negative Bacteria (CR-GNB) acquisition X non-acquisition;**	**Gram-negative bacilli**

According to Table 3, many studies only evaluate a small amount of data on a single healthcare data set. These data (good case/control and bad cases/cases) are often less imbalanced than those in real life.

The studies have predominantly used the natural distribution of the imbalanced classes, ignoring the various approaches developed in data mining and the evidence of the impaired accuracy caused by imbalanced data. A practice that was widely adopted is matching case and control by some specific condition or randomly under-sampling, where proportional numbers of cases and controls are used for model development by excluding examples of the majority class (CHANG et al., 2011; DANTAS et al., 2019; FERREIRA et al., 2017; HU et al., 2016; PARK et al., 2011; PATEL et al., 2014; PLAYFORD; CRAIG; IREDELL, 2007; ROMANELLI et al., 2009; SCHWABER et al., 2008; SURASARANG et al., 2007; TACCONELLI et al., 2008; TUMBARELLO et al., 2011a, 2011b; WILLMANN et al., 2014; YANG et al., 2016). Cut-off points were selected in (CHANG et al., 2011; DANTAS et al., 2019; FALCONE et al., 2018; GOODMAN et al., 2019; KENGKLA et al., 2016; LEE et al., 2017; SONG; JEONG, 2018; TACCONELLI et al., 2008; TSENG et al., 2017; TUMBARELLO et al., 2011a; VASUDEVAN et al., 2014; WILLMANN et al., 2014; YANG et al., 2016).

The balancing strategies appear only in Li et al. (2016) and Tseng et al. (2017). Hence, the issue of which classification technique to use for CR-GNB detecting, particularly with a small number of observations in a group (imbalanced), is a problem that needs to be addressed and combined.

Regarding the independent variables, Jung et al. (2010) analyzed 93 different variables - the largest among the selected studies - followed by Alexiou et al. (2012), Marchenay et al. (2015) e Vardakas et al. (2015). As shown in Table 3, our thesis analyzes 114 variables - the largest among these works. Moreover, we apply and compare 13 data balancing strategies, introducing new knowledge to the area.

Almost 40% of the studies aimed to predict colonization or infection by carbapenem-resistant bacteria (AN et al., 2017; DEBBY et al., 2012; HU et al., 2016; HUANG et al., 2012; KIDDEE et al., 2018; MARCHENAY et al., 2015; PLAYFORD; CRAIG; IREDELL, 2007; ROMANELLI et al., 2009; ROUTSI et al., 2013; SCHWABER et al., 2008; SONG; JEONG, 2018; VARDAKAS et al.,

2015; YANG et al., 2016). Of the 35 studies, 11 included and analyzed all Gram-negative bacteria collectively. The Enterobacteriaceae family, *A. baumannii*, and *P. aeruginosa* were also objects of study.

Table 4 - Overview of empirical studies selected by missing value analysis, feature selection, best results, and risk factors.

Authors	Missing value analysis	Feature selection methods	Best Results	Developed predictive model	Important features
(DANTAS et al., 2019)	uninformed	variable selection with backward elimination	Acc=0.891; Sens=0.875; Spec=0.895; PPV=0.718; NPV=0.959; AUC=0.914	YES	Increased Simplified Acute Physiology Score 3, patients with severe chronic obstructive pulmonary disease and exposure to hemodialysis catheter, central venous catheter, or mechanical ventilation.
(GOODMAN et al., 2019)	uninformed	variable selection with backward elimination and automatic feature selection	AUC=0.87; Sens=0.49; Spec=0.99; PPV=0.95; NPV=0.92	YES	no available
(FALCONE et al., 2018)	uninformed	stepwise variable selection	AUC=0.74; Sens=0.98; Spec=0.06	YES	Transfer from long-term care facility, hospitalization in the last three months, urinary catheter use, antibiotic therapy, and age more than 75 years.
(GOMILA et al., 2018)	uninformed	stepwise variable selection	AUC=0.80	YES	Male gender, acquisition of cUTI in a medical care facility, presence of an indwelling urinary catheter, urinary tract infection within the previous year, and antibiotic treatment within the last 30 days.
(KIDDEE et al., 2018)	uninformed	variable selection with backward elimination	-	NO	Use of an enteral feeding tube, hospitalization within the previous six months, antibiotic usage within the last three months.
(SONG; JEONG, 2018)	uninformed	automatic feature selection and variable selection with forwarding elimination	AUC=0.8; many cut-off points	YES	Isolation of multidrug-resistant organisms, ≥15 days of cephalosporin administration, ≥15 days of carbapenem administration, score ≥21 on Acute Physiology and Chronic Health Evaluation II.
(AN et al., 2017)	uninformed	variable selection with backward elimination	Sens=0.84; Spec=0.90; PPV=0.47; NPV=0.98	YES	Isolation and enhanced contact precaution.
(FERREIRA et al., 2017)	uninformed	variable selection with forwarding elimination	-	NO	Male, being aged >50 years and having an insertion of a central venous line during the hospital stay.
(LEE et al., 2017)	imputation	variable selection with backward elimination	AUC=0.92; Sens=0.85; Spec=0.92; PPV=0.40; NPV=0.99	YES	Recent antimicrobial use, recent invasive procedures, nursing home residents, and frequent ED user
(TAN et al., 2017)	uninformed	none and automatic feature selection	AUC=0.84; Sens=0.85; Spec=0.82	YES	Exposure to TB patients, family with financial difficulties, history of other chronic respiratory diseases, and smoking.
(TSENG et al., 2017)	uninformed	variable selection with backward elimination	AUC=0.80; Sens=0.57; Spec=0.85	YES	Age, residence in a long-term-care facility, history of cerebrovascular accidents, hospitalization within 1 month, and recent antibiotic exposure.
(HU et al., 2016)	uninformed	stepwise variable selection	-	NO	Previous carbapenem exposure
(KENGKLA et al., 2016)	excluded	variable selection with backward elimination	AUC=0.77; Sens=0.74; Spec=0.66; PPV=0.73; NPV=0.68; Acc=0.70	YES	Male gender, age, healthcare-associated infection, hospital-acquired infection, sepsis, prolonged hospitalization, history of ESBL infection, and prior use of antibiotics.
(LI; TANG; HE, 2016)	excluded	none and automatic feature selection	AUC=0.71	NO	no available
(YANG et al., 2016)	uninformed	none	AUC=0.90;Sens=0.85;Spec=0.68	YES	Age, male gender, cardiovascular disease, hospital stay, recent admission to the intensive care unit, indwelling urinary catheter, and mechanical ventilation.
(MARCHENAY et al., 2015)	uninformed	variable selection with backward elimination	-	NO	Duration of previous treatments with piperacillin-tazobactam.
(VARDAKAS et al., 2015)	uninformed	none	-	NO	No independent risk factors for the development of CRKP infections were identified.
(CHAISATHAPHOL; CHAYAKULKEEREE , 2014)	uninformed	none	-	NO	Admission to medical wards, respiratory tract origin, and hospital-onset infection.
(PATEL et al., 2014)	uninformed	none	-	NO	An immunocompromised state and exposure to amikacin, levofloxacin, or trimethoprim-sulfamethoxazole.

Authors	Missing value analysis	Feature selection methods	Best Results	Developed predictive model	Important features
(VASUDEVAN et al., 2014)	uninformed	variable selection with forwarding elimination	AUC=0.77	YES	Surgery during hospitalization, dialysis with end-stage renal disease; prior use of carbapenems; and stay in the ICU for more than five days.
(WILLMANN et al., 2014)	uninformed	variable selection with backward elimination	AUC=0.83	YES	Presence of a central venous catheter, presence of a urinary catheter, and ciprofloxacin administration.
(ROUTSI et al., 2013)	uninformed	variable selection with backward elimination	-	NO	Presence of Ventilator-Associated Pneumonia (VAP) and additional intravascular devices, and the duration of exposure to carbapenems and colistin.
(DEBBY et al., 2012)	uninformed	stepwise variable selection	-	NO	Recent surgical procedures and patient severity.
(HUANG et al., 2012)	uninformed	none	-	NO	Hematological malignancy, previous use of cefepime, and use of total parenteral nutrition.
(ALEXIOU et al., 2012)	uninformed	variable selection with forwarding elimination	-	NO	Use of special treatments during hospitalization, such as immunosuppressive therapies, use of metronidazole, and carbapenems.
(CHANG et al., 2011)	uninformed	variable selection with backward elimination	AUC=0.87; Sens=0.83; Spec=0.81; Acc=0.99 (external validation)	YES	Foley catheterization, central venous catheterization, arterial line, nasogastric tube, hemodialysis, stress ulcer prophylaxes and systemic glucocorticosteroids.
(PARK et al., 2011)	uninformed	variable selection with forwarding elimination	-	NO	Mechanical ventilation and APACHE II score.
(TUMBARELLO et al., 2011a)	excluded	variable selection with backward elimination	AUC=0.83; many cut-off points; calibration curve	YES	Recent hospitalization, transfer from another health care facility, Charlson comorbidity score, recent -lactam and/or fluoroquinolone treatment, recent urinary catheterization, and age.
(TUMBARELLO et al., 2011b)	uninformed	variable selection with backward elimination	-	NO	Presence of central venous catheter (CVC), previous antibiotic therapy, and corticosteroid therapy.
(JUNG et al., 2010)	uninformed	none	-	NO	Infection and respiratory failure at the time of ICU admission, maintenance of mechanical ventilation, and endotracheal tube maintenance instead of switching to a tracheostomy, recent central venous catheter insertion, bacteremia caused by another microorganism after colonization, and prior antimicrobial therapy.
(ROMANELLI et al., 2009)	uninformed	none	-	NO	Prior infection and mechanical ventilation.
(SCHWABER et al., 2008)	uninformed	stepwise variable selection	-	NO	Poor functional status, intensive care unit (ICU) stay, and receipt of antibiotics, particularly fluoroquinolones.
(TACCONELLI et al., 2008)	uninformed	none	McFadden R^2 = 0.70/McFadden R^2=0.65; many cut-off points	YES	Charlson index, previous methicillin-resistant Staphylococcus aureus isolation, and b-lactam use were independent risk factors for both. Bedridden status and previous ICU admission were associated only with colonization, while the presence of a CVC and surgery were related to infection.
(SURASARANG et al., 2007)	uninformed	none	-	NO	Prolonged admission of more than two weeks, use of devices, and prior treatment with certain antimicrobials.
(PLAYFORD; CRAIG; IREDELL, 2007)	uninformed	variable selection with backward elimination	Sens=0.91; NPV=0.8	YES	Prevalence of ICU colonized patients and ICU antibiotic use.
Dantas's thesis (2020) – screening model	**Imputation**	**Recursive Feature Elimination (RFE) with random forest, Selection by Filter (SBF), Class Decomposition with filter, and Class Decomposition with random forest.**	**AUC=0.75;Sens=0.92;Spec=0.39;PPV=0.14;NPV=0.98 MCC=0.20;**	**YES**	**Prior use of antibiotics, duration, and use of invasive devices, such as central venous or arterial catheters, mechanical ventilation, and urinary catheters; length of stay before the test; Admission Source; Admission Reason; Chronic Health Status; Saps 3; MFI point; Charlson Comorbidity Index; Chronic Atrial Fibrillation; Diabetes Uncomplicated; Dementia; Age; Stroke Sequelae; Neurological Coma Stupor Obtunded Delirium; Alcoholism.**

Authors	Missing value analysis	Feature selection methods	Best Results	Developed predictive model	Important features
Dantas's thesis (2020) – CR-GNB acquisition risk model	**Imputation**	**Recursive Feature Elimination (RFE) with random forest**	**Brier score = 0.152; MCC = 0.327**	**YES**	**Prior use of antibiotics, duration, and use of invasive devices, such as central venous or arterial catheters, mechanical ventilation, and urinary catheters; length of stay before the test; Admission Source; Admission Reason; Chronic Health Status; Saps 3; History of pneumonia; Digestive Acute Abdomen; Stroke Sequelae.**

Legend: AUC – Area Under Curve; Sens = Sensitivity; Spec = Specificity; PPV = Positive Predictive Value; NPV = Negative Predictive Value; Acc = Accuracy.

Table 4 presented some peculiarities of each study concerning the missing values analysis, feature selection, evaluation metrics, and relevant risk factors.

Most studies did not mention the handling of missing values (~90%), and when it is said, most of them directly exclude the records (KENGKLA et al., 2016; LI; TANG; HE, 2016; TUMBARELLO et al., 2011a). Only one applied the imputation method (LEE et al., 2017). Regarding the selection of factors, the only method employed was the stepwise variable selection with backward or forward elimination into multiple logistic regression in ~70% of studies.

As shown in Table 4, the evaluation metrics were the AUC value and the parameters resulting from the confusion matrix, such as accuracy, sensitivity, specificity, PPV, and NPV. However, the prediction evaluation was not carried out in 48% of cases; these studies only used the traditional statistical method to analyze risk factors. Since the dependent variables and inclusion criteria are different among the studies, we will not discuss the results.

Some risk factors are similar among the studies. The use of invasive devices (LEE et al., 2017; SURASARANG et al., 2007), such as a central venous catheter (CVC) (CHANG et al., 2011; DANTAS et al., 2019; FERREIRA et al., 2017; JUNG et al., 2010; TACCONELLI et al., 2008; TUMBARELLO et al., 2011b; WILLMANN et al., 2014), the arterial catheter (CHANG et al., 2011), mechanical ventilation (DANTAS et al., 2019; PARK et al., 2011; ROMANELLI et al., 2009; YANG et al., 2016), urinary catheters (FALCONE et al., 2018; GOMILA et al., 2018; TUMBARELLO et al., 2011a; WILLMANN et al., 2014; YANG et al., 2016), and hemodialysis (CHANG et al., 2011; DANTAS et al., 2019) were most likely to result in isolation of multidrug-resistant organisms or infection. Previous use of antibiotics was significant for about 60% of the studies (ALEXIOU et al., 2012; GOMILA et al., 2018; HU et al., 2016; JUNG et al., 2010; KENGKLA et al., 2016; KIDDEE et al., 2018; LEE et al., 2017; MARCHENAY et al., 2015; PATEL et al., 2014; PLAYFORD; CRAIG; IREDELL, 2007; ROUTSI et al., 2013; SCHWABER et al., 2008; SONG; JEONG, 2018; SURASARANG et al., 2007; TSENG et al., 2017; TUMBARELLO et al., 2011a, 2011b; VASUDEVAN et al., 2014; WILLMANN et al., 2014).

The severity of illness (DANTAS et al., 2019; DEBBY et al., 2012; PARK et al., 2011; TUMBARELLO et al., 2011a), and respiratory disease (DANTAS et al., 2019; JUNG et al., 2010; TAN et al., 2017) are also considered as risk factors. All the essential features can be seen in Table 4. Vardakas et al. (2015) did not identify any risk factor, and significant variables were not reported (not available) in two (GOODMAN et al., 2019; LI; TANG; HE, 2016) of 35 articles.

Our thesis calculates the Brier score value to assess the model's prediction in addition to the classification objective. We also compare four feature selection techniques to improve the performance of the algorithms.

Briefly, we describe some interesting findings from some articles cited in this review.

Goodman et al. (2019) explored decision tree and logistic regression methods, finding similar results. They compared the strengths and limitations of both classification methods. Chang et al. (2011) analyzed and concluded that both NN and LR models displayed excellent discrimination using external validations.

Li et al. (2016) focused on developing a warning system, which could early evaluate TB patients' risk converting to MDRTB using machine-learning methods (LI; TANG; HE, 2016). They used the CART as a classification method to compare three different imbalanced sampling strategies - CART with Bagging, CART + Under Sampling + Bagging (CART-USBagg), and CART + Easy Ensemble. The results showed that the best prediction could be obtained by adopting the CART-USBagg classification model, but they did not report risk factors.

Alexiou et al. (2012) analyzed postoperative infections caused by MDR-GNB in surgical patients. Patients who received antibiotics had 3.8 times higher odds of acquiring an infection caused by MDR-GNB.

In addition to the studies mentioned in Table 2, Table 3, and Table 4, some other resistance literature findings are exciting and deserve to be said ((BRAGA et al., 2018; COSGROVE, 2006; MAULDIN et al., 2010; PELEG; HOOPER, 2010). Cosgrove et al. (2006) studied the relationship between antimicrobial resistance and mortality, length of hospital stay, and health care costs. Mauldin et al. (2010) determine the additional total hospital cost and LOS attributable to HAIs caused by Gram-negative pathogens.

Peleg and Hooper (2010) explain the mechanisms of resistance in Gram-negative bacteria and the affected antibiotics, the evidence-based guidelines for preventing HAI, the risk factors for healthcare-associated infections and infection with drug-resistant bacteria, and the recommended empirical therapy to cover Gram-negative organisms that cause HAI.

Braga et al. (2018) provide an up-to-date picture of the extent, etiology, risk factors, and patterns of infections in ICUs of 28 Brazilian hospitals of different sizes. They found that the overall prevalence of ICU-acquired infections in Brazilian hospitals was higher than that reported in most European countries and the USA. Non-fermenting Gram-negative bacteria cause the highest proportion of infections.

Willmann et al. (2014) and Kiddee et al. (2018) were the only ones who worked with screening models. The first aimed to build a screening culture strategy using a conditional logistic regression model and a clinical risk score conducted by a matched case-control study for nosocomial colonization with extensively drug-resistant *P. aeruginosa*. The second one analyzed the screening for CR-GNB at ICU admission and discharge.

We found four systematic reviews about Gram-negative bacilli (BURILLO; MUÑOZ; BOUZA, 2019; FALAGAS; KOPTERIDES, 2006; MOHD SAZLLY LIM et al., 2019; RAMAN et al., 2018).

Burillo et al. (2019) reviewed some articles about risk factors for colonization or infection by MDR-GNB. They found that the patients colonized with an MDR-GN pathogen are older, previously exposed to antibiotics, have advanced comorbidities, a poor functional status, have prolonged hospital stays, or have been subjected to invasive procedures, such as CVC, mechanical ventilation, hemodialysis catheter.

Burillo et al. (2019) also present a review about predicting which patients carry a higher risk of colonization or infection. According to them, the accuracy of the model for predicting colonization is low, and ICU patients infected with MDR-GN have a worse critical than those infected by non-resistant microorganisms.

Falagas and Kopterides (2016) systematically reviewed the risk factors for the isolation of multi-drug-resistant *A. baumannii* and *P. aeruginosa*. The study

shows that prior use of carbapenems, third-generation cephalosporins, or fluoroquinolones is an independent risk factor for acquiring MDR *A. baumannii*.

Raman et al. (2018) conducted a review examining risk factors of the acquisition of resistant *P. aeruginosa*. They concluded that ICU admission, previous antibiotics, and prior hospital or ICU stay were the most significant variables.

Mohd Sally Lim et al. (2019) reviewed four existing clinical prediction models for extended-spectrum b-lactamase-producing Enterobacteriaceae (ESBL-EKP) colonization or infection. These studies apply only logistic regression. The most included predictors were the previous antibiotic use, prior hospitalization, transfer from another healthcare facility, and previous procedures.

In short, the existing analysis methods for predicting MDR-GNB or CR-GNB are poorly interpretable and little extensible in the literature. Thus, we used the literature found in section 2.1 to understand the learning methods most used in the healthcare context and how they were developed, aiming to bring these applications into the multiresistant context. On the other hand, the works cited in this section were useful for understanding the entire context, defining multiresistant, colonization, and infection, selecting inclusion criteria, study design evaluation, which the most significant variables, and how to treat them.

Our analysis adds to the current studies in four aspects: machine learning techniques, balancing strategies, feature selection, and performance evaluation. Our real-world data set is imbalanced, containing a much smaller number of observations in the class of antibiotic-resistant bacteria than in the non-colonized patient class.

2.3. Data Mining and Machine Learning in Health Care

The Knowledge Discovery in Databases (KDD) process can be commonly defined as a simplified process of (1) Selection, (2) Pre-processing (data cleaning, data integration, data reduction), (3) Transforming (normalization), (4) Data Mining and (5) Interpretation/evaluation. Many works treat data mining as a synonym for KDD, while others view data mining merely as an essential step in the

process of knowledge discovery (HAN; KAMBER; PEI, 2011). We consider both data mining and KDD as synonyms.

Data mining is the process of applying methods to discover patterns hidden in large data sets (HAN; KAMBER; PEI, 2011; KANTARDZIC, 2011), involving six standard classes of tasks: Anomaly detection, Association rule learning, Clustering, Classification, Regression, and Summarization (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Data Mining can explain some phenomena that are happening in data. Before data mining algorithms can be used, preprocess is essential to analyze and clean up the selected data set.

Machine learning uses the principles of data mining to build models that can predict future outcomes. They often intersect or are confused with each other, but in short, Data Mining explains patterns, and Machine Learning predicts with models (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Figure 1 demonstrates the association between KDD, Data Mining, and Machine Learning.

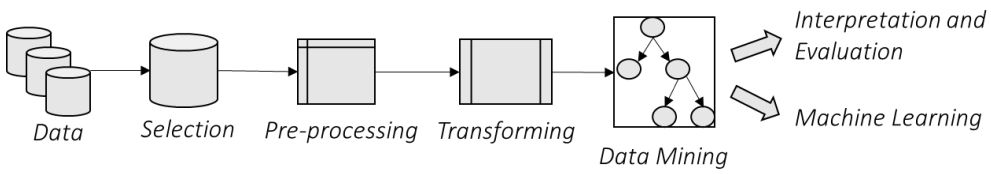


Figure 1 - Knowledge Discovery in Databases (adapted by (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996))

Machine Learning (ML) was initially described as a program that learns to perform a task or decide automatically from data. The ML algorithms attempt to identify a suitable model among the several possibilities, finding out which variables are essential among many individual measurements collected (BEAM; KOHANE, 2018).

To date, the industries have been the most beneficiaries in the availability of big data, ML, and data science, since they collected data and hire staff for that. The learning methods developed in industries also offer the potential for medical research and discovery, as most providers increasingly employ Electronic Health Record (EHR) (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019).

ML is a natural extension to the traditional statistical approaches and is a valuable and increasingly necessary tool for the modern health care system. Considering the vast amounts of information, a physician may need to evaluate the data (such as the patient's personal history, familial diseases, medications, among others) to find insight that guides the clinical decisions (BEAM; KOHANE, 2018).

The difference between statistical and ML techniques is poorly defined. Given the commonalities shared between them, Sidey-Gibbons and Sidey-Gibbons (2019) separate these approaches considering their primary goal. Statistical methods aim for inference, i.e., understanding the relationships between variables. On the other hand, ML focuses on the accurate prediction of real-life outcomes. Thus, models are developed not to infer the relationships between factors but rather to produce useful predictions from original data. However, prediction and inference are not mutually exclusive. ML is used when the relationship among features is complicated, usually nonlinear (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019).

ML is a computational method for automatic learning from experience (KAUR; KUMARI, 2018), offering new alternatives in the within-hospital surveillance and between-hospital comparisons (RABHI; JAKUBOWICZ; METZGER, 2019). In machine learning, classification aims to learn a system capable of predicting the new output class of a previously unseen instance with a good generalization ability (GALAR et al., 2012).

The ML field is concerned with developing algorithms and techniques that allow computers to learn and gain intelligence based on experience. It is a branch of Artificial Intelligence (AI) and is closely related to statistics. By learning, the system can identify and understand the input data to make decisions and predictions based on it (KAUR; KUMARI, 2018; SIDEY-GIBBONS; SIDEY-GIBBONS, 2019).

In section 2.4., we will explain in detail some classification techniques already previously applied and capable of developing good predictive models, such as Logistic Regression (LR), LR with regularization, Linear Discriminant Analysis (LDA), Nearest Shrunken Centroids (NSC), linear and radial Support Vector Machines (SVM), Neural Networks (NN), k-Nearest Neighbors (kNN), Naive

Bayes (NB), decision trees (C4.5, CART, and C50), Random Forest (RF), a Gradient Boosting Machine (GBM), Bagging, and AdaBoost.

In real-world applications, the number of cases where classes are imbalanced and overlapped is frequent, making it difficult for many learning algorithms (YANG; GAO, 2013). Since neither of the algorithms deals with the imbalance problem directly (GALAR et al., 2012), it has to be changed or combined with balancing strategies. The balancing strategies will be explained in detail in section 2.5.

2.4. Overview of classification techniques

There are supervised and unsupervised learning techniques that can be used to implement the machine learning process. Supervised learning techniques are used when the historical data (inputs and responses) are available for a specific problem. This approach includes, for example, decision trees, SVM, ANN, NB classifier, among others. On the other hand, the unsupervised learning technique is used when the available training data is unlabeled; i.e., the class level is unknown. The algorithm must explore and identify the patterns from the available records to make decisions or predictions. Unsupervised approaches include, for example, k-means clustering, hierarchical clustering, PCA, and the Hidden-Markov model.

We selected supervised machine learning algorithms for predicting whether a patient acquired CR-GNB or not, comparing the performance of a wide range of classification techniques within a healthcare context, thereby assessing class imbalance. Below, we present a brief explanation of each one of the classification methods applied in this thesis. We divided this topic into Linear Classification Models, Nonlinear Classification Models, and Classification Trees, as shown the Figure 2.

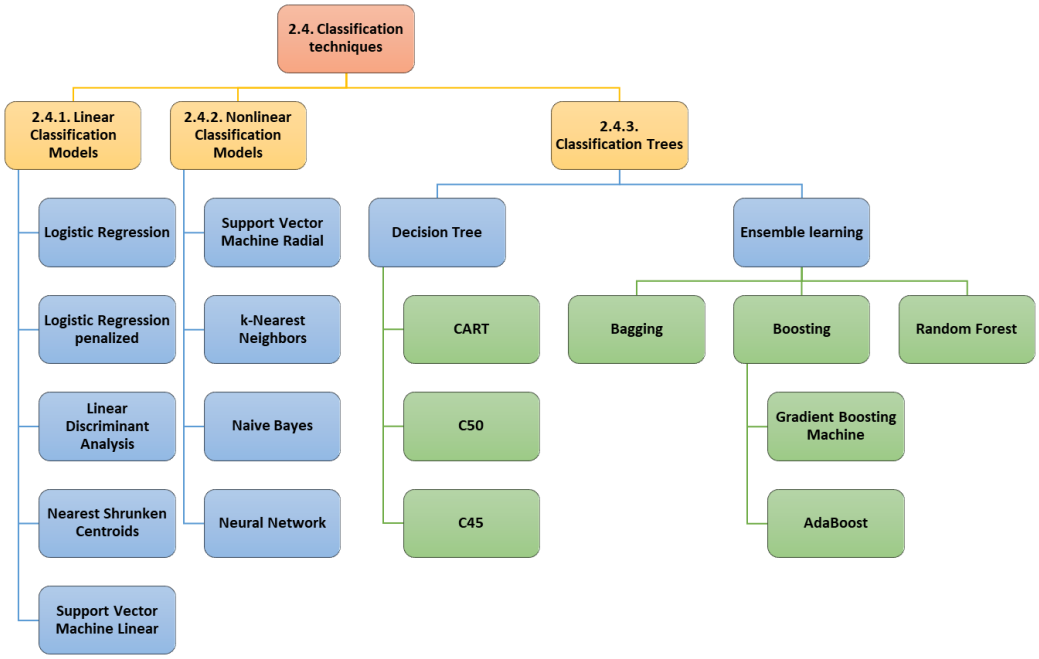


Figure 2 - Illustration of the supervised learning techniques used in this work.

2.4.1.
Linear Classification Models

Linear classifiers classify data into labels based on the value of a linear combination of input features (JAMES et al., 2013).

2.4.1.1.
Logistic Regression

In this thesis, we focus on the binary response of whether a test is positive or not and on the patient's risk of acquiring CR-GNB. For the first screening model, the response variable, y, can take on one of two possible values, i.e., y = 1 if the patient has a positive test, y = 0 if he/she had a negative test. For the acquisition risk model, y has a probability value between 0 and 1. The logistic regression model then takes the form of eq. (1).

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \propto + \beta^{T}x \qquad (1)$$

where $\propto$ is the intercept parameter, x is a column vector of M independent variables, and β^{T} contains the variable coefficients (HOSMER; LEMESHOW,

2000), representing the relationship between the variables and the reference level. The term p is the response probability of being modeled (likelihood of an event or a specific class), and the values range between 0 and 1, where $p = Pr(y = 1|x)$.

The term $\frac{p}{1-p}$ is known as the odds of the CR-GNB risk (Odds Ratio - OR) and can take on any value between 0 and ∞. Values of the odds close to 0 and ∞ indicate very low and very high probabilities of acquiring CR-GNB, respectively.

Logistic regression seems to be still the standard technique to predict CR-GNB, as seen in the sections 2.1 and 2.2. We believe that this happens due to two reasons. First, logistic regression is known for its straightforward interpretability, a property of great importance in medical applications. Second, logistic regression is probably the most established multivariate prediction technique for binary outcomes in statistics (SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019).

When the database has many factors, it is better to use logistic regression with regularization, also known as penalized logistic regression. This regression imposes a penalty to the logistic model for having many variables, reducing the fewer significant variables' coefficients toward zero. It reduces the model's variance, avoiding overfitting (KASSAMBARA, 2018; SIDEY-GIBBONS; SIDEY-GIBBONS, 2019).

The penalized regression includes Ridge, Lasso, and Elastic net regression (KASSAMBARA, 2018):

- Ridge regression: all the variables are incorporated into the model, but less significant variables have coefficients close to zero.
- Lasso regression: only the most contributive variables are kept in the final model, being the other coefficients forced to zero.
- Elastic net regression: a combination of ridge and lasso regression. It reduces some coefficients toward zero (like ridge regression) and some to zero (like lasso regression).

Penalized regression works like a feature selector that picks out the essential factors, i.e., most predictive (and have the lowest p-values).

2.4.1.2. Linear Discriminant Analysis

Discriminant Analysis (DA) is used to predict the probability of a case that belongs to a given category based on one or multiple predictor variables (KASSAMBARA, 2018). This technique assigns an observation of the response, $y(y \in \{0,1\})$, with the highest posterior probability; i.e., classify into class 0 if $p(0|x) > p(1|x)$, or category 1 if the opposite is true. According to Bayes' theorem, these probabilities are given by eq. (2) (BROWN; MUES, 2012):

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{2}$$

The Linear Discriminant Analysis (LDA) algorithm divides the space of predictor factors into regions. The regions are labeled by class and have linear boundaries. The model predicts the category of a new instance according to which part it lies in. The model predicts that all the cases within an area belong to the same group.

This algorithm aims to find directions that maximize the separation between classes, using these directions to predict individuals' class. These directions, called linear discriminants, are linear combinations of predictor variables. LDA assumes that predictors usually are distributed (Gaussian distribution) and that the different classes have class-specific means and equal variance/covariance (KASSAMBARA, 2018).

The scale in which predictor variables are measured also can affect this method. So, it is generally recommended to normalize continuous predictors before the analysis. Besides, if we normalize, the discriminator weights (coefficients of the scoring function) can measure variable importance for feature selection (KASSAMBARA, 2018).

LDA uses the input data to derive the coefficients of a scoring function for each category. Each function takes as arguments the numeric predictor variables of a case. It then scales each variable according to its category-specific coefficients and outputs a score. The LDA model looks at each function's score and uses the nearest score to allocate a case to a category (prediction). We call these scoring functions the discriminant functions (HOARE, 2020).

2.4.1.3.
Nearest Shrunken Centroids

The Nearest Shrunken Centroids (NSC) method computes a centroid of the data for each class by taking each predictor's average value in the training set. Suppose a predictor does not contain much information for a particular category. In that case, the centroid for that class is likely to be close to the overall centroid, computed using the training dataset (KUHN; JOHNSON, 2013).

The NSC method takes the profile of a new sample and compares it to each centroid. The class which centroid that is closest to is the predicted class for that new sample. This method has only one tuning parameter, the shrinkage threshold, and it is essential to normalize the predictors (KUHN; JOHNSON, 2013).

2.4.2.
Nonlinear Classification Models

2.4.2.1.
Support Vector Machine

Support Vector Machines (SVM) are powerful supervised learning techniques proposed by Cortes and Vapnik (CORTES; VAPNIK, 1995), mainly used for data whose distribution is unknown. It can be efficient processing in a linear and nonlinear data structure.

In the SVM model, data points are represented in a multi-dimensional space and are categorized into groups, where points with similar properties fall into the same group. The principle behind an SVM classifier algorithm is to build a hyperplane separating data for different classes in some transformed feature space. The focus while constructing the hyperplane is maximizing the hyperplane's distance to either category's nearest data point. These nearest data points are known as Support Vectors. The hyperplane has a maximum distance to any class's support vectors (FARQUAD; BOSE, 2012; HEARST et al., 1998; LI; LIU; HU, 2010).

SVM has high performance, deriving accurate predictions in situations where the relationship between features and the outcome is nonlinear. It uses the kernel

substitution principle to transform the feature space (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019).

An acceptable error, which represents C, and kernel are the only parameters for training SVM. For example, when we modify C to higher values, the decision boundary moves more towards majority class instances, in turn, misclassifying majority class instances as minority class instances (FARQUAD; BOSE, 2012).

In linear SVM, the given data set is considered a p-dimensional vector separated by a maximum of p-1 planes called hyper-planes (HEARST et al., 1998). We can compare linear SVM with linear regression, while nonlinear SVM is comparable to logistic regression. So, SVM can also be used for nonlinear data without requiring any assumptions about its functional form.

SVMs are not naturally probabilistic, but methods exist to turn them into probabilistic classifiers. SVM can detect and ignore outliers.

2.4.2.2. k-Nearest Neighbors

The k-Nearest Neighbors (kNN) algorithm classifies a new sample using K-closest samples from the training set (KUHN; JOHNSON, 2013). That is, given a new point y_0, we find the k training points (y(x), x=1,…,k)) closest in the distance to y_0, and then classify this new sample by taking a majority vote among the k neighbors (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). The closeness measure is determined by a distance metric (DUDANI, 1976; KUHN; JOHNSON, 2013), like Euclidean, Minkowski, Hamming, or Manhattan distance.

This thesis used the Euclidean distance, eq. (3), between the two points in feature space.

$$d_{(i)} = \left|\left|y_{(i)} - y_0\right|\right| \tag{3}$$

However, to allow each predictor to contribute equally to the distance calculation, it is necessary to normalize all quantitative predictors, reducing the effects of widely different scales of the numerical variables. Moreover, categorical values need to be converted into binary dummy variables.

Another question to the basic kNN algorithm is that not all neighbors are equally effective. Therefore, some works include weighing each of the k neighbors' contributions according to their distance, giving higher weight to closer neighbors (BATISTA; PRATI; MONARD, 2004).

This algorithm is not a black-box; that is, the neighbors can explain the classification result (DREISEITL; OHNO-MACHADO, 2002).

2.4.2.3.
Naive Bayes

Naive Bayes (NB) is based on Bayes' rule (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997). This classifier learns the conditional probability of each attribute given the category label from the training dataset. NB approach assumes that all descriptors are statistically independent, considering each of them individually. Bayes' theorem finds the probability of an event occurring given the likelihood of another event that has already happened. For example, the possibility of a patient to be in one or the other category depends on the ratio of members in each of the classes that share the descriptor value. The overall probability of activity is computed by the individual probabilities product (PERIWAL et al., 2011).

In short, the Bayesian Classifier estimates the parameters of a probability distribution, assuming that inputs are independent. Subsequently, the method calculates the posterior probability of validation samples belonging to each class. The samples are classified according to the highest posterior probability (CATENI; COLLA; VANNUCCI, 2014). Since NB requires a strong assumption of independent predictors, dependent predictors may lead to poor model performance (YANG, 2019).

2.4.2.4.
Neural Network

Neural Network (NN) has its structure inspired by the brain, and it is considered one of the most efficient techniques. A neural network is a collection of connected input/output units. The input layer connects with hidden layers not

visible to the external systems, relating to the output layer. These connections between layers are weighted, and the weights are adjusted during the learning step, helping the network predict the correct class label of the input. Therefore, the hidden layers perform nonlinear transformations of the inputs through an activation function and direct them as the output (HAN; KAMBER; PEI, 2011).

In short, NNs are built from units called perceptron (ptron). A perceptron has one or more inputs, a bias, an activation function, and an output. The perceptron receives data, multiplies them by some weight, and then passes them into an activation function to produce an output. Your output value is based upon the values of their inputs (AKWEI, 2017). The outcome is compared to a known label, and its weights adjust. This process repeats until we have reached a maximum number of allowed iterations or an acceptable error rate (KUHN; JOHNSON, 2013). To the NN method, data normalization is essential to speed up the learning and reduce the possibility of being stuck in local minimum (BROWNLEE, 2019a).

The advantages of neural networks include their high tolerance for noisy data and their ability to classify patterns on which they have not been trained. They can be used when you may have little knowledge of the relationships between attributes and classes and are well suited for continuous-valued inputs and outputs, unlike most decision tree algorithms. However, it is complex, and the net is essentially a black box, i.e., we can see the results of a neural network but cannot understand much about the fitting, the weights, and the model. The algorithm can be computationally expensive and not get better results than simple methods (HAN; KAMBER; PEI, 2011).

2.4.3.
Classification Trees

2.4.3.1.
Decision Tree

The decision tree is a type of supervised learning algorithm used in both regression and classification problems. It uses a recursive splitting mechanism to grow a tree, in which each split at a node is chosen to maximize information gain

or minimize entropy. The partitions creation process is repeated until some stop condition is met (depth of the tree, no more information gain, etc.) (PATEL; UPADHYAY, 2012).

Each non-leaf node represents a feature property test in a decision tree. Each branch represents the output of this feature property within a range, and each leaf node represents a category. A decision tree can be formally described as follows in eq. (4) (LI; TANG; HE, 2016):

$$(X, Y) = (x_1, x_2, x_3, \dots, x_k, Y) \tag{4}$$

The dependent variable, $Y \in \{-1, 1\}$, is the target variable, where $Y = 1$ represents a positive class, and $Y = -1$ means a negative class, with a probability between 0 or 1. The vector X is composed of all input variables, and each one of them consists of the patient's basic information, medication record, or clinical data.

The decision tree starts with all instances in the same group from the root node, then splits the data based on attributes until each case is classified, arriving at the leaf node (LI; TANG; HE, 2016). We can use the complexity parameter (cp) to control the size and select the optimal tree. If adding another variable to the tree from the current node does not decrease the error associated (cost) to the model, at least the value of cp, then the tree building does not continue; that is, this new split does not include. The larger the cp, the less complicated the decision tree.

One can produce different results from different decision tree algorithms. Besides that, other criteria are used to judge "best" in different algorithms. Here, we discuss three: CART, C4.5, and C5.0.

CART (Classification and Regression Trees) is based on Hunt's algorithm (BREIMAN et al., 1984). It uses cost complexness pruning to remove the branches that increase the tree's size but do not decrease the model (weak branches) and use Gini Index/Gini Impurity as a criterion for doing the binary splits.

Gini impurity represents the probability of a randomly chosen sample to be wrongly labeled in a subset. Suppose $i \in \{1, 2, \dots, m\}$ and m is the number of target categories, the Gini impurity ($\mathrm{Gini}(p)$) can be described as the following form of eq. (5):

$$\text{Gini}\,(p) = \sum_{i=1}^{m} p_i(1 - p_i) = \sum_{i=1}^{m} (p_i - p_i^2) = \sum_{i=1}^{m} p_i - \sum_{i=1}^{m} p_1^2 = 1 - \sum_{i=1}^{m} p_i^2 \qquad (5)$$

where p_i represents the probability of a sample is chosen as a category i. The maximum value of Gini impurity is $1 - 1/m$, and the minimum value of it is 0 when each instance in the node has a single target category.

Unlike CART, the C4.5 decision tree usages Information Gain and Entropy (QUINLAN, 2014), eq. (6), to build trees.

$$\text{Entropy} = -P_0 \log_2(P_0) - P_1 \log_2(P_1) \qquad (6)$$

where P_0 and P_1 are the proportions of the class values (-1/1) in the sample. The algorithm considers the difference in entropy (normalized information gain) as a splitting criterion, being the attribute with the highest gain chosen to decide. It adopts a strategy of post pruning.

Quinlan (2014) made improvements to C4.5 and called it C5.0. The C5.0 algorithm is faster, requires less memory, and gets results similar to C4.5 but smaller decision trees. When there are many alternatives to variables, a C5.0 mechanism called "winnowing" can select a subset of the variables used in the tree (KUHN; JOHNSON, 2013).

Decision trees are the most susceptible algorithms to overfitting, and significant pruning can reduce this likelihood. Overfitting happens when a model memorizes the training data so well that it is a learning noise. Pruning procedures, known as pre-pruning or post-pruning, reduce decision trees' size by removing nodes of the tree that do not improve sorting the instances.

Decision tree-based machine learning algorithms have several advantages, such as (KUHN; JOHNSON, 2013):

• Easy to understand even for people from a non-analytical background. It does not require any statistical knowledge to read and interpret them;

• Identify risk factors because they are immune to multicollinearity by design, i.e., clearly show which properties are more important among all features;

• Less data cleaning - outliers and missing values do not influence it to a fair degree;

• The data type is not a constraint: It can deal with numerical and categorical variables and does not require normalization and dummy variables.

However, decision trees also have some disadvantages, such as overfitting, losing information when categorizing continuous variables, and the predictive accuracy not quite robust. By aggregating many decision trees with methods like Bagging, random forests, and Boosting, the predictive performance can be substantially improved.

2.4.3.2.
Ensemble learning

Ensemble methods use multiple learning algorithms aiming to obtain better predictive performance. As an example of ensemble learning, the random forest algorithm combines random decision trees with Bagging to achieve high classification accuracy (BREIMAN, 2001). Two of the most popular techniques are Bootstrap aggregating (also called Bagging) and Boosting (LI; TANG; HE, 2016).

2.4.3.2.1.
Bagging and Boosting

Breiman introduced the concept of Bootstrap aggregating to construct ensembles (BREIMAN, 1996). Instead of fitting the model based on a single sample of the population, models provide different random subsamples with replacement. The Bagging algorithm consists of training different classifiers with bootstrapped replicas from the original training dataset. After, it is necessary to aggregate these models by using a voting system. The generated models are independent of each other and have equal weight. It aims to improve the performance of simple models and reduce the overfitting of more complex models.

Unlike Bagging, Boosting is an ensemble technique that attempts to create a robust classifier from some weak classifiers (GALAR et al., 2012). Constructing a model from the training data creates a second model that attempts to correct the prior model's errors. Other models are added in sequence, trying to fix the predecessor's mistakes, until the training set predicts correctly, or a maximum

number of models be added. Two Boosting representative algorithms are AdaBoost (Adaptive Boost) and Gradient Boosting Machine (GBM).

AdaBoost (FREUND; SCHAPIRE, 1997) was the first Boosting algorithm developed for binary classification. The idea is to find a weak classifier, calculate errors between predicted and real values, use the mistakes to find the reweigh of each observation, repeat this process many times, and output a combination of all classifier outputs. The Gradient Boosting (GB) (FRIEDMAN, 2002) is an algorithm similar to AdaBoost. The GB tries to fit the new predictor using the previous predictor's residual errors, not the instance weights.

2.4.3.2.2. Random Forests

Random forests (RF) use multiple models to achieve better performance than a single tree model (BREIMAN, 2001). This technique improves predictive accuracy by generating many trees based on random feature selection and bootstrap samples from the training data. The training observations may differ slightly while sampling, but the overall population remains the same. A case is classified by combining the results of all the trees generated in this new forest (an average in regression, a majority vote in classification). These tree-voting procedures are collectively defined as random forests. There are two parameters used to tune this technique: the number of trees and the number of attributes used to grow each tree (BREIMAN, 2001).

These multiple models can improve decision trees' performance on the test set, eventually avoiding overfitting. Since each tree grows out entirely, they each overfit, but in different ways. Thus, the mistakes one makes are averaged out over them all. We refer the reader to Breiman (2001) for additional details about RF and how to train them.

The random forest has been a preferred choice, especially in many biomedical applications, because it not only shows excellent prediction performance but is also known for its ability to tune and identify the most important variables (BREIMAN, 2001). However, RF is a predictive and non-descriptive modeling tool. It means that if one is looking for a description of the relationships in data, we should choose

other approaches, such as a decision tree. Besides, it is not easy to interpret since the models are created from many different trees.

Neither of these algorithms by itself deals with the imbalance problem directly. It has to be changed or combined with another technique (GALAR et al., 2012).

2.5. Methods for Imbalanced Learning

Several aspects may influence the performance achieved by existing learning systems. One of these aspects is related to a class imbalance in which cases belonging to one class far outweigh the other class's instances. It means that some categories have many more examples than others. This situation often happens in the real world. The classifiers' bias to the majority class tends to ignore the minority class, considering the minority class as noise (LI; LIU; HU, 2010; YANG; GAO, 2013).

We can find this problem, for example, in medical record databases regarding a rare disease, where there is a large number of patients who do not have that disease (BATISTA; PRATI; MONARD, 2004; CATENI; COLLA; VANNUCCI, 2014). Researchers also have reported difficulties to learn from unbalanced datasets in other domains, such as in fraud, telecommunications management, and detection of oil spills in satellite images (CHAWLA et al., 2002), detection of credit scoring to loan applicants (BROWN; MUES, 2012), financial problem (LIAO et al., 2014), caravan car policy (FARQUAD; BOSE, 2012), among others.

Most Machine Learning (ML) algorithms are not prepared to cope with a vast difference between the amounts of instances belonging to each class. Rule- and tree-based techniques, SVMs, and NN have often cited examples of inducers that suffer from this issue. The resulting classifier emphasizes the majority class instances at the expense of neglecting minority class examples, while the latter is usually the phenomenon of interest (VANHOEYVELD; MARTENS, 2018).

However, some domains show that class imbalance is not the only problem responsible for decreasing learning algorithms' performance. In addition to the problem related to learning with too few minority class examples, there is the

presence of other complicating factors, such as the degree of data overlapping among the classes (BATISTA; PRATI; MONARD, 2004; DENIL; TRAPPENBERG, 2010; HOANG; BOUZERDOUM; LAM, 2009; PRATI; BATISTA; MONARD, 2004).

The overlap problem means that the same region contains a similar number of training data for each class (DENIL; TRAPPENBERG, 2010). Conventionally, an example is considered to be overlapped if it possesses the equal probability of belonging to two different categories in a data distribution (LI; LIU; HU, 2010). That is, overlap occurs when the data of each class share the same area.

Overlapping and imbalanced data can lead to ineffective learning. The solutions to both problems are somehow inter-correlated since the final goal is to build up a reasonable decision boundary between the classes, providing good learning to models (DEVI; BISWAS; PURKAYASTHA, 2019).

To illustrate these problems, we consider a k-neighbor closest classifier (kNN) and decision trees. In the case of the kNN, the algorithm may incorrectly classify many instances of the minority class since the probability of the closest neighbors of these cases belonging to the majority class is higher. Many branches need to be created to distinguish the instances among classes regarding decision trees due to overlap. It usually leads to overfitting, making it necessary to prune the decision tree. When pruned, branches considered very specialized are removed, and there is a high probability that the majority class is the dominant class of the remaining nodes.

Some papers have discussed both problems (PRATI; BATISTA; MONARD, 2004), finding that the class overlapping has an even stronger role than the class imbalance in the concept of induction. They develop a systematic study using artificially generated datasets, aiming to show that class overlapping strongly correlates with class imbalance. As data complexity increases, the class imbalance factor affects the classifiers' ability (JAPKOWICZ; STEPHEN, 2002). Thus, dealing only with class imbalance problems does not always help classifiers' performance improvement.

Currently, most studies consider the two problems (imbalanced rate and overlap) separately. To the class-imbalance problem, several solutions were

proposed both at the data and algorithm levels. At the data level, sampling methods modify the training set by resampling (SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019). At the algorithmic level, solutions create or change the algorithms to adjust the classes (GALAR et al., 2012).

To the overlapping problem, data cleaning techniques have often been used, aiming at removing noise and the overlap that is introduced from the sampling schemes. The sampling can be combined with a possible data cleaning technique (HAIBO HE; GARCIA, 2009; VANHOEYVELD; MARTENS, 2018; YANG; GAO, 2013).

In addition to these approaches, other techniques are considered when using ensemble techniques. Ensemble learning algorithms do not solve the imbalanced data problem when directly applied, but their combination with other methods has led to positive results (GALAR et al., 2012).

Our next subsections cover the techniques most widely used in this work to cope with these issues, divided into sampling level (external approach), data cleaning methods, algorithm level (internal strategy), and ensemble learning, illustrated in Figure 3.

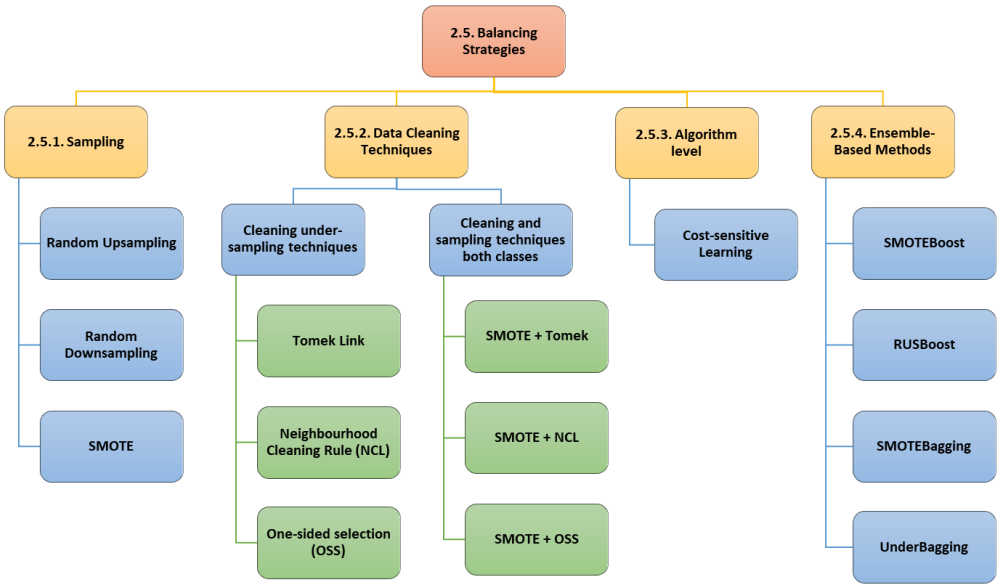


Figure 3 - Illustration of the methods for imbalanced learning used in this work.

In addition to these techniques, some other procedures have been introduced, cited in recent reviews (FARQUAD; BOSE, 2012; SHELKE; DESHMUKH; SHANDILYA, 2017; VANHOEYVELD; MARTENS, 2018). Vanhoeyveld and

Martens (2018) investigate the effects of resampling, cost-sensitive learning, and Boosting techniques. Shelke et al. (2017) present different methods of oversampling and undersampling. Farquad and Bose (2012) provide an overview of various researchers' balancing strategies and introduce a new approach for handling imbalanced data distribution. However, experiments must be conducted on each real-world data set to verify which method applies to them. The conclusions and results drawn from previous studies dealing with imbalanced learning cannot be generalized to specific data types.

2.5.1. Sampling

Researchers' most effective and most straightforward approaches to deal with unbalanced datasets are to resize the training samples (FARQUAD; BOSE, 2012).

The sampling methods consist of modifying an imbalanced dataset to provide a balanced distribution (VANHOEYVELD; MARTENS, 2018; YANG; GAO, 2013), and they are considered data level solutions. These techniques generally consist of oversampling (up-sampling) the minority class, under-sampling (down-sampling) the majority class, or a combination.

Down-sampling happens when some larger class instances are disregarded, aiming to create a sample with the same amount of cases from the smaller group. Many authors do not advise under-sampling once they consider it a potential loss of information (CRONE; FINLAY, 2012). Up-sampling (LING; LI, 1998) is a method that aims to balance class distribution through the random replication of minority class examples (BATISTA; PRATI; MONARD, 2004). More advanced techniques will introduce synthetic samples (explained below), in which the classifier creates more significant and less specific decision regions (CHAWLA et al., 2002). The sample methods are procedures independent of resampling methods such as bootstrapping and cross-validation. Crone and Finlay (2012) and Batista et al. (2004) concluded that oversampling usually performs better than under-sampling, mainly when small samples were involved.

Most of the time, we subsample the training set before model fitting. However, this approach generates samples that may not reflect the real class

imbalance, leading to overly optimistic performance estimates. Besides that, the results can differ under different subsample. The alternative to avoid these problems and improve the prediction is to include the subsampling inside the usual resampling procedure. The disadvantages are increased computational times and possible complications in some analyzes (ESTABROOKS; JO; JAPKOWICZ, 2004).

Balancing the class before or during the classifier's training does not necessarily improve this classifier's performance. According to Crone and Finlay (2012), for some methods, such as logistic regression, there is no benefit to building sample balance. On another side, discriminant analysis and decision trees are sensitive to the class imbalance, with a balanced sample performing better than the imbalanced one, mainly in the decision tree approach. However, it depends on each type of data.

Chawla et al. (2002) proposed a Synthetic Minority Over-sampling Technique (SMOTE). It is an over-sampling method that aims to create a new minority class by interpolating several minority class examples that lie together. The SMOTE algorithm selects k-nearest neighbors for each instance in the minority class. It creates synthetic samples along the lines between the minority class cases and their k-nearest neighbors (YANG; GAO, 2013). The overfitting problem is avoided but causes the minority class's decision boundaries to spread further into the majority class space (BATISTA; PRATI; MONARD, 2004), increasing overlapping between types.

Batista et al. (2004) also proposed two new methods trying to resolve the balancing and overlapping problems, SMOTE Tomek and SMOTE ENN, allying the SMOTE with the data cleaning methods Tomek links (KUBAT; MATWIN, 1997) and Edited Nearest Neighbor (LAURIKKALA, 2001; WILSON, 1972), respectively. The aim is to balance the training data and find better-defined class clusters, removing noisy instances lying on the wrong side of the decision border. They concluded that the worst problem happens when the class imbalance is allied to highly overlapped classes, decreasing the number of minority class examples classified correctly. Therefore, many actual questions add data cleaning techniques to the sampling methods.

2.5.2.
Data Cleaning Techniques

Some cleaning techniques have been created to solve overlap problems and removing noise in the learning system, such as Tomek links, Neighborhood Cleaning Rule (NCL), Edited Nearest Neighbor (ENN), Condensed Nearest Neighbor Rule (CNN), and One Side Selection (OSS). These data cleaning techniques have been used with sampling techniques (BATISTA; PRATI; MONARD, 2004).

2.5.2.1.
Cleaning under-sampling techniques

Tomek links method (KUBAT; MATWIN, 1997) can be used as a subsampling and cleaning method, removing noisy and borderline majority class examples until all minimally distanced nearest-neighbor pairs are of the same class. It resolves the overlap problem between the type of categories (BATISTA; PRATI; MONARD, 2004).

In the ENN method, noisy samples from the majority class are removed to under-sample the data (WILSON, 1972). It removes instances whose belonging level differs from at least half of its nearest k neighbors. Later, based on Wilson's ENN method, Laurikkala (2001) proposes the NCL. This technique modifies the ENN method by increasing the role of data cleaning. Firstly, NCL removes most examples whose class label differs from the class of at least two of its three nearest neighbors. After that, the neighbors of each minority example are found, and the ones belonging to the majority class are removed.

CNN rule aims to eliminate the examples from the majority class distant from the decision border since these cases might be considered less relevant for the learning system (BATISTA; PRATI; MONARD, 2004). “CNN searches for a consistent subset of the provided dataset, i.e., a subset that is enough for correctly classifying the rest of instances using 1-NN. To do so, CNN stores the first instance and goes for a first sweep over the dataset, adding to the stored bag those instances which are not correctly classified by 1-NN, taking the stored bag as the training set.

Then, the process is iterated until all non-stored cases are correctly classified" (HART, 1968).

The technique OSS aims to create a training set consisting of safe cases from application Tomek links followed by CNN's application (BATISTA; PRATI; MONARD, 2004). For that, noisy, borderline, and redundant majority class cases should be eliminated while leaving untouched all instances from the minority class (KUBAT; MATWIN, 1997).

2.5.2.2.
Cleaning and sampling techniques both classes

In this topic, instead of removing only the majority class examples that form Tomek links, samples from both levels are removed to create better-defined class clusters. First, the original data set is oversampled with SMOTE, and then Tomek links are identified and removed, producing a balanced data set with well-defined class clusters. This technique is known as SMOTE + Tomek (BATISTA; MONARD; BAZZAN, 2004). In the same year, it was created another similar method, known as SMOTE + ENN. ENN tends to remove more examples than the Tomek links, so it is expected to provide more in-depth data cleaning (BATISTA; PRATI; MONARD, 2004).

2.5.3.
Algorithm level

At the algorithmic level, solutions for imbalanced datasets include adjusting the weights of the classes or the probabilistic estimate, known as Cost-sensitive Learning (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006; PROVOST; FAWCETT, 2001). This technique defines a cost function against misclassification, giving different weights to specific types of errors (THAI-NGHE; GANTNER; SCHMIDT-THIEME, 2010; VANHOEYVELD; MARTENS, 2018). It incorporates costs during model training.

Unlike previous approaches, unequal costs can affect the model parameters. Therefore, class probabilities cannot be generated for the built models by cost-

sensitive learning algorithms and then estimating the ROC curve is impossible. Since we use the ROC curve to optimize the best hyperparameters set and choose the optimal cut-off for the classification model (in detail in Chapter 3), we cannot consider this balancing approach in the present work.

2.5.4. Ensemble-Based Methods for Class Imbalance Problem

The ensemble of classifiers is known to increase single classifiers, but as said earlier, these techniques alone cannot solve the class imbalance problem. Galar et al. (2012) reviewed state of the art on ensemble techniques for imbalanced data and found promising behavior approaches, which combine under-sampling techniques with Bagging or Boosting ensembles. They review many ensemble learning approaches proposed in the literature.

Here, we briefly describe four of these ensemble-based methods: SMOTEBoost (CHAWLA et al., 2003), RUSBoost (SEIFFERT et al., 2009), UnderBagging (BARANDELA; VALDOVINOS; SANCHEZ, 2003), and SMOTEBagging (WANG; YAO, 2009). Such methods combine ensemble methods (Boosting or Bagging) and data sampling techniques to improve model performance in the presence of class imbalance problems.

SMOTEBoost and RUSBoost are sampling approaches combined with Adaboost. AdaBoost iteratively builds an ensemble of weak learners, adjusting the weights of misclassified instances during each iteration. A misclassified sample has a higher weight to increase the probability that it appears in the next weak learner (CHAWLA et al., 2003). The difference between SMOTEBoost and RUSBoost is that the first one is based on the SMOTE algorithm and the other on the random undersampling at each Boosting iteration.

SMOTEBagging and UnderBagging are sampling approaches combined with Bagging. Instead of performing a random sampling of the whole dataset, minority class cases are selected by the SMOTE preprocessing algorithm and random undersampling, respectively, before training each classifier. The Bagging algorithm consists of training the classifiers with bootstrapped from the training dataset.

Even though many works are adopting only the balancing strategies to solve overlapping class issues, these alternatives might not be enough since this problem can also be caused by irrelevant or redundant features (ALI; SHAMSUDDIN; RALESCU, 2015). Thus, feature selection is also an important strategy used to address this issue.

2.6. Feature Selection

Determining which predictors are associated with finding the best model is referred to as feature selection (JAMES et al., 2013).

When data is presented with high dimensionality, usually, there is an increased risk of overfitting. Some models, notably SVM and NN, are sensitive to irrelevant predictors. However, even when an algorithm is insensitive, it makes sense to include the minimum possible set that provides acceptable results. Removing predictors can reduce the cost of acquiring data or improve the software's operation used to make predictions. Irrelevant features cause unnecessary expansion of model space, increasing training time, and reducing information provided by informative features (KUHN; JOHSON, 2019).

Feature Selection methods help with these problems by reducing the dimensions without a significant loss of the complete information. These methods are based on Filter Methods, Wrapper Methods, or Embedded Methods. The first one defines the relevance of features and filters out irrelevant features before learning, considering each predictor separately. The wrapper approach evaluates feature subsets using procedures that add or remove predictors to find the optimal combination that maximizes model performance, capturing interactions among multiple features (FREITAS, 2001; KUHN, 2011). Embedded (intrinsic) methods use algorithms with built-in feature selection, such as the tree- and rule-based models and regularization models. A good strategy is to combine an intrinsic non-linear model with a wrapper method (KUHN, 2011; KUHN; JOHSON, 2019).

However, significant bias may be introduced when dealing with the highly imbalanced dataset since the selected features may favor the majority class, not suitable for predicting the rare level (YANG et al., 2013; YIN et al., 2013). Some

papers had proposed different approaches for feature selection from data with a highly imbalanced class distribution. Yang et al. (2013) aimed to create multiple balanced datasets from the original imbalanced dataset via sampling, and for each balanced dataset, evaluate feature subsets using an ensemble classifier. Yin et al. (2013) provide a new approach based on class decomposition. Other studies for feature selection can be seen in the literature (CHEN; WASIKOWSKI, 2008; FORMAN, 2003; GROBELNIK, 1999; KIRA; RENDELL, 1992).

In this work, we used four different techniques: Recursive Feature Elimination (RFE) with random forest (non-linear model with a wrapper method), Selection by Filter (SBF), Class Decomposition with filter, and Class Decomposition with random forest. These techniques will be applied as described in our Methodology in Chapter 3.

3
Case Study Setting and Methodology

This chapter presents the study overview, database settings, material, and methods. We developed a methodological framework that can be used in other settings. This framework will be applied in Chapters 4 and 5, aiming to build a screening model that reliably detects who does not need to be tested and develops a risk model that can predict ICU patients' probability of acquiring CR-GNB during hospitalization, respectively. The methodological differences and particularities of each specific objective will be summarized at the end of this chapter.

For this study, our focus is on classification techniques on datasets with a large original class imbalance. The methods described in this work compute a continuous output between 0 and 1, representing a probability concerning the acquisition of CR-GNB. Depending on the goal, this probability can be classified in a binary (0 or 1) output through a cut-off.

3.1.
Study Overview

The study involves the hospitalized patients in ICUs of five hospitals at a sizeable Brazilian network. Nowadays, these hospitals' protocol performs weekly culture tests in all inpatients, known as a screening process, to detect the existence of Carbapenem-Resistant Gram-negative bacteria.

In the past, the patients were included in-hospital surveillance based on risk factors considered at the time of admission or at any time of hospitalization (old protocol detailed in Supplementary Appendix). However, starting in 2017, the protocol changed, recommending week culture screening for all patients hospitalized, independently of the risk of colonization or unit.

As stated in the introduction, since colonized patients are prone to spread these bacteria by contact without any symptoms, there is good evidence supporting this recommendation. Screening allows isolating these patients and, if necessary, to treat them before compromising other patients and workers. However, despite the benefit of screening, there has been an increase in hospital costs and laboratory waiting times since hospitals and practitioners dedicate a significant amount of time and resources to the surveillance of these infections.

That said, we aim to build a screening model that reliably detects ICU patients who do not need to be tested since the high cost of surveillance testing can be avoided for some specific patients. Therefore, the model reduces the budget of culture tests and laboratory and time spent by the laboratory staff, doctors, and nurses to collect and process exams. This predictive model can be included in the hospital system and followed by the surveillance activities to identify the patients that do not need to be screening.

Since we use different machine learning techniques, balancing strategies, and feature selection methods to classify the patients, we compare the predictive model's performance for discrimination regarding their ability to detect non-acquisition. Moreover, we evaluate the trade-off between model performance and computational time.

In addition to screening tests, there are other clinical exams needed during hospitalization, ordered by physicians for specific clinical reasons. Some of them also test the existence of Carbapenem-Resistant Gram-negative bacteria.

Thus, we also aim to develop a risk model that can estimate ICU patients' probability of acquiring CR-GNB, considering patients with screening or clinic tests. We call it the "acquisition risk model" since it includes colonized and infected patients. Unlike the previous objective, we assess the acquisition probability by measuring the calibration of the predictions. Moreover, we evaluate the importance of factors and use the association rules to identify the factors that often occur together. We aim to predict a risk model and find the critical prognostic factors influencing the acquisition, transforming them into decision support tools in the medical domain.

3.2.
Database settings

Our database gathers different data types from five Brazilian hospitals from May 8th, 2017 to August 31st, 2019, involving hospitalized patients in 24 adult ICUs. They have ~320 ICU beds and are tertiary care facilities with ~17,500 annual ICU admissions. Table 5 presents the structure of each hospital.

Table 5 - Structure of each hospital

Hospital	# ICUs	# ICU Beds	# Annual ICU admission
A	1	~10	~600
B	2	~26	~1400
C	5	~52	~4500
D	9	~140	~5700
E	7	~92	~5300

We obtained patients' data from the Epimed Monitor System®, antibiotic data from the Business Intelligence (BI) System, and microbiology data from the REAL system (see Appendix C). The Epimed Monitor System® database identifies all ICU admissions and the demographic and clinical variables of patients. It is a private electronic database of critically ill patients in Brazil. The BI database includes information related to the antibiotics used. Data from the microbiology laboratory was used to identify all patients with a positive or negative test for carbapenem-resistant GNB (*A. baumannii*, *P. aeruginosa*, or Enterobacteriaceae). Positive results include colonized or infected patients. The microbiology service provides data about the exams with species identification and phenotypic antimicrobial susceptibility testing (AST). Phenotypic AST was performed on the Vitek 2 platform (bioMérieux, Marcy l'Etoile, France) and interpreted according to the Clinical and Laboratory Standards Institute (CLSI) reference tables (PATEL; COCKERILL, 2017). Carbapenem resistance was defined phenotypically by CLSI criteria and MDR by a standard definition (MAGIORAKOS et al., 2012; PATEL; COCKERILL, 2017). All tests were requested without researchers' interference, and data were anonymized before being provided to us for analysis.

This research has been approved by the Research Ethics Committee (Comitê de Ética em Pesquisa – CEP) of the Plataforma Brasil, permission number CAE 15054519.3.0000.5249.

3.3.
A framework to machine learning analysis

This section describes the step-by-step of the learning process to develop the predictive models of this work. Figure 4 shows a framework created by us from other studies (HAN; KAMBER; PEI, 2011; KASSAMBARA, 2018; KUHN, 2011; KUHN; JOHNSON, 2013; KUHN; JOHSON, 2019) about "how to conduct a machine learning analysis" that can be replicated in different healthcare settings.

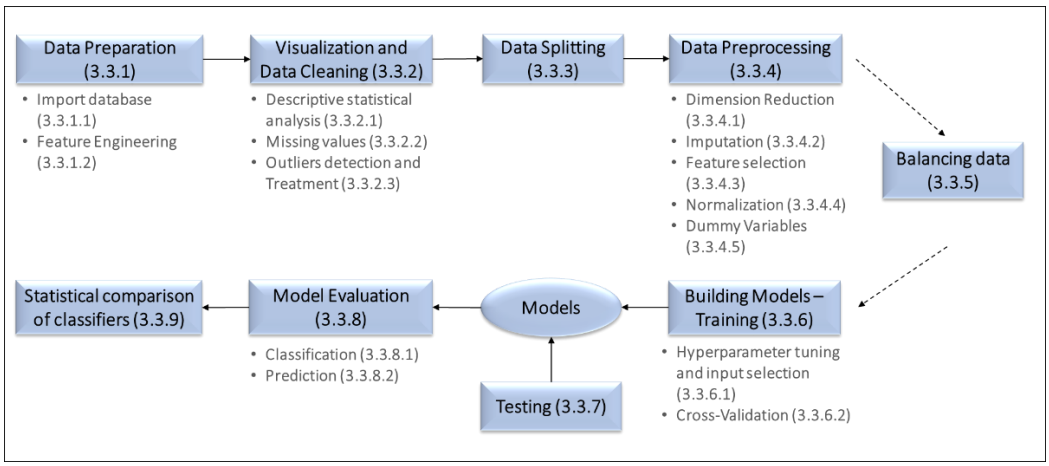


Figure 4 - The essential process to conduct a machine learning analysis (created by the author).

The learning process started with data extraction from unlinked EHRs, followed by the structuration of these data, feature engineering, missing data analysis, and outlier detection. We then split that base into training and test data and prepared the database through preprocessing from the training data. The preprocessing step can include analyses of variance and correlations, imputation, normalization, variables transformation, among other issues, depending on the algorithm.

Next, we built the models using supervised machine learning techniques based upon different algorithms through hyperparameter tuning and cross-validation. Since colonization is a rare event, we can balance the classes by applying also balancing strategies.

Once the models are built, we test them to calculate system performance metrics according to the predictive model's objective. Finally, since we trained many different models, we compared them. The following subsections explain each step applied in detail for our problem.

3.3.1.
Data Preparation

3.3.1.1.
Import database

The three databases collected (EPIMED, BI, and REAL - see section 3.2) were merged by hospital record and admission date and represented in a matrix where the n columns include n-1 features and one outcome. The features are the variables used to predict our model (for example, age, gender, admission source), and the result is the dependent variable (positive or negative test). Appendix C presents all variables available in each dataset.

In summary, our variables include patient information (age, BMI, gender, etc.), SAPS 3, Charlson Index, MFI points, underlying and associated diseases (such as hepatic failure, immunosuppression, steroid use, among others), reasons and source of admission, duration and previous exposure to invasive procedures, Length of Stay (LOS), and antibiotics use. These variables were selected based on their availability in the hospital electronic record system and through literature review (see Table 4).

The previous use of invasive procedures was considered if it occurred between 24 hours and 15 days before the test date and the previous antibiotic use if it occurred between 24 hours and 30 days before the test.

3.3.1.2.
Feature Engineering

Feature engineering is an essential step during the machine learning process, using the data to create appropriate predictors. We make 44 new input attributes, more informative, from the existing ones (Appendix C), such as length of stay in ICUs before the test (LOS_ICU_before_test), the total duration of vesical catheter use before the test (VesDURTOTAL), number of times that the vesical catheter was changed between one test and another (VesTIMESMORE), duration of the use of invasive procedures between tests, among others.

The categorical features with sparse classes were grouped since they have few observations and may cause a problem for some algorithms due to overfitting. For example, we combined two categories of the variable "Admission Source" for “Others” since the number of observations was less than 15. We also removed unused or redundant features from the dataset. Table 6 describes all features used in this work by category and type.

Table 6 - The description of each feature included in this work by category and type.

Category	Features	Description	Type
Laboratory tests	RESULT	Test result for carbapenem-resistant Gram-negative bacteria	binary
	tests_before	Amount of previous culture tests	numerical
Patient information	Age	Patient's age at hospital admission	numerical
	Gender	Sex of the patient	binary
	BMI	The measure of body fat based on height and weight known as Body Mass Index	numerical
ICU information	LOS_ICU_before_test	Length of stay before the culture test	numerical
Hospital information	LOS_hospital_before_test	Length hospital stay prior test date	numerical
	Hospital	Place where the patient was admitted.	categorical
Index	CharlsonIndex	Aggregate measure for prognosticating comorbidities	numerical
	MFIpoints	Modified Frailty Index	numerical
	FrailPatientMFI	Frailty (Yes or No) assessed using the MFI.	binary
	Saps3Points	Patient severity index based on physiological data	numerical
	SofaScore	Sequential Organ Failure Assessment score	numerical
Comorbidities	ChronicHealthStatus	Type of assistance to the patient	categorical
	IsChfNyhaClass23, IsChfNyhaClass4, IsCrfNoDialysis,IsCrfDialysis, IsCirrhosisChildAB, IsCirrhosisChildC, IsHepaticFailure, IsSolidTumorLocoregiol, IsSolidTumorMetastatic, IsHematologicalMalignancy, IsImmunossupression, IsSevereCopd, IsSteroidsUse, IsAids, IsArterialHypertension, IsAsthma, IsDiabetesUncomplicated, IsDiabetesComplicated, IsAngina, IsPreviousMI, IsCardiacArrhythmia, IsDeepVenousThrombosis, IsPeripheralArteryDisease, IsChronicAtrialFibrilation, IsRheumaticDisease, IsStrokeSequelae, IsStrokeNoSequelae, IsDementia, IsTobaccoConsumption, IsAlcoholism, IsPsychiatricDisease, IsMorbidObesity, IsMalnourishment, IsPepticDisease, Transplant, IsHypothyroidism, IsHyperthyroidism, IsDyslipidemias, IsChemotherapy, IsRadiationTherapy, IsHistoryOfPneumonia	If the patient has the comorbidity before arriving the unit	binary
Invasive device during hospitalization	VesDURTOTAL	The total duration of vesical catheter use before the culture test	numerical
	VesDURMORE	Time of vesical catheter use between one test and another	numerical
	VesTIMESTOTAL	Number of times vesical catheter was changed before the test	numerical
	VesTIMESMORE	Number of times vesical catheter was changed between one test and another	numerical
	VESICAL	Use of vesical catheter between 24h and 15 days before the test	binary
	CVCDURTOTAL	The total duration of central venous catheter use before the culture test	numerical
	CVCDURMORE	Time of central venous catheter use between one test and another	numerical
	CVCTIMESTOTAL	Number of times central venous catheter was changed before the test	numerical
	CVCTIMESMORE	Number of times central venous catheter was changed between one test and another	numerical
	CVC	Use of central venous catheter between 24h and 15 days before the test	binary
	DiaDURTOTAL	The total duration of hemodialysis catheter use before the culture test	numerical
	DiaDURMORE	Time of hemodialysis catheter use between one test and another	numerical
	DiaTIMESTOTAL	Number of times hemodialysis catheter was changed before the test	numerical
	DiaTIMESMORE	Number of times hemodialysis catheter was changed between one test and another	numerical
	DIALYSIS	Use of hemodialysis catheter between 24h and 15 days before the test	binary
	MVDURTOTAL	The total duration of mechanical ventilation use before the culture test	numerical
	MVDURMORE	Time of mechanical ventilation use between one test and another	numerical
	MVTIMESTOTAL	Number of times mechanical ventilation was changed before the test	numerical
	MVTIMESMORE	Number of times mechanical ventilation was changed between one test and another	numerical
	MV	Use of mechanical ventilation between 24h and 15 days before the test	binary
	PerDURTOTAL	The total duration of peripheral catheter use before the culture test	numerical
	PerDURMORE	Time of peripheral catheter use between one test and another	numerical
	PerTIMESTOTAL	Number of times peripheral catheter was changed before the test	numerical
	PerTIMESMORE	Number of times peripheral catheter was changed between one test and another	numerical
	PERIPHERAL	Use of peripheral catheter between 24h and 15 days before the test	binary
	ArtDURTOTAL	The total duration of arterial catheter use before the culture test	numerical
	ArtDURMORE	Time of arterial catheter use between one test and another	numerical
	ArtTIMESTOTAL	Number of times arterial catheter was changed before the test	numerical
	ArtTIMESMORE	Number of times arterial catheter was changed between one test and another	numerical
	ARTERIAL	Use of arterial catheter between 24h and 15 days before the test	binary
Reasons for ICU admission	AdmissionSource	Where the patient was before arriving in the ICU (Emergency; Hemodynamic Room; Operation Room; Other ICU from the hospital; Semi-Intensive Unit; Transfer from another hospital; Ward/Room; Others)	categorical

Category	Features	Description	Type
	AdmissionReason	Admission reason in the ICU (Cardiovascular/Shock; Elective Surgery; Emergency surgery; Endocrine/Metabolic/Renal; Infection/Sepsis; Liver and Pancreas/Gastrointestinal; Neurological; Non-surgical trauma; Oncological/Hematological; Respiratory; Others)	categorical
	Priority	1 - critical patients; 2 - intensive monitoring; 3 – vital and low probability of recovery; 4 - the low possibility of recovery with intensive monitoring; 5 - terminally ill	categorical
	IsNeurologicalComaStuporObtundedDelirium, IsNeurologicalSeizures, IsNeurologicalFocalNeurologicDeficit, IsNeurologicalIntracranialMassEffect, IsCardiovascularHypovolemicHemorrhagicShock, IsCardiovascularSepticShock, IsCardiovascularRhythmDisturbances, IsCardiovascularAphylacticMixedUndefinedShock, IsDigestiveAcuteAbdomen, IsDigestiveSeverePancreatitis, IsLiverFailure, IsTransplantSolidOrgan, IsTraumaMultipleTrauma, IsCardiacSurgery, IsNeurosurgery	If the patient entered the unit for these specific reasons	binary
Antibiotic use	Antibiotic	Use of any antibiotic between 24 hours and 30 days before testing	binary
	J01A, J01C, J01D, J01E, J01F, J01G, J01M, J01X, or J04A	Use of specific antibiotics between 24 hours to 30 days before the test, according to Anatomical Therapeutic Chemical Code (ATCC)	binary

3.3.2.
Visualization and Data Cleaning

After we have arranged our dataset into a suitable format with all possible variables, we may begin the visualization, cleaning, and pre-processing of data. These steps are essential before training our algorithms.

Our initial analysis summarizes the essential characteristics of data to a better understanding. First, we obtain a summary of the data through descriptive statistical detecting missing values. Then, we carefully visualize the data to detect outliers and errors. In these steps, we can identify and remove irrelevant and inconsistent data.

3.3.2.1.
Descriptive statistical analysis

Comparative analyses were performed between the positive and negative result tests. Continuous variables were matched between patient groups using Student's t-test or Wilcoxon signed-rank test, and categorical variables were compared using chi-square or Fisher's exact tests, as appropriate.

3.3.2.2.
Missing values

Missing clinical data values is unavoidable, and since the statistical analysis needs complete data, most studies usually exclude the patients with values not available (NA). It is known as "complete case analysis' (CCA) and is the number one strategy in the ICUs literature (VESIN et al., 2013). However, these excluded patients can lead to bias and loss of precision, affecting predictive models' performance. Most studies on MDR prediction, for example, have not mentioned the handling of missing values (Table 4), and when it is said, most of them exclude these patients.

There is no general rule on how much missing data is acceptable. In our case, if the variable had more than 10% missing values and a similar proportion between classes, we excluded it to avoid a negative impact on the actual data distribution. The remaining variables must be analyzed to understand the missing values'

randomness before making any decisions (SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019).

The risk of bias due to missing data depends on the reasons why data is missing, commonly classified as Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) (LITTLE; RUBIN, 2019; RUBIN, 1976; STERNE et al., 2009).

To determine if our data were MCAR, we used the statistical test Little´s MCAR test, which tests the null hypothesis that data is completely missing at random (NIELS, 2020). Since the test requires that the variables be normally distributed, we use the Box-Cox transformation to approximate normality.

If the null hypothesis is rejected, we must show that our data could be MAR through visualization of the missingness pattern (ZHANG, 2015). If MAR, we can affirm that the missing values depend on other variables (CHEVRET; SEAMAN; RESCHE-RIGON, 2015).

When the data is MAR, but not MCAR, analyses based only on complete cases may be biased. In this case, imputation methods (see subsection 3.3.4.2) should be used, replacing the missing data with values. They allow patients with incomplete data to be included in studies (CHEVRET; SEAMAN; RESCHE-RIGON, 2015; STERNE et al., 2009). Since MAR does not depend on the missing variables, we could model them using the observed data.

3.3.2.3.
Outlier Detection and Treatment

In data mining and machine learning, outliers refer to those samples that are different from most of the examples in datasets. Outlier detection is the task of finding outliers in some individual datasets according to specific rules (YANG; GAO, 2013).

According to Osborne and Overbay, outliers (or extreme scores) can be caused by a different type of errors, such as intentional misreporting of the participants; selecting of a distinct population from the rest of the sample; research methodology error; incorrect assumptions about the distribution of the data; and

human error in data collection, recording, or entry. However, outliers also can occur due to the inherent variability of the data. In this case, the outliers are legitimate cases sampled from the population (OSBORNE; OVERBAY, 2004).

We need to decide what to do with the identified outliers - whether to remove them or not. Here, we applied the tools of Overlaid Density Plots, Boxplot, and Histograms to detect outliers and data inconsistencies.

The overlaid density plots consider only numeric variables and display where values are concentrated over the interval. They are better at determining the distribution shape. Box plot represents the variation of observed data of the numerical variables through quartiles. Finally, we used multiple histograms to define the categorical variables.

3.3.3. Data Splitting

We divided the data into a training set (80%) and a testing set (20%), keeping the same proportion of majority and minority classes among subsamples. The training set creates predictive models, and the remaining validate the proposed model.

Balancing strategies (if necessary) are only applied to the training set since we cannot balance the test set artificially. When testing our developed model, we must represent the correct population proportions from which the sample was taken. Therefore, the test set remains unchanged throughout the whole analysis.

3.3.4. Data Preprocessing

Since we have several attributes, one of the first steps is to understand and transform the original data structure into an ideal form for the algorithm.

We have different pre-processing rules for each machine learning technique, as can be seen in

Table 7. We developed this table considering only the methods and algorithms used in this thesis.

Table 7 - Summary of the classification methods.

Method	Problem Type	Results interpretable	Algorithm {libraries in R}	Parameter tuning	Normalization	Dummy variable
LINEAR CLASSIFICATION MODELS						
Logistic Regression	Classification	Yes	Glm {-}	none	No	No
Logistic Regression with regularization	Classification	Yes	glmnet {glmnet, Matrix}	2 (alpha; lambda)	Yes	Yes
Linear Discriminant Analysis (LDA)	Classification	Yes	lda {MASS}	none	Yes	Yes
Nearest Shrunken Centroids (NSC)	Classification	A little	pam {pamr}	1 (threshold)	Yes	Yes
Support Vector Machine (SVM) - Linear	Either	Yes	svmLinear { kernlab}	1 (cost)	Yes	Yes
NONLINEAR CLASSIFICATION MODELS						
Neural Network	Either	No	Nnet {nnet}	2 (size; decay)	Yes	Yes
Support Vector Machine (SVM) - Radial	Either	No	svmRadial { kernlab}	2 (sigma/gamma; cost)	Yes	Yes
k-Nearest Neighbors (kNN)	Either	Yes	kNN {-}	1 (k)	Yes	Yes
Naive Bayes	Classification	A little	naive_bayes {naivebayes}	3 (laplace; adjust; usekernel)	No	No
CLASSIFICATION TREES						
Decision Tree C45	Either	Yes	J48 {RWeka}	2 (M, C)	No	No
Decision Tree CART	Either	Yes	Rpart {rpart}	1 (cp)	No	No
Decision Tree C50	Either	Yes	C5.0 {C50, plyr}	3 (trials; model; winnow)	No	No
Random Forest (RF)	Either	No	Rf {randomForest}	1 (mtry)	No	No
Gradient Boosting Machines (GBM)	Either	No	Gbm {gbm, plyr}	4 (interaction.depth; n.trees; shrinkage; n.minobsinnode)	Yes	Yes
Bagging	Either	No	Treebag {ipred, plyr, e1071}	none	No	No
AdaBoost	Either	No	AdaBoost.M1 {adabag, plyr}	3 (mfinal; maxdepth; coeflearn)	No	No

In short, the step by step of pre-processing works as follows: identification of zero-variance and near-zero-variance predictors; between-predictor correlations analysis; imputation to predict the missing values; normalization; and creation of dummy variables.

It is important to emphasize that the pre-processing must be done separately for the training and test sets. Once the model is created, we apply the same preprocessing parameters of the test set's training set as though the test set did not exist before. Thus, the test set does not influence the model training.

3.3.4.1.
Dimension Reduction

3.3.4.1.1.
Zero- and Near Zero-Variance Predictors

In some situations, data generation can create zero-variance (zv) predictors with a single unique value. It may cause failures or instability for the statistical models, except for those based on trees (KUHN, 2011).

Besides, factors may also become zero-variance predictors after data splitting into cross-validation sub-samples (see subsection 3.3.6.2). These predictors are known as near-zero-variance (nzv) predictors and need to be identified and eliminated before model construction.

To identify zv and nzv predictors, we calculated two metrics: the frequency ratio of the most common value over the second most frequent value and the number of unique values divided by the total number of samples (times 100). If the frequency ratio and individual value percentage are higher or less than a pre-specified threshold, respectively, we may consider a predictor to be near-zero-variance (KUHN; JOHSON, 2019). We should be careful with this identification and elimination analysis. Before eliminating any factor, we need to analyze its patterns in each class – this factor can be an "nzv" factor but a good predictor of a specific category.

3.3.4.1.2.
Between-Predictor Correlations Analysis

Correlation between features can negatively impact some model's stability and affect the prediction (KUHN; JOHNSON, 2013). Removing correlated predictors reduces multicollinearity and thus allow some types of models to be applied. Models like linear or logistic regression can be affected by correlated predictors.

The Pearson method (BENESTY et al., 2009) was used to calculate the correlation between continuous variables, reducing the number of predictors so that no pair has an absolute association higher than 0.75 (correlation considered strong) (RAFTER et al., 2003).

For categorical variables, the correlation concept can be understood in terms of effect size (strength of association) and significance test. Effect size indicates the power of the relationship. Goodman and Kruskal's tau is an appropriate and preferred measure for nominal variables. We used the association greater than 0.40 to consider a high association (VAUS; VAUS, 2013). For all pairs with a high association, we did the significance test. According to the significance test, when the p-value is less than the cut-off value (in this case, 0.05), we can reject the null hypothesis (H0: The two variables are independent) in favor of the alternative hypothesis (H1: The two variables are dependent), affirming that the variables are correlated to each other (SALKIND, 2010).

3.3.4.2. Imputation

We imputed the missing values of the variables by Multivariate Imputation by Chained Equations (MICE). MICE assumes that the lost data is MAR. It is an iterative algorithm that uses an imputation model specified separately for each variable and involving the other variables as predictors (STERNE et al., 2009). For example, suppose we have Y_1, Y_2, …, Y_k variables. If Y_1 has missing values, we use the Y_2 to Y_k variables as independent variables to predict the missing values in Y_1 (ANALYTICS VIDHYA, 2016). This imputation technique uses the observed data and then replaces the missing values with predicted values from a regression model.

Appropriate models were specified for different types of variables, such as Predictive Mean Matching (PMM) for imputing a continuous variable, Logistic Regression for imputing binary variables, and Bayesian Polytomous Regression for imputing factor variables (≥ 2 levels) (KUHN, 2011). MICE is fast and efficient on small datasets. It is better than kNN or means/mode methods (SCHMITT; MANDEL; GUEDJ, 2015).

Once the choice of imputation method has been made, we need to decide which variables will be used as predictors for imputation. Using every variable in the dataset to estimate the missing values can be problematic because variables that do not correlate with the variable attributed only adds noise to the estimation. Therefore, we built a predictor matrix with the predictors related to at least 0.2 with the target-variable (variables with missing values). Variables weakly correlated are left out. Besides, the estimates can be biased if too many auxiliary variables were included (MOONS et al., 2006).

Once imputed, the new database is then compared to the original one without missing values (only complete records). We visually checked the imputations, inspecting the initial distributions and the imputed data using scatter and density plots. The aim is to detect significant differences between observed and imputed data since both datasets should give similar results.

As said early, the test set cannot be influenced by the training set, and so both sets should not be preprocessed together.

3.3.4.3. Feature Selection

After mining, cleaning, and pre-processing, we may reduce the space's dimensionality by removing the irrelevant variables. Since the feature selection may change its importance after balancing or depending on the algorithm, we decide to select it before.

We used four different selection techniques: Recursive Feature Elimination (RFE) with random forest, Selection by Filter (SBF), Class Decomposition with filter, and Class Decomposition with random forest. These techniques are applied as described below, and the best feature subset is chosen by comparing some metrics and classifiers. We did not use stepwise selection because this procedure is appropriate only to linear models. It should be avoided due to the inflation of false-positive findings and model overfitting (KUHN; JOHSON, 2019).

Firstly, we used the wrapper method based on Recursive Feature Elimination in a random forest, known as the Random-Forest-Recursive Feature Elimination (RF-RFE) algorithm (GREGORUTTI; MICHEL; SAINT-PIERRE, 2017). RFE is

an approach that can be combined with any model to identify a useful subset of features with optimal performance for the model of interest (GUYON; WESTON; BARNHILL, 2002). It is probably the most used method for feature selection (KUHN, 2011). In our case, the random forest model conducted the backward selection, and the model's importance scores were used to rank the predictors.

Another approach is to use simple univariate statistical to select the variables. For that, we used the function SBF with univariate methods, such as t-test, Wilcoxon test, chi-square, or fisher's exact tests, as appropriate. This function considers a cross-validation approach in the search (KUHN, 2011). Filter methods do not consider the impact of multiple features together.

Aiming to deal with the highly imbalanced dataset, we propose to apply an approach based on class decomposition (YIN et al., 2013), as follows: firstly, we decomposed the majority class into i balanced pseudo-subclasses (i=1,2,…, T) by clustering, where the number of clusters K(i) corresponds to the ratio between the two classes. Then, we change each case's label to the name i by clustering, forming a multi-class dataset with $\sum_{i=1}^{T} K(i)$ subclasses. For the minority class, K(i)=1. We ranked the features according to the calculated scores using the filter method and random forest by the pseudo-labels. We named the first method of D.SBF and the second one of D.RF. We select the best features and then turn back to original labels.

For our problem, we used Mixed Data Types as the clustering method and Gower distance as the distance measure since there are categorical and continuous variables. The data were partitioned into ten clusters around medoids, known as Partitioning Around Medoids (PAM) - a more robust version and less sensitive to outliers than K-means (KAUFMAN; ROUSSEEUW, 1990).

Then, to evaluate the feature selection methods proposed (RF-RFE, SBF, D.SBF, and D.RF), we used four different algorithms (C4.5, SVM, kNN, and LR) as classifiers to better reflect the usefulness of the selected features.

We compared the performance of all ROC values and the Average Ranked (AR) performances on each classifier. We used ten-fold cross-validation to evaluate these measures and choose the approach that achieves better AR performance. Moreover, the Friedman test considered if each classifier's performance was

significantly different using different feature selection methods. As a result, we have some selected features.

3.3.4.4.

Normalization

Considering that the dimension between numerical features (including age, BMI, length of stay, etc.) is different and their range varies, it is necessary to normalize the original data before applying some methods, such as kNN, SVM, among others (

Table 7). Feature scaling, eq. (7), was used to this normalization, scaling all values into the range [0,1]. These transformations only change the data range, not the distribution (KUHN; JOHSON, 2019).

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (7)$$

, where x is an original value and $x_{normalized}$ is the normalized value.

3.3.4.5.

Dummy variables

Some classification methods are adaptive to apply unordered categorical variables, but others can only be used to continuous numerical data. For example, since the kNN is based on the Euclidean distance, it cannot be applied directly to categorical data. Therefore, we should convert categorical variables into binary dummy variables.

The mathematical function required to do this translation is known as a contrast or parameterization function. One of the categories of the predictor is left unaccounted for in the resulting dummy variables. For example, if we have five possible values and convert them, the contrast function would create only four dummy variables. It happens because since we know the four dummy variables' values, the fifth can be inferred (KUHN; JOHSON, 2019). The factor representing the corresponding categorical feature has a value of 1, and the other resulting factors have values of 0.

Table 7 shows to what methods this transformation is required.

3.3.5.
Balancing data

In this work, there is a significant data imbalance and class overlapping. We can balance the classes for the learning phase by applying balancing strategies or select data from potential controls matched. For the screening model (Chapter 4), we use different balancing strategies to demonstrate and compare the performance classification methods. Table 8 presents these techniques, described earlier in section 2.5. For the acquisition risk model (Chapter 5), we performed a matched case-control study by hospital and admission date.

The group of data with a more significant number of instances is called the majority class or negative class. In contrast, the group of data with the smallest number of cases is called the minority class or positive class.

Table 8 - Chronological overview of the balancing strategies used in this work.

Balancing approaches	References
Random downsampling (or undersampling)	-
Random upsampling (or oversampling)	-
SMOTE	(CHAWLA et al., 2002)
Tomek Links	(TOMEK, 1976)
Neighbourhood Cleaning Rule (NCL)	(LAURIKKALA, 2001; WILSON, 1972)
One-sided selection (OSS)	(KUBAT; MATWIN, 1997)
SMOTE + Tomek	(BATISTA; MONARD; BAZZAN, 2004)
SMOTE + NCL	(YONG SUN; FENG LIU, 2016)
SMOTE + OSS	Proposed by us
SMOTEBoost	(CHAWLA et al., 2003)
RUSBoost	(SEIFFERT et al., 2009)
SMOTEBagging	(WANG; YAO, 2009)
UnderBagging	(BARANDELA; VALDOVINOS; SANCHEZ, 2003)

We proposed the SMOTE + OSS strategy, following the same logic of the SMOTE + Tomek Link and SMOTE + NCL, but considering the OSS strategy.

The first nine strategies presented in Table 8 (downsampling, upsampling, SMOTE, Tomek Link, NCL, OSS, SMOTE + Tomek, SMOTE + NCL, and SMOTE + OSS) were applied to all 16 techniques analyzed.

We specified five learning algorithms to train weak learners within the ensemble model for the four ensemble-based strategies (SMOTEBoost, RUSBoost,

SMOTEBagging, and UnderBagging), as follows: SVM radial, NB, CART, C50, and RF.

Since these approaches can generate samples that may not reflect the real class imbalance, we include the usual resampling procedure strategies. That is, at each resampling into cross-validation, the data set is balanced and consequently changed.

In addition to the model's development, we want to analyze the most robust balancing strategy among our dataset and classification objective. That is, given a large variety of strategy, which one is the most capable of present an overall good (better) performance. Besides, we want to know how the balancing strategy influences the model's good result according to the different methods.

3.3.6.
Building Models - Training

We apply the following classification methods: Logistic Regression with and without regularization; Linear Discriminant Analysis (LDA); Nearest Shrunken Centroids (NSC); linear kernel Support Vector Machine (SVM-linear); Artificial Neural Network (ANN); radial basis kernel Support Vector Machine (SVM-RBF); k-Nearest Neighbor (kNN); Naive Bayes (NB); Decision Tree (CART, C45, and C50); Random Forest (RF); Stochastic Gradient Boosting (GBM); Bagging; and AdaBoost.

In short, we use 16 algorithms, classified into three families as follows: linear classification models, nonlinear classification models, and classification trees.

3.3.6.1.
Hyperparameter tuning and input selection

We used a tuned grid search to determine sets of hyperparameters that optimize each model fit. To improve the pruning strategy's reproducibility, we repeatedly varied the hyperparameters configurations and obtained the metric estimates using 10-fold cross-validation (subsection 3.3.6.2). We ran all models with the values of hyperparameters. The ones that maximize prediction based on

AUC (or Brier score) are stored and used in the final training model. After the hyperparameters are established, it remains fixed through the training process.

In the regularized logistic regression, the grid search mechanism is used to define the optimal value of lambda (λ), minimizing the cost function, and the type of regularization to be applied (α). Nearest Shrunken Centroids searches in the grid the best threshold value for the centroid shrinkage.

For the SVM classifier, linear and radial kernels were chosen. The constraint violation (C) cost is the parameter need for both techniques and the sigma/gamma parameter just for the radial SVM. A higher value of parameter C means a small margin hyperplane, and the smaller the margin, the smaller the misclassification. However, a lower misclassification in the training dataset does not mean a lower on testing data. The gamma parameter determines the reach of a training instance. High gamma values indicate that the SVM decision boundary is dependent on just the points that are closest to the border, ignoring cases that are farther away (BEN-HUR; WESTON, 2010).

The neural network classifiers are trained after selecting the best number of hidden layers (size) and the weight (decay). The "maxit" parameter sets the maximum number of 500 iterations used during training. We use the logistic sigmoid function for hidden layer activation.

The k-Nearest Neighbors technique applies a range from 3 to 10 neighbors (k) to choose the best result.

Naïve Bayes uses the grid search mechanism to the Laplace hyperparameter. The Laplace is a value incorporated into all probability estimates aiming that no probability be precisely zero. When a class and a feature never occur together in the training data, and the frequency-based probability estimate is zero, it wipes out all information in the other probabilities when multiplied (DOMINGOS; PAZZANI, 1997). We also add the kernel as a distribution type because we have non-binary predictors. The “adjust” hyperparameter allows us to tune the kernel density.

About the decision tree, we considered the post-pruning strategy of Minimum Error, in which the tree is cut back to the point where the cross-validated error is minimum. This happens after the tree has been built. For the pruning strategy of C4.5, we varied the Minimum Instances per Leaf (M) and the Confidence Level (C)

to decide whether to replace an internal node. The strategy of CART uses the complexity parameter (cp), explained in section 2.4. We also analyzed the maximum depth of the final tree (maxdepth) and the minimum number of observations in a node (minsplit). The decision tree by the C50 algorithm needs tuning the number of Boosting iterations (trials), model type (rule or tree), and if we want or not a feature selection (winnow).

The number of attributes (mtry) used to grow each tree is a parameter for the Random Forest technique. So, a range of different randomly selected attributes per tree has been assessed.

Four parameters have to be set for the GBM: the maximum depth of each tree (interaction.depth), the total number of trees to fit (n.trees), the learning rate (shrinkage), and the minimum number of observations in the terminal nodes of the trees (n.minobsinnode).

The AdaBoost classification trees need to choose three parameters: the number of iterations for which Boosting is run (mfinal), maximum depth (maxdepth), and the algorithm (outlearn).

The LDA, LR, and Bagging classification techniques require no parameter tuning.

Each algorithm has other particularities controlled, such as the minimum number of observations in a node (CART) and the method for stopping (Boosting).

For the balancing approaches, each learning algorithm in the ensemble strategies trained ten weak learners.

Table 9 shows the configuration hyperparameters that we have used to run the algorithms. The hyperparameter ranges were chosen according to the literature concerning what makes sense for our data. There was no specific rule.

Table 9 - Hyperparameter ranges.

Method	Hyperparameters
Logistic Regression	none
Logistic Regression with regularization	.alpha = c(0, .2, .4, .6, .8, 1); .lambda = seq(.01, 1, length = 20)
Linear Discriminant Analysis (LDA)	none
Nearest Shrunken Centroids (NSC)	.threshold = c(0:5)
Support Vector Machine (SVM) - Linear	.C = 2^(-3:3)
Neural Network	size = 1:5; .decay = c(0, .1, .5, 1, 1.5, 2)

Method	Hyperparameters
Support Vector Machine (SVM) - Radial	.sigma=2^(-3:3); .C = 2^(-3:3)
k-Nearest Neighbors (kNN)	k=c(3:10)
Naive Bayes	laplace=c(seq(1, 5, length = 40)); usekernel = TRUE; adjust=c(seq(1, 5, length = 40)
Decision Tree C45	.M=c(10,15,20,25,30,35,40), .C=c(0.5,0.4,0.3,0.2,0.1,0.05,0.025,0.01,0.005,0.001)
Decision Tree CART	cp=c(0.01,0.05,0.1)
Decision Tree C50	.winnow = c(TRUE,FALSE); .trials=c(5,10,15); .model="tree"
Random Forest (RF)	mtry=c(5:20)
Stochastic Gradient Boosting (GBM)	interaction.depth = c(3, 5, 10, 15); n.trees = (5:10)*30; shrinkage = c(0.1, 0.01, 0.001); n.minobsinnode = c(10,20)
Bagging	None
AdaBoost	mfinal = (1:5)*10; maxdepth = c(8, 10, 12); coeflearn = c("Breiman")

3.3.6.2.
Cross-validation

Since any method with tuning parameters can be prone to overfitting, we must take the cross-validation or resampling approach to determine each technique's parameters' optimal value. We choose to follow with cross-validation, in which all observations are used for both training and validation once (LI et al., 2019). A large variance in performance between folds may be indicative of overfitting.

The K-fold cross-validation technique avoids overfitting and assesses how well the models can predict different data subsets (JAMES et al., 2013). The 10-fold cross-validation is the process from which data is randomly partitioned into ten same-sized subsets. Among these subsets, a single subset is retained as the validation data for testing the model, and the other nine are used as training data. The process is repeated ten times, with each of the ten subsets used precisely once as validation (LI et al., 2019). We choose 10-folds to guarantee a statistically significant measure of the mean and standard deviation (RABHI; JAKUBOWICZ; METZGER, 2019).

We used the cross-validation method to computes the best hyperparameters of each model. Once it is chosen, we train the model, using all the training set.

3.3.7.
Testing

Once training is concluded, the final models are applied to the features in the testing set. The algorithm's predictions are compared to the testing dataset's known outcomes to establish model performance. The algorithm must generalize well to new data (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019). Based on the differences between the proposed model´s predicted and observed, some performance measures evaluate the model's ability.

3.3.8.
Model Evaluation

We must decide how to evaluate the developed models. First, we want to know which metric to consider each model and then device the best model. Moreover, there is a difference between the models' outcomes, depending on whether we assess the discrimination (classification) or prediction (ALBA et al., 2017).

3.3.8.1.
Classification

The default metrics used for classification and regression problems are usually Accuracy and Root Mean Squared Error (RMSE), respectively. However, when the class probabilities are quite different, the use of Accuracy measures might lead to misleading results since they are strongly biased to favor the majority class.

This subsection shows how to use some evaluation metrics for classification based on the confusion matrix analysis and the Receiver Operating Characteristic (ROC) curve. Table 10 illustrates a confusion matrix for a two-class problem with positive and negative class values.

Table 10 - Confusion matrix for a two-class problem (adapted by (KUHN; JOHNSON, 2013)).

		True Value	
		Positive	**Negative**
Predictive Value	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

From this matrix, we extract some metrics widely used for measuring the performance of learning systems, such as Sensitivity (eq. (8)), Specificity (eq. (9)), PPV (eq. (10)), NPV (eq. (11)), and MCC (eq.(12)). These equations can be seen below.

$$\text{Sensitivity} = \text{Recall} = \frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Negative}} \quad (8)$$

$$\text{Specificity} = \frac{\sum \text{True Negative}}{\sum \text{False Positive} + \sum \text{True Negative}} \quad (9)$$

$$\text{Positive Predictive Value(PPV)} = \frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Positive}} \quad (10)$$

$$\text{Negative Predictive Value (NPV)} = \frac{\sum \text{True Negative}}{\sum \text{True Negative} + \sum \text{False Negative}} \quad (11)$$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (12)$$

MCC metric is a balanced measure among the TP, TN, FP, and FN and can be useful even if the classes are of quite different sizes.

One common problem with these evaluation metrics presented above (eq.8-12) is that they are only dependent on the choice of the TP, FP, FN, and TN, which are based on a preset score threshold (CHEN; WASIKOWSKI, 2008; RABHI; JAKUBOWICZ; METZGER, 2019). One single limit pattern cannot tell us which parameters and feature set is better.

In general, the greater the specificity of a test, the lower its sensitivity (or vice versa); therefore, we must consider the thresholds not as single values but rather curves. The analysis of the Area Under the ROC Curve (AUC) addresses this point, and it can be used to better evaluate the model behavior. It is a graphic of the trade-off between the TP rate (sensitivity) and the FP rate (1-specificity), where each threshold value produces a different point in the ROC space. Because the method scans over all possible thresholds, it is independent of a specific cut-off value (VANHOEYVELD; MARTENS, 2018). For a well-performing classifier, the ROC curve needs to be as near to the top-left corner as possible.

Following the same idea, we have the Precision-Recall AUC (prAUC) metric. It also does not depend on the threshold value, but it analyzes the trade-off between sensitivity and PPV. This metric is chosen over AUC when the model has not priority the true negatives.

Ferri et al. (2009) analyzed different qualitative and quantitative performance metrics' behavior, finding that AUC measures perform relatively well and are preferable.

We use AUC values as the primary methods to assess each experiment's hyperparameters and compare cross-validation results. However, when necessary to discriminate the value (Chapter 4), a cut-off threshold must be decided. Choosing cut-off points arbitrarily or using non-optimal criteria can lead to unnecessary misclassification. Thus, we explore an optimal cut-off value based on the Youden index statistics given by the ROC curve.

However, since the binormal assumption does not hold, i.e., the two distributions' variances are not equal, and error costs are not the same, the classifier's best choice does not only be established by the ROC curve. We need to consider additional information to choose the optimal threshold, such as the relative cost of false-negative classification (compared with a false positive classification) and the proportion of positive cases in the dataset. The best cutoff value for the final model's prediction was determined based on the Youden index statistics by the training set, considering a weight two times higher for false-negative records, since missing an infected patient is worse than screening a healthy patient.

Youden's J statistic chooses the optimal cut-off that maximizes the distance to the diagonal line, that is, maximize the sum of sensitivity and specificity (YOUDEN, 1950). Including the relative cost of a False Negative classification, the optimality criteria are modified (PERKINS; SCHISTERMAN, 2006) in eq. (13).

$$\max(\text{sensitivities} + r * \text{specificities}) \tag{13}$$

with $r = \frac{1-\frac{\text{n.cases}}{\text{n.controls}+\text{n.cases}}}{\text{cost}*\frac{\text{n.cases}}{\text{n.controls}+\text{n.cases}}}$.

The cut-off value is the limit for deciding whether the patient should be tested or not. Since our classification goal is to minimize the number of false negatives

and maximize the number of true negatives, we compare the models by the NPV. We also evaluate the MCC metric, a balanced measure among TP, TN, FP, and FN.

3.3.8.2. Prediction

Classification is best used with deterministic outcomes that occur frequently, and not when two individuals with identical inputs can easily have different results due to choosing thresholds. For the latter, modeling probabilities is the best way. Instead of predicting class values directly for a classification problem, it can be convenient to predict the likelihood of an observation belonging to each possible class (HARRELL, 2017).

A proper accuracy scoring rule must assess the predicted probabilities. These measures are useful when evaluating whether models have good predictions, not only if they failed the classification (FERRI; HERNÁNDEZ-ORALLO; MODROIU, 2009). The two metrics most commonly used are the Brier score (BS) or the logarithmic scoring rule (HARRELL, 2017). We apply the Brier score since it has a more straightforward interpretation and can be used for binary outcomes.

The Brier score, eq. (14), is the squared residuals (i.e., quadratic error measure). It can be considered as a measure of the "calibration" of the probabilistic predictions. Therefore, the lower the Brier score, the better the predictions are calibrated. It takes on a value between zero and one (ROULSTON, 2007).

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{14}$$

, where N is the population, f_t is the forecast probability, and o_t is the observed/real outcome (1 or 0).

Unfortunately, no measure simultaneously combines the threshold and the estimated probability (FERRI; HERNÁNDEZ-ORALLO; MODROIU, 2009).

In short, besides classifying patients in positive or negative into the screening test, we also aim to provide a probabilistic prediction for each patient through an acquisition risk model (Chapter 5) that could be used as input to a formal decision independent of any possible error costs (SPIEGELHALTER, 1986). For example,

if the probability of CR-GNB acquisition is 5% by the predictive model, then inappropriate antibiotic treatment can be avoided since patients at low risk of multidrug resistance. If, in this case, the real value is “negative,” the likelihood of an error is 0.05. On the other hand, a probability of 60% may lead the physician to start adequate therapy as early as possible, or infection control policies can be established to control these bacteria's spread. So, these probabilities can aid decision making. We assessed the prediction model performances by Brier score and interpreted the result by the calibration curves.

3.3.8.2.1.
Calibration

The calibration measures the degree of consistency between observed outcome and estimated outcome (LIN; HU; KONG, 2019). The aim is to show if the models overestimated or underestimated the patient's colonization, validating the model's reliability.

The calibration curves building happens as follows. Firstly, the Hosmer-Lemeshow test is applied, which divides data into N groups, defining the number of observed and expected events for each group (Y = 1/positive and Y = 0/negative) (HOSMER; LEMESHOW, 2000). These data are then fitted to a linear model, assuming a polynomial relationship that defines the calibration curve. After that, a likelihood ratio test is used to generate a confidence region around the calibration curve, known as the calibration belt. This test also gives us a p-value, indicating if the model is miscalibrated or not (NATTINO; FINAZZI; BERTOLINI, 2016). For our study, a model is well-calibrated if the predicted probabilities accurately match the response's observed proportions considering a confidence level of 0.05.

In this study, we used the package “givitiR” in R. It assesses the calibration of binary outcome models with the GiViTI (*Gruppo Italiano per la valutazione degli interventi in Terapia Intensiva*, Italian Group for the Evaluation of the Interventions in Intensive Care Units) calibration belt.

3.3.9.
Statistical comparison of classifiers

We used Friedman's test (FRIEDMAN, 1940) to compare the different classifiers. The Friedman test statistic is based on the Average Ranked (AR) performances of the classification techniques on each dataset/strategy, calculated by eq. (15) (BROWN; MUES, 2012):

$$\chi_F^2 = \frac{12D}{K(K+1)}\left[\sum_{j=1}^{K} AR_j^2 - \frac{K(K+1)^2}{4}\right], where\ AR_j = \frac{1}{D}\sum_{i=1}^{D} r_i^j \quad (15)$$

D denotes the number of strategies used in our study; K is the total number of classifiers and r_i^j is the rank of classifier j on data set i. χ_F^2 is distributed according to the Chi-square distribution with K-1 degrees of freedom. If the value of χ_F^2 is large, then the null hypothesis (Ho: there no difference between the techniques) can be rejected. The Friedman statistic is less susceptible to outliers (FRIEDMAN, 1940).

The Nemenyi test (NEMENYI, 1963) is applied to report any significant differences between individual classifiers. This post hoc test affirms that two or more classifiers' performances are significantly different if their average ranks differ by at least the Critical Difference (CD), given by eq. (16).

$$CD = q_{\alpha,\infty,K}\sqrt{\frac{K(K+1)}{12D}} \quad (16)$$

In this formula, the value $q_{\alpha,\infty,K}$ is based on the studentized range statistic (NEMENYI, 1963).

According to the Nemenyi posthoc test (NEMENYI, 1963) for multiple joint samples, a strategy or method is "highly" different from another when $p < 0.01$ and differs significantly when $p < 0.05$. In this study, a p-value bigger than 0.05 is not significant. Some studies also use the Nemenyi test and Friedman's test to compare classifiers (BROWN; MUES, 2012; DEMSAR; DEMSAR, 2006).

In addition to these tests, we used descriptive statistics to compare and discuss all possible combinations between the balancing strategies and machine learning techniques. We compare the models by NPV and MCC for discrimination and by Brier score for prediction.

3.3.10.
Running

The experiments were performed on an Intel® Core ™ i7 processor with 16GB of RAM and R 4.0.2 software. We used the CARET framework (correspondent algorithms cited in

Table 7), imbalanced-learn packages, and others. We have adapted the functions of balancing strategies in CARET. Before all the techniques were run, we preprocessed the data according to the specific method.

The R Statistical Programming Language is an open-source tool for statistics and programming, computationally efficient and understandable without specialized computer science training. The Machine Learning and Statistical Learning task view list almost 100 packages dedicated to ML (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019).

We provide our code written in R Statistical Programming Environment in GitHub by the link "https://github.com/leiladantas/PredictionMDR," easily applied to other classification problems. This acts as a framework upon which researchers can develop their ML studies. The code and models may be fitted to diverse types of data. Figure 5 summarizes the flow of the model building and evaluating process.

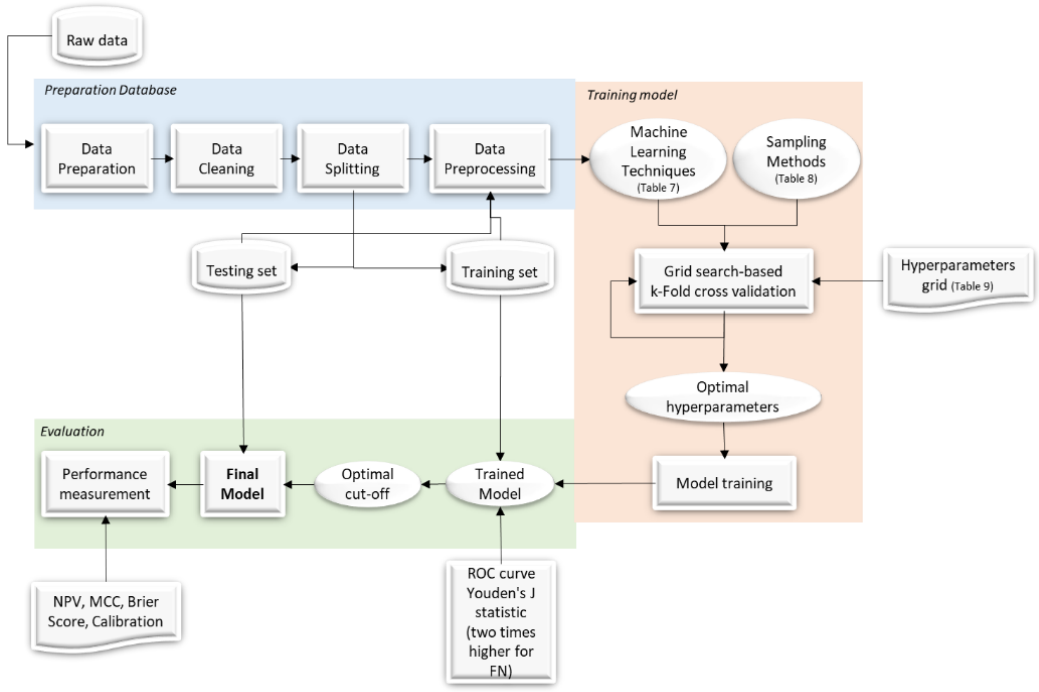


Figure 5 - Model building and evaluating process.

3.4.
Important Factors

In addition to the prediction model, we are interested in finding the explanatory variables expressing what patients and clinical characteristics are associated with the acquisition of CR-GNB. Thus, we discuss the most significant predictors based on Information Gain (IG). It looks at each feature in isolation, computes its information gain, and measures how important and relevant it is to the class label (ALHAJ et al., 2016). The variable that maximizes the information gain minimizes the entropy (eq (6)) and best splits the dataset into groups for correct classification (BROWNLEE, 2019b).

3.5.
Association Rules Mining

Association rules aim to find frequent itemsets from a transaction dataset and derive association rules (WU et al., 2009), i.e., identify the items that often occur together. We use Association Rule Mining (ARM) to automatically detect what interesting patterns and if-then rules could be found in the binarized data (SAARELA; RYYNÄNEN; ÄYRÄMÖ, 2019). We aim to find rules of strongly associated features in our data that indicate patterns that can better help clinicians in the decision-making process.

We used the Apriori algorithm, proposed for frequent itemset mining (AGRAWAL et al., 1994). The steps followed in this algorithm are: Join, where it generates (k+1) candidate itemsets from k-itemsets by joining each item with itself, and Prune, checking if each of the candidate itemsets meets minimum support.

Before applying association rules, we must discretize all the datasets, converting numeric vectors into factors with categories having approximately the same number of data points (based on a training set). Moreover, it is necessary to convert to transactions for creating items.

The metrics that we used for finding frequent itemset were support, confidence, and lift. Support is the transaction percentage in the item set that occurs {Sup(A→B)=Sup(AUB)}. Once the itemset is obtained, we generate association rules with minimum confidence. The confidence denotes the proportion of data items containing B in all items containing A {Conf(A→B)=Sup(AUB)/Sup(A)}. Its value indicates how reliable this rule is (WU et al., 2009). The lift is the ratio of the confidence of the rule and the expected confidence of the rule. It refers to how A increases the frequency of B {Lift(A→B)=Conf(A→B)/Sup(B)}.

Our main goal is to find what factors combined influenced the acquisition of CR-GNB: strongly associated features in our data indicate that a patient is at risk of acquiring these pathogens. In this case, we used a classification approach based on ARM, where the class positive is considered in RHS (Right-Hand Side). The higher the lift value, the better the rule.

3.6. Differences between the models

Table 11 summarizes the difference between the screening and acquisition risk models according to the evaluation metrics used, study population, and objectives. They are presented and discussed in the following chapters.

Table 11 - Difference between the screening and acquisition risk models.

Model	Screening	Acquisition risk
Type	Discrimination/Classification	Prediction
Study Population	All Screenings Tests	Screening Tests and Clinical Exams
Unit of Analysis	Test	Patient
Main Objective	To detect those who do NOT need testing	To find the probability of each patient to acquire the bacteria
Sampling method	Different Balancing Strategies	Matched Case-control Study
Hyperparameter Tuning Metric	AUC	Brier score
Evaluation Metric	MCC and NPV	Brier score
Interpretation	Error analysis (confusion matrix)	Calibration Belt
Comparison of the techniques' performances	Yes	No
Computational Time Analysis	Yes	No
Analysis of the difference between hospitals	No	Yes
Importance Factors	No	Yes
Association Rules Mining	No	Yes

4
Screening Model

Chapter 4 presents the screening model's development to detect ICU patients who do not need to be tested, following the methodology presented in Chapter 3. We evaluate different machine learning techniques, balancing strategies, and feature selection techniques.

4.1.
Setting and study population

Our database gathers different types of data collected from five Brazilian hospitals. However, when analyzing the microbiology data, we noticed that the hospital E (Table 5) did not follow the protocol and did not weekly culture tests for ICUs. The data from these hospitals are not included in this analysis.

A cohort design compared two groups of screening test results. The Positive Group includes the positive tests, i.e., tests that detected Carbapenem-Resistant Gram-Negative Bacteria identified by culture tests in ICUs after 48 hours of hospital admission. The Negative Group consists of the negative tests, i.e., tests that had no detection of CR-GNB in that exam.

It should be noted that all the screening tests per patient for CR-GNB were taken into account for this study; that is, if an inpatient had five negative surveillance culture tests on different dates, the five tests would be considered. However, only the first episode of bacterial isolation was evaluated for each subject, i.e., we did not include any test made after a positive culture. Figure 6 illustrates this approach. The unit of analysis is the test, not the patient.

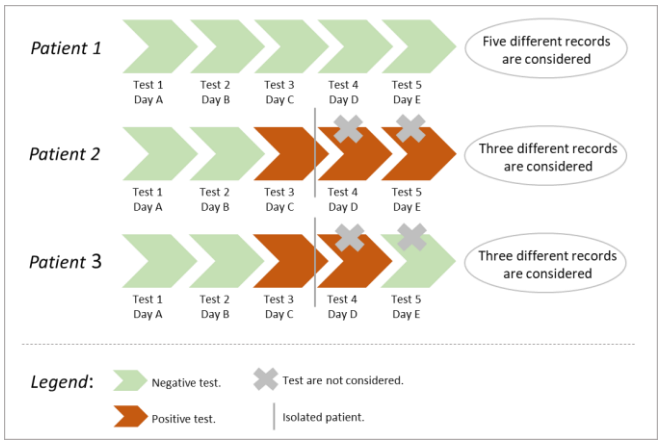


Figure 6 - Illustration of how the screening tests were selected. We considered only the first episode of carbapenem-resistant Gram-negative bacterial isolation for each patient.

The screening involved rectal, nasal, and pharyngeal swabs. We consider the patient positive for colonization if at least one swab indicates Carbapenem-Resistant Gram-negative bacteria (*A. baumannii, P. aeruginosa* Enterobacteriaceae). The culture result is negative if all specimens are negative for detecting CR-GNB. Other clinical exams, which doctors have ordered for specific clinical reasons, are not part of this chapter and will be included only in Chapter 5.

The data selection included all tests matching inclusion criteria during the study period: screening culture made in adult ICUs; testing in patients with admission date after May 8th, 2017; patients aged ≥18 years old; tests realized between 48h and 60days after patient admission. 48h criterium was selected because it represents the local hospitals' definition for community-acquired colonization.

After applying the inclusion criteria, as shown in Appendix D, we have a total of 394 positive screening cultures and 3,517 negative cultures between May 8th, 2017 and August 31st, 2019, resulting in 2,306 patients with at least one screening culture – a minimum of one and maximum of 9 screenings per patient during 60 days. There were 2,097 patients with only negative cultures and 209 with only positive cultures for CR-GNB during their hospitalizations. Thus, the number of medical records analyzed was 3911, unequally distributed concerning the presence/absence of Carbapenem-Resistant Gram-negative bacteria. All information mentioned above can also be found in Table 12 for each hospital. It was considered only the first episode of isolation for each subject (see Figure 6).

Table 12 - The number of patients and culture tests in each hospital.

Hospital	#Screening Tests	#Positive Tests	#Negative Tests	% Positive Tests	#Patients with at least one screening	#Patients with negative cultures	#Patients with only positive cultures	#Maximum tests by a patient
A	310	60	250	19.4%	206	175	31	7
B	806	57	749	7.1%	445	420	25	9
C	1081	81	1000	7.5%	568	533	35	9
D	1714	196	1518	11.4%	1087	969	118	8
All	3911	394	3517	11.3%	2306	2097	209	-

We have about 11% of culture-positive tests, ranging from 7.1% to 19.4%, depending on the hospital. Although each hospital has particularities, our goal (aligned with the physicians) was to develop a screening model applied to any of the network's four hospitals. Thus, we decided to use the "Hospital" to consider possible differences in baseline risk by each one on the main outcome.

4.2.
Conducting a machine learning analysis

We initialized our analysis by predicting CR-GNB non-acquisition using supervised learning techniques, considering the negative test's reference level. As early pointed out, we are especially interested in some classifiers' performance that reliably detects ICU patients who do not need to be tested.

Since the dataset is imbalanced, containing a smaller number of observations in positive antibiotic-resistant tests, we combined the different supervised techniques with balancing strategies to reduce this problem, pre-processing our data according to each method.

Our original dataset includes the patient, ICU, and hospital information, indexes (such as SAPS3 and Charlson), presence of comorbidities, use of the invasive devices during hospitalization, reasons for ICU admission, antibiotic use, and laboratory test results (see Appendix C). The variables used in this work are described in detail in Table 6, Chapter 3. It includes 112 independent variables and one dependent variable.

4.2.1.

Visualization and Data Cleaning

The statistical information of CR-GNB (positive-culture) and non-CR-GNB (negative-culture) groups for all the 112 features are presented in Appendix E. Table 13 summarizes the 57 significant variables from univariate analysis for a Confidence Interval (CI) of 90% ($p \leq 0.10$).

Table 13 - Descriptive statistical analysis comparing the negative and positive culture tests.

Variables	Negative-culture tests (n=3,517)	Positive-culture tests (N=394)	p-value
Laboratory tests			
tests_before			
Mean (SD)	1.38 (1.47)	1.49 (1.41)	0.032
Median [Min, Max]	1.00 [0, 10.0]	1.00 [0, 7.00]	
Hospital Information			
Hospital			
A	250 (7.1%)	60 (15.2%)	<0.001
B	749 (21.3%)	57 (14.5%)	
C	1,000 (28.4%)	81 (20.6%)	
D	1,518 (43.2%)	196 (49.7%)	
LOS_hospital_before_test			
Mean (SD)	14.8 (12.3)	19.2 (13.7)	<0.001
Median [Min, Max]	10.0 [3.00, 60.0]	15.0 [3.00, 60.0]	
ICU Information			
LOS_ICU_before_test			
Mean (SD)	13.0 (11.9)	16.4 (12.7)	<0.001
Median [Min, Max]	9.00 [0, 60.0]	13.0 [0, 60.0]	
Index			
CharlsonIndex			
Mean (SD)	1.77 (1.96)	2.02 (2.06)	0.007
Median [Min, Max]	1.00 [0, 12.0]	2.00 [0, 12.0]	
Missing	2 (0.1%)	0 (0%)	
FrailPatientMFI			
NO	2,876 (81.8%)	308 (78.2%)	0.094
YES	641 (18.2%)	86 (21.8%)	
Saps3Points			
Mean (SD)	52.8 (12.9)	57.0 (13.8)	<0.001
Median [Min, Max]	52.0 [8.00, 104]	56.0 [19.0, 104]	
SofaScore			
Mean (SD)	1.75 (2.91)	2.97 (3.81)	<0.001
Median [Min, Max]	1.00 [0, 17.0]	1.00 [0, 17.0]	
Missing	1,108 (31.5%)	124 (31.5%)	
Priority			
Priority 1	419 (11.9%)	81 (20.6%)	0.001
Priority 2	1,073 (30.5%)	109 (27.7%)	
Priority 3	2 (0.1%)	0 (0%)	
Priority 4	4 (0.1%)	0 (0%)	
Priority 5	12 (0.3%)	1 (0.3%)	
Missing	2,007 (57.1%)	203 (51.5%)	
Comorbidities			
ChronicHealthStatus			
Independent	1,872 (53.2%)	179 (45.4%)	0.019
Need for assistance	812 (23.1%)	106 (26.9%)	
Restricted / bedridden	824 (23.4%)	105 (26.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsSevereCopd			
FALSE	3,122 (88.8%)	335 (85.0%)	0.08
TRUE	386 (11.0%)	55 (14.0%)	
Missing	9 (0.3%)	4 (1.0%)	
IsAsthma			
FALSE	3,402 (96.7%)	371 (94.2%)	0.069
TRUE	106 (3.0%)	19 (4.8%)	
Missing	9 (0.3%)	4 (1.0%)	
IsAngina			
FALSE	3,269 (92.9%)	376 (95.4%)	0.019
TRUE	239 (6.8%)	14 (3.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsDeepVenousThrombosis			
FALSE	3,346 (95.1%)	359 (91.1%)	0.006
TRUE	162 (4.6%)	31 (7.9%)	
Missing	9 (0.3%)	4 (1.0%)	
IsStrokeSequelae			
FALSE	3,374 (95.9%)	357 (90.6%)	<0.001
TRUE	134 (3.8%)	33 (8.4%)	

Variables	Negative-culture tests (n=3,517)	Positive-culture tests (N=394)	p-value
Missing	9 (0.3%)	4 (1.0%)	
IsChemotherapy			
FALSE	3,367 (95.7%)	366 (92.9%)	0.064
TRUE	141 (4.0%)	24 (6.1%)	
Missing	9 (0.3%)	4 (1.0%)	
IsHistoryOfPneumonia			
FALSE	3,317 (94.3%)	360 (91.4%)	0.088
TRUE	191 (5.4%)	30 (7.6%)	
Missing	9 (0.3%)	4 (1.0%)	
Invasive Device during Hospitalization			
VesDURTOTAL			
Mean (SD)	6.72 (8.90)	10.9 (9.79)	<0.001
Median [Min, Max]	3.00 [0, 57.0]	9.00 [0, 52.0]	
VesDURMORE			
Mean (SD)	1.61 (3.41)	2.67 (3.98)	<0.001
Median [Min, Max]	0 [0, 56.0]	0 [0, 24.0]	
VesTIMESTOTAL			
Mean (SD)	0.928 (0.940)	1.30 (0.900)	<0.001
Median [Min, Max]	1.00 [0, 7.00]	1.00 [0, 5.00]	
VesTIMESMORE			
Mean (SD)	0.0836 (0.311)	0.124 (0.345)	0.003
Median [Min, Max]	0 [0, 5.00]	0 [0, 2.00]	
VESICAL			
NO	1,231 (35.0%)	59 (15.0%)	<0.001
YES	2,286 (65.0%)	335 (85.0%)	
ArtDURTOTAL			
Mean (SD)	3.88 (6.75)	7.83 (9.25)	<0.001
Median [Min, Max]	0 [0, 59.0]	5.00 [0, 53.0]	
ArtDURMORE			
Mean (SD)	0.803 (2.40)	1.88 (3.66)	<0.001
Median [Min, Max]	0 [0, 39.0]	0 [0, 22.0]	
ArtTIMESTOTAL			
Mean (SD)	0.606 (0.868)	1.08 (1.06)	<0.001
Median [Min, Max]	0 [0, 6.00]	1.00 [0, 5.00]	
ArtTIMESMORE			
Mean (SD)	0.0427 (0.227)	0.109 (0.336)	<0.001
Median [Min, Max]	0 [0, 3.00]	0 [0, 2.00]	
ARTERIAL			
NO	2,073 (58.9%)	143 (36.3%)	<0.001
YES	1,444 (41.1%)	251 (63.7%)	
DiaDURTOTAL			
Mean (SD)	1.07 (4.31)	2.84 (6.91)	<0.001
Median [Min, Max]	0 [0, 41.0]	0 [0, 42.0]	
DiaDURMORE			
Mean (SD)	0.273 (1.54)	0.665 (2.48)	<0.001
Median [Min, Max]	0 [0, 31.0]	0 [0, 21.0]	
DiaTIMESTOTAL			
Mean (SD)	0.150 (0.542)	0.378 (0.827)	<0.001
Median [Min, Max]	0 [0, 5.00]	0 [0, 6.00]	
DiaTIMESMORE			
Mean (SD)	0.0199 (0.159)	0.0431 (0.203)	<0.001
Median [Min, Max]	0 [0, 2.00]	0 [0, 1.00]	
DIALYSIS			
NO	3,187 (90.6%)	304 (77.2%)	<0.001
YES	330 (9.4%)	90 (22.8%)	
CVCDURTOTAL			
Mean (SD)	6.35 (8.73)	11.1 (10.2)	<0.001
Median [Min, Max]	3.00 [0, 60.0]	9.00 [0, 51.0]	
CVCDURMORE			
Mean (SD)	1.47 (3.34)	2.95 (4.59)	<0.001
Median [Min, Max]	0 [0, 56.0]	0 [0, 32.0]	
CVCTIMESTOTAL			
Mean (SD)	0.849 (0.975)	1.38 (1.09)	<0.001
Median [Min, Max]	1.00 [0, 6.00]	1.00 [0, 6.00]	
CVCTIMESMORE			
Mean (SD)	0.0893 (0.320)	0.193 (0.455)	<0.001
Median [Min, Max]	0 [0, 5.00]	0 [0, 3.00]	
CVC			
NO	1,560 (44.4%)	81 (20.6%)	<0.001
YES	1,957 (55.6%)	313 (79.4%)	
MVDURTOTAL			
Mean (SD)	4.19 (8.75)	8.51 (11.0)	<0.001
Median [Min, Max]	0 [0, 57.0]	5.00 [0, 49.0]	
MVDURMORE			
Mean (SD)	0.978 (2.79)	2.35 (4.61)	<0.001
Median [Min, Max]	0 [0, 56.0]	0 [0, 33.0]	
MVTIMESTOTAL			
Mean (SD)	0.400 (0.649)	0.766 (0.782)	<0.001
Median [Min, Max]	0 [0, 4.00]	1.00 [0, 5.00]	
MVTIMESMORE			
Mean (SD)	0.0205 (0.151)	0.0381 (0.192)	0.013
Median [Min, Max]	0 [0, 2.00]	0 [0, 1.00]	
MV			
NO	2,379 (67.6%)	159 (40.4%)	<0.001

Variables	Negative-culture tests (n=3,517)	Positive-culture tests (N=394)	p-value
YES	1,138 (32.4%)	235 (59.6%)	
Reasons for ICU admission			
AdmissionSource			
Emergency	2,020 (57.4%)	188 (47.7%)	<0.001
Hemodynamic Room	55 (1.6%)	3 (0.8%)	
Operation Room	364 (10.3%)	45 (11.4%)	
Other ICU from hospital	419 (11.9%)	71 (18.0%)	
Others	24 (0.7%)	7 (1.8%)	
Semi Intensive Unit	201 (5.7%)	28 (7.1%)	
Transfer from another hospital	33 (0.9%)	10 (2.5%)	
Ward/Room	392 (11.1%)	38 (9.6%)	
Missing	9 (0.3%)	4 (1.0%)	
AdmissionReason			
Cardiovascular / Shock	846 (24.1%)	48 (12.2%)	<0.001
Elective Surgery	253 (7.2%)	29 (7.4%)	
Emergency surgery	182 (5.2%)	18 (4.6%)	
Endocrine / Metabolic / Renal	85 (2.4%)	10 (2.5%)	
Infection / Sepsis	1,200 (34.1%)	170 (43.1%)	
Liver and Pancreas / Gastrointestinal	193 (5.5%)	16 (4.1%)	
Neurological	303 (8.6%)	43 (10.9%)	
Non-surgical trauma	80 (2.3%)	10 (2.5%)	
Oncological / Hematological	67 (1.9%)	8 (2.0%)	
Others	60 (1.7%)	8 (2.0%)	
Respiratory	239 (6.8%)	30 (7.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsNeurologicalComaStuporObtundedDelirium			
FALSE	2,968 (84.4%)	301 (76.4%)	<0.001
TRUE	540 (15.4%)	89 (22.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsNeurologicalSeizures			
FALSE	3,350 (95.3%)	364 (92.4%)	0.074
TRUE	158 (4.5%)	26 (6.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsNeurologicalFocalNeurologicDeficit			
FALSE	3,435 (97.7%)	373 (94.7%)	0.008
TRUE	73 (2.1%)	17 (4.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCardiovascularHypovolemicHemorrhagicShock			
FALSE	3,470 (98.7%)	381 (96.7%)	0.063
TRUE	38 (1.1%)	9 (2.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCardiovascularSepticShock			
FALSE	3,335 (94.8%)	344 (87.3%)	<0.001
TRUE	173 (4.9%)	46 (11.7%)	
Missing	9 (0.3%)	4 (1.0%)	
Antibiotic use			
J01A			
FALSE	3,406 (96.8%)	350 (88.8%)	<0.001
TRUE	111 (3.2%)	44 (11.2%)	
J01C			
FALSE	1,337 (38.0%)	114 (28.9%)	<0.001
TRUE	2,180 (62.0%)	280 (71.1%)	
J01D			
FALSE	1,625 (46.2%)	88 (22.3%)	<0.001
TRUE	1,892 (53.8%)	306 (77.7%)	
J01E			
FALSE	3,395 (96.5%)	368 (93.4%)	0.003
TRUE	122 (3.5%)	26 (6.6%)	
J01F			
FALSE	2,378 (67.6%)	233 (59.1%)	<0.001
TRUE	1,139 (32.4%)	161 (40.9%)	
J01G			
FALSE	3,300 (93.8%)	337 (85.5%)	<0.001
TRUE	217 (6.2%)	57 (14.5%)	
J01X			
FALSE	2,442 (69.4%)	163 (41.4%)	<0.001
TRUE	1,075 (30.6%)	231 (58.6%)	
Antibiotic			
FALSE	495 (14.1%)	10 (2.5%)	<0.001
TRUE	3,022 (85.9%)	384 (97.5%)	

We can see in Table 13 that the patients with a high length of stay in hospital or ICU are more likely to be colonized. The positive test group had higher severity indices, such as the Charlson Comorbidity Index, Saps 3 Points, and Sofa Score. The age of patients ranges from 18 to 105 years old.

The colonized patients by CR-GNB received more antibiotics and used more invasive devices than those who did not acquire these pathogens. The antibiotics use of codes J01A, J01C, J01D, J01E, J01F, J01G, and J01X between 24 hours to 30 days before the test increase significantly the likelihood of a positive result.

The invasive device's use duration is quite different between the groups. A more prolonged use time of mechanical ventilation, arterial, vesical, central venous, and hemodialysis catheters increases the probability of acquiring the pathogen. The length of time that a procedure is used between one test and another, the number of times they were changed, and the use between 24h and 15 days before the test are also significant. The peripheral catheter, on the other hand, does not seem to have a significant relationship.

The CR-GNB group has more likely reasons for ICU admission, such as neurological coma stupor obtunded delirium, neurological seizures, focal neurological deficit, cardiovascular hypovolemic hemorrhagic shock, and cardiovascular septic shock. Patients admitted to the ICUs for any of these reasons are more likely to obtain a positive test for the pathogen's acquisition.

The CR-GNB acquisition was higher in patients admitted from sepsis/infection or neurological disease and for those having as admission source the operation room or other ICU from the hospital.

Patients who presented at the admission time some comorbidities such as severe COPD, asthma, deep venous thrombosis, stroke sequela, chemotherapy, and history of pneumonia, had a higher likelihood of obtaining positive tests. On the other hand, patients with angina presented a lower probability of acquisition. There was no significant difference among the groups in gender, age, hospital readmission, most of the comorbidities, and some reasons for admission.

We can see that some variables have a "Missing" category in Appendix E and Table 13. Since some algorithms do not work with missing values, we must rectify these records. Altogether, our dataset has less than 2% of missing data. However, 61% of the variables are incomplete. These missing values occur in 2658 (~67%) of the observations. That means that only for ~33% of the patients, we have values for all variables. On the other hand, many features have only a couple or no missing values. Table 14 shows the frequency of missing values for each variable.

Table 14 - Number and percentage of missing values for each category and variable.

Variables with missing	Negative (N=3,517)	Positive (N=394)	Overall (N=3,911)
BMI	901 (25.6%)	84 (21.3%)	985 (25.2%)
AdmissionSource	9 (0.3%)	4 (1.0%)	13 (0.3%)
Admission Reason	9 (0.3%)	4 (1.0%)	13 (0.3%)
CharlsonIndex	2 (0.1%)	0 (0%)	2 (0.1%)
MFIpoints	60 (1.7%)	14 (3.6%)	74 (1.9%)
SofaScore	1,108 (31.5%)	124 (31.5%)	1,232 (31.5%)
Priority type	2,007 (57.1%)	203 (51.5%)	2,210 (56.5%)
***Each* comorbidities**	9 (0.3%)	4 (1.0%)	13 (0.3%)
***Each* admission reasons in ICUs**	9 (0.3%)	4 (1.0%)	13 (0.3%)

Since we aimed to exclude variables that had more than 10% missing values, the variables BMI (25.2%), Sofa Score (31.5%), and Priority type (56.5%) were eliminated in the study, as can be seen in Table 14. After removing these variables, the missing data were reduced to only 3% of the records.

The remaining variables were analyzed to understand the randomness of the missing values. To determine if our data are Missing Completely at Random (MCAR), we used the statistical test Little´s MCAR test, in which the null hypothesis was rejected (p-value <0.001). That is, we cannot simply delete the missing records. After that, we showed that our data could be Missing at Random (MAR) by visualization of the missingness pattern. The missingness pattern is explained in detail in Appendix F. These remaining missing records will go through the imputation process later, replacing the missing data with values.

The next step was to identify outliers and inconsistent data. For this, we used the Overlaid Density Plots and Boxplot, shown in Appendix G. Visually, we can locate many "outliers" values. However, since they were considered legitimate data cases, we decided to keep them.

4.2.2. Data Splitting

After preparation and cleaning the data, we have the final database, including 109 independent variables. The data set has been divided into two parts (training and testing). We trained our model with 80% of the data and tested it with 20% remaining data.

4.2.3.
Data Preprocessing

We apply to preprocess steps, such as dimension reduction, imputation, feature selection, normalization, and variables transformation. The test set must not be influenced by the training set during these steps.

Firstly, we reduce the data dimension by analyzing the correlation between features and zero- and near zero-variance predictors. Both can negatively impact the models.

The results of the zero-variance analysis are presented in Appendix H. The pre-specified thresholds to frequency ratio and individual value percentages were 50 and 5, respectively. According to this analysis, we removed 23 variables (IsChfNyhaClass4, IsCirrhosisChildAB, IsCirrhosisChildC, IsHepaticFailure, IsHematologicalMalignancy, IsAids, IsRheumaticDisease, IsMalnourishment, IsPepticDisease, IsHyperthyroidism, DiaDURTOTAL, DiaTIMESMORE, MVTIMESMORE, PerDURMORE, IsNeurologicalIntracranialMassEffect, IsCardiovascularHypovolemicHemorrhagicShock, IsDigestiveSeverePancreatitis, IsCardiovascularAphylacticMixedUndefinedShock, IsLiverFailure, IsTransplantSolidOrgan, IsCardiacSurgery, IsNeurosurgery, J04A).

Using the Pearson method to calculate the correlation between continuous variables, we found and eliminated four predictors with an association higher than 0.75. The variables removed include CVCDURTOTAL, ArtDURTOTAL, LOS_ICU_before_test, and PerTIMESTOTAL. Appendix H visually shows the correlation matrix, where deep colors highlight greater values.

For categorical variables, we used the Goodman and Kruskal's tau (or lambda) measure to indicate the strength of the relationship between the factors. All pairs with an association higher than 0.40 were considered a suggestive association, and then a significance test was performed. Appendix H shows a table with the correlation values for each pair. We affirm that the following variables have a strong association: IsChemotherapy and Immunosuppression; CVC and ARTERIAL; MV and ARTERIAL.

We concluded that everyone who did chemotherapy had immunosuppression, and usually, the patient who uses arterial catheter uses CVC and/or MV. So, we removed "IsChemotherapy" and "ARTERIAL."

We followed with 80 explanatory variables. Of these 80 factors, 42 have some missing value. We attributed the lost values of the 42 variables by imputation, assuming that the missing data are missing randomly (MAR).

We did five times the imputation process, calculating five different datasets. However, since we use different algorithms and need a solid database, we did not consider using multiple imputations (MICE allows this imputation type). Thus, we selected from the five imputed datasets that gave us the smallest deviation rate using a generalized linear model via the lasso penalty. Its metric measures the deviance of the fitted model to a perfect model. The results of the deviation rate can be seen in Table 15.

Table 15 - Deviance rate between the fitted model and the perfect model from each imputed dataset and original dataset using a generalized linear model via the lasso penalty.

Dataset	Deviation rate
Imputation 1	0.0932
Imputation 2	0.0931
Imputation 3	0.0929
Imputation 4	0.0932
Imputation 5	0.0942
Original (removing incomplete cases)	0.0961

The deviation rates were similar, showing us that the five datasets have similar imputed records. Since we must choose a database, we decided to select the new dataset "Imputation 3", which obtained the lowest rate. After that, we evaluated each variable, comparing the original dataset (removing incomplete cases) and the imputed dataset, and we did not find any value outside the range. The average and median were similar.

Once imputed, we reduce the space's dimensionality by removing the irrelevant variables using feature selection methods. We evaluated four different approaches (RF-RFE, SBF, D.SBF, and D.RF) for selecting factors, aiming to choose the best among some classifiers (C4.5, SVM, kNN, and LR). Table 16 shows all methods' performance on each classifier by AUC values and Average Ranked (AR).

Table 16 - Comparison of all methods' performance on each classifier by AUC values, Average Ranked (AR), and the number of variables.

	Mean AUC values				AR	Number of variables
	C45	SVM Radial	KNN	LR		
RF-RFE	0.632	0.690	0.642	0.713	**1.25**	35
SBF	0.624	0.674	0.625	0.709	**2.75**	42
D.SBF	0.568	0.658	0.641	0.702	**3.75**	76
D.RF	0.607	0.687	0.658	0.708	**2.25**	24
Friedman test (p-value)			**0.007**			

Legend: RF-RFE - Recursive Feature Elimination with random forest; SBF - Selection by Filter; D.SBF - Class Decomposition with filter; D.RF - Class Decomposition with random forest; LR – Logistic Regression; kNN – k-Nearest Neighbors; SVM – Support Vector Machine; AR – Average Ranked; AUC – Area Under the Curve.

The Friedman test result (p-value=0.007) indicates that the classifiers' performance is significantly different using distinct feature selection methods. The approach that achieved better AR performance was the RF-RFE, with the following variables: Hospital, J01X, VesDURTOTAL, AdmissionSource, Saps3Points, AdmissionReason, VesTIMESTOTAL, tests_before, MFIpints, Age, PerDURTOTAL, MVDURTOTAL, LOS_hospital_before_test, CVCTIMESTOTAL, J01D, J01G, ChronicHealthStatus, IsNeurologicalComaStuporObtundedDelirium, CVCDURMORE, DiaTIMESTOTAL, VesDURMORE, IsHistoryOfPneumonia, CharlsonIndex, J01C, ArtDURMORE, ArtTIMESTOTAL, IsAlcoholism, DIALYSIS, MVTIMESTOTAL, IsChronicAtrialFibrilation, IsStrokeSequelae, IsDiabetesUncomplicated, PERIPHERAL, IsDementia, and MV. The RF-RFE method obtained the best AUC values for three of the four classifiers analyzed.

Appendix I shows the variables selected by each feature selection method. We can observe similarities between the AUC values and the variables chosen by D.RF and RF-RFE.

We can also conclude that the D.SBF method has the worst performance, including almost all variables in the models (76 out of 80). It further emphasizes the importance of having a selection of factors before model training.

If the goal were to compare the "Sensitivity" metric rather than "AUC" values, our proposed method, "D.RF," would be chosen as the best method. It makes sense since the decomposition approach aims to emphasize the positive classes. However, since we need to be concerned with the true negatives, we used the AUC for comparing.

Before all the algorithms were run, the new database with the 35 selected variables undergoes different normalization processes and transformations depending on the machine learning technique, as seen in

Table 7. At the end of the preprocessing steps, we follow on to the training process. Figure 7 shows the number of variables remaining after each step of data cleaning and pre-processing.

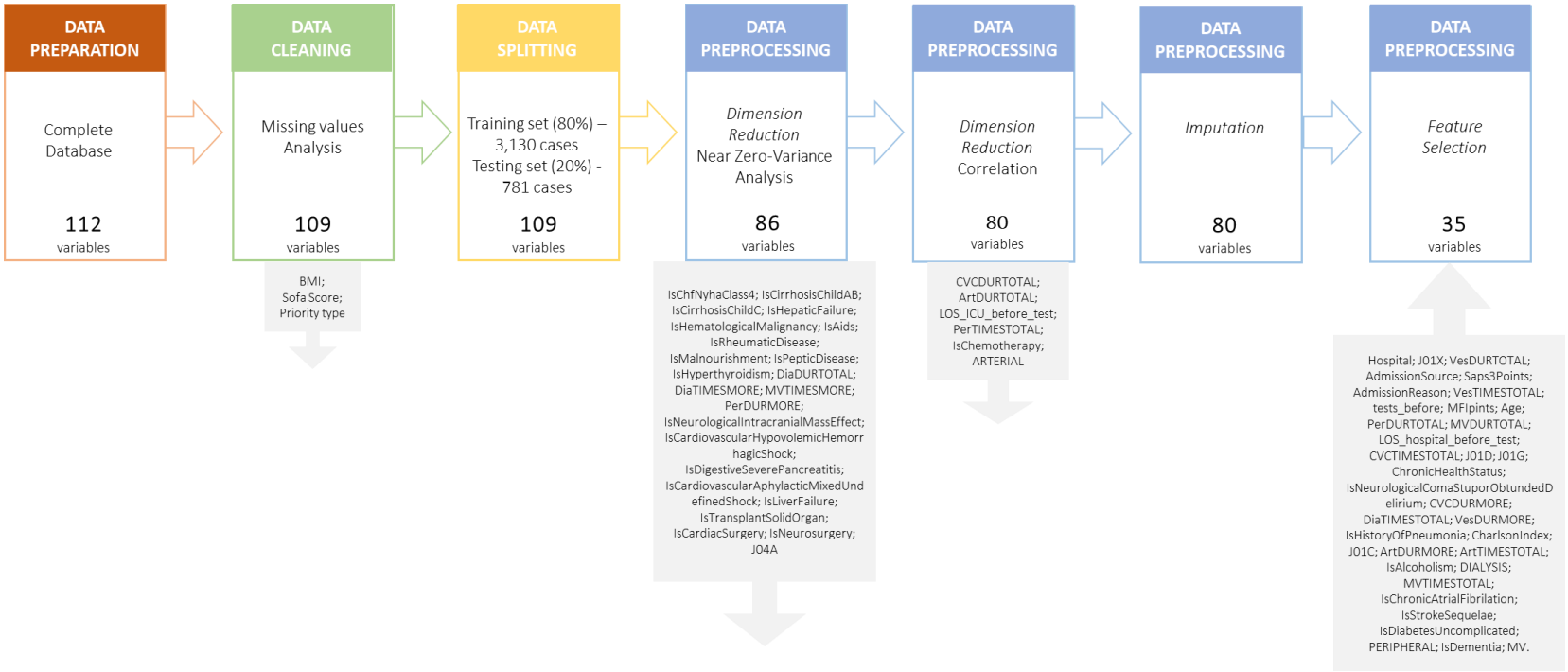


Figure 7 - Exclusion of variables during the process. Of the 112 initials, only 35 remain on the final base.

4.2.4.
Building models - Training

During the training process, we implemented and compared 16 different algorithms, as follows: LR; LR with regularization; LDA; NSC; SVM linear and radial; NN; kNN; NB; decision trees (C4.5, CART, and C50); RF; GBM; Bagging; and AdaBoost.

Besides that, since we consider two-class problems, positive and negative, where the examples from the negative class far outnumber the cases from the positive level, we have to solve the data imbalance and overlap problem. We implemented the balancing strategies discussed in Chapter 3, Table 8: data sampling (random downsampling, random upsampling, SMOTE); data cleaning methods (Tomek links, NCL, and OSS); ensemble-based methods (SMOTEBoost, RUSBoost, SMOTEBagging, and UnderBagging); and data cleaning with sampling (SMOTE+Tomek, SMOTE+NCL, SMOTE + OSS). In short, we performed combinations between the imbalanced and machine learning techniques to find good models.

We started building models without using any balancing strategy. In this work, we call this strategy "none." After that, we applied the sampling strategies (random downsampling, random upsampling, SMOTE) and data cleaning methods (Tomek links, NCL, and OSS). However, since the cleaning strategies alone did not show good results, we decided to add the SMOTE strategy to Tomek, NCL, and OSS (SMOTE+Tomek, SMOTE+NCL, SMOTE + OSS). Finally, we follow by applying ensemble-based methods (SMOTEBoost, RUSBoost, SMOTEBagging, and UnderBagging).

That said, we first show the combination of the first six balancing strategies mentioned above (data sampling and cleaning methods), combining them with all machine learning techniques and resulting in 96 combinations (16x6).

We ran all these combinations over our training set. We used grid-search hyperparameter optimization with 10-fold cross-validation to choose the best performing combination of hyperparameters, avoiding the problem of overfitting. The hyperparameters are discussed in section 3.3.6.1. After selecting the best combinations, the models are refit on the full training data set, building our final models. Balancing strategies are included in the resampling procedure.

We evaluate the performance of all the combinations by parameters PPV, NPV, sensitivity, specificity, and AUC/ROC. PPV has been used to determine the ability of

classifiers to provide CR-GNB positive tests correctly. On the other hand, the NPV measures the correct prediction of non-acquisition. Sensitivity (SENS) finds the proportion of actual positive cases correctly identified by the classifier. Specificity (SPEC) determines the classifier's capability to detect negative instances. The Receiver Operating Characteristic (ROC) curve addresses the trade-off between the sensitivity and the false-positive rate (1-specificity), where each threshold value produces a different point in the space.

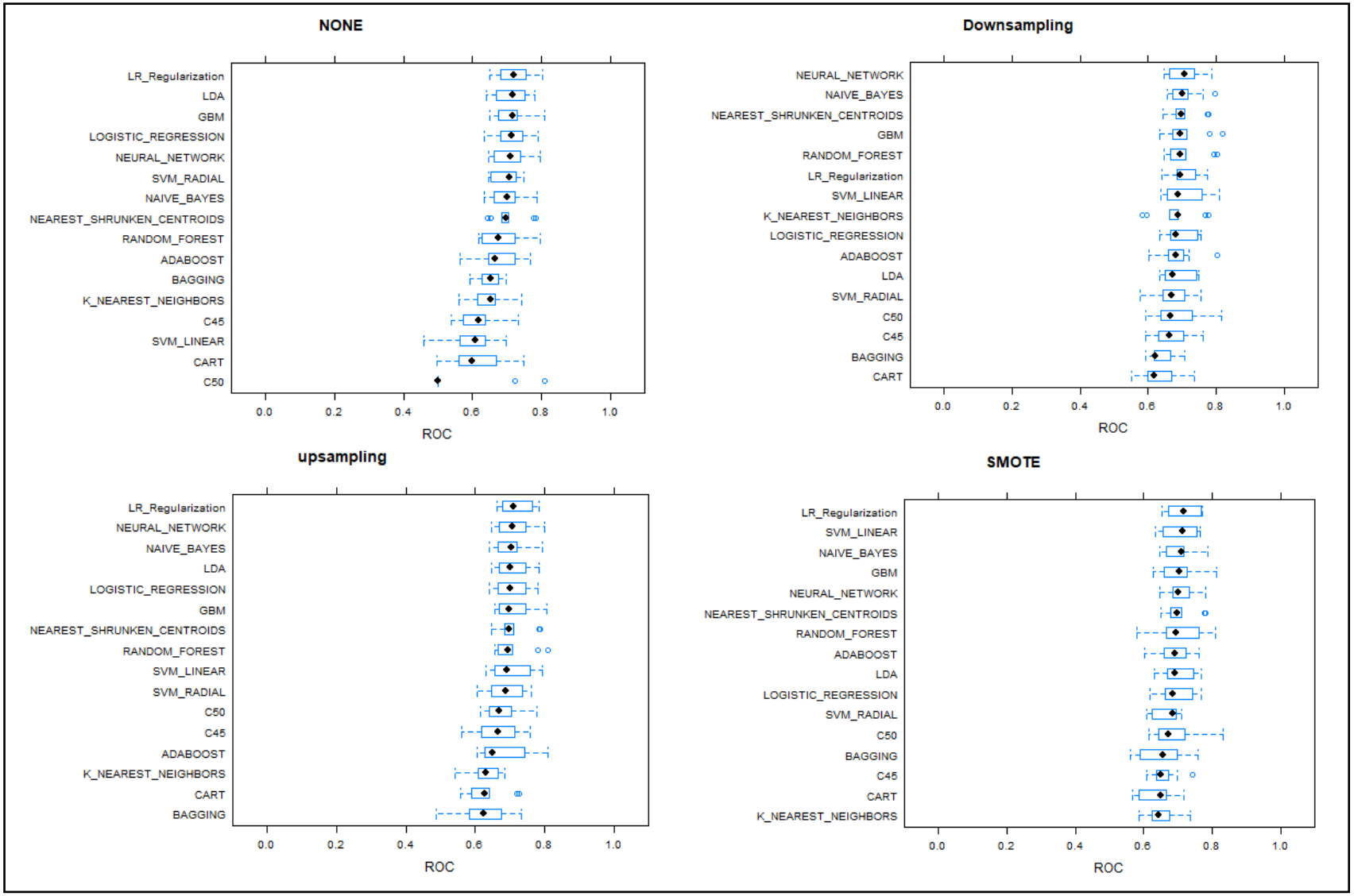
Appendix J shows, for each combination, the best hyperparameters values and the data representation that lead to the best AUC value, since this metric scans over all possible thresholds, and it is independent of a specific cut-off value. Since the focus is to compare the optimized learning algorithms, we include the results related to each metric's best model's ten folds.

For a better interpretation of the data, we compress Appendix J's results, presenting only the cross-validation averages for each combination of strategy and method in Table 17. Figure 8 uses box plots to represent the ROC median, extreme values, and interquartile methods. Boxplots to the other metrics (PPV, NPV, sensitivity, and specificity) can also be seen in Appendix J.

Table 17 - Average of the metric estimates using 10-fold cross-validation for the best hyperparameters based on AUC values.

METHODS	NONE					Downsampling					Upsampling					SMOTE				
	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV
LOGISTIC_REGRESSION	0.71	0.04	0.99	0.42	0.90	0.69	0.64	0.65	0.17	0.94	0.71	0.60	0.70	0.18	0.94	0.69	0.52	0.75	0.19	0.93
LR_Regularization	**0.72**	0.01	1.00	0.44	0.90	**0.71**	0.63	0.68	0.18	0.94	**0.72**	0.61	0.70	0.19	0.94	**0.72**	0.50	0.78	0.21	0.93
LDA	0.71	0.09	0.98	0.38	0.91	0.69	0.63	0.66	0.17	0.94	0.71	0.60	0.70	0.18	0.94	0.70	0.52	0.76	0.19	0.93
NEAREST_SHRUNKEN_CENTROIDS	0.70	0.09	0.97	0.23	0.90	0.70	0.61	0.70	0.19	0.94	0.70	0.61	0.70	0.19	0.94	0.70	0.54	0.75	0.20	0.94
SVM_LINEAR	0.59	0.00	1.00	NA	0.90	**0.71**	0.63	0.68	0.18	0.94	0.71	0.61	0.70	0.19	0.94	0.71	0.50	0.78	0.20	0.93
NEURAL_NETWORK	0.71	0.00	1.00	NA	0.90	**0.71**	0.68	0.65	0.18	0.95	**0.72**	0.63	0.71	0.19	0.94	0.71	0.51	0.77	0.20	0.93
SVM_RADIAL	0.70	0.00	1.00	0.00	0.90	0.67	0.93	0.19	0.11	0.96	0.69	0.00	0.98	0.00	0.90	0.66	0.04	0.94	0.09	0.90
K_NEAREST_NEIGHBORS	0.64	0.01	1.00	0.21	0.90	0.68	0.59	0.70	0.18	0.94	0.63	0.62	0.58	0.14	0.93	0.65	0.53	0.67	0.15	0.93
NAIVE_BAYES	0.70	0.22	0.92	0.23	0.91	**0.71**	0.54	0.74	0.19	0.93	0.71	0.50	0.77	0.20	0.93	0.71	0.55	0.73	0.19	0.93
C45	0.61	0.04	0.99	0.22	0.90	0.67	0.65	0.64	0.17	0.94	0.66	0.54	0.71	0.18	0.93	0.66	0.27	0.90	0.23	0.92
CART	0.61	0.09	0.95	0.16	0.90	0.63	0.59	0.67	0.17	0.94	0.63	0.60	0.65	0.16	0.93	0.64	0.13	0.95	0.27	0.91
C50	0.55	0.02	1.00	0.75	0.90	0.68	0.64	0.65	0.17	0.94	0.68	0.12	0.96	0.27	0.91	0.69	0.25	0.91	0.23	0.91
RANDOM_FOREST	0.68	0.00	1.00	NA	0.90	**0.71**	0.68	0.64	0.17	0.95	0.71	0.64	0.65	0.17	0.94	0.70	0.07	0.96	0.20	0.90
GBM	**0.72**	0.01	1.00	0.30	0.90	**0.71**	0.66	0.65	0.18	0.95	0.71	0.60	0.72	0.19	0.94	0.70	0.24	0.92	0.24	0.92
BAGGING	0.65	0.08	0.98	0.31	0.90	0.64	0.56	0.61	0.14	0.93	0.62	0.10	0.95	0.19	0.90	0.65	0.26	0.88	0.20	0.91
ADABOOST	0.68	0.03	0.98	0.17	0.90	0.69	0.67	0.63	0.17	0.94	0.68	0.03	0.98	0.20	0.90	0.69	0.28	0.89	0.21	0.92

METHODS	Tomek Links					Neighbourhood Cleaning Rule (NCL)					One-Sided Selection (OSS)				
	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV
LOGISTIC_REGRESSION	**0.72**	0.07	0.98	0.34	0.90	**0.72**	0.08	0.98	0.38	0.90	**0.72**	0.07	0.99	0.42	0.90
LR_Regularization	**0.72**	0.02	1.00	0.31	0.90	**0.72**	0.04	0.99	0.26	0.90	**0.72**	0.02	1.00	0.57	0.90
LDA	0.71	0.13	0.97	0.38	0.91	0.71	0.15	0.96	0.31	0.91	0.71	0.12	0.97	0.38	0.91
NEAREST_SHRUNKEN_CENTROIDS	0.70	0.12	0.95	0.23	0.91	0.70	0.17	0.94	0.24	0.91	0.70	0.12	0.96	0.23	0.91
SVM_LINEAR	0.61	0.00	1.00	NA	0.90	0.64	0.00	1.00	NA	0.90	0.58	0.00	1.00	NA	0.90
NEURAL_NETWORK	0.71	0.01	1.00	0.23	0.90	0.71	0.01	1.00	0.28	0.90	**0.72**	0.01	1.00	0.30	0.90
SVM_RADIAL	0.70	0.01	0.97	0.02	0.90	0.68	0.03	0.94	0.06	0.90	0.70	0.01	0.96	0.02	0.90
K_NEAREST_NEIGHBORS	0.65	0.04	0.99	0.28	0.90	0.65	0.08	0.98	0.30	0.90	0.65	0.04	0.99	0.36	0.90
NAIVE_BAYES	0.70	0.00	1.00	NA	0.90	0.70	0.00	1.00	0.00	0.90	0.70	0.00	1.00	1.00	0.90
C45	0.60	0.10	0.96	0.22	0.90	0.61	0.12	0.96	0.24	0.91	0.59	0.11	0.96	0.23	0.91
CART	0.60	0.05	0.98	0.21	0.90	0.62	0.20	0.93	0.22	0.91	0.60	0.05	0.98	0.21	0.90
C50	0.50	0.00	1.00	NA	0.90	0.57	0.01	1.00	0.18	0.90	0.55	0.01	1.00	0.20	0.90
RANDOM_FOREST	0.67	0.00	1.00	NA	0.90	0.68	0.00	1.00	NA	0.90	0.68	0.00	1.00	NA	0.90
GBM	0.71	0.02	1.00	0.39	0.90	0.71	0.04	0.99	0.41	0.90	0.71	0.01	1.00	0.34	0.90
BAGGING	0.67	0.08	0.97	0.21	0.90	0.64	0.08	0.96	0.19	0.90	0.65	0.09	0.97	0.26	0.90
ADABOOST	0.68	0.07	0.97	0.24	0.90	0.68	0.06	0.97	0.18	0.90	0.67	0.04	0.98	0.18	0.90



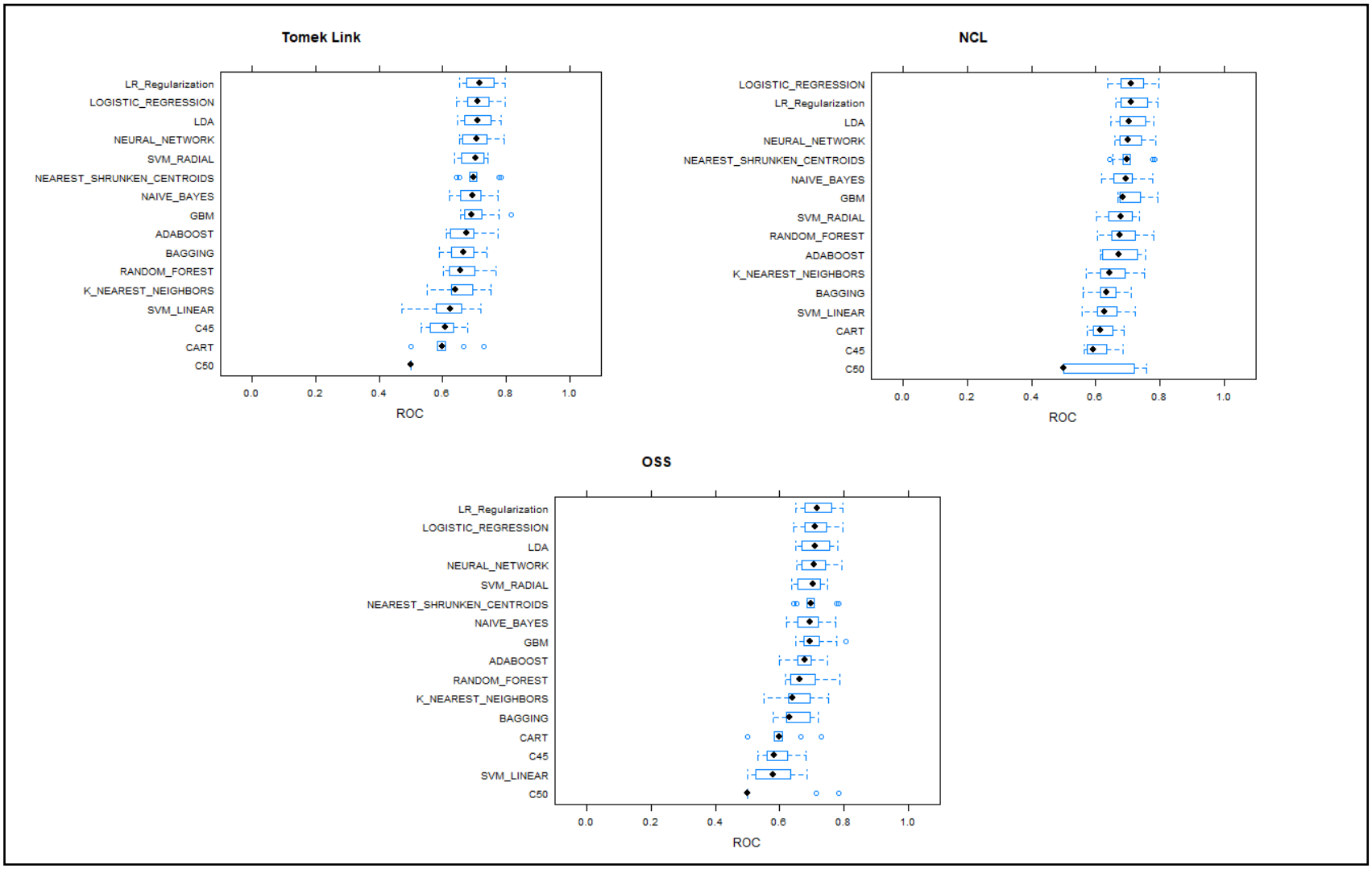


Figure 8 - Boxplots representing the ROC values from the cross-validation process for each strategy and method.

Table 17 shows the data-cleaning techniques have better Specificity values than the sampling methods. However, the sensitivity is low, and some PPV values do not exist (NaN). It happens when the model predicts all records as negative and none as positive. Besides, the NPV values are less or equal to 0.90. Since the negative class proportion is already about 90%, these models do not have predictive value.

Looking at Figure 8, we can see that the Logistic Regression penalized obtained the best ROC median for almost all approaches. On the other hand, the decision tree-based algorithms are among the worst.

Since the sampling methods gave us better NPV results than the "OSS," "NCL," and "Tomek link" strategies, we decided to add the SMOTE strategy to data cleaning techniques (SMOTE + NCL, SMOTE + Tomek, SMOTE + OSS), resulting in another 48 combinations (16x3). We also used the 10-fold cross-validation to evaluate these combinations. The results can be seen in Appendix J.

Besides, we decided to include some balancing strategies that have obtained good results in previous studies, such as the ensemble-based methods (SMOTEBoost, RUSBoost, SMOTEBagging, and UnderBagging), which apply SMOTE or undersampling using the AdaBoost technique. The four ensemble-based strategies were employed to SVM radial, NB, CART, C50, and RF methods, totaling 20 combinations. For each ensemble-based algorithm, we used a size of 10 weak learners.

Table 18 shows the results from cross-validation for each new combination. We used only the training set to train and build our models.

Table 18 - Mean of the best model to AUC, PPV, NPV, sensitivity, and specificity by cross-validation for each new combination.

METHODS	SMOTE + Tomek					SMOTE + NCL					SMOTE + OSS				
	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV
LOGISTIC_REGRESSION	0.69	0.54	0.72	0.18	0.93	0.69	0.59	0.69	0.18	0.94	0.69	0.55	0.72	0.18	0.93
LR_Regularization	0.72	0.53	0.77	0.21	0.94	0.71	0.56	0.73	0.19	0.94	0.72	0.53	0.77	0.21	0.94
LDA	0.70	0.53	0.75	0.19	0.93	0.70	0.56	0.71	0.18	0.94	0.70	0.53	0.75	0.19	0.93
NEAREST_SHRUNKEN_CENTROIDS	0.70	0.54	0.75	0.20	0.94	0.70	0.58	0.73	0.19	0.94	0.70	0.54	0.75	0.20	0.94
SVM_LINEAR	0.71	0.52	0.77	0.21	0.93	0.71	0.57	0.74	0.20	0.94	0.71	0.51	0.78	0.21	0.93
NEURAL_NETWORK	0.71	0.52	0.76	0.20	0.93	0.71	0.56	0.73	0.19	0.94	0.71	0.52	0.76	0.20	0.93
SVM_RADIAL	0.66	0.05	0.94	0.10	0.90	0.65	0.07	0.92	0.09	0.90	0.66	0.05	0.94	0.09	0.90
K_NEAREST_NEIGHBORS	0.65	0.53	0.66	0.15	0.93	0.66	0.59	0.64	0.15	0.93	0.65	0.53	0.66	0.15	0.93
NAIVE_BAYES	0.68	0.57	0.71	0.18	0.94	0.69	0.58	0.71	0.18	0.94	0.69	0.74	0.54	0.16	0.95
C45	0.68	0.43	0.80	0.19	0.93	0.68	0.48	0.79	0.21	0.93	0.68	0.43	0.80	0.19	0.93
CART	0.66	0.39	0.80	0.19	0.92	0.65	0.46	0.77	0.19	0.93	0.66	0.41	0.81	0.20	0.92
C50	0.69	0.34	0.88	0.24	0.92	0.70	0.38	0.85	0.22	0.92	0.70	0.34	0.88	0.25	0.92
RANDOM_FOREST	0.70	0.49	0.78	0.21	0.93	0.70	0.49	0.77	0.19	0.93	0.70	0.50	0.77	0.19	0.93
GBM	0.70	0.33	0.89	0.27	0.92	0.71	0.40	0.85	0.24	0.93	0.70	0.34	0.89	0.27	0.92
BAGGING	0.67	0.34	0.84	0.20	0.92	0.66	0.39	0.82	0.20	0.92	0.66	0.33	0.84	0.18	0.92
ADABOOST	0.68	0.34	0.87	0.23	0.92	0.68	0.32	0.86	0.20	0.92	0.69	0.28	0.88	0.21	0.92

METHODS	SMOTEBoost					RUSBoost					SMOTEBagging					UnderBagging				
	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV	AUC	SENS	SPEC	PPV	NPV
SVM_RADIAL	0.66	0.15	0.93	0.21	0.91	0.66	0.13	0.94	0.20	0.91	0.68	0.43	0.79	0.19	0.92	0.63	0.04	0.98	NA	0.90
NAIVE_BAYES	0.68	0.33	0.86	0.22	0.92	0.65	0.13	0.95	0.26	0.91	0.69	0.55	0.72	0.18	0.94	0.69	0.39	0.81	0.19	0.92
CART	0.67	0.02	0.99	0.19	0.90	0.64	0.02	0.99	0.39	0.90	0.66	0.42	0.80	0.21	0.92	0.57	0.00	1.00	NA	0.90
C50	0.63	0.05	0.97	0.14	0.90	0.63	0.06	0.97	0.21	0.90	0.66	0.12	0.94	0.20	0.90	0.65	0.00	1.00	NA	0.90
RANDOM_FOREST	0.68	0.07	0.98	0.28	0.90	0.69	0.03	0.99	NA	0.90	0.69	0.06	0.97	0.31	0.90	0.69	0.02	1.00	NA	0.90

Table 18 exposes the sensitivity and NPV metrics, which improved after a combination of sampling and data-cleaning methods. Consequently, Specificity and PPV were reduced. When comparing the ensemble-based methods, the Bagging look provides little better AUC results than the Boosting techniques.

After building 114 models with the best hyperparameters, we use the test set to choose and propose the best screening model for our problem. The test set was untouched during the entire training and parameter optimization, ensuring it will be used only for the final models' evaluation.

4.2.5. Model Evaluation and Comparison

According to Ferri et al. (2009) and our best knowledge, no measure simultaneously combines the classification threshold and the estimated probability. Moreover, the overall accuracy is not a suitable metric to evaluate the classification performance on an imbalanced dataset. Therefore, we had to choose which metrics to use for performance evaluation.

Since our primary goal is to minimize the number of false negatives and maximize the number of true negatives, i.e., to predict negative tests correctly, we evaluate and compare the models by the NPV. However, the results depend on the choice of the threshold. For this, we used the threshold method to set the best cut-off values based on the Youden index statistics, considering a weight two times higher for false-negative records when compared with a false-positive. To avoid bias in the model, we used the training data to choose the best threshold and the test data for model testing.

In addition to the NPV, we also evaluate the MCC metric, a balanced measure among TP, TN, FP, and FN. Our goal is to propose the hospital's decision-maker two models: one more conservative (choose by NPV) and the other moderate (choose by MCC).

We compared and discussed the possible combinations using descriptive statistics. Moreover, the Friedman and Nemenyi tests examine the MCC and NPV of the different classifiers and strategies, reporting any significant differences.

Appendix K presents the ML models' performance computed from the independent test set, showing the Sensitivity, Specificity, PPV, NPV, AUC, MCC, and Brier score values for each combination, changing the cut-off value. We focus first on NPV analysis and then on MCC.

Table 19 reports the NPV of all 16 classifiers on ten different balancing and descriptive analysis strategies with mean, median, maximum, minimum, standard deviation (sd), and interquartile range (IQR) for each strategy and method. The highest NPV is highlighted. To not disturb the study, we did not include the bad strategies which can predict all negative cases: "none," "Tomek," "NCL," and "OSS." Figure 9 and Figure 10 present the boxplots with the NPV for all strategies and methods, respectively. The greater the NPV, the better the model. We compared and discussed 116 possible combinations.

Table 19 - NPV of all 16 classifiers on 11 different balancing strategies, the Average Ranked (AR) among sampling approaches, and the descriptive analysis for each strategy and method. The highest NPV for each strategy is highlighted.

Methods	Downsampling	Upsampling	SMOTE	SMOTE_Tomek	SMOTE_NCL	SMOTE_OSS	SMOTEBoost	RUSBoost	SMOTEBagging	UnderBagging	Coluna1	n	min	max	median	iqr	mean	sd	AR
LOGISTIC_REGRESSION	0.97	0.97	0.95	0.95	0.95	0.95						6	0.95	0.97	0.95	0.015	0.957	0.01	3.83
LR_Regularization	**0.98**	0.96	0.96	0.96	0.96	0.96						6	0.96	0.98	0.96	0	0.963	0.008	**1.83**
LDA	0.97	0.97	0.95	0.95	0.95	0.95						6	0.95	0.97	0.95	0.015	0.957	0.01	3.83
NEAREST_SHRUNKEN_CENTROIDS	0.96	0.96	0.95	0.96	0.95	0.95						6	0.95	0.96	0.955	0.01	0.955	0.005	4.17
SVM_LINEAR	0.96	0.96	0.95	0.96	0.96	0.96						6	0.95	0.96	0.96	0	0.958	0.004	3.00
NEURAL_NETWORK	0.96	0.96	0.96	0.96	0.95	0.96						6	0.95	0.96	0.96	0	0.958	0.004	3.00
SVM_RADIAL	0.9	0.9	0.9	0.9	0.89	0.9	0.95	0.95	0.95	0.94		10	0.89	0.95	0.9	0.047	0.918	0.026	15.17
K_NEAREST_NEIGHBORS	0.94	0.92	0.95	0.95	0.95	0.95						6	0.92	0.95	0.95	0.008	0.943	0.012	7.17
NAIVE_BAYES	0.97	0.97	0.97	0.94	0.94	0.96	0.96	0.95	**0.98**	0.97		10	0.94	0.98	0.965	0.018	0.961	0.014	4.00
C45	0.95	0.93	0.93	0.93	0.94	0.93						6	0.93	0.95	0.93	0.007	0.935	0.008	10.33
CART	NA	0.95	0.92	0.92	0.95	0.93	0.97	0.96	0.98	NA		8	0.92	0.98	0.95	0.035	0.948	0.023	10.83
C50	0.95	0.9	0.93	0.91	0.92	0.92	0.91	0.91	0.91	0.96		10	0.9	0.96	0.915	0.017	0.922	0.019	12.00
RANDOM_FOREST	**0.98**	0.97	0.95	0.96	0.94	0.96	0.9	0.9	0.9	0.9		10	0.9	0.98	0.945	0.06	0.936	0.033	**2.83**
GBM	0.96	0.96	0.93	0.94	0.94	0.94						6	0.93	0.96	0.94	0.015	0.945	0.012	8.17
BAGGING	0.94	0.9	0.92	0.91	0.92	0.92						6	0.9	0.94	0.92	0.008	0.918	0.013	12.83
ADABOOST	0.91	0.9	0.91	0.91	0.91	0.9						6	0.9	0.91	0.91	0.007	0.907	0.005	14.17
n	15	16	16	16	16	16	5	5	5	4									
min	0.90	0.90	0.90	0.90	0.89	0.90	0.90	0.90	0.90	0.90									
max	0.98	0.97	0.97	0.96	0.96	0.96	0.97	0.96	0.98	0.97									
median	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95									
iqr	0.03	0.05	0.02	0.04	0.02	0.03	0.05	0.04	0.07	0.03									
mean	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.93	0.94	0.94									
sd	0.02	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.03									

Legend: n - the number of individuals; min - minimum; max - maximum; sd - standard deviation of the mean; IQR - interquartile range; AR - Average Ranked.

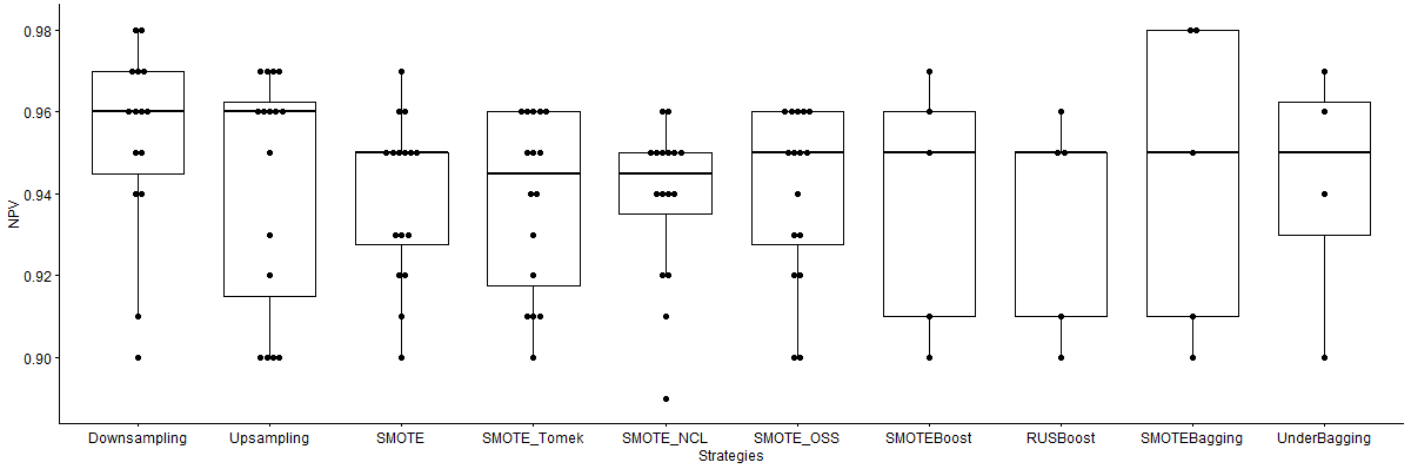


Figure 9 - Boxplots representing the NPV of each method (points) for all strategies.

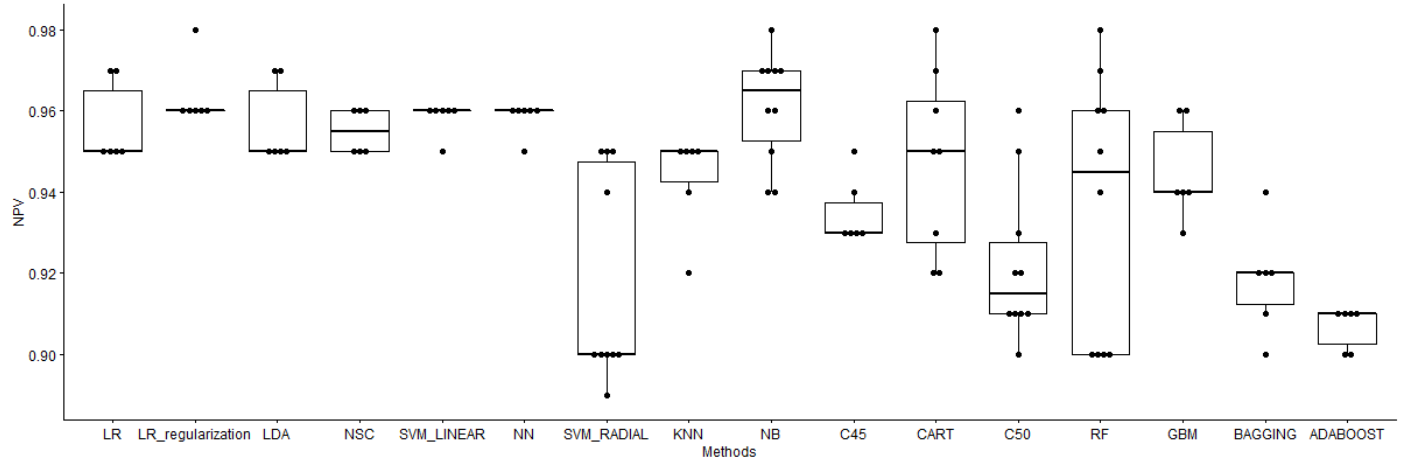


Figure 10 - Boxplots representing the NPV of each strategy (points) for all methods.

We can see in Table 19, Figure 9, and Figure 10 that the medians of the strategies are similar, but the NPV varies from 0.89 to 0.98 depending on the method for the same strategy. Moreover, some methods perform poorly depending on the strategy, even using the same hyperparameters for searching, such as SVM radial.

The four strategies based on ensemble "UnderBagging," "RUSBoost," "SMOTEBoost," and "SMOTEBagging" did not obtain the best NPV means since they did not work well for tree-based methods. However, they are good at using radial SVM and NB as weak learners. To our classification problem, sampling methods obtained the best NPV averages.

Since some methods have similar min, max, and mean of NPV, we decided to use the Average Ranked (AR) metric to select the best technique considering only the strategies applied to all methods (downsampling, upsampling, SMOTE, SMOTE+Tomek, SMOTE + NCL, SMOTE + OSS) - The lower the AR, the better the model. The Logistic Regression with regularization presented better results, and the NB technique the highest median.

We also saw that decision tree algorithms did not work well to classify CR-GNB non-acquisition. On the other hand, if we compared the Brier score (Appendix K), we know that these methods are the best for predicting the probability. Therefore, depending on the study's objective and the metric, the chosen models may be different. This metric will be better discussed in our acquisition risk model in the next chapter.

We used the Friedman and Nemenyi test for significance analysis. The Friedman test statistic was significant among the balancing strategies (Friedman chi-squared = 18; p-value = 0.003) and among the methods (Friedman chi-squared = 60.6; p-value < 0.001). Post hoc Nemenyi tests were then applied to verify significant differences among the approaches with a 95% confidence level. The results are summarized in Appendix L.

Looking at Appendix L, there is no difference in the balance of our data using down, upsampling, or SMOTE (p-value > 0.568). Also, there is no difference between the data-cleaning approaches (p-value = 1) or the ensemble approaches (p-value $\geq$ 0.247). However, this test shows that, in general, the ensemble approach's models obtained significantly better results than the data cleaning and sampling approaches (p-value < 0.001), especially SMOTEBagging and UnderBagging.

According to Figure 10 and Appendix L, the radial SVM has significantly worse results than the linear methods and neural network (p-value < 0.05). Moreover, for the classification, we can see that the more straightforward, linear techniques such as LR with regularization also give a relatively good performance, which is not significantly different from the more complex classifiers, such as NB and RF. There was no difference in the strategies for linear methods.

The maximum NPV (0.98) was found by the Naive Bayes using the SMOTEBagging strategy and by the combination of logistic regression regularized or random forest with downsampling approach. We selected these models aiming to detect whether a patient needs a culture test. We analyze scenarios and false negatives in the next sections.

The NPV must be as high as possible since our purpose is to detect the negatives instances correctly while controlling the number of false-negative notifications. However, we should not exclude the importance of sensitivity - we cannot have nonexistent or low values for this metric, identifying the proportion of actual positive cases correctly.

Thus, after comparing the results for discrimination through the NPV, we reached the models by MCC. The higher the MCC, the better the model. Table 20 reports the MCC of all 16 classifiers on different balancing and descriptive analysis strategies with mean, median, standard deviation, interquartile range, maximum, and minimum value for each approach. Figure 11 and Figure 12 present the boxplots with the MCC for all strategies and techniques, respectively.

Table 20 - MCC of all 16 classifiers on 11 different balancing and descriptive analysis strategies for each strategy and method. The highest MCC for each strategy is highlighted.

Methods	Downsampling	Upsampling	SMOTE	SMOTE_Tomek	SMOTE_NCL	SMOTE_OSS	SMOTEBoost	RUSBoost	SMOTEBagging	UnderBagging	n	min	max	median	iqr	mean	sd	AR
LOGISTIC_REGRESSION	0.18	0.21	0.17	0.18	0.17	0.18					6	0.17	0.21	0.18	0.008	0.182	0.015	7.33
LR_Regularization	0.17	0.21	0.19	0.20	0.19	0.20					6	0.17	0.21	0.195	0.01	0.193	0.014	**4.83**
LDA	0.17	0.21	0.18	0.19	0.18	0.20					6	0.17	0.21	0.185	0.018	0.188	0.015	6.00
NEAREST_SHRUNKEN_CENTROIDS	0.20	0.18	0.18	0.18	0.17	0.17					6	0.17	0.2	0.18	0.008	0.18	0.011	8.33
SVM_LINEAR	0.15	0.19	0.17	0.19	0.19	0.19					6	0.15	0.19	0.19	0.015	0.18	0.017	7.50
NEURAL_NETWORK	0.17	0.17	0.23	**0.24**	0.22	**0.24**					6	0.17	0.24	0.225	0.055	0.212	0.033	**3.83**
SVM_RADIAL	0.00	-0.04	-0.07	-0.06	-0.07	-0.06	0.21	**0.24**	0.22	**0.24**	10	-0.07	0.24	-0.02	0.277	0.061	0.145	15.83
K_NEAREST_NEIGHBORS	0.14	0.12	0.16	0.16	0.16	0.16					6	0.12	0.16	0.16	0.015	0.15	0.017	11.67
NAIVE_BAYES	0.16	0.19	0.21	0.14	0.15	0.20	0.18	0.13	0.16	0.20	10	0.13	0.21	0.17	0.045	0.172	0.028	8.33
C45	0.11	0.13	0.22	0.15	0.23	0.14					6	0.11	0.23	0.145	0.07	0.163	0.05	9.00
CART	0.00	0.20	0.17	0.10	0.19	0.14	0.17	0.22	0.21	0.00	10	0	0.22	0.17	0.088	0.14	0.082	11.00
C50	0.20	0.08	0.22	0.12	0.17	0.17	0.13	0.16	0.09	0.22	10	0.08	0.22	0.165	0.07	0.156	0.05	8.33
RANDOM_FOREST	0.20	0.18	0.21	0.19	0.19	0.21	0.00	-0.02	-0.02	0.09	10	-0.02	0.21	0.185	0.175	0.123	0.1	**4.83**
GBM	0.18	0.21	0.17	0.22	0.23	0.20					6	0.17	0.23	0.205	0.032	0.202	0.023	**4.33**
BAGGING	0.20	-0.01	0.13	0.12	0.14	0.19					6	-0.01	0.2	0.135	0.055	0.128	0.075	11.00
ADABOOST	0.11	-0.04	0.09	0.16	0.16	0.09					6	-0.04	0.16	0.1	0.057	0.095	0.073	13.67
n	16.00	16.00	16.00	16.00	16.00	16.00	5.00	5.00	5.00	5.00								
min	0.00	-0.04	-0.07	-0.06	-0.07	-0.06	0.00	-0.02	-0.02	0.00								
max	0.20	0.21	0.23	0.24	0.23	0.24	0.21	0.24	0.22	0.24								
median	0.17	0.18	0.18	0.17	0.18	0.19	0.17	0.16	0.16	0.20								
iqr	0.05	0.09	0.04	0.06	0.03	0.05	0.05	0.09	0.12	0.13								
mean	0.15	0.14	0.16	0.16	0.17	0.16	0.14	0.15	0.13	0.15								
sd	0.06	0.09	0.07	0.07	0.07	0.07	0.08	0.10	0.10	0.10								

Legend: n - the number of individuals; min - minimum; max - maximum; sd - standard deviation of the mean; IQR - interquartile range; AR - Average Ranked.

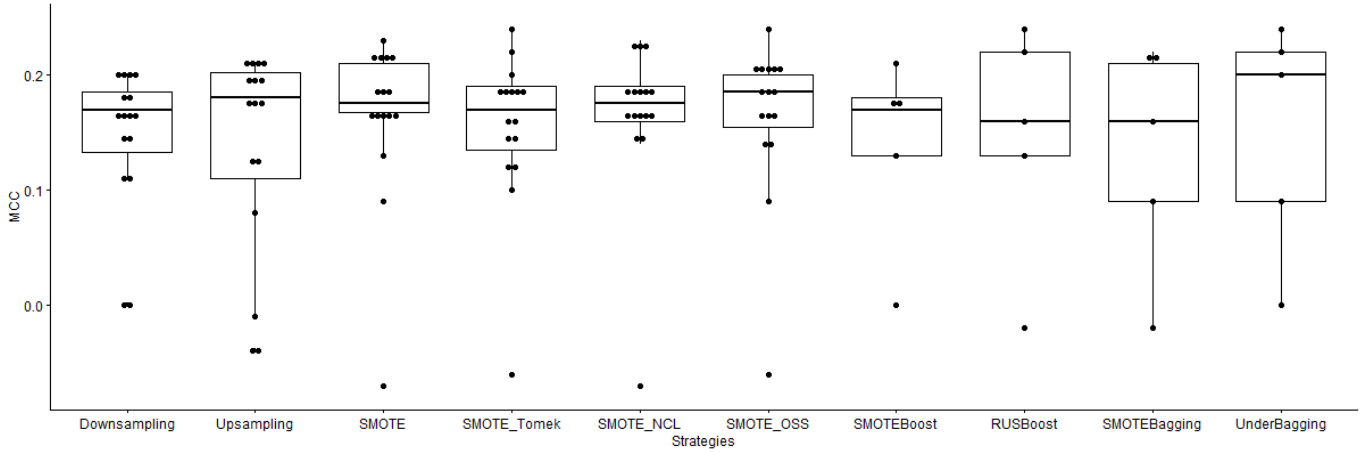


Figure 11 - Boxplots representing the MCCs of each method (points) for all strategies.

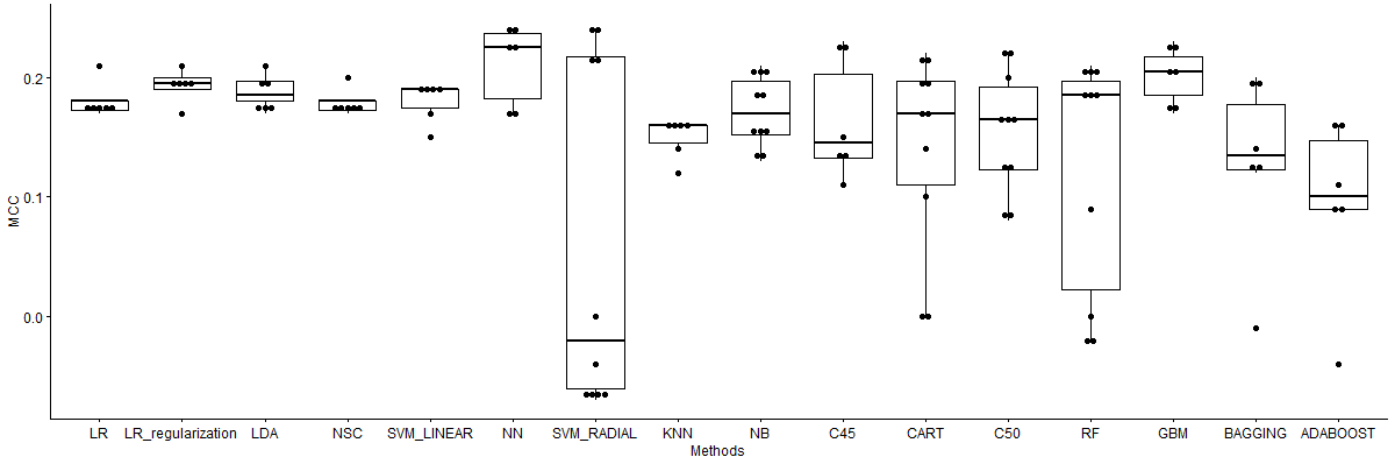


Figure 12 - Boxplots representing the MCCs of each strategy (points) for all methods.

Like NPV analyses, the balancing strategies based on the ensemble also did not work well for tree-based methods. SMOTE+OSS and UnderBagging obtained the best medians.

Comparing the machine learning techniques, we can see that the neural network (AR=3.83) and gradient Boosting (AR=4.33) have the best average ranked, followed by logistic regression penalized and random forest (AR=4.83). The SVM Radial method, combined with ensemble approaches, gives a good MCC (0.24) but with a high variance and IQR. High MCC also can be seen in combinations of the NN with the data cleaning approaches. Looking at the boxplots, the linear models seem to provide a more stable performance. In general, there was no difference between linear and non-linear methods.

It is essential to note that there are high standard deviation values both for methods and strategies. Thus, it is likely that both can explain the differences in MCC. The Friedman test statistic was significant when we compared the balancing approach and methods, and the post hoc Nemenyi tests are summarized in Appendix L. The behavior is like the NPV since we also use FN and TN to calculate the MCC.

The overview comparison shows that most techniques yielded classification performances that are quite competitive with each other, like in Brown and Mues (2012). Even though the differences between the classifiers are small, it is essential to note that in an infection context, an increase in the prediction ability, even a low percentage, may save lives and reduce costs. We also concluded that there are differences between some balancing strategies, and they give us better models than the original without balancing.

In short, there is no generic rule to choose a single best method or strategy. The choice depends on each problem, database, and evaluation metric used. Table 21 and Table 22 present the best performance of balanced data sets for all methods and classification metrics to facilitate understanding. To our problem, sampling strategies, in general, presented better results.

These values result from using AUC to find the best combination of hyperparameters, considering twice the weight for false negatives.

Table 21 - The best strategy for each metric and method.

Methods	Sens	Spec	PPV	NPV	BrierScore	AUC	MCC
LOGISTIC_REGRESSION	DOWNSAMPLING	SMOTE + TOMEK LINK	UPSAMPLING; SMOTE + TOMEK LINK	DOWNSAMPLING; UPSAMPLING	SMOTE	DOWNSAMPLING; UPSAMPLING	UPSAMPLING
LR_Regularization	DOWNSAMPLING	SMOTE + TOMEK LINK; SMOTE + OSS	UPSAMPLING; SMOTE + TOMEK LINK; SMOTE + OSS	DOWNSAMPLING	SMOTE; SMOTE + TOMEK LINK; SMOTE + OSS	UPSAMPLING	UPSAMPLING
LDA	DOWNSAMPLING	SMOTE + NCL	SMOTE + OSS	DOWNSAMPLING; DOWNSAMPLING	SMOTE; SMOTE + TOMEK LINK; SMOTE + OSS	UPSAMPLING	UPSAMPLING
NEAREST_SHRUNKEN_CENTROIDS	DOWNSAMPLING; DOWNSAMPLING	SMOTE	DOWNSAMPLING	DOWNSAMPLING; UPSAMPLING	SMOTE; SMOTE + TOMEK LINK; SMOTE + OSS	UPSAMPLING	DOWNSAMPLING
SVM_LINEAR	DOWNSAMPLING	SMOTE	SMOTE; SMOTE + TOMEK LINK + SMOTE +NCL; SMOTE +OSS	DOWNSAMPLING; UPSAMPLING; SMOTE+NCL, SMOTE+TOMEK, SMOTE+OSS	SMOTE; SMOTE + TOMEK LINK; SMOTE + OSS	UPSAMPLING; SMOTE + OSS	UPSAMPLING; SMOTE + TOMEK LINK; SMOTE + NCL; SMOTE + OSS
NEURAL_NETWORK	DOWNSAMPLING	SMOTE; SMOTE + OSS	SMOTE; SMOTE + TOMEK LINK; SMOTE +OSS	DOWNSAMPLING; UPSAMPLING; SMOTE; SMOTE+TOMEK, SMOTE+OSS	UnderBagging	RUSBoost	SMOTE + TOMEK LINK; SMOTE + OSS
SVM_RADIAL	SMOTEBagging	DOWNSAMPLING	RUSBoost; UnderBagging	SMOTEBagging; SMOTEBoost	UPSAMPLING	DOWNSAMPLING; UPSAMPLING	RusBoost; UnderBagging
K_NEAREST_NEIGHBORS	SMOTE + NCL	UPSAMPLING	UPSAMPLING	SMOTE; SMOTE+NCL, SMOTE+TOMEK, SMOTE+OSS	DOWNSAMPLING	SMOTE + NCL	SMOTE + OSS; UnderBagging
NAIVE_BAYES	SMOTEBagging	SMOTE + TOMEK LINK	SMOTE; SMOTE + OSS	DOWNSAMPLING; UPSAMPLING; SMOTE	RUSBoost	DOWNSAMPLING; UPSAMPLING	SMOTE
C45	DOWNSAMPLING	SMOTE	SMOTE	DOWNSAMPLING	SMOTE	SMOTE + NCL	SMOTE
CART	DOWN; UnderBagging	SMOTE	SMOTE	SMOTEBagging	UnderBagging	RUSBoost; SMOTEBagging	SMOTEBagging
C50	UnderBagging	UPSAMPLING	UPSAMPLING	SMOTE + NCL	UnderBagging	UPSAMPLING; SMOTE; SMOTE + OSS	SMOTE; UnderBagging
RANDOM_FOREST	DOWNSAMPLING	SMOTEBagging	UnderBagging	DOWNSAMPLING	UnderBagging	DOWNSAMPLING; UPSAMPLING; UnderBagging	SMOTE; SMOTE + OSS
GBM	DOWNSAMPLING	SMOTE	SMOTE + TOMEK LINK; SMOTE + NCL	DOWNSAMPLING; UPSAMPLING	SMOTE	UPSAMPLING	SMOTE + NCL
BAGGING	DOWNSAMPLING	UPSAMPLING	SMOTE + OSS	DOWNSAMPLING	UPSAMPLING	DOWNSAMPLING	DOWNSAMPLING
AdaBoost	DOWNSAMPLING; SMOTE + TOMEK	UPSAMPLING	SMOTE + NCL	DOWNSAMPLING; SMOTE; SMOTE+TOMEK; SMOTE+NCL	UPSAMPLING	SMOTE + OSS	SMOTE + TOMEK LINK; SMOTE + NCL

Table 22 - The best performance for each metric and methods.

Methods	Sens	Spec	PPV	NPV	brierScore	AUC	MCC
LOGISTIC_REGRESSION	0.88	0.58	0.16	0.97	0.17	0.73	0.21
LR_Regularization	0.94	0.56	0.16	0.98	0.17	0.75	0.21
LDA	0.88	0.60	0.17	0.97	0.17	0.73	0.21
NEAREST_SHRUNKEN_CENTROIDS	0.81	0.53	0.16	0.96	0.18	0.74	0.20
SVM_LINEAR	0.83	0.60	0.16	0.96	0.17	0.72	0.19
NEURAL_NETWORK	0.86	0.65	0.19	0.96	0.17	0.73	0.24
SVM_RADIAL	0.72	0.97	0.22	0.95	0.09	0.72	0.24
K_NEAREST_NEIGHBORS	0.74	0.75	0.16	0.95	0.20	0.70	0.16
NAIVE_BAYES	0.94	0.55	0.16	0.98	0.12	0.74	0.21
C45	0.83	0.86	0.25	0.95	0.11	0.69	0.23
CART	1.00	0.88	0.23	0.98	0.09	0.74	0.22
C50	0.82	0.99	0.29	0.96	0.08	0.74	0.22
RANDOM_FOREST	0.92	1.00	0.26	0.98	0.08	0.75	0.21
GBM	0.85	0.79	0.21	0.96	0.11	0.75	0.23
BAGGING	0.62	1.00	0.25	0.94	0.10	0.71	0.20
ADABOOST	0.19	0.99	0.29	0.91	0.10	0.71	0.16
The Best	1.00	1.00	0.29	0.98	0.08	0.75	0.24

SMOTEBagging and UnderBagging performed well for various metrics among the methods in which they were applied (NB, SVM Radial, CART, C50, and RF). A hypothesis for this performance is because these ensemble approaches do not consider only a limited set for training, but different samples from the dataset.

According to traditional methods, they were as good or better than data cleaning techniques. Cleaning methods tend to remove extreme instances, but this did not seem to improve our dataset significantly.

In section 4.2.6, we will analyze each false-negative case in the selected classifiers' confusion matrix, aiming to know these records and why the algorithm failed. The models chosen by NPV were Naïve Bayes with SMOTEBagging, Logistic Regression Regularized with downsampling, and Random Forest with downsampling. MCC's selected ones were Neural Network with SMOTE+OSS, Neural Network with SMOTE+Tomek, and Support Vector Machine Radial with RUSBoost. Moreover, some models can be better evaluated than others but fail to objectives such as interpretability and computational time (CRONE; FINLAY, 2012). Thus, we will also compare the computational time spent for each model built and the interpretation capability.

4.2.5.1.
Computational Time

The purpose of this subsection is to present the computational time spent both on the tuned grid search to determine the sets of hyperparameters (Timings Everything) and the final model (Timings Final Model).

Table 23 shows that the sampling strategies have the lowest medians, followed by data cleaning strategies. As expected, tree-based strategies take longer to build the final model. Regarding methods (see Table 24), we can see that the linear models are more efficient computationally, followed by decision trees. The SVM times are longer than average was, and Adaboost is the slowest.

It is worth mentioning that these timings consider the hyperparameters and algorithms used for this work, depending on the number of combinations and the algorithm type. Thus, this comparison applies only to that specific job.

Table 23 - Summary of computational time for each strategy in ascending order by the median.

Timing Everything (min)	Strategies	n	min	max	median	IQR	mean	sd
	Downsampling	16	1.1	404.39	24.145	49.805	59.396	103.714
	Upsampling	16	1.8	5394.5	40.855	370.162	555.838	1339.765
	OSS	16	6.82	935	43.285	453.69	216.056	285.345
	SMOTE	16	3.72	880.16	67.01	185.46	168.763	245.224
	Tomek	16	4.28	1562.22	74.41	239.12	244.732	412.757
	UnderBagging	5	5.11	21084.61	75.25	887.95	4435.108	9315.584
	NCL	16	10.17	1718	121.35	431.915	343.222	468.258
	SMOTE_Tomek	16	6.14	1092.17	125.13	286.562	234.591	310.299
	SMOTE_OSS	16	8.03	1123.66	135.21	336.685	256.974	320.841
	SMOTE_NCL	16	11.55	1324.91	157.575	433.882	341.888	411.452
	RUSBoost	5	26.88	47882.63	1001.92	6325.73	11093.356	20736.03
	SMOTEBoost	5	59.44	57859.35	1402.73	9546.61	13860.67	24922.239
	SMOTEBagging	5	82.21	129886.99	1474.25	1937.78	26782.062	57644.029

Timing Final Model (min)	Strategies	n	min	max	median	IQR	mean	sd
	Downsampling	16	0	4.73	0.1	0.13	0.439	1.171
	SMOTE	16	0.21	10.05	0.385	0.45	1.322	2.542
	Upsampling	16	0.02	12.37	0.455	1.898	2.428	4.075
	Tomek	16	0.3	10.11	0.475	0.873	1.487	2.47
	SMOTE_Tomek	16	0.47	12.94	0.7	1.138	1.932	3.137
	OSS	16	0.53	14.7	0.9	0.803	2.052	3.5
	SMOTE_OSS	16	0.65	11.11	0.92	1.025	1.964	2.647
	NCL	16	0.81	12.46	1.27	0.808	2.227	2.837
	SMOTE_NCL	16	0.99	16.14	1.3	1.235	2.739	3.742
	UnderBagging	5	0.2	46.16	2.44	5.89	11.094	19.754
	RUSBoost	5	2.12	100.47	6.5	14.22	26.55	41.825
	SMOTEBoost	5	4.54	126.67	9.28	17.29	35.02	51.885
	SMOTEBagging	5	3.85	287.8	10.29	7.89	64.458	124.914

Table 24 - Summary of computational time for each method in ascending order by the median.

Timing Everything (min)	Methods	n	min	max	median	IQR	mean	sd
	LDA	9	1.1	13.91	6.14	4.31	6.494	4.408
	LR	9	2.06	11.55	6.43	4.56	6.307	3.397
	NSC	9	1.79	14.11	6.5	4.89	6.953	4.391
	CART	13	2.53	82.21	7.56	10.08	18.08	24.72
	BAGGING	9	2.31	17.29	9.75	5.86	9.961	4.532
	C50	13	12.28	264.54	39.2	31.98	72.952	79.521
	KNN	9	5	96.14	44.8	35.73	46.199	31.291
	LR_regularization	9	21.09	92.46	51.14	27.81	54.592	24.044
	SVM_LINEAR	9	12.64	1104.38	180.75	99.51	246.304	330.484
	RF	13	30.77	2202.32	292.42	727.78	589.355	634.919
	NN	9	50.63	486.31	296.96	122.72	282.594	130.749
	NB	13	27.2	9764.22	319.83	360.16	1528.683	3018.846
	C45	9	57.23	886.97	413.45	326.6	436.77	282.664
	ADABOOST	9	403.63	538.77	451.75	61.77	457.282	46.105
	GBM	9	159.05	1059.32	815.48	193.09	768.184	285.097
	SVM_RADIAL	13	36.89	129886.99	1562.22	19992.44	20768.202	38065.737

Timing Final Model (min)	Methods	n	min	max	median	IQR	mean	sd
	NSC	9	0.02	1.13	0.47	0.49	0.494	0.403
	KNN	9	0	1.14	0.5	0.47	0.494	0.413
	LR	9	0.06	0.99	0.5	0.45	0.483	0.332
	LDA	9	0.03	1.27	0.51	0.46	0.549	0.433
	C45	9	0.05	1.27	0.56	0.49	0.598	0.424
	CART	13	0.02	6.23	0.56	0.85	1.352	1.903
	LR_regularization	9	0.1	1.43	0.64	0.66	0.708	0.45
	NB	13	0.03	25.95	0.75	0.99	4.148	8.304
	BAGGING	9	0.1	1.53	0.85	0.48	0.836	0.436
	NN	9	0.13	1.83	1	0.57	1.006	0.53
	C50	13	0.12	10.29	1.55	1.49	2.784	3.202
	RF	13	0.14	14.12	1.95	4.86	3.768	4.129
	GBM	9	0.24	3.28	2.43	0.77	2.259	1.098
	SVM_RADIAL	13	0.22	287.8	3.64	43.7	45.698	83.749
	SVM_LINEAR	9	1	5.94	4.22	3.44	3.272	1.999
	ADABOOST	9	4.73	16.14	11.77	2.83	11.557	3.256

4.2.6.
Model Analysis

In this section, we discuss the results found in the confusion matrices of the best-classifiers selected by NPV (NB, RF, and LR regularized) and by MCC (NN and SVM Radial) and analyze the false-negative cases, aiming to find out which (and why) records were mispredicted.

We analyzed the classifiers mentioned using the 781 records as a reference (78 positives and 703 negatives). The confusion matrix and its metrics for each of these methods choose by NPV can be seen in Table 25, considering twice the weight for false negatives. These models are considered conservative since they have few false-negative cases.

Table 25 - Confusion matrix and its metrics for Naïve Bayes, Random Forest, and Logistic Regression regularized methods. The values predicted as false negatives are highlighted in red and true negatives in green.

NB (SMOTEBagging)						RF (downsampling)						LR_regularization (downsampling)					
Sens	Spec	PPV	NPV	AUC		Sens	Spec	PPV	NPV	AUC		Sens	Spec	PPV	NPV	AUC	
0.94	0.30	0.13	0.98	0.73		0.92	0.39	0.14	0.98	0.75		0.94	0.32	0.13	0.98	0.73	
		Reference						Reference						Reference			
		Pos	Neg					Pos	Neg					Pos	Neg		
Predicted	Pos	73	490			Predicted	Pos	72	429			Predicted	Pos	73	480		
	Neg	5	213				Neg	6	274				Neg	5	223		

Table 25 shows the values predicted as false negatives highlighted in red and true negatives in green. For example, the Random Forest model classified 280 cases as negative, of which 6 cases were wrongly predicted (false negatives) and 274 cases correctly predicted (true negatives). These six cases are positive reference cases. Thus, we have an NPV of 98% and a Sensitivity of 92%.

If we decide to use this model to determine who will be screened, we will not perform 280 tests of the 781 proposed. Of these, six patients colonized or infected with carbapenem-resistant bacteria will not be isolated, and 274 unnecessary tests would be avoided (Specificity of 39%). This saving directly affects the laboratory´s work charge, reduce the hospital's budget, and save time on collecting exams. However, it is essential to note that six patients will not be detected and isolated at that time.

On the other hand, if we choose regularized logistic regression as our final prediction model, we obtain results with higher sensitivity (94%), less specificity (32%), and similar NPV (98%). The False Negatives (FN) decrease to 5, and the

number of unnecessary negative tests drop to 223, with a difference of 7% in specificity (a decline from 39% to 32%).

The Naïve Bayes also has an NPV similar to the two techniques previously mentioned, but with specificity worse than both and PPV worse than the RF model.

I addition to the NPV, we also evaluated the MCC metric. The confusion matrix and its metrics for each of the methods choose by MCC can be seen in Table 26. These models are moderate since they have a more significant reduction in culture tests but a consequent increase in false negatives.

Table 26 - Confusion matrix and its metrics for the methods Neural Network and SVM Radial. The values predicted as false negatives are highlighted in red and true negatives in green.

NN (SMOTE+OSS)					**SVM Radial (RUSBoost)**					**NN (SMOTE+Tomek)**				
Sens	Spec	PPV	NPV	MCC	Sens	Spec	PPV	NPV	MCC	Sens	Spec	PPV	NPV	MCC
0.74	0.65	0.19	0.96	0.24	0.62	0.75	0.22	0.95	0.24	0.76	0.64	0.19	0.96	0.24

		Reference	
		Pos	Neg
Predicted	Pos	58	247
	Neg	**20**	**456**

		Reference	
		Pos	Neg
Predicted	Pos	48	175
	Neg	**30**	**528**

		Reference	
		Pos	Neg
Predicted	Pos	59	251
	Neg	**19**	**452**

Table 26 shows the values predicted for the three best models by MCC. For example, the Neural Network (SMOTE + Tomek) classified 471 cases as negative, of which 19 cases were wrongly predicted (false negatives) and 452 cases correctly predicted (true negatives). Thus, we have an NPV of 96% and a Sensitivity of 76%, and both decrease when compared to the previous model. The specificity and PPV increase by about 25% and 5%, respectively, reducing unnecessary negative tests by approximately 64%. The Neural Network (SMOTE+OSS) works similarly, but the SVM Radial has an awful sensitivity despite having an equal MCC value, with 30 false-negative cases (Sens = 62%).

If one decides to use the NN (SMOTE + Tomek) model to determine who should be screened, we will not perform 471 tests of the 781 proposed. However, 19 patients with carbapenem-resistant bacteria will not be isolated. This model can be useful for hospitals that need to decrease costs more grossly or even for those who do not use the screening protocol. The test would be done for only 40% of patients using this model, which is better than not testing anyone.

It is essential to know that no single performance measure can always be best than others (LORETO; LISBOA; MOREIRA, 2020). The aim of clinical

application determines which performance measure is more important, and thus which model is more suitable for each application scenario. For example, if the clinicians prefer to detect one more patient positive than reduce 51 culture tests, the "LR Regularization" model may have more advantages than the RF model based on the performance comparison of sensitivity. However, if it is more important to reduce the tests due to limited laboratory resources, the NN is the choice.

We analyzed each model aiming to know the reason the algorithm failed to classify each positive instance. Regarding them, the errors found are related to the following.

According to the first three algorithms, the five false-negative cases did not use mechanical ventilation. They had zero duration in the critical variables "CVCTIMESTOTAL," "CVCDURMORE," "MVTIMESTOTAL," and "ArtDURMORE." In Table 13, we can see that positive cases are more likely to use these invasive devices. Besides, the "Emergency" source reason is a more likely reason for a negative case. Aiming to find some characteristic in common between the records that may explain the positive reference, we observed that all these patients use antibiotics of class J01D and/or J01C, which is more likely in the group of positives. It is difficult to identify the combinations of terms that caused the record to be classified as negative by the algorithm. The same false-negative cases appeared in the NB, RF, and LR. In the next section, we interpret the model using the logistic regression coefficients.

4.2.6.1. Interpretability

In some models, the results are directly interpreted, but others are challenging to comprehend, as shown in

Table *7*. For example, we can understand and demonstrate the predictive logistic regression model's practical use using the coefficients calculated for each factor. However, the random forest or neural network, known as a black-box model, does not allow interpretation; we cannot precisely know how the patients were classified. Thus, we will show the predictive model's practical application using the

regularized logistic regression as one of the best models to interpret the screening model results.

From the developed model, we can estimate the CR-GNB acquiring individual by feeding the model with the values of independent variables associated with the patient. Depending on the likelihood, the patient can be classified as a positive or negative acquisition. eq. (17) indicates a general form of the logistic regression model.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta X \tag{17}$$

Where p is the acquisition probability, β_0 is a constant, X is the matrix of independent variables used for predicting the risk, and β is the vector of coefficients corresponding to X, representing the relationship between the variables and the reference level. The term p/1-p is known as the odds of the acquisition (Odds Ratio - OR).

The regularized logistic regression adds a penalty to model fit during the training, changing the estimate coefficients. However, these coefficients are interpreted in the same way that the pattern logit model. To demonstrate the predictive model's practical use, we show all coefficients and ORs in Table 27.

Table 27 - The output from regularized logistic regression, including coefficients (β) and odds ratio (OR).

-	-	β	OR = exp(β)	-	-	β	OR = exp(β)
	(Intercept) β0	-0.912	0.402				
X1	HospitalB	-0.104	0.901	X23	CVCTIMESTOTAL	0.232	1.262
X2	HospitalC	-0.415	0.660	X24	J01DTRUE	0.331	1.392
X3	HospitalD	-0.098	0.907	X25	J01GTRUE	0.283	1.327
X4	J01XTRUE	0.386	1.471	X26	ChronicHealthStatusNecessidade de assistencia	0.200	1.222
X5	VesDURTOTAL	0.371	1.449	X27	ChronicHealthStatusRestrito / acamado	0.171	1.187
X6	AdmissionSourceOperation Room	0.040	1.041	X28	IsNeurologicalComaStuporObtunded DeliriumTRUE	0.279	1.321
X7	AdmissionSourceOther ICU from hospital	0.076	1.079	X29	CVCDURMORE	0.165	1.179
X8	AdmissionSourceSemi Intensive Unit	-0.069	0.933	X30	DiaTIMESTOTAL	0.429	1.536
X9	AdmissionSourceWard/Room	-0.027	0.973	X31	VesDURMORE	0.166	1.181
X10	Saps3Points	-0.059	0.942	X32	IsHistoryOfPneumoniaTRUE	0.397	1.487
X11	AdmissionReasonElective Surgery	0.078	1.081	X33	CharlsonIndex	-0.353	0.703
X12	AdmissionReasonInfection / Sepsis	0.190	1.210	X34	J01CTRUE	0.059	1.061
X13	AdmissionReasonLiver and Pancreas / Gastrointestinal	-0.364	0.695	X35	ArtDURMORE	0.405	1.499
X14	AdmissionReasonNeurological	0.295	1.344	X36	ArtTIMESTOTAL	0.234	1.263
X15	AdmissionReasonRespiratory	0.169	1.184	X37	DIALYSISYES	0.379	1.461
X16	VesTIMESTOTAL	0.419	1.521	X38	MVTIMESTOTAL	0.461	1.586
X17	tests_before	-0.865	0.421	X39	IsChronicAtrialFibrilationTRUE	-0.115	0.891
X18	MFIpoints	0.184	1.202	X40	IsStrokeSequelaeTRUE	0.397	1.487
X19	Age	0.220	1.247	X41	IsDiabetesUncomplicatedTRUE	-0.043	0.958
X20	PerDURTOTAL	-1.421	0.242	X42	PERIPHERALYES	0.000	1.000
X21	MVDURTOTAL	0.205	1.228	X43	IsDementiaTRUE	0.039	1.039
X22	LOS_hospital_before_test	-0.110	0.896	X44	MVYES	0.201	1.223

The coefficient (β) of the level selected for each factor X is imputed and calculated. The coefficient is zero if the chosen class is the reference level.

Let us consider an example. A 90-year-old patient admitted to the ICU by Ward/Room at Hospital C with admission reason "Respiratory" did the test 6 days after in hospital. This man had SAPS 3 of 70, MFIpoint of 2, Charlson Comorbidity Index of 7, used drugs from J01C family, chronic health status "independent," without any invasive procedure use, and presented Diabetes Uncomplicated. Using the coefficients from Table 27, the non-acquisition rate calculation for this patient is the following (eq. (18)).

$$ln\frac{\hat{p}}{1-\hat{p}} = -0.912 - 0.415 - 0.027 + (-0.059 * 0.6458) + 0.169 + (0.184 * 0.25) + (0.220 * 0.8275) + (-0.110 * 0.052) + (-0.353 * 0.583) + 0.059 - 0.043 = -1.19$$

$$\frac{\hat{p}}{(1-\hat{p})} = e^{-1.19} = 0.304 \rightarrow \hat{p} = 0.304(1-\hat{p}) \rightarrow \hat{p} = 0.233 \quad (18)$$

Thus, resulting in an acquisition CR-GNB probability of 23.3%. That is, this patient has an approximately 23.3% chance to be a positive test.

It is important to remember that since the regularized logistic regression fits the model using normalized data, then the numerical values used in the equation are normalized, non-integer. For example, the age of 90 years old has a value of 0.8275 after normalization.

Using the cut-off of p = 0.34 for a positive case (see Appendix K), this patient is predicted Negative for our model (sensitivity of 0.94, a specificity of 0.32, PPV of 0.13, and NPV of 0.98). In this case, the sensitivity and specificity are the percentages of "true acquisition" and "true non-acquisition," respectively. Incorporating our model in the hospital's screening system, this patient would not be tested.

In this case, OR < 1 predicts a lower likelihood of colonization by CR-GNB when compared to the reference level, whereas OR > 1 predicts a greater chance of a positive test. Table 27 shows that Mechanical Ventilation use and drug use before the culture test increases the probability of acquisition. The patient with an admission source "Operation Room" has 1.04 times more likely to be a positive test than those admitted from Emergency (reference level). This method uses algorithms with built-in feature selection.

Taking the logistic regression model as an example, we illustrate in Figure 13 how a screening culture strategy would work during hospitalization.

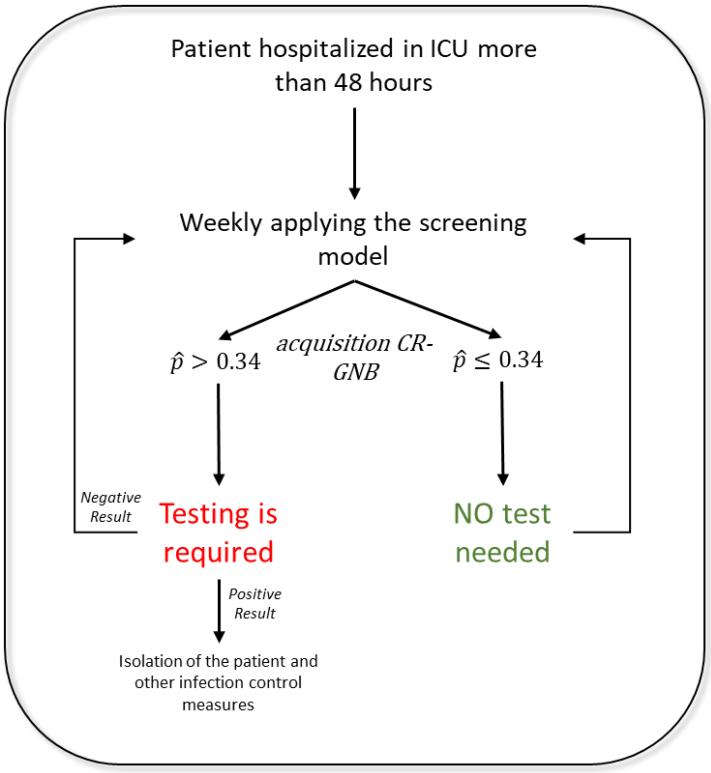


Figure 13 - Flowchart for a screening culture strategy.

Patients once classified as “testing is required” should do the screening. If the result is negative, the model must be applied again in the following week; if positive, the patient must be isolated.

5
Risk Model for the acquisition of CR-GNB

This Chapter presents the risk model development to estimate ICU patients' probability of acquiring CR-GNB, following the methodology presented in Chapter 3. We evaluate the likelihood of being colonized by measuring the predictions' calibration and Brier score. The best model is validated using data from other hospitals still not included in this work. In addition to the general model for all hospitals, we also develop an individual model for each hospital. We discuss the factors' importance and use association rule mining to identify those that often occurred together. Figure 14 summarizes the sections in this chapter.

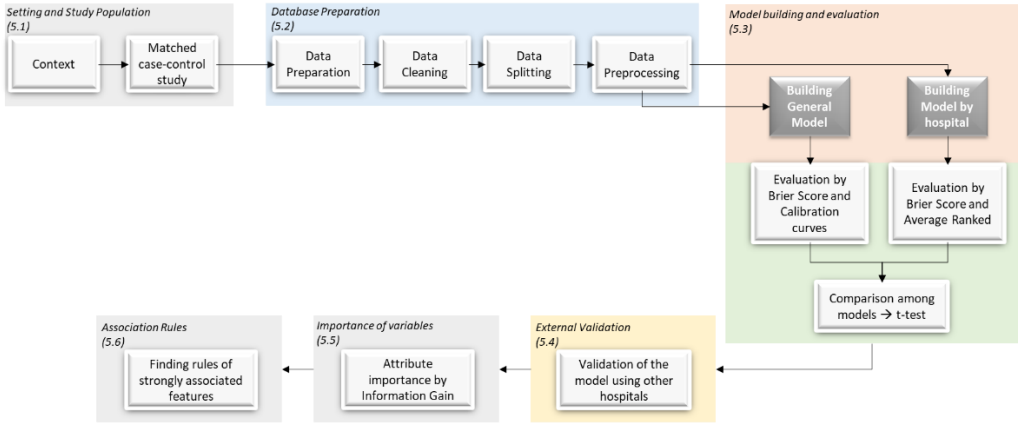


Figure 14 - Summary of this Chapter steps.

5.1.
Setting and study population

In addition to the screening tests, we also include other clinical exams (uroculture, blood culture, general culture) ordered by doctors during the inpatient hospitalization to detect CR-GNB. Thus, regardless of which test was performed, all patients tested for CR-GNB are included; that is, if an inpatient did no screening test but did a clinical exam to CR-GNB, this patient is considered. However, the rule that only the first episode of Carbapenem-Resistant Gram-negative bacterial

isolation is considered continues, i.e., we did not include in our selection any test made after isolation (positive test). Since the study unit is the patient, Hospital E is included in this analysis, totaling five hospitals.

We used the same inclusion criteria as before, namely: exams made in adult ICUs; in patients with admission date after May 8th, 2017; patients aged ≥18 years old; and tests realized between 48h and 60 days after patient admission. After applying these criteria, as shown in Appendix M, and considering only the first episode of isolation for each subject, we have a total of 527 positive and 7,462 negative exams between May 8th, 2017 and August 31st, 2019 from five hospitals. It results in 3,604 patients with at least one test with a minimum of one and a maximum of 16 tests per patient. There were 3,425 patients with only negative exams and 179 with only positive exams for CR-GNB during their hospitalizations. Table 28 shows the information for each hospital.

Table 28 - The number of patients and tests in each hospital.

Hospital	# Tests	# Negative Tests	# Positive Tests	# Patients	%Positive Patients	# Patients with negative tests	# Patients with only positive tests	# Maximum tests by a patient
A	404	341	63	214	29.4	187	27	13
B	1,039	971	68	469	14.5	444	25	12
C	1,540	1,452	88	611	14.4	586	25	16
D	3,849	3,616	233	1,658	14.1	1,583	75	16
E	1,157	1,082	75	652	11.5	625	27	11
All	7,989	7,462	527	3,604	14.6	3,425	179	-

We have about 14.6% of positive patients, ranging from 14.1% to 29.4%, depending on the hospital. This number is high when compared to published data. The World Health Organization (WHO, 2018) states that 7% of hospitalized patients in high-income countries acquire some infection (resistant or not) during hospitalization, raising this proportion to 10% in low-income countries.

Since there is a high difference between the hospitals and the general model, we also develop an individual model. The post-hoc for t-test identifies whether there is a statistically significant difference between the models.

In this Chapter, the unit of analysis is the patient, not the test. That said, we select only one test per patient. If the patient has more than one test, the selection is as follows: if all tests are negative, we randomly select only one; if the patient has positive and negative tests, we choose the first positive test. Figure 15 illustrates this approach.

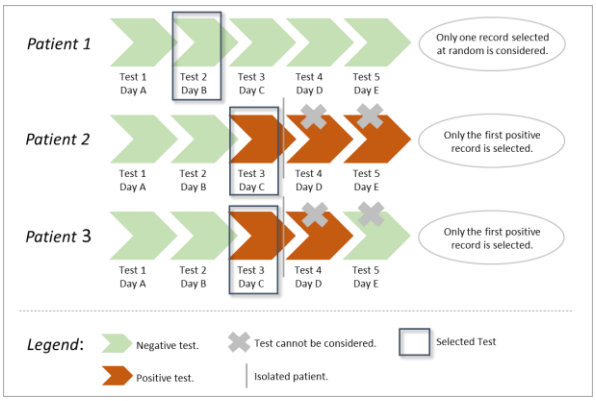


Figure 15 - Illustration of how the tests for each patient were selected.

Since the dataset is imbalanced, containing a smaller number of patients with positive antibiotic-resistant, we did a percentage reduction in the negative observations using a matched case-control design to compare the two groups of patients. Cases were defined as patients colonized or infected by Carbapenem-Resistant Gram-negative bacteria identified by a positive test. The controls were defined as patients who had no detection of MDRGN bacteria. They were randomly selected from potential controls matched by the hospital and admission date, altering the dataset to give an about 3:1 class distribution (control: case).

That said, we followed our analysis after the matched case-control selection with 2,070 data (527 positives and 1,543 negatives), according to Table 29. Hospital A was the only one that could not obtain a 3:1 distribution because they only had 151 different patients without a positive test.

Table 29 - The number of patients considered by the hospital after the matching process.

Hospital	# Patients	# Patients After Matched	# Positive/Case	# Negative/Control
A	214	214	63	151
B	469	272	68	204
C	611	352	88	264
D	1658	932	233	699
E	652	300	75	225
All	3604	2070	527	1543

5.2.
Database Preparation

Our dataset includes the variables previously described in Table 6 about the patient, ICU and hospital information, indexes (such as SAPS3 and Charlson), comorbidities, the use of invasive devices during hospitalization, and reasons for ICU admission, antibiotics use, and laboratory test. However, since we are

considering only one test per patient, the variables "test_before," "VesDURMORE," "VesTIMESMORE," "CVCDURMORE," "CVCTIMESMORE," "DiaDURMORE," "DiaTIMESMORE," "MVDURMORE," "MVTIMESMORE," "PerDURMORE," "PerTIMESMORE," "ArtDURMORE," and "ArtTIMESMORE" were excluded in this objective. These variables describe the difference between one test and another, so it does not make sense to remain. The variable "Hospital" was also excluded since it is used in the pairing. Therefore, we started our analysis with 98 of the 112 variables available.

The statistical information of the positive and negative patients for all the 98 features is presented in Appendix N. Table 30 summarizes the 51 significant variables from univariate analysis for a Confidence Interval (CI) of 90% ($p \leq 0.10$).

Table 30 - Descriptive statistical analysis comparing the Positive and Negative patients.

Variables	Negative (N=1543)	Positive (N=527)	P-value
Hospital Information			
LOS_hospital_before_test			
Mean (SD)	11.7 (11.2)	17.8 (12.6)	<0.001
Median [Min, Max]	8.00 [3.00, 60.0]	14.0 [3.00, 60.0]	
ICU Information			
LOS_ICU_before_test			
Mean (SD)	9.90 (10.5)	15.5 (12.0)	<0.001
Median [Min, Max]	6.00 [0, 60.0]	12.0 [0, 60.0]	
Index			
CharlsonIndex			
Mean (SD)	1.68 (1.85)	2.05 (2.01)	<0.001
Median [Min, Max]	1.00 [0, 11.0]	2.00 [0, 12.0]	
MFIpoints			
Mean (SD)	2.14 (1.38)	2.39 (1.47)	0.002
Median [Min, Max]	2.00 [0, 7.00]	2.00 [0, 7.00]	
Missing	50 (3.2%)	14 (2.7%)	
FrailPatientMFI			
NO	1303 (84.4%)	409 (77.6%)	<0.001
YES	240 (15.6%)	118 (22.4%)	
Saps3Points			
Mean (SD)	51.4 (12.8)	57.9 (13.7)	<0.001
Median [Min, Max]	51.0 [16.0, 104]	56.0 [23.0, 97.0]	
SofaScore			
Mean (SD)	1.71 (2.65)	2.79 (3.67)	<0.001
Median [Min, Max]	1.00 [0, 16.0]	1.00 [0, 17.0]	
Missing	390 (25.3%)	140 (26.6%)	
Priority			
Priority 1	167 (10.8%)	98 (18.6%)	<0.001
Priority 2	516 (33.4%)	127 (24.1%)	
Priority 3	1 (0.1%)	0 (0%)	
Priority 4	2 (0.1%)	0 (0%)	
Priority 5	10 (0.6%)	1 (0.2%)	
Missing	847 (54.9%)	301 (57.1%)	
Comorbidities			
ChronicHealthStatus			
Independent	905 (58.7%)	235 (44.6%)	<0.001
Need for assistance	360 (23.3%)	136 (25.8%)	
Restricted / bedridden	274 (17.8%)	152 (28.8%)	
Missing	4 (0.3%)	4 (0.8%)	
IsHematologicalMalignancy			
FALSE	1517 (98.3%)	509 (96.6%)	0.091
TRUE	22 (1.4%)	14 (2.7%)	
Missing	4 (0.3%)	4 (0.8%)	
IsSevereCopd			
FALSE	1405 (91.1%)	455 (86.3%)	0.006

Variables	Negative (N=1543)	Positive (N=527)	P-value
TRUE	134 (8.7%)	68 (12.9%)	
Missing	4 (0.3%)	4 (0.8%)	
IsAsthma			
FALSE	1497 (97.0%)	498 (94.5%)	0.032
TRUE	42 (2.7%)	25 (4.7%)	
Missing	4 (0.3%)	4 (0.8%)	
IsAngina			
FALSE	1455 (94.3%)	506 (96.0%)	0.057
TRUE	84 (5.4%)	17 (3.2%)	
Missing	4 (0.3%)	4 (0.8%)	
IsDeepVenousThrombosis			
FALSE	1479 (95.9%)	480 (91.1%)	<0.001
TRUE	60 (3.9%)	43 (8.2%)	
Missing	4 (0.3%)	4 (0.8%)	
IsChronicAtrialFibrilation			
FALSE	1335 (86.5%)	437 (82.9%)	0.082
TRUE	204 (13.2%)	86 (16.3%)	
Missing	4 (0.3%)	4 (0.8%)	
IsStrokeSequelae			
FALSE	1494 (96.8%)	483 (91.7%)	<0.001
TRUE	45 (2.9%)	40 (7.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsDementia			
FALSE	1283 (83.1%)	416 (78.9%)	0.055
TRUE	256 (16.6%)	107 (20.3%)	
Missing	4 (0.3%)	4 (0.8%)	
IsPepticDisease			
FALSE	1538 (99.7%)	519 (98.5%)	0.022
TRUE	1 (0.1%)	4 (0.8%)	
Missing	4 (0.3%)	4 (0.8%)	
IsHistoryOfPneumonia			
FALSE	1477 (95.7%)	484 (91.8%)	0.003
TRUE	62 (4.0%)	39 (7.4%)	
Missing	4 (0.3%)	4 (0.8%)	
Invasive Device during Hospitalization			
VesDURTOTAL			
Mean (SD)	4.64 (7.35)	10.2 (9.28)	<0.001
Median [Min, Max]	2.00 [0, 58.0]	8.00 [0, 52.0]	
VesTIMESTOTAL			
Mean (SD)	0.735 (0.808)	1.23 (0.861)	<0.001
Median [Min, Max]	1.00 [0, 7.00]	1.00 [0, 5.00]	
VESICAL			
NO	665 (43.1%)	81 (15.4%)	<0.001
YES	878 (56.9%)	446 (84.6%)	
ArtDURTOTAL			
Mean (SD)	2.51 (5.50)	7.26 (8.37)	<0.001
Median [Min, Max]	0 [0, 59.0]	5.00 [0, 53.0]	
ArtTIMESTOTAL			
Mean (SD)	0.437 (0.743)	1.05 (1.01)	<0.001
Median [Min, Max]	0 [0, 6.00]	1.00 [0, 5.00]	
ARTERIAL			
NO	1046 (67.8%)	185 (35.1%)	<0.001
YES	497 (32.2%)	342 (64.9%)	
DiaDURTOTAL			
Mean (SD)	0.804 (3.63)	2.29 (6.13)	<0.001
Median [Min, Max]	0 [0, 33.0]	0 [0, 42.0]	
DiaTIMESTOTAL			
Mean (SD)	0.107 (0.432)	0.326 (0.770)	<0.001
Median [Min, Max]	0 [0, 4.00]	0 [0, 6.00]	
DIALYSIS			
NO	1433 (92.9%)	420 (79.7%)	<0.001
YES	110 (7.1%)	107 (20.3%)	
CVCDURTOTAL			
Mean (SD)	4.55 (7.30)	10.6 (9.71)	<0.001
Median [Min, Max]	0 [0, 60.0]	9.00 [0, 54.0]	
CVCTIMESTOTAL			
Mean (SD)	0.655 (0.838)	1.36 (1.07)	<0.001
Median [Min, Max]	0 [0, 5.00]	1.00 [0, 6.00]	
CVC			
NO	815 (52.8%)	111 (21.1%)	<0.001
YES	728 (47.2%)	416 (78.9%)	
MVDURTOTAL			
Mean (SD)	2.30 (6.60)	7.95 (10.0)	<0.001

Variables	Negative (N=1543)	Positive (N=527)	P-value
Median [Min, Max]	0 [0, 57.0]	5.00 [0, 50.0]	
MVTIMESTOTAL			
Mean (SD)	0.260 (0.525)	0.765 (0.749)	<0.001
Median [Min, Max]	0 [0, 4.00]	1.00 [0, 5.00]	
MV			
NO	1195 (77.4%)	204 (38.7%)	<0.001
YES	348 (22.6%)	323 (61.3%)	
PerDURTOTAL			
Mean (SD)	1.31 (2.92)	0.962 (2.59)	0.006
Median [Min, Max]	0 [0, 29.0]	0 [0, 26.0]	
PerTIMESTOTAL			
Mean (SD)	0.512 (1.06)	0.361 (0.871)	0.005
Median [Min, Max]	0 [0, 9.00]	0 [0, 8.00]	
PERIPHERAL			
NO	1118 (72.5%)	411 (78.0%)	0.015
YES	425 (27.5%)	116 (22.0%)	
Reasons for ICU admission			
AdmissionSource			
Emergency	922 (59.8%)	253 (48.0%)	<0.001
Hemodynamic Room	17 (1.1%)	4 (0.8%)	
Operation Room	183 (11.9%)	63 (12.0%)	
Other ICU from hospital	139 (9.0%)	85 (16.1%)	
Others	12 (0.8%)	11 (2.1%)	
Semi Intensive Unit	80 (5.2%)	35 (6.6%)	
Transfer from another hospital	17 (1.1%)	17 (3.2%)	
Ward/Room	169 (11.0%)	55 (10.4%)	
Missing	4 (0.3%)	4 (0.8%)	
AdmissionReason			
Cardiovascular / Shock	402 (26.1%)	74 (14.0%)	<0.001
Elective Surgery	147 (9.5%)	39 (7.4%)	
Emergency surgery	63 (4.1%)	27 (5.1%)	
Endocrine / Metabolic / Renal	47 (3.0%)	13 (2.5%)	
Infection / Sepsis	499 (32.3%)	216 (41.0%)	
Liver and Pancreas / Gastrointestinal	90 (5.8%)	22 (4.2%)	
Neurological	131 (8.5%)	55 (10.4%)	
Non-surgical trauma	31 (2.0%)	14 (2.7%)	
Oncological / Hematological	34 (2.2%)	13 (2.5%)	
Others	22 (1.4%)	10 (1.9%)	
Respiratory	73 (4.7%)	40 (7.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsNeurologicalComaStuporObtundedDelirium			
FALSE	1269 (82.2%)	380 (72.1%)	<0.001
TRUE	268 (17.4%)	143 (27.1%)	
Missing	6 (0.4%)	4 (0.8%)	
IsNeurologicalSeizures			
FALSE	1489 (96.5%)	492 (93.4%)	0.006
TRUE	48 (3.1%)	31 (5.9%)	
Missing	6 (0.4%)	4 (0.8%)	
IsNeurologicalFocalNeurologicDeficit			
FALSE	1512 (98.0%)	496 (94.1%)	<0.001
TRUE	25 (1.6%)	27 (5.1%)	
Missing	6 (0.4%)	4 (0.8%)	
IsCardiovascularSepticShock			
FALSE	1434 (92.9%)	455 (86.3%)	<0.001
TRUE	103 (6.7%)	68 (12.9%)	
Missing	6 (0.4%)	4 (0.8%)	
Antibiotic use			
J01A			
FALSE	1471 (95.3%)	466 (88.4%)	<0.001
TRUE	72 (4.7%)	61 (11.6%)	
J01C			
FALSE	714 (46.3%)	143 (27.1%)	<0.001
TRUE	829 (53.7%)	384 (72.9%)	
J01D			
FALSE	813 (52.7%)	128 (24.3%)	<0.001
TRUE	730 (47.3%)	399 (75.7%)	
J01E			
FALSE	1503 (97.4%)	497 (94.3%)	0.001
TRUE	40 (2.6%)	30 (5.7%)	
J01F			
FALSE	1092 (70.8%)	305 (57.9%)	<0.001
TRUE	451 (29.2%)	222 (42.1%)	
J01G			

Variables	Negative (N=1543)	Positive (N=527)	P-value
FALSE	1483 (96.1%)	466 (88.4%)	<0.001
TRUE	60 (3.9%)	61 (11.6%)	
J01X			
FALSE	1138 (73.8%)	249 (47.2%)	<0.001
TRUE	405 (26.2%)	278 (52.8%)	
Antibiotic			
FALSE	306 (19.8%)	15 (2.8%)	<0.001
TRUE	1237 (80.2%)	512 (97.2%)	

Table 30 shows that positive patients have the highest severity indices, such as the Charlson Comorbidity Index, Saps 3 Points, MFI point, and Sofa Score. The hospital's length of stay is longer for positive (mean: 17.8; median: 14.0) than negative patients (mean:11.7; median: 8.0). The same happens for the length of stay in ICUs: median of 6 days for patients who did not acquire the bacteria and 12 days for those who did.

As seen earlier, the colonized patients by CR-GNB received more antibiotics and invasive devices and for a longer duration than those who non-acquired these pathogens. A more prolonged use time of mechanical ventilation, arterial, vesical, central venous, and hemodialysis catheters increases the probability of acquiring the pathogen. The peripheral catheter use, on the other hand, has an inverse relationship. Table 30 also shows the frequency of missing values for each variable.

We developed our analysis following the same framework of Figure 4, except for the data balancing step, since we decided to use the paired case-control study. The cleaning, splitting, and data pre-processing follow the same rules used in Chapter 4. We summarize the results of each step in Figure 16.

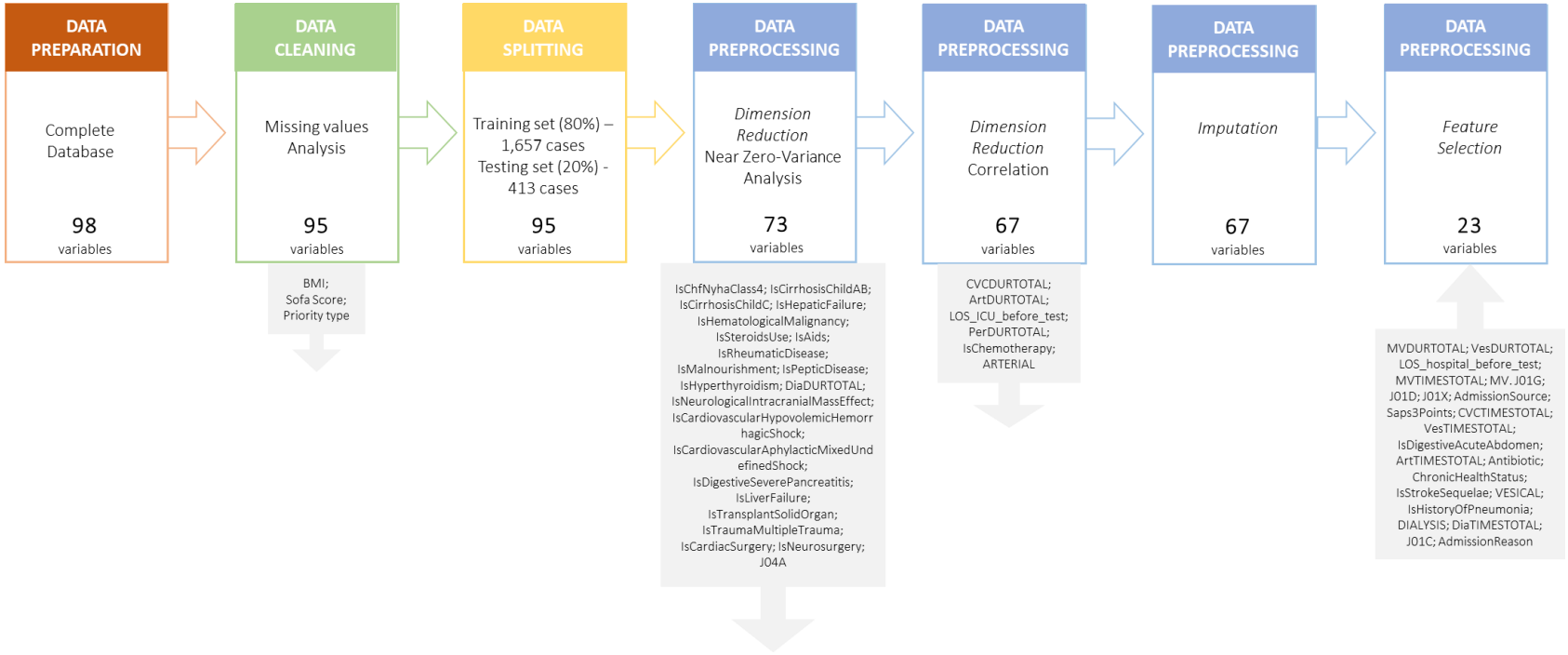


Figure 16 - Cleaning, Splitting, and Preprocessing data.

We removed three variables with more than 10% missing values, 22 after the near-zero variance analysis, and six due to correlation analysis. After imputation, we selected 23 of 67 the factors using Recursive Feature Elimination with random forest, as follows: MVDURTOTAL, VesDURTOTAL, LOS_hospital_before_test, MVTIMESTOTAL, MV, J01G, J01D, J01X, AdmissionSource, Saps3Points, CVCTIMESTOTAL, VesTIMESTOTAL, IsDigestiveAcuteAbdomen, ArtTIMESTOTAL, Antibiotic, ChronicHealthStatus, IsStrokeSequelae, VESICAL, IsHistoryOfPneumonia, DIALYSIS, DiaTIMESTOTAL, J01C, AdmissionReason. The new database with the 23 selected variables undergoes different normalization processes and transforming depending on the machine learning technique.

The data set has been divided into two parts (training and testing). We trained our model with 80% of the data (1,657 patients) and tested with 20% remaining data (413 patients).

5.3.
Model building and evaluation

During the training process, we implemented and compared the 16 different algorithms in

Table *7*, as follow: LR; LR with regularization; LDA; NSC; SVM linear and radial; NN; kNN; NB; decision trees (C4.5, CART, and C50); RF; GBM; Bagging; and AdaBoost.

We used grid-search hyperparameter optimization with 10-fold cross-validation to choose the best performing combination of hyperparameters (see Section 3.3.6.1) for each model. The hyperparameters that maximize prediction based on the Brier score metric are stored and used in the final training model. The lower the Brier score, the better the predictions are calibrated.

Once the training is concluded, the final models are applied to the features in the testing set. The model's overall performance is assessed by the average squared deviation between the predicted probability and the actual outcome via the Brier score and calibration curve. We divide the result into two subsections. Firstly, we

develop a general model for all hospitals. After that, we present a model for each one.

5.3.1.
General model

We ran all the algorithms over the training set by cross-validation. We analyzed the metrics estimates' average to avoid overfitting and evaluating whether the models can predict different subsets. Figure 17 shows the boxplot to represent the Brier score median and extreme values. The low interquartile value of the models for the Brier score confirmed that there was no overfitting. The best hyperparameters values, other data representation, and boxplots to the other metrics can be seen in Appendix O.

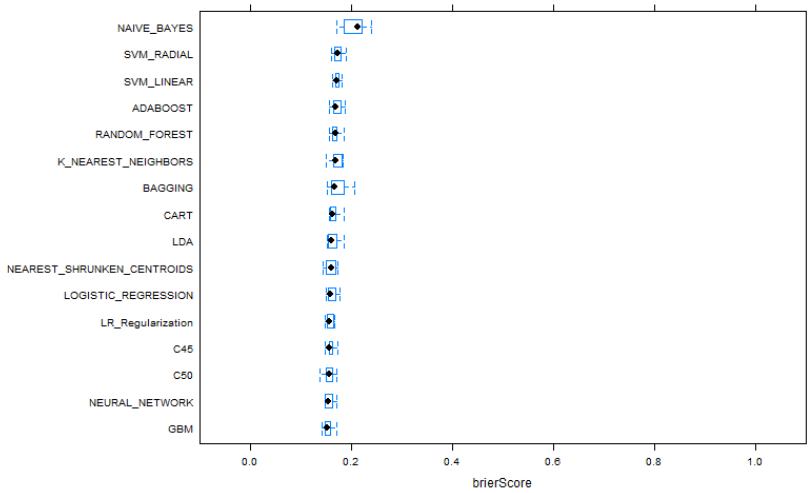


Figure 17 - Boxplots representing the Brier score from the cross-validation process for each method.

Since our goal is to estimate the acquisition probability, not to classify, we use our trained models and the testing set to evaluate the Brier score and compare techniques' calibration performance, which measures the consistency between observed outcome and estimated probability. Prediction performance results are shown in Table 31. The MCC values were also computed, but only for interpreting the model.

Table 31 - Brier score and MCC of all 16 methods. The best values of the Brier score are highlighted.

Methods	Brier score	MCC
LOGISTIC_REGRESSION	0.163	0.338
LR_Regularization	**0.155**	0.318

Methods	Brier score	MCC
LDA	**0.159**	0.327
NEAREST_SHRUNKEN_CENTROIDS	**0.152**	0.327
SVM_LINEAR	0.177	0.345
NEURAL_NETWORK	0.160	0.335
SVM_RADIAL	0.171	0.109
K_NEAREST_NEIGHBORS	0.173	0.296
NAIVE_BAYES	0.196	0.339
C45	0.165	0.383
CART	0.167	0.379
C50	0.160	0.399
RANDOM_FOREST	0.176	0.326
GBM	**0.159**	0.312
BAGGING	0.183	0.308
ADABOOST	0.172	0.295

We can see that the Nearest Shrunken Centroids (0.152), Logistic Regression with regularization (0.155), Linear Discrimination Analysis (0.159), and Gradient Boosting Machine (0.159) have the best/lowest Brier scores.

The Naïve Bayes technique has better discrimination power (MCC=0.339) than others but has the biggest Brier score value (0.196). This technique can be appropriate for classifying a patient as colonized and non-colonized but does not perform very well to estimate the probability.

Interestingly the various machine learning algorithms yielded close Brier score. Still, since we want to predict the associated probability, we need to know if this probability gives us confidence in the prediction. Not all classifiers provide well-calibrated probabilities, some being over-confident, while others being under-confident (PEDREGOSA et al., 2011). Thus, a calibration analysis of predicted probabilities is often desirable as postprocessing. Figure 18 illustrates the calibration belt of the best model. The plot is obtained by plotting the predicted CR-GNB acquisition risk against the observed cases.

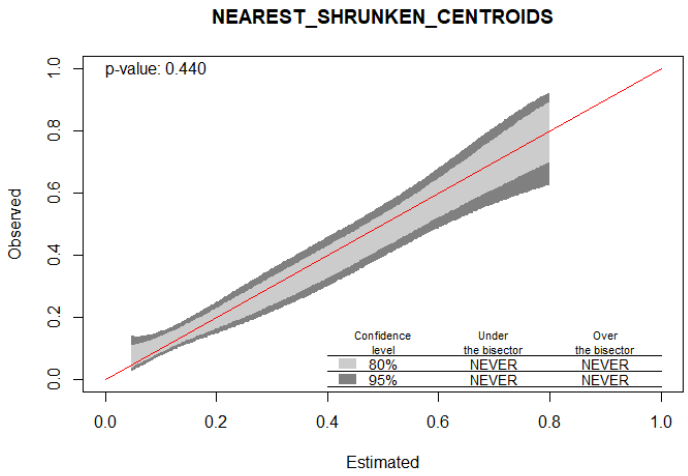


Figure 18 - Calibration belts for the Nearest Shrunken Centroids at two confidence levels. CI:0-80% (light shaded area) and CI:0-95% (dark shaded area).

The better calibrated the prediction, the closer the curve appears along the main diagonal. The position of the curve relative to the diagonal helps us to interpret the reliability of the probabilities. Points below the line present probabilities are too large, overestimated (Observed < Estimated). On the other hand, points above the line show probabilities too small, underestimated (Observed > Estimated).

The table on the bottom-right side of the figure reports the ranges of the predicted probabilities. To estimate the degree of uncertainty around the calibration curve, we must compute the curve's confidence belt. The belt significantly deviates from the bisector to the 80% and 95% confidence level. The values that appear in this table present the range estimated probabilities outside confidence level; that is, the model significantly overestimates (under the bisector) or underestimated (over the bisector) the colonization risk patients. The word "NEVER" means that there is no deviation outside the confidence level. The model's overall calibration is synthesized into the likelihood ratio test´s p-value, reported in the figure's top-left corner. Accordingly, the p-value suggests if the calibration of the model is acceptable or not.

For the NSC, no evidence of the lack of calibration emerges from the calibration belt. The belt covers almost the whole diagonal line, and the likelihood-ratio test gives a p-value of 0.440, suggesting that the model calibration on the development is acceptable. Considering a confidence level of 95%, we cannot reject

the hypothesis that the model is calibrated, and the fit of the curve is good. It encompasses the bisector about in the 0-0.8 range.

We performed the same calibration belt analysis for the other models. Our goal is to understand if they are also suitable for prediction. Table 32 presents the p-values and the ranges of the predicted probabilities where the belt significantly deviates from the bisector. The calibration belt plots can be seen in Appendix P.

Table 32 - P-values and the ranges of the predicted probabilities where the belt significantly deviates from the bisector (under and over) to the 80% and 95% confidence level using the testing set, in decreasing order of p-value.

Methods	p-value	Confidence level	Under the bisector	Over the bisector
NSC	0.440	80%	NEVER	NEVER
		95%	NEVER	NEVER
GBM	0.384	80%	NEVER	NEVER
		95%	NEVER	NEVER
CART	0.141	80%	NEVER	NEVER
		95%	NEVER	NEVER
LR	0.124	80%	NEVER	NEVER
		95%	NEVER	NEVER
LR regularized	0.097	80%	NEVER	0.43-0.66
		95%	NEVER	NEVER
LDA	0.074	80%	NEVER	NEVER
		95%	NEVER	NEVER
SVM RADIAL	0.028	80%	NEVER	0.08-0.15
		95%	NEVER	0.08-0.11
C45	0.014	80%	0.67-0.79	0.16-0.48
		95%	0.74-0.79	0.19-0.32
NN	0.005	80%	NEVER	0.02-0.10
		95%	NEVER	0.02-0.08
ADABOOST	<0.001	80%	0.13-0.69	NEVER
		95%	0.18-0.61	NEVER
C50	<0.001	80%	NEVER	0.01-0.04
		95%	NEVER	0.02-0.03
kNN	<0.001	80%	NEVER	0.00-0.14
		95%	NEVER	0.00-0.07
RF	<0.001	80%	0.60-0.92	0.00-0.37
		95%	0.70-0.92	0.00-0.33
BAGGING	<0.001	80%	0.31-1.00	0.00-0.15
		95%	0.35-0.98	0.00-0.10
SVM LINEAR	<0.001	80%	0.14-0.21/0.60-0.79	0.23-0.54
		95%	0.14-0.21/0.62-0.79	0.24-0.52
NB	<0.001	80%	0.48-1.00	0.00-0.35
		95%	0.52-1.00	0.00-0.32

According to the likelihood-ratio test, the NSC (p=0440), GBM (p=0.384), CART (p=0.141), LR (p=0.124), LR regularized (p=0.097), and LDA (p=0.074) are calibrated models, suitable for prediction, considering a confidence level of 95%; hence these methods could be used as predictive tools to these hospitals. Looking at the LR regularized, we can see that the model is acceptable in almost the whole range but lacks calibration over the bisector in the 0.43-0.66. The conclusion is that the model slightly underestimates the acquisition of medium-risk patients.

In general, the NSC is the best model to estimate CR-GNB acquisition risk. It presents the lowest Brier score, one of the largest MCC, comprises almost the whole of the diagonal, has low uncertainty, and the highest p-value.

We can see that the likelihood-ratio tests of the other methods (SVM radial, C45, NN, Adaboost, C50, KNN, RF, Bagging, SVM linear, and NB) give a p-value less than 0.05, suggesting that the models are unacceptable and miscalibrated. They are the ones with the worst Brier score values. The deviates from the bisector show the models' heavy overestimates and underestimates colonization in patients with low, medium, and high risk, and the belt does not cover most of the diagonal line (Appendix P).

For example, NB, Bagging, and RF overestimate the colonization for medium and high-risk patients and underestimates low-risk patients. The SVM linear model overestimates the risk for low and high-risk patients and underestimates medium-risk patients. In addition to presenting the worst Brier score, the NB is out almost the whole diagonal line (over 0.00-0.35/under 0.48-1.00).

Although these models have good discrimination according to the MCC, they have the worst Brier score and an inadequate calibration curve. These models can be appropriate for classifying a patient but not to estimate the acquisition probability.

Figure 19 compares how well NB and NSC's probabilistic predictions are estimated to understand the difference between the worst and best models. The x-axis represents the predicted probability in each bin. The y-axis is the frequency of likelihood.

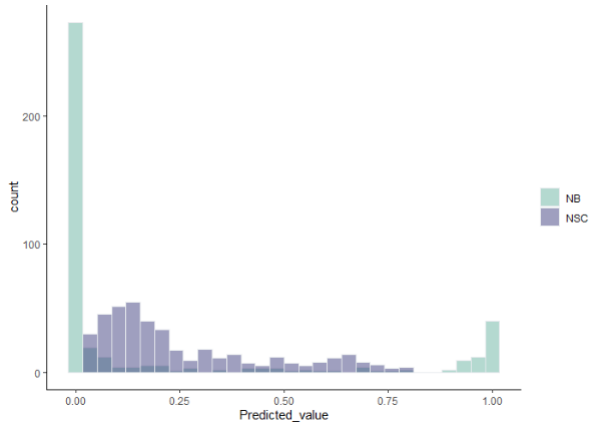


Figure 19 - Predicted values by NB and NSC.

We can see from Figure 19 that the NB model estimates many patients with a zero or one probability, while the NSC has more balanced values ranging from 0.10 to 0.8. Under certain conditions, NSC's soft shrinkage is equivalent to a LASSO penalty (CHOI; BAIR; LEE, 2017). On the other hand, Naive Bayes tends to push probabilities to 0 or 1, mainly because it assumes that features are conditionally independent. However, some factors may interfere with others, returning bad-calibrated predictions (PEDREGOSA et al., 2011).

External validation of the model in other settings will be developed in section 5.4. After that, the prognostic factors influencing the CR-GNB acquisition will be present in Section 5.5. and they are validated by association rules in Section 5.6. They are useful and could be translated into decision support tools in the medical domain.

5.3.2.
Model by hospital

Since we have a high difference in the proportion of the positive tests among hospitals, we decided to develop individual models. Thus, the data were stratified by hospital, and the process to conduct a machine learning analysis was repeated to build each model. We added a summary of the clinical characteristics of patients for each hospital in Appendix Q.

The objective is to understand if the built general model can be used for all hospitals and if there is a significant difference between it and the five individual models through t-test.

We used the average ranked (AR) performances of the classification techniques on each data set to compare the different classifiers.

Table 33 reports the Brier score and Table 34, the Average Ranked (AR) performances of all 16 algorithms on the five data sets, and general data. The lowest Brier score and AR on each data set are highlighted.

Table 33 - Brier score result for each model using test data.

BRIER SCORE	Hospital A	Hospital B	Hospital C	Hospital D	Hospital E	All hospitals	Mean

						(General Model)	
LOGISTIC_REGRESSION	0.082	0.160	0.154	0.177	0.194	0.163	0.155
LR_Regularization	0.103	0.158	0.143	0.157	0.185	0.155	0.150
LDA	0.082	0.162	0.153	0.176	0.203	0.159	0.156
NEAREST_SHRUNKEN_CENTROIDS	0.118	0.164	0.131	0.151	0.161	0.152	**0.146**
SVM_LINEAR	0.107	0.148	0.157	0.167	0.194	0.177	0.158
NEURAL_NETWORK	0.089	0.158	0.143	0.155	0.176	0.160	**0.147**
SVM_RADIAL	0.129	0.160	0.167	0.175	0.191	0.171	0.166
K_NEAREST_NEIGHBORS	0.157	0.163	0.138	0.163	0.193	0.173	0.165
NAIVE_BAYES	0.172	0.218	0.180	0.218	0.199	0.196	0.197
C45	0.145	0.163	0.166	0.167	0.179	0.165	0.164
CART	0.117	0.163	0.186	0.158	0.188	0.167	0.163
C50	0.108	0.157	0.136	0.159	0.191	0.160	0.152
RANDOM_FOREST	0.118	0.178	0.138	0.151	0.194	0.176	0.159
GBM	0.098	0.151	0.146	0.150	0.195	0.159	0.150
BAGGING	0.114	0.197	0.184	0.176	0.211	0.183	0.178
ADABOOST	0.163	0.187	0.166	0.180	0.191	0.172	0.176
Mean	**0.119**	**0.168**	**0.155**	**0.168**	**0.190**	**0.168**	**0.161**

Table 34 - Average Ranked for each model using test data.

ORDER	Hospital A	Hospital B	Hospital C	Hospital D	Hospital E	All hospitals (General Model)	AR
LOGISTIC_REGRESSION	2	7	9	14	11	7	8.3
LR_Regularization	5	5	5	5	4	4	**4.7**
LDA	**1**	8	8	12	15	15	9.8
NEAREST_SHRUNKEN_CENTROIDS	11	12	**1**	2	**1**	**1**	**4.7**
SVM_LINEAR	6	**1**	10	9	10	10	7.7
NEURAL_NETWORK	3	4	6	4	2	2	**3.5**
SVM_RADIAL	12	6	13	11	7	7	9.3
K_NEAREST_NEIGHBORS	14	9	4	8	9	9	8.8
NAIVE_BAYES	16	16	14	16	14	14	15.0
C45	13	10	12	10	3	3	8.5
CART	9	10	16	6	5	5	8.5
C50	7	3	2	7	8	8	5.8
RANDOM_FOREST	10	13	3	3	12	12	8.8
GBM	4	2	7	**1**	13	13	6.7
BAGGING	8	15	15	13	16	16	13.8
ADABOOST	15	14	11	15	6	6	11.17

Hospital A was the one that obtained the best Brier score mean (0.119) among all techniques, followed by Hospital C (0.155), B (0.168), D (0.168), and E (0.190). The average of the general model was 0.168, near to most models.

Table 34 informs which reference technique is the best for that specific data set in ascending order. Thus, the best models (Order=1) to predict colonization/infection were as follows: LDA (Hospital A), Linear SVM (Hospital B), NSC (Hospital C, E and All), and GBM (Hospital D). Some techniques worked very well for a specific data set but not for another, such as the LDA, which presented the best result for Hospital A but was close to the worst for Hospital E. Figure 20 shows the ranges and the Average Ranked (AR) for each technique.

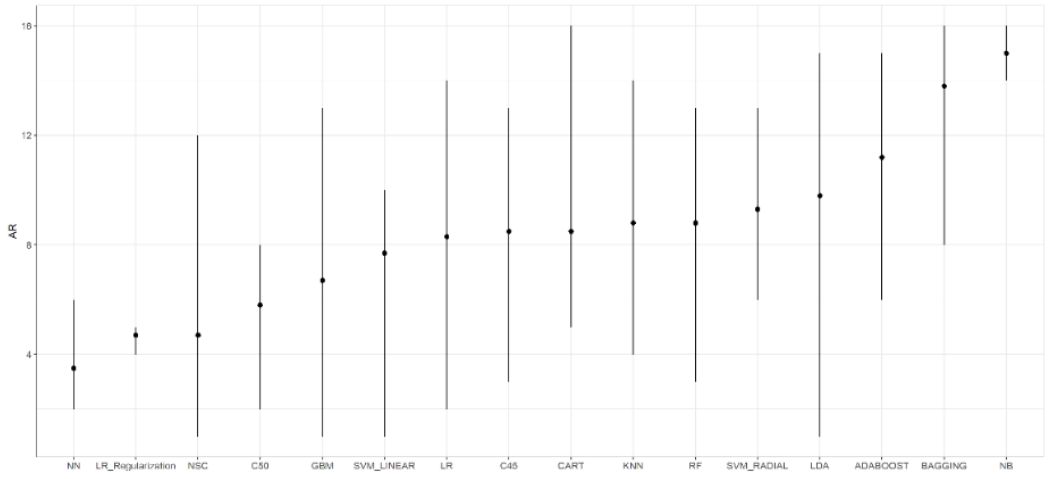


Figure 20 - Classifiers´ mean rank across datasets. The point corresponds to the AR.

These experiments show that the Neural Network seemed to have good results for the individual model, adding the best mean AR among the techniques. The regularized LR, NSC, C50, and GBM, which had previously been the best techniques for the general model, also performed well in most datasets, presenting good average ranks. Moreover, the method NB obtained the worst AR, confirming the results meeting in the general model. Remembering that the lower the average ranked, the better the overall performance of the technique.

The t-test evaluates the null hypothesis that there is no difference between the models. We chose the developed models with the lowest Brier scores and compared them using the training set and cross-validation. The t-test result can be seen in Table 35. The upper diagonal values estimate the difference and the lower diagonal the p-value for {H0: difference = 0}. Since the p-value is about 1, we can affirm no discrepancies between the models by Brier scores. Therefore, we can use the general model for all hospitals without losing performance, being adequate to predict the probability of being colonized.

Table 35 - Comparison of models using t-test.

Best Model		ALL HOSPITALS	A	B	C	D	E
NSC	**ALL HOSPITALS**		-0.0004964	-0.010764	0.0083301	0.0026707	0.0010038
LDA	**A**	1		-0.010268	0.0088265	0.0031671	0.0015002
SVM Linear	**B**	1	1		0.0190945	0.0134351	0.0117682
NSC	**C**	1	1	1		-0.0056594	-0.0073263
GBM	**D**	1	1	1	1		-0.0016669
NSC	**E**	1	1	1	1	1	

If we compared the NB instead of the NSC, we had a significant difference for all hospitals with p-values <0.10.

We also used the t-test to analyze the difference between the methods for each hospital. Hospitals "A," "B," and "C" showed no significant difference between the models. However, some methods have a difference in Hospital D, as follows: Naive Bayes and GBM; c50 and radial SVM; AdaBoost and radial SVM. CART is statistically different from logistic regression, SVM linear, NN, and KNN for Hospital E.

5.4. External Validation

Our final risk model for the acquisition of CR-GNB is the NSC. As previously stated, it presents the lowest Brier score, and the calibration belt covers almost the whole diagonal line, suggesting that the model is acceptable and could be used as predictive tools to the hospitals. However, this model was trained and well-calibrated for the specific set of five hospitals.

We performed an external validation using data from two other hospitals in the same network, named Hospital F and Hospital G. We extracted and prepared them using the same selection criteria as before: exams made in adult ICUs; in patients with admission date between May 8th, 2017 and August 31st, 2019; patients aged ≥18 years old; and tests realized between 48h and 60 days after patient admission, considering only the first episode of isolation for each subject.

External validation of the final NSC model was developed for each hospital, and the information and results can be seen in Table 36, along with the general model information. We have a total of 624 different patients (73 positives and 551 negatives). Hospital F has 14.6% of CR-GNB positive and Hospital G has 9.52%.

Table 36 – Information and results of external validation.

Hospital	# Patient	#Positive	#Negative	% Positive Tests	Brier Score	MCC
F	267	39	228	14.61%	0.128	0.261
G	357	34	323	9.52%	0.079	0.261

According to Table 36, the NSC model does not classify well the non-acquisition of CR-GNB (MCC = 0.261). On the other hand, if we compared the Brier score (0.128 and 0.079), we know that this model can predict the probability of acquiring CR-GNB from both hospitals. Following the acquisition risk model

objective, the overall performance is assessed by the likelihood of being colonized/infected via the Brier score. To graphically represent the model's goodness of fit, we constructed the calibration belts following the same approach (see Figure 21).

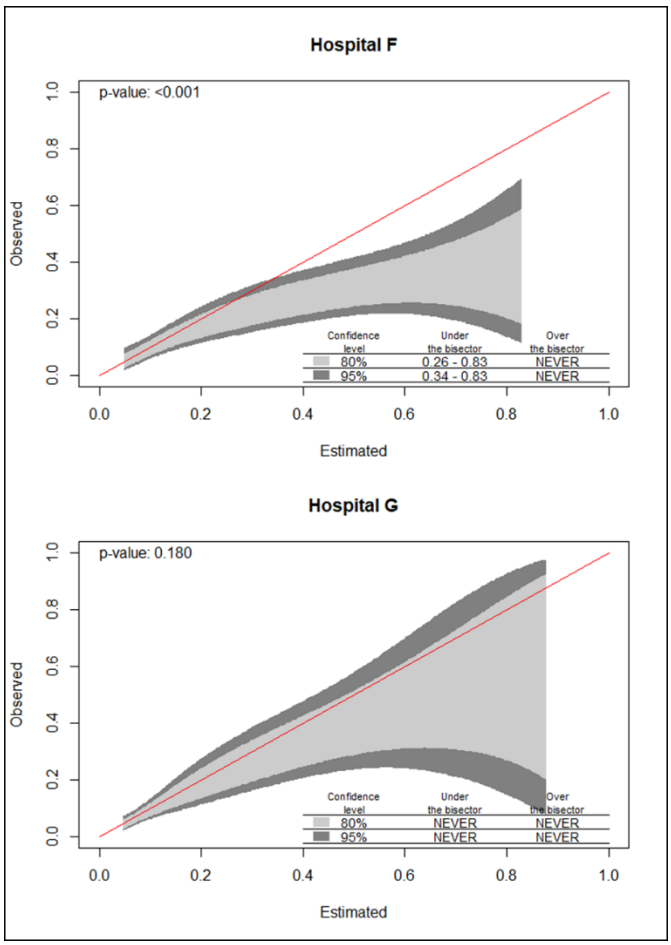


Figure 21 - Calibration belts for the Nearest Shrunken Centroids at two confidence levels. CI:0-80% (light shaded area) and CI:0-95% (dark shaded area) using external validation data.

We can see that the likelihood-ratio tests give a p-value less than 0.01 to Hospital F and 0.18 to Hospital G, suggesting that the model is miscalibrated to application in the first hospital and is acceptable to the other. The deviates from the Hospital F calibration belt bisector show the model overestimates colonization in patients, and the belt does not cover most of the diagonal line. On the other hand, Hospital G does not present deviation outside the confidence level.

We concluded that the model is well-calibrated and acceptable to be introduced at Hospital G. However, for Hospital F, although the model also has a

low Brier score value, it overestimates the colonization of patients, suggesting a need for better calibration for future implementation.

5.5.
Importance of variables

After evaluating the models, we identified the attribute importance by Information Gain to assess the attributes' power in predicting CR-GNB acquisition. It measures the expected reduction in uncertainty (entropy), calculating the class's degree of purity, and considering only the feature and the class.

Firstly, we analyze the important attributes using the database with all hospitals. Figure 22 shows the top 20 risk factors, according to their information gain.

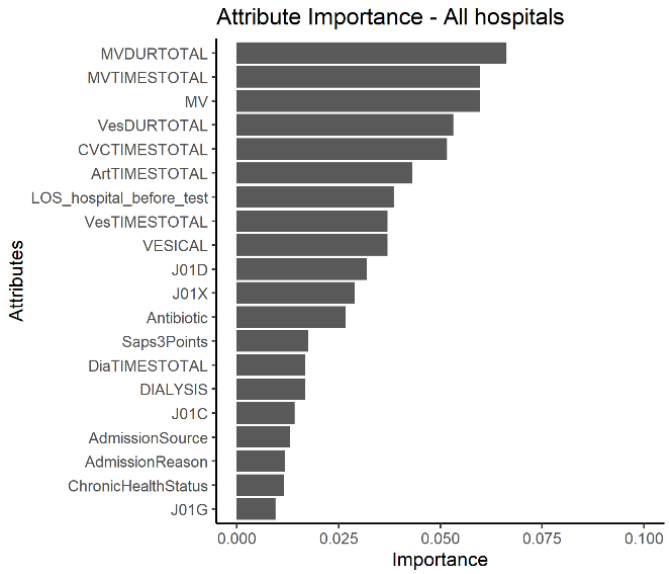


Figure 22 - Top 20 attributes ranked by their Information Gain for all hospitals.

The variables related to the duration of the use of invasive devices, especially mechanical ventilation, and the number of times this dispositive was changed before the test are essential. The antibiotics groups are also important features, including the J01D, J01X, J01C, and J01G families. The criticality index Saps3, admission reason, and admission source obtained good information gains. The length of stay before the test also has high predictive power. On the other hand, comorbidity and other criticality indexes do not have a high capacity to predict CR-GNB acquisition.

To ensure that we considered the most critical variables in our model, we also did the information gain analysis for the complete base of 67 variables before feature selection. Of the top 20 variables of the whole database, 19 appear in Figure 22, except CVC. It may have lost its importance when adjusting to other variables. Thus, we can conclude that our factors selected on preprocessing were the ones with the most significant gain in information. The list of importance can be found in Appendix R.

We also performed the ranking of the attributes of each hospital by IG. Figure 23 shows the ten most discriminative features for each hospital.

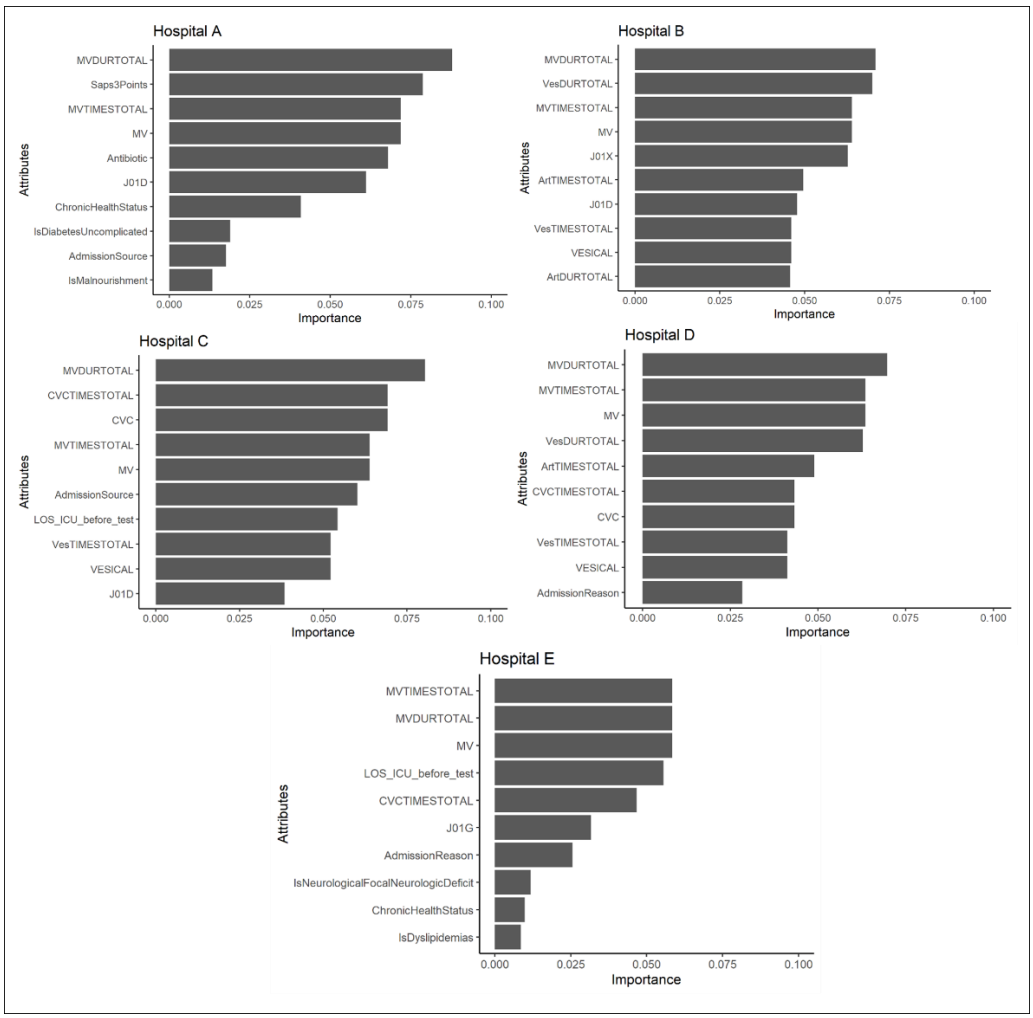


Figure 23 - Top 10 attributes ranked by their Information Gain.

The total duration of mechanical ventilation use is the most critical variable for all hospitals. Some features appear on the top 10 of all of them, such as mechanical ventilation use and the number of times this dispositive was changed before the test. The bladder catheter, use of antibiotics of the J01D family,

admission source, admission reason, CVC, and length of stay before test also appeared as the most important in at least two of the five hospitals. It is noticed that the most important attributes are associate with the use of some invasive device. Comorbidities are not among the most important factors. A complete list of the importance of variables per hospital can be found in Appendix R.

5.6. Association Rules

The variables importance analysis tells us which variables, alone, have the most significant predictive power but not if there is an important association between them. Therefore, we decided to apply association rules mining to find rules of strongly associated features in our data that indicate that a patient is at risk of being colonized.

Firstly, we discretized the dataset, converting numeric vectors into factors and the database to transactions for creating items. Once itemsets are obtained, we extracted a list of 49,161 rules. Since it is impossible to analyze all these rules, we selected some criteria for extracting and interpreting the best ones. Moreover, some of these rules have complementary or repeated conditions.

Since our goal is to find rules strongly associated with the at-risk CR-GNB acquisition, we specified the labels of the items for predictive a positive patient ({Condition} → {RESULT=pos}). Moreover, we considered only rules with support (frequency) more significant than 10% of cases, the confidence more than 0.5, and a length of more than two variable-value pairs. That said, we extracted a list of 157 association rules with predictive value "positive," seen in Appendix S. Table 37 presents the top 20 rules (conditions) and their respective measures, ordered by the lift.

Table 37 - List of the 20 rules with higher lift generated from the association rule mining.

#	Rules	Support	Confidence	Lift
1	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
2	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.100	0.575	2.257
3	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
4	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01D=TRUE} => {RESULT=pos}	0.100	0.575	2.257
5	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
6	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.100	0.575	2.257
7	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
8	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01D=TRUE} => {RESULT=pos}	0.100	0.575	2.257
9	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257

#	Rules	Support	Confidence	Lift
10	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.100	0.575	2.257
11	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
12	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE} => {RESULT=pos}	0.100	0.575	2.257
13	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
14	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.100	0.575	2.257
15	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
16	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE} => {RESULT=pos}	0.100	0.575	2.257
17	{MVDURTOTAL=[4,57],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
18	{MVDURTOTAL=[4,57],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
19	{MVDURTOTAL=[4,57],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
20	{MVDURTOTAL=[4,57],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234

In Table 37, the 16 first rules have the highest confidence level (0.575) and lift (2,257). This means that, for example, in 57.5% of cases in which the patient has the condition "{MVDURTOTAL=[4,57] + VesDURTOTAL=[6,58] + J01D=TRUE + Antibiotic=TRUE + VESICAL=YES}", they are positive. Moreover, since the positive case proportion is ~25.5%, the expected confidence is 0.255. Thus, the lift (#1) is 2.257 (0.575/0.255): the more lift, the better the rule. A lift value greater than 1 indicates that the rule appears more often together than expected. It means that the condition's occurrence has a positive effect on the occurrence of a positive result. This rule appears in 10% of cases (Support = 0.10). We can conclude that whether a patient is hospitalized in the ICU with these conditions, this patient has a 57.5% probability of acquisition; and just five variables can give us relevant information.

When analyzing the 157 rules, we noticed that some have almost the same conditions, with the same lift and confidence. For example:

- {MVDURTOTAL=[4,57], VesDURTOTAL=[6,58], J01D=TRUE, Antibiotic=TRUE, VESICAL=YES} => {RESULT=pos}
- {MVDURTOTAL=[4,57], VesDURTOTAL=[6,58], J01D=TRUE, Antibiotic=TRUE} => {RESULT=pos}

These rules have the same confidence of 0.575. The addition of the VESICAL variable did not change the rule's reliability. Thus, we can use the rule without this variable and have the same result.

We can understand the high importance of some variables that appear in Figure 22. For example, 96.2% of the rules include MVDURTOTAL or VESDURTOTAL, and 86.0% the antibiotic use. All the conditions selected include some information about invasive dispositive use, showing the relevance of these

variables in the decision criterion. These strongly associated features enable an indication of patterns that can help clinicians in the decision-making process. Regarding antibiotics use, the strong predictive power can be explained by early clinical diagnosis since the provider can start using antibiotics before confirming the infection by clinical examination. This will be discussed in future studies.

We also evaluate the similarity between items transforming the correlation into distances. The dendrogram in Figure 24 indicates the (dis)similarity between observations. It reinforces the associations shown in Appendix S. The higher the height of the fusion, the less similar it is.

Some antibiotics, such as J01D and J01C, have a high correlation (close distance). It means that the use of these antibiotics is common for the same patient. Regarding the use of an invasive device, we realized the proximity between the use of mechanical ventilation, its high duration (MVDURTOTAL=[4,57]), and device change (MVTIMESTOTAL=[1,5]). The high “Saps3 [58, 104]” and the admission reason "Infection/Sepsis" also appear together in the dendrogram, as well as a medium “Saps3 [48, 58[” and the admission source "Emergency." A low “Saps3 [16, 48[” is strongly correlated with independent chronic health status. Some variables are not close to others, such as antibiotics and admission sources or chronic health status.

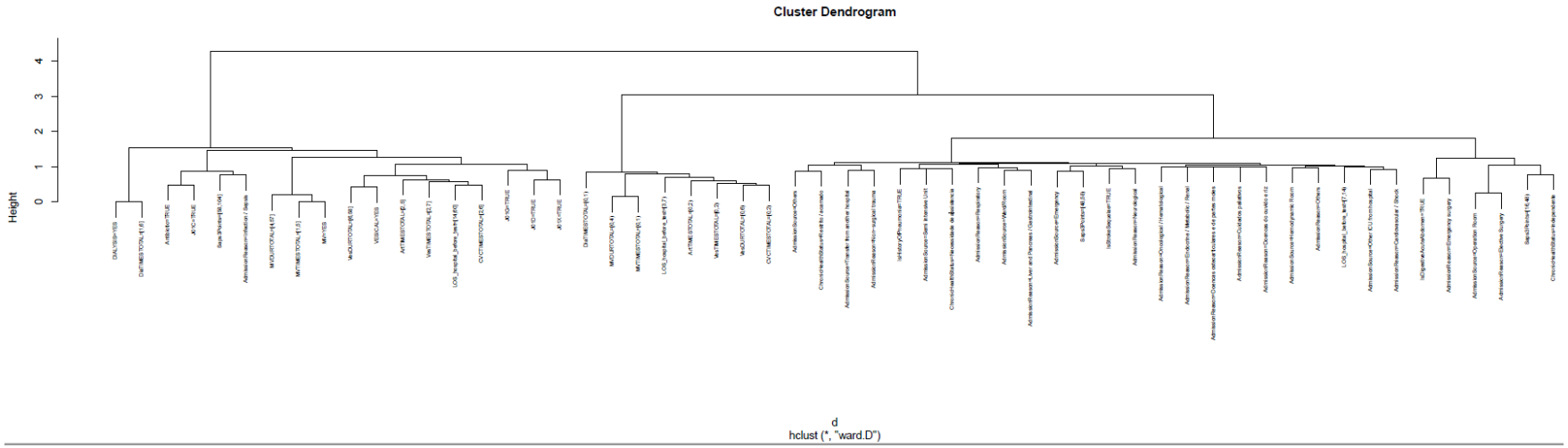


Figure 24 - Dendrogram: similarity between items.

6
Discussion

Carbapenem-Resistant Gram-Negative Bacteria is a global threat and a major concern for infection control around the world. Early diagnosis of colonization by these pathogens can help avoid transmission risks and decrease future infection rates. Some Brazilian hospitals perform weekly culture tests in all inpatients, known as a screening process. However, despite the benefit of screening, there is an increase in hospital costs and laboratory waiting times. Thus, this thesis aimed to develop a comprehensive and systematic approach to apply machine-learning techniques to build screening models that detect unneeded tests. In addition to the screening model, we developed a risk model that estimates ICU patients' probability of acquiring CR-GNB. We also identify the factors and rules of strongly associated features that indicate that a patient is at risk.

In short, we proposed the hospital's decision-maker two screening models, one more conservative and the other moderate, a risk model for the acquisition of CR-GNB, and an overall insight into the most significant factors. These predictive models can be included in the hospital system. The risk factors and association rules can increase the discussion on the topic and help clinicians make decisions. The predictors used are available in the clinical setting, ensuring that these variables' values can be collected in the moment of clinical decision.

Recently, there has been a myriad of information about the mechanism of resistance to carbapenems and articles reviewing specific agents. However, although previous studies have tried to find factors associated with MDR-GNB or CR-GNB and some methods have already proven their efficiency in applying health services based on the EHR, we have not seen research well-structured in this area using machine learning techniques or evaluation methods. Few original studies considered acquisition predictors in low- and middle-income countries' (LMIC)

hospitals. We did not find any CR-GNB screening models applying ML techniques considering all weekly culture tests or duration variables between the two tests.

Generally, in the healthcare literature, the existing analysis methods for predicting are poorly interpretable. Our analysis adds to the current studies in four respects: machine learning techniques, balancing strategies, feature selection, and performance evaluation. We introduced data mining and machine-learning concepts within a context that medical researchers find familiar and accessible, presenting a methodological framework applicable in other settings.

In the following subsections, we will summarize the most fundamental findings of the thesis, followed by comparing works existing in the literature. Finally, we will include limitations, future researches, final considerations, and a list of my publications.

6.1. Main Findings

Our database gathers the patients, antibiotic, and microbiology data from five Brazilian hospitals from May 8th, 2017 to August 31st, 2019, involving hospitalized patients in 24 adult ICUs. Information from the laboratory was used to identify all patients with a positive or negative test for carbapenem-resistant GNB, *A. baumannii, P. aeruginosa*, or Enterobacteriaceae. We have a total of 539 positive and 7,462 negative tests, resulting in 3,604 patients with at least one exam after 48 hours hospitalized.

In the five hospitals analyzed, about 14.6% are positive patients, ranging from 14.1% to 29.4%, depending on the hospital. The culture-positive test varies from 7.1% to 19.4%. The WHO (WHO, 2018) states that about 10% of patients acquire some infection during hospitalization in low-income countries.

Aiming to deal with the highly imbalanced dataset for the screening problem, we proposed to apply an approach based on class decomposition, combining feature selection and cluster techniques, the Class Decomposition with random forest (D.RF). This method was the second best to our aim, losing to Recursive Feature Elimination with random forest (RF-RFE). However, it was the best method to discriminate the positive classes and comparing the Sensitivity. These two

algorithms achieved results better than those that include almost all variables. Thus, models trained with a limited number of features can obtain a good evaluation.

We performed machine learning techniques to find suitable models and compare their performance over various balancing strategies. Friedman's test and Nemenyi's post hoc tests were applied to determine whether the differences were statistically significant. In short, there is no generic rule to choose a single best method or strategy. The choice depends on each problem, database, and evaluation metric used.

Our test shows that the SMOTEBagging and UnderBagging approaches obtained significantly better results than the data cleaning. However, in general, common sampling strategies were suitable for most techniques.

The comparison shows that most techniques yielded classification performances that are quite competitive with each other. Still, even though the differences between the classifiers are small, it is essential to note that in an infection context, an increase in the prediction ability, even a low percentage, may save lives and reduce costs. We also concluded that balancing strategies give us better models than the original models without balancing.

The results also showed that the more straightforward linear techniques, such as LR with regularization, give a relatively good performance, which is not significantly different from the more complex classifiers, such as NB and RF. There was no difference in linear methods strategies, and these models seem to provide a more stable performance. Comparing the computational time spent for each model, we can see that the linear models also are more efficient, followed by the decision tree. Moreover, LR has greater interpretation capability. Our data usefully demonstrates an essential principle of ML cited by Sidey-Gibbons and Sidey-Gibbons (2009): "more complex algorithms do not necessarily beget more useful predictions."

The best models by NPV were Naïve Bayes with SMOTEBagging, Logistic Regression Regularized with downsampling, and Random Forest with downsampling. MCC's selected ones were Neural Network with SMOTE+OSS, Neural Network with SMOTE+Tomek, and Support Vector Machine Radial with a RUSBoost strategy.

We proposed to the hospital's decision-maker two screening models, one more conservative and the other moderate. If the decision-makers decide to use the random forest's conservative model to determine who should do the weekly screening, we avoided approximately 39% of the unnecessary tests, increasing the laboratory's speed of response and reducing costs, as previously mentioned. However, the model lets that about six of 78 positive patients (8%) do not take the culture test (Sensitivity = 92%).

On the other hand, if one decided by the moderate model using the Neural Network, the unnecessary test is avoided 64% (25% more than the previous model), but 19 positive tests are misclassified (Sensitivity = 76%). This model can be useful for hospitals that need to decrease costs more assertively or even for those who do not use the screening protocol. The test would be done for only 40% of patients using this model, which is better than not testing anyone.

The hospital unit manager must make the final decision, and a future recommendation application can be developed based on these models. The first model could decrease the need for culture screenings for at least 36% of patients hospitalized in ICUs. In determining which patients to screen, we can balance the need to identify the colonized patient and available laboratory resources. It is important to remember that the threshold (cut-off point) can be changed to reduce false negatives.

In addition to the screening model, we include other clinical exams and developed a risk model to estimate ICU patients' probability of acquiring CR-GNB, measuring the predictions' calibration and Brier score. We developed a general model for all hospitals and an individual model for each one.

For the general model, the Nearest Shrunken Centroids, Gradient Boosting Machine, CART, Logistic Regression with or without regularization, and Linear Discrimination Analysis have the best Brier scores. The calibration analysis suggests that the calibration of these models is acceptable since the belt covers almost the whole diagonal line and the likelihood ratio test gives a p-value more than 0.05; that is, considering a confidence level of 95%, we cannot reject the hypothesis that the model is calibrated. The NSC is the best model to estimate CR-GNB acquisition risk. It presents the lowest Brier score, one of the largest MCC,

comprises almost the whole of the diagonal, has low uncertainty, and the highest p-value.

The Naïve Bayes technique has better discrimination power than others but has the worst Brier score value. This technique can be appropriate for classifying a patient as colonized and non-colonized but does not perform very well to estimate the probability. This model overestimates the acquisition for medium and high-risk patients and underestimates low-risk patients. Complex nonlinear algorithms do not directly make probabilistic predictions and instead use approximations. A classifier can produce excellent rankings, but probabilities might differ from the actual chances (FERRI; HERNÁNDEZ-ORALLO; MODROIU, 2009).

We developed a model for each hospital and used the average ranked (AR) performances to compare the different classifiers. Hospital A was the one that obtained the best Brier score mean among all methods, followed by Hospital C, B, D, and E, respectively. The average of the general model was 0.168, near to most hospitals´ models. Some techniques worked very well for a specific data set but flawed for another, such as the LDA, which presented the best result for Hospital A but was almost the worst for Hospital E and the general model.

These experiments show that the Neural Network seemed to have good results for the individual model, adding the best mean AR among the techniques. The regularized LR, NSC, C50, and GBM, which had previously been the best techniques for the general model, also performed well in most datasets, presenting good average ranks. Moreover, the method NB obtained the worst AR, confirming the results found in the general model. The t-test confirmed that the null hypothesis that there is no difference between the models. Therefore, we can use the general model for all hospitals without losing performance, being adequate to predict the probability of being colonized/infected.

We performed an external validation using data from two other hospitals in the same network. Although the model had obtained good brier score values for both, we concluded that the model is well-calibrated and acceptable to be introduced at Hospital G but overestimates the colonization of patients in Hospital F.

After evaluating the models, we identified the importance of attributes by Information Gain to assess the factors' power in predicting CR-GNB acquisition. The variables related to the duration of the use of invasive devices, especially mechanical ventilation, and the number of times this dispositive changed are essential. The antibiotics groups, the criticality index Saps3, admission reason, and admission source, obtained good information gains. The length of stay before the test also has high predictive power. On the other hand, comorbidity and other criticality indexes do not have a high strength to predict. The total duration of mechanical ventilation use is the most important variable for all hospitals.

Applying association rules mining, we find that 96.2% of the rules selected include MVDURTOTAL or VESDURTOTAL, and 86.0% the antibiotic use. Moreover, all conditions include some information about invasive dispositive use, showing the relevance of these variables in the decision criterion. Using this collection of strongly associated features enables us to indicate patterns that can help clinicians in the decision-making process. Efforts can be made to mitigate acquisition by implementing care bundles, removing devices earlier, using alternative procedures, or reducing their utilization. In some developing countries, the frequency of infections associated with CVC, ventilators, and other invasive devices can be up to 19 times higher than those reported from Germany and the USA (WHO, 2011).

6.2. Comparing Related Works

The results for risk factors are consistent with previous studies in which the use of invasive devices was most likely to result in the isolation of multidrug-resistant organisms, such as a central venous catheter (CVC) (CHANG et al., 2011; DANTAS et al., 2019; FERREIRA et al., 2017; JUNG et al., 2010; TACCONELLI et al., 2008; TUMBARELLO et al., 2011b; WILLMANN et al., 2014), the arterial catheter (CHANG et al., 2011), mechanical ventilation (DANTAS et al., 2019; PARK et al., 2011; ROMANELLI et al., 2009; YANG et al., 2016), urinary catheters (FALCONE et al., 2018; GOMILA et al., 2018; TUMBARELLO et al., 2011a; WILLMANN et al., 2014; YANG et al., 2016), and hemodialysis (CHANG

et al., 2011; DANTAS et al., 2019). Mechanical ventilation was the most significant to our work, followed by the urinary, central venous, arterial, and hemodialysis catheters.

Like our study, most also found previous use of antibiotics as significant (ALEXIOU et al., 2012; GOMILA et al., 2018; HU et al., 2016; JUNG et al., 2010; KENGKLA et al., 2016; KIDDEE et al., 2018; LEE et al., 2017; MARCHENAY et al., 2015; PATEL et al., 2014; PLAYFORD; CRAIG; IREDELL, 2007; ROUTSI et al., 2013; SCHWABER et al., 2008; SONG; JEONG, 2018; SURASARANG et al., 2007; TSENG et al., 2017; TUMBARELLO et al., 2011a, 2011b; VASUDEVAN et al., 2014; WILLMANN et al., 2014). Charlson comorbidity index also showed significance to Tacconelli et al. (2008) and Tumbarello et al. (2011a).

Burillo et al. (2019) reviewed articles about risk factors for colonization or infection by MDR-GNB. They found that the patients colonized with an MDR-GN pathogen are older, previously exposed to antibiotics, have advanced comorbidities, have low functional status, have prolonged hospital stays, or have been subjected to invasive procedures CVC, mechanical ventilation, hemodialysis catheter. All these factors are in line with our results.

Although previous studies have analyzed MDR factors, we have not found well-structured research in this area using different techniques or evaluation methods. Moreover, the predictive models found in the literature on multidrug-resistant are restricted to a specific type of infection, as in Chang et al. (2011), or aim only to analyze significance. Even though it was not possible to directly compare our predictive model's performance analysis with these studies, we explored some of the techniques' performance.

We found only a few works that developed other techniques than logistic regression to predict multiresistant bacteria (CHANG et al., 2011; GOODMAN et al., 2019; LI; TANG; HE, 2016; SONG; JEONG, 2018; TAN et al., 2017). Goodman et al. (2019) explored decision tree and logistic regression methods, finding similar results among these two. Chang et al. (2011) analyzed and concluded that both NN and LR models displayed excellent discrimination. Song

and Jeong (2018) and Tan et al. (2017) applied the decision tree and logistic regression but did not compare them.

According to these studies, we also showed that linear techniques, such as logistic regression, mainly regularized, give good discrimination power and prediction, which is not significantly different from the more complex classifier. Neural networks also presented promising results for the forecast. The tree-based models did not give us the best results for discrimination, but they are not statistically different from those already mentioned. Like our work, Tumbarello et al. (2011a) analyzed the Logistic Regression model's calibration, which displayed good calibration. No paper about MDR evaluated the model via the Brier score to the best of our knowledge.

Willmann et al. (2014) and Kiddee et al. (2018) were the only ones who worked with screening models. The first aimed to build a screening culture strategy to extensively drug-resistant P. aeruginosa (XDR-PA) using a conditional logistic regression model and a clinical risk score conducted by a matched case-control study. It cited an AUC of 0.83 but did not assess or discuss performances. The second one analyzed the screening for CR-GNB at ICU admission and discharge, using logistic regression and identifying the most significant factors from the p-value, but did not develop a model, not consider all weekly culture tests, nor including duration variables between two tests.

Some studies compare more than one technique for other purposes, as shown in Table 1. In many of them, the random forest showed the best performance compared to other methods. Kang et al. (2020) found that the random forest model showed the highest AUC for ICU mortality, followed by artificial neural networks and extreme gradient boost models. Ganggayah et al. (2019) compared the algorithms such as decision tree, RF, NN, extreme boost, LR, and SVM. Both model accuracy and calibration measure produced comparable outcomes. The lowest value was obtained from the decision tree and the highest from the random forest to detect breast cancer's survival rate.

Among the adopted machine learning algorithms for predicting hospital-acquired pneumonia in Kuo et al. (2019), random forest and decision tree exhibited a better predictive accuracy than the remaining algorithms (SVM, LR, NB, KNN).

Saarela et al. (2019), Keltch et al. (2014), and Periwal et al. (2011) affirm that random forest outperformed others, such as logistic regression. Loreto et al. (2020) found random forest as the best model for most metrics using cost-sensitive learning, and Saarela et al. (2019) and Bach et al. (2017) showed that the highest efficiency was achieved while using the SMOTE approach. On the other hand, Hartvigsen et al. (2018) found that SVM and LR outperform RF to support early recognition of MRSA infection by estimating risk at several time points during hospitalization. Looking at our work, we concluded that the random forest with the downsampling strategy was the best for our conservative screening culture model.

Lin et al. (2019) built a mortality prediction model using the RF algorithm and found the same effect that our calibration plot. The RF model slightly overestimates mortality in patients with low risk and underestimates mortality in patients with high risk.

Regarding the decision tree, the findings are divergent. The tree model works well in some works (KUO et al., 2019; TAN et al., 2017) but shows the worst performance in others (GOODMAN et al., 2019; LORETO; LISBOA; MOREIRA, 2020). Our studies have also not shown promising results for decision tree algorithms. Like our study, Kang et al. (2020) also show KNN with lousy performance.

Among the most common applied ML algorithms to the prediction outcomes, Shillan et al. (2019) found neural networks, support vector machines, and classification trees in their literature review. However, the choice of the most appropriate algorithm depends on many parameters, including the types of data collected, the size of the data samples, the time limitations, and the objectives (KOUROU et al., 2015). For example, Li et al. (2014) compared liver fibrosis prediction techniques and showed significant variability in accuracy, sensitivity, and specificity. Although neural network methods showed the highest sensitivity and specificity, the Logistic regression and naïve Bayes methods were the best in PPV. Our results from Table 22 shows the difference in the performance of the techniques for each metric.

Li et al. (2016) used different methods based on trees to compare different imbalanced sampling strategies and predict multidrug-resistant tuberculosis (MDR-

TB). The results showed that the best prediction could be obtained by adopting the Under Sampling + Bagging. It is in line with our results since we concluded that SMOTEBagging and UnderBagging got significantly better results than the data cleaning and sampling approaches.

Batista (2004) tested some alternative techniques in dealing with class imbalances, such as Tomek Link, CNN, OSS, CNN + Tomek links, NCL, SMOTE, SMOTE + Tomek links, and SMOTE + ENN. Like our results, he affirms that random sampling methods are very competitive to these more complex methods. Batista also analyzed under-sampling and over-sampling strategies. The findings suggested that, generally, over-sampling provides more accurate results than under-sampling methods considering the area under the ROC curve (AUC). However, this result seems to contradict the results previously published in the literature (BATISTA; PRATI; MONARD, 2004). Our work did not find a significant difference between sampling methods, but the downsampling approach performed better for most techniques and metrics.

6.3. Limitations

This study has some limitations. These results cannot be directly extrapolated to other healthcare institutions, given the study hospitals' case-mix specificity. We used the same clinical datasets to train the model and test the different methods, limiting our findings' generalizability. Brazil is a country with a high prevalence of MDR organisms, and our findings cannot be generalized to other countries. Still, the methods described and the analytical process can be adapted or extended. Heterogeneous Gram-negative bacteria were included and analyzed collectively, but patients colonized by these bacteria may have different risk factors and prognosis.

Since colonization or infection is a positive test result, we do not know precisely how the patient acquired the bacteria. Moreover, clinical data may be conflicting since patients with the same conditions may have different types and timing of observations. We did not have access to organizational variables for each

unit. All the data were manually inputted into the system; then, some records may be lost due to data imputation human errors.

6.4.
Future Researches

We propose some future studies on CR-GNB acquisition. To generalize the screening model, we need to perform an external validation using the best model in new periods and hospitals. Moreover, one can develop time series models considering variable changes over time, such as the device's duration. We will use the association rules without defining the RHS to find the strong relationships between the explanatory variables.

Considering the pandemic moment we are going through, another possibility would be to compare the relationship between antibiotics use and the positive cases of CR-GNB between the periods before and during the pandemic, trying to find any connection between the increase or decrease in cases and the antibiotic doses. One also analyzes the relationship between the use of carbapenem drugs in positive patients.

Finally, we are interested in analyzing the influence of acquisition by CR-GNB for the patient's outcome within 30 days after a positive test using a survival model. Moreover, we propose to examine the impact of the most significant variables on patient survival.

6.5.
Final Consideration

We believe that identifying risk factors and developing a model that estimates ICU patients' probability of acquiring CR-GNB may benefit them. The models for predicting resistance can offer utility where rapid diagnostics are unavailable or resource impractical. In clinical practice, usually, it is necessary to wait for 48h or more for the test results. In this interval of time, clinicians should identify patients at high risk for MDR-GNB and start adequate therapy as early as possible. Our predictive model also can help avoid inappropriate antibiotic treatment in patients

at low risk of multidrug resistance. Infection control policies can be established to control the spread of these bacteria.

Moreover, identifying patients who don´t need a weekly culture test decreases hospital costs and laboratory waiting times. We concluded that our models present good performance after our experiments and seem sufficiently reliable to predict a patient with this pathogen. These predictive models can be included in the hospital system and applied to each patient during hospitalization.

The invasive procedures use, mainly mechanical ventilation, are the most essential and significant attributes for the acquisition of CR-GNB. Knowledge of risk predictors and their combinations could provide a valuable instrument for the clinician to decide to initiate or not a broad-spectrum antibiotic therapy covering potentially carbapenem-resistant pathogens. Moreover, efforts can be made to mitigate acquisition by implementing care bundles, removing devices earlier, using alternative procedures, or reducing their utilization.

Finally, the framework on how to conduct a machine learning analysis and the code developed for this work is designed to be reusable and easily adaptable so that other researchers may apply these techniques to their datasets.

6.6. Publications

*h-index by Scopus = **4***
*h-index by Web of Science = **3***

Articles in Scientific Journals (6 documents)

ANTUNES, B. B. P. ; PERES, I. T. ; BAIAO, F. A. ; RANZANI, O. T. ; BASTOS, L. S. L. ; SILVA, A. A. B. ; SOUZA, G. F. G. ; MARCHESI, J. F. ; **DANTAS, L.F.** ; VARGAS, S. A. ; MACAIRA, P. ; HAMACHER, S. ; BOZZA, F. A. . Progression of confirmed COVID-19 cases after the implementation of control measures. RBTI, v. 1, p. 12-22, 2020. (Cited by Scopus: **2**)

PRADO, M. F. ; ANTUNES, B. B. P. ; BASTOS, L. S. L. ; PERES, I. T. ; SILVA, A. A. B. ; **DANTAS, L.F.** ; BAIAO, F. A. ; MACAIRA, P. ; HAMACHER, S. ; BOZZA, F.A. . Analysis of COVID-19 under-reporting in Brazil. RBTI, v. 00, p. 1-5, 2020. (Cited by Scopus: **4**)

DANTAS, L.F.; DALMAS, B.; ANDRADE, R.M.; HAMACHER, S.; BOZZA, F.A. Predicting acquisition of carbapenem-resistant Gram-negative pathogens in

intensive care units. JOURNAL OF HOSPITAL INFECTION, v. 103, p. 121-127, 2019. (Cited by Scopus: **4**)

DANTAS, L. F.; MARCHESI, J. F.; PERES, I. T.; HAMACHER, S.; BOZZA, F. A.; QUINTANO NEIRA, R. A. Public hospitalizations for stroke in Brazil from 2009 to 2016. PLoS One, v. 14, p. e0213837, 2019. (Cited by Scopus: **3**)

DANTAS, L. F.; HAMACHER, S.; CYRINO OLIVEIRA, F. L.; BARBOSA, S. D. J.; VIEGAS, F. Predicting Patient No-show Behavior: a Study in a Bariatric Clinic. OBESITY SURGERY, v. 29, p. 40-47, 2018. (Cited by Scopus: **4**)

DANTAS, L. F.; FLECK, J. L.; CYRINO OLIVEIRA, F. L.; HAMACHER, S. No-shows in Appointment Scheduling - a Systematic Literature Review. HEALTH POLICY, v. 122, p. 412-421, 2018. (Cited by Scopus: **57**)

Complete works published in proceedings of conferences in the last 5 years (8 documents)

ROCHA, N. G. ; **DANTAS, L. F.** ; HAMACHER, S. ; FIORENCIO, L. ; MARIANI, B. L. ; SOUSA, P. H. . Mineração de Processos aplicada à logística de uma empresa de óleo e gás. In: SBPO, 2019, Limeira. LI Simpósio Brasileiro de Pesquisa Operacional, 2019.

MATTOS, L. M. ; **DANTAS, L. F.** ; OLIVEIRA, F. L. C. . Análise estatística dos fatores que afetam o no-show de pacientes em agendamentos clínicos. In: SBPO, 2017, Blumenau. XLIX Simpósio Brasileiro de Pesquisa Operacional, 2017.

CUNHA, V. A. M. C. ; **DANTAS, L. F.** ; BREMENKAMP, L. H. ; PESSOA, L. S. . Dimensionamento de mão de obra e roteamento através de um algoritmo VND: Estudo de caso em uma empresa de medição de consumo de energia. In: SBPO, 2017, Blumenau. XLIX Simpósio Brasileiro de Pesquisa Operacional, 2017.

DANTAS, L. F.; et al. Stairway to value: mining the loan application process. In: International Conference on Business Process Management, 2017, Barcelona. Stairway to value: mining the loan application process, 2017.

DANTAS, L. F. ; OLIVEIRA, F. L. C. ; PERES, I. T. . Simulação de Eventos Discretos com balanceamento de linha de produção: uma aplicação na manufatura. In: Simpósio Brasileiro de Pesquisa Operacional, 2016, Vitória. XLVIII Simpósio Brasileiro de Pesquisa Operacional - SBPO 2016, 2016.

PERES, I. T. ; OLIVEIRA, F. L. C. ; **DANTAS, L. F.** ; PESSOA, L. S. . Simulação de políticas de agendamento de pacientes em serviços ambulatoriais: uma aplicação em um consultório de ortodontia. In: Simpósio Brasileiro de Pesquisa Operacional, 2016, Vitória. XLVIII Simpósio Brasileiro de Pesquisa Operacional - SBPO 2016, 2016.

BREMENKAMP, L. H. ; MONTEIRO, N. J. ; REPOLHO, H. M. V. ; CUNHA, V. A. M. C. ; **DANTAS, L. F.** . Aplicação da heurística de Clarke & Wright para um problema de roteirização de veículos homogêneos em uma distribuidora. In: ENEGEP, 2016, João Pessoa. Aplicação da heurística de Clarke & Wright para um problema de roteirização de veículos homogêneos em uma distribuidora, 2016.

CUNHA, V. A. M. C. ; **DANTAS, L. F.** ; REPOLHO, H. M. V. ; PESSOA, L. S. . Solução heurística para o problema de dimensionamento de mão de obra e roteirização através de um algoritmo Clarke e Wright. In: ANPET, 2016, Rio de Janeiro. ANPET, 2016.

Awards and Titles

2017 Best report Academic Category, BPI Challenge, Business Process Intelligence Workshop.

Expanded Summary published in proceedings of conferences

PERES, I. T.; MARQUESI, J.; **DANTAS, L. F.;** HAMACHER, S. Incidence and mortality of public hospitalizations for stroke in Brazil between 2009 and 2015. In: INFORMS Healthcare, 2017, Roterdã. INFORMS Healthcare, 2017.

7
References

AGÊNCIA BRASIL. **No Brasil, taxa de infecções hospitalares atinge 14% das internações**. Disponível em: <http://agenciabrasil.ebc.com.br/saude/noticia/2019-05/no-brasil-taxa-de-infeccoes-hospitalares-atinge-14-das-internacoes>. Acesso em: 31 jan. 2020.

AGRAWAL, R. et al. Fast Algorithms for Mining Association Rules. p. 32, 1994.

AKWEI, J. **RPubs - ContextBase Deep Learning**. Disponível em: <https://www.rpubs.com/johnakwei/311897>. Acesso em: 2 fev. 2020.

ALBA, A. C. et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. **JAMA**, v. 318, n. 14, p. 1377, 10 out. 2017.

ALEXIOU, V. G. et al. Multi-drug-resistant gram-negative bacterial infection in surgical patients hospitalized in the ICU: a cohort study. **European Journal of Clinical Microbiology & Infectious Diseases**, v. 31, n. 4, p. 557–566, abr. 2012.

ALHAJ, T. A. et al. Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation. **PLOS ONE**, v. 11, n. 11, p. e0166017, 28 nov. 2016.

ALI, A.; SHAMSUDDIN, S. M. H.; RALESCU, A. L. Classification with class imbalance problem: a review. **Int. J. Advance Soft Compu. Appl.**, v. 7, n. 3, p. 176–204, 2015.

AN, J. H. et al. Active surveillance for carbapenem-resistant Acinetobacter baumannii in a medical intensive care unit: Can it predict and reduce subsequent infections and the use of colistin? **American Journal of Infection Control**, v. 45, n. 6, p. 667–672, 1 jun. 2017.

ANALYTICS VIDHYA. **Tutorial on 5 Powerful Packages used for imputing missing values in R**, 4 mar. 2016. Disponível em: <https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/>. Acesso em: 31 jan. 2020

ANVISA. Programa Nacional de Prevenção e Controle de Infecções Relacionadas a Assistência a Saúde. p. 38, 2016.

BACH, M. et al. The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. **Information Sciences**, v. 384, p. 174–190, abr. 2017.

BAESENS, B. et al. Benchmarking state-of-the-art classification algorithms for credit scoring. **Journal of the Operational Research Society**, v. 54, n. 6, p. 627–635, jun. 2003.

BARANDELA, R.; VALDOVINOS, R. M.; SANCHEZ, J. S. New Applications of Ensembles of Classifiers. **Pattern Analysis & Applications**, v. 6, n. 3, p. 245–256, 1 dez. 2003.

BATISTA, G. E. A. P. A.; MONARD, M. C.; BAZZAN, A. L. C. Improving Rule Induction Precision for Automated Annotation by Balancing Skewed Data Sets. In: LÓPEZ, J. A.; BENFENATI, E.; DUBITZKY, W. (Eds.). . **Knowledge Exploration in Life Science Informatics**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. v. 3303p. 20–32.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20, 1 jun. 2004.

BEAM, A. L.; KOHANE, I. S. Big Data and Machine Learning in Health Care. **JAMA**, v. 319, n. 13, p. 1317, 3 abr. 2018.

BENESTY, J. et al. Pearson Correlation Coefficient. In: COHEN, I. et al. (Eds.). . **Noise Reduction in Speech Processing**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. v. 2p. 1–4.

BEN-HUR, A.; WESTON, J. A User's Guide to Support Vector Machines. **Methods in molecular biology (Clifton, N.J.)**, v. 609, p. 223–39, 1 jan. 2010.

BONTEN, M. J. Colonization pressure: a critical parameter in the epidemiology of antibiotic-resistant bacteria. **Critical Care**, v. 16, n. 4, p. 142, ago. 2012.

BOUZBID, S. et al. Automated detection of nosocomial infections: evaluation of different strategies in an intensive care unit 2000–2006. **Journal of Hospital Infection**, v. 79, n. 1, p. 38–43, set. 2011.

BRAGA, I. A. et al. Multi-hospital point prevalence study of healthcare-associated infections in 28 adult intensive care units in Brazil. **Journal of Hospital Infection**, v. 99, n. 3, p. 318–324, jul. 2018.

BREIMAN, L. et al. **Classification and Regression Trees.** 1st. ed. New York: CRC press, 1984.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, ago. 1996.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BROWN, I.; MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. **Expert Systems with Applications**, v. 39, n. 3, p. 3446–3453, fev. 2012.

BROWNLEE, J. **How to use Data Scaling Improve Deep Learning Model Stability and PerformanceMachine Learning Mastery**, 3 fev. 2019a. Disponível em: <https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/>. Acesso em: 8 maio. 2020

BROWNLEE, J. **Information Gain and Mutual Information for Machine LearningMachine Learning Mastery**, 15 out. 2019b. Disponível em: <https://machinelearningmastery.com/information-gain-and-mutual-information/>. Acesso em: 17 nov. 2020

BURILLO, A.; MUÑOZ, P.; BOUZA, E. Risk stratification for multidrug-resistant Gram-negative infections in ICU patients: **Current Opinion in Infectious Diseases**, v. 32, n. 6, p. 626–637, dez. 2019.

CARDOSO, T. et al. Microbiology of healthcare-associated infections and the definition accuracy to predict infection by potentially drug resistant pathogens: a systematic review. **BMC Infectious Diseases**, v. 15, n. 1, p. 565, dez. 2015.

CATENI, S.; COLLA, V.; VANNUCCI, M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. **Neurocomputing**, v. 135, p. 32–41, jul. 2014.

CHAISATHAPHOL, T.; CHAYAKULKEEREE, M. Epidemiology of infections caused by multidrug-resistant gram-negative bacteria in adult hospitalized patients at Siriraj Hospital. **Journal of the Medical Association of Thailand = Chotmaihet Thangphaet**, v. 97 Suppl 3, p. S35-45, mar. 2014.

CHANG, Y.-J. et al. Predicting Hospital-Acquired Infections by Scoring System with Simple Parameters. **PLoS ONE**, v. 6, n. 8, p. e23137, 24 ago. 2011.

CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 1 jun. 2002.

CHAWLA, N. V. et al. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: LAVRAČ, N. et al. (Eds.). . **Knowledge Discovery in Databases: PKDD 2003**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. v. 2838p. 107–119.

CHEN, X.; WASIKOWSKI, M. **FAST: a roc-based feature selection metric for small samples and imbalanced data classification problems**. . In: PROCEEDING OF THE 14TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. Las Vegas, Nevada, USA: ACM Press, 2008

CHEVRET, S.; SEAMAN, S.; RESCHE-RIGON, M. Multiple imputation: a mature approach to dealing with missing data. **Intensive Care Medicine**, v. 41, n. 2, p. 348–350, fev. 2015.

CHOI, B. Y.; BAIR, E.; LEE, J. W. Nearest shrunken centroids via alternative genewise shrinkages. **PLoS ONE**, v. 12, n. 2, 15 fev. 2017.

CORTES, C.; VAPNIK, V. Support-vector networks. v. 20, p. 25, 1995.

COSGROVE, S. E. The Relationship between Antimicrobial Resistance and Patient Outcomes: Mortality, Length of Hospital Stay, and Health Care Costs. **Clinical Infectious Diseases**, v. 42, n. Supplement_2, p. S82–S89, 15 jan. 2006.

CRONE, S. F.; FINLAY, S. Instance sampling in credit scoring: An empirical study of sample size and balancing. **International Journal of Forecasting**, v. 28, n. 1, p. 224–238, jan. 2012.

DANTAS, L. F. et al. Predicting acquisition of carbapenem-resistant Gram-negative pathogens in intensive care units. **Journal of Hospital Infection**, v. 103, n. 2, p. 121–127, out. 2019.

DEBBY, B. D. et al. Epidemiology of carbapenem resistant Klebsiella pneumoniae colonization in an intensive care unit. **European Journal of Clinical Microbiology & Infectious Diseases**, v. 31, n. 8, p. 1811–1817, ago. 2012.

DELAHANTY, R. J. et al. Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis. **Annals of Emergency Medicine**, v. 73, n. 4, p. 334–344, abr. 2019.

DEMSAR, J.; DEMSAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. **Journal of Machine Learning Research**, v. 7, p. 30, 2006.

DENG, H. et al. Understanding the importance of key risk factors in predicting chronic bronchitic symptoms using a machine learning approach. **BMC Medical Research Methodology**, v. 19, n. 1, dez. 2019.

DENIL, M.; TRAPPENBERG, T. Overlap versus Imbalance. In: FARZINDAR, A.; KEŠELJ, V. (Eds.). . **Advances in Artificial Intelligence**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. v. 6085p. 220–231.

DEVI, D.; BISWAS, S. K.; PURKAYASTHA, B. Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique. **Connection Science**, v. 31, n. 2, p. 105–142, 3 abr. 2019.

DOMINGOS, P.; PAZZANI, M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. **Machine Learning**, v. 29, n. 2–3, p. 103–130, 1997.

DREISEITL, S.; OHNO-MACHADO, L. Logistic regression and artificial neural network classification models: a methodology review. **Journal of Biomedical Informatics**, v. 35, n. 5–6, p. 352–359, out. 2002.

DUALE, N. et al. Using prediction models to identify miRNA-based markers of low dose rate chronic stress | Elsevier Enhanced Reader. **Science of the Total Environment**, v. 717, p. 1–14, 2020.

DUDANI, S. A. The Distance-Weighted k-Nearest-Neighbor Rule. **IEEE Transactions on Systems, Man, and Cybernetics**, v. SMC-6, n. 4, p. 325–327, abr. 1976.

EHRENTRAUT, C. et al. Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. **Health Informatics Journal**, v. 24, n. 1, p. 24–42, 1 mar. 2018.

ESCOLANO, S. et al. A multi-state model for evolution of intensive care unit patients: prediction of nosocomial infections and deaths. **Statistics in Medicine**, v. 19, n. 24, p. 3465–3482, 2000.

ESTABROOKS, A.; JO, T.; JAPKOWICZ, N. A Multiple Resampling Method for Learning from Imbalanced Data Sets. **Computational Intelligence**, v. 20, n. 1, p. 18–36, 2004.

FALAGAS, M. E. et al. Pandrug-resistant Klebsiella pneumoniae, Pseudomonas aeruginosa and Acinetobacter baumannii infections: Characteristics and outcome in a series of 28 patients. **International Journal of Antimicrobial Agents**, v. 32, n. 5, p. 450–454, nov. 2008.

FALAGAS, M. E.; KOPTERIDES, P. Risk factors for the isolation of multi-drug-resistant Acinetobacter baumannii and Pseudomonas aeruginosa: a systematic review of the literature. **Journal of Hospital Infection**, v. 64, n. 1, p. 7–15, set. 2006.

FALCONE, M. et al. Predicting resistant etiology in hospitalized patients with blood cultures positive for Gram-negative bacilli. **European Journal of Internal Medicine**, v. 53, p. 21–28, jul. 2018.

FARQUAD, M. A. H.; BOSE, I. Preprocessing unbalanced data using support vector machine. **Decision Support Systems**, v. 53, n. 1, p. 226–233, abr. 2012.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

FERREIRA, E. et al. Risk factors for health care–associated infections: From better knowledge to better prevention. **American Journal of Infection Control**, v. 45, n. 10, p. e103–e107, out. 2017.

FERRI, C.; HERNÁNDEZ-ORALLO, J.; MODROIU, R. An experimental comparison of performance measures for classification. **Pattern Recognition Letters**, v. 30, n. 1, p. 27–38, jan. 2009.

FORMAN, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. **Journal of Machine Learning Research**, v. 3, p. 1289–1305, 2003.

FREITAS, A. A. Understanding the Crucial Role of Attribute Interaction in Data Mining. **Artificial Intelligence Review**, v. 16, n. 3, p. 177–199, 2001.

FREUND, Y.; SCHAPIRE, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **J. Comput. Syst. Sci.**, v. 55, p. 119–139, 1997.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational Statistics & Data Analysis**, v. 38, n. 4, p. 367–378, fev. 2002.

FRIEDMAN, M. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. **The Annals of Mathematical Statistics**, v. 11, n. 1, p. 86–92, 1940.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian Network Classifiers. p. 33, 1997.

GALAR, M. et al. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 42, n. 4, p. 463–484, jul. 2012.

GANGGAYAH, M. D. et al. Predicting factors for survival of breast cancer patients using machine learning techniques. **BMC Medical Informatics and Decision Making**, v. 19, n. 1, dez. 2019.

GIRARD, R. et al. WORLD HEALTH ORGANIZATION. p. 72, 2002.

GOMILA, A. et al. Predictive factors for multidrug-resistant gram-negative bacteria among hospitalised patients with complicated urinary tract infections. **Antimicrobial Resistance & Infection Control**, v. 7, n. 1, p. 111, dez. 2018.

GOODMAN, K. E. et al. A Clinical Decision Tree to Predict Whether a Bacteremic Patient Is Infected With an Extended-Spectrum β-Lactamase–Producing Organism. **Clinical Infectious Diseases**, v. 63, n. 7, p. 896–903, 1 out. 2016.

GOODMAN, K. E. et al. A methodological comparison of risk scores versus decision trees for predicting drug-resistant infections: A case study using extended-spectrum beta-lactamase (ESBL) bacteremia. **Infection Control & Hospital Epidemiology**, v. 40, n. 4, p. 400–407, abr. 2019.

GREGORUTTI, B.; MICHEL, B.; SAINT-PIERRE, P. Correlation and variable importance in random forests. **Statistics and Computing**, v. 27, n. 3, p. 659–678, maio 2017.

GROBELNIK, M. **Feature selection for unbalanced class distribution and naive bayes**. . In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING. 1999

GUYON, I.; WESTON, J.; BARNHILL, S. Gene Selection for Cancer Classification using Support Vector Machines. **Machine Learning**, v. 46, p. 34, 2002.

HAIBO HE; GARCIA, E. A. Learning from Imbalanced Data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, set. 2009.

HALEY, R. W. et al. THE EFFICACY OE INFECTION SURVEILLANCE AND CONTROL PROGRAMS IN PREVENTING NOSOCOMIAL INFECTIONS IN US HOSPITALS. **American Journal of Epidemiology**, v. 121, n. 2, p. 182–205, fev. 1985.

HALPERN, N. A.; PASTORES, S. M. Critical care medicine beds, use, occupancy and costs in the United States: a methodological review. **Critical care medicine**, v. 43, n. 11, p. 2452–2459, nov. 2015.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. [s.l.] Elsevier, 2011.

HARRELL, F. **Classification vs. Prediction**. Disponível em: <https://fharrell.com/post/classification/>. Acesso em: 31 jan. 2020.

HART, P. The condensed nearest neighbor rule. **IEEE Transactions on Information Theory**, v. 18, p. 515–516, 1968.

HARTVIGSEN, T. et al. **Early Prediction of MRSA Infections using Electronic Health Records:** Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies. **Anais**...Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018Disponível em: <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006599601560167>. Acesso em: 18 fev. 2020

HASHEM, E. M.; MABROUK, M. S. A Study of support vector machine algorithm for liver disease diagnosis. v. 4, n. 1, p. 9–14, 2014.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. Second Edition ed. [s.l.] Springer Science & Business Media, 2009.

HEARST, M. A. et al. Support vector machines. **IEEE Intelligent Systems and their Applications**, v. 13, n. 4, p. 18–28, jul. 1998.

HIDRON, A. I. et al. Antimicrobial-Resistant Pathogens Associated With Healthcare-Associated Infections: Annual Summary of Data Reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2006–2007. **Infection Control & Hospital Epidemiology**, v. 29, n. 11, p. 996–1011, nov. 2008.

HOANG, G.; BOUZERDOUM, A.; LAM, S. Learning Pattern Classification Tasks with Imbalanced Data Sets. In: YIN, P.-Y. (Ed.). . **Pattern Recognition**. [s.l.] InTech, 2009.

HOARE, J. **Linear Discriminant Analysis in R: An IntroductionDisplayr**, 2020. Disponível em: <https://www.displayr.com/linear-discriminant-analysis-in-r-an-introduction/>. Acesso em: 31 jan. 2020

HOSMER, D.; LEMESHOW, S. **Applied Logistic Regression**. Second Edition ed. [s.l.] A Wiley-Interscience Publication, 2000.

HU, Y. et al. A retrospective study of risk factors for carbapenem-resistant Klebsiella pneumoniae acquisition among ICU patients. **Journal of Infection in Developing Countries**, v. 10, n. 3, p. 208–213, 31 mar. 2016.

HUANG, S.-T. et al. Risk factors and clinical outcomes of patients with carbapenem-resistant Acinetobacter baumannii bacteremia. **Journal of Microbiology, Immunology, and Infection**, v. 45, n. 5, p. 356–362, out. 2012.

JAMES, G. et al. **An Introduction to Statistical Learning**. New York, NY: Springer New York, 2013. v. 103

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study1. **Intelligent Data Analysis**, v. 6, n. 5, p. 429–449, 15 nov. 2002.

JARRELL, A. S. et al. Factors associated with in-hospital mortality among critically ill surgical patients with multidrug-resistant Gram-negative infections. **Journal of Critical Care**, v. 43, p. 321–326, fev. 2018.

JUNG, J. Y. et al. Risk factors for multi-drug resistant Acinetobacter baumannii bacteremia in patients with colonization in the intensive care unit. **BMC infectious diseases**, v. 10, p. 228, 30 jul. 2010.

KANG, M. W. et al. Machine learning algorithm to predict mortality in patients undergoing continuous renal replacement therapy. **Critical Care**, v. 24, 6 fev. 2020.

KANTARDZIC, M. **Data Mining: Concepts, Models, Methods, and Algorithms**. [s.l.] John Wiley & Sons, 2011.

KASSAMBARA, A. **Machine Learning Essentials: Practical Guide in R**. [s.l.] sthda, 2018.

KAUFMAN, L.; ROUSSEEUW, P. **Finding Groups in Data: An Introduction to Cluster Analysis**. [s.l.] Wiley Interscience, 1990.

KAUR, H.; KUMARI, V. Predictive modelling and analytics for diabetes using a machine learning approach. **Applied Computing and Informatics**, p. S221083271830365X, dez. 2018.

KELTCH, B.; LIN, Y.; BAYRAK, C. Comparison of AI Techniques for Prediction of Liver Fibrosis in Hepatitis Patients. **Journal of Medical Systems**, v. 38, n. 8, p. 60, ago. 2014.

KENGKLA, K. et al. Clinical risk scoring system for predicting extended-spectrum β-lactamase-producing Escherichia coli infection in hospitalized patients. **Journal of Hospital Infection**, v. 93, n. 1, p. 49–56, maio 2016.

KIDDEE, A. et al. Risk Factors for Gastrointestinal Colonization and Acquisition of Carbapenem-Resistant Gram-Negative Bacteria among Patients in Intensive Care Units in Thailand. **Antimicrobial Agents and Chemotherapy**, v. 62, n. 8, 11 jun. 2018.

KIM, S.; KIM, W.; PARK, R. W. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. **Healthcare Informatics Research**, v. 17, n. 4, p. 232, 2011.

KIRA, K.; RENDELL, L. **The Feature Selection Problem: Traditional Methods and a New Algorithm**. Aaai. **Anais**...1992

KOLLEF, M. H.; FRASER, V. J. Antibiotic Resistance in the Intensive Care Unit. **Annals of Internal Medicine**, v. 134, n. 4, p. 298, 20 fev. 2001.

KOTSIANTIS, S. B.; KANELLOPOULOS, D.; PINTELAS, P. E. **Handling imbalanced datasets: A review**. GESTS International Transactions on Computer Science and Engineering. **Anais**...2006

KOUROU, K. et al. Machine learning applications in cancer prognosis and prediction. **Computational and Structural Biotechnology Journal**, v. 13, p. 8–17, 2015.

KUBAT, M.; MATWIN, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. **Icml**, v. 97, p. 179–186, 1997.

KUHN, M. **The caret Package**. Vienna, Austria: R Foundation for Statistical Computing, 2011.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York, NY: Springer New York, 2013.

KUHN, M.; JOHSON, K. **Feature Engineering and Selection: A Practical Approach for Predictive Models**. [s.l.] CRC press, 2019.

KUO, K. M. et al. Predicting hospital-acquired pneumonia among schizophrenic patients: a machine learning approach. **BMC Medical Informatics and Decision Making**, v. 19, n. 1, dez. 2019.

LAURIKKALA, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. In: QUAGLINI, S.; BARAHONA, P.; ANDREASSEN, S. (Eds.). . **Artificial Intelligence in Medicine**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. v. 2101p. 63–66.

LEE, C.-H. et al. A simple scoring algorithm predicting extended-spectrum β-lactamase producers in adults with community-onset monomicrobial Enterobacteriaceae bacteremia: Matters of frequent emergency department users. **Medicine**, v. 96, n. 16, p. e6648, abr. 2017.

LI, C. et al. Prediction of Length of Stay on the Intensive Care Unit Based on Least Absolute Shrinkage and Selection Operator. **IEEE Access**, v. 7, p. 110710–110721, 2019.

LI, D.-C.; LIU, C.-W.; HU, S. C. A learning method for the class imbalance problem with medical data sets. **Computers in Biology and Medicine**, v. 40, n. 5, p. 509–518, maio 2010.

LI, S.; TANG, B.; HE, H. An Imbalanced Learning based MDR-TB Early Warning System. **Journal of Medical Systems**, v. 40, n. 7, p. 164, jul. 2016.

LIAO, J.-J. et al. An ensemble-based model for two-class imbalanced financial problem. **Economic Modelling**, v. 37, p. 175–183, fev. 2014.

LIN, K.; HU, Y.; KONG, G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model | Elsevier Enhanced Reader. **International journal of medical informatics**, v. 125, p. 55–61, 2019.

LING, C. X.; LI, C. Data Mining for Direct Marketing: Problems and Solutions. **Kdd**, v. 98, p. 73–79, 1998.

LITTLE, R. J. A.; RUBIN, D. **Statistical Analysis with Missing Data.** [s.l.] John Wiley & Sons, 2019. v. 793

LORETO, M.; LISBOA, T.; MOREIRA, V. P. Early prediction of ICU readmissions using classification algorithms | Elsevier Enhanced Reader. **Computers in Biology and Medicine**, v. 118, 2020.

LUNA, C. M. et al. Gram-Negative Infections in Adult Intensive Care Units of Latin America and the Caribbean. **Critical Care Research and Practice**, v. 2014, p. 1–12, 2014.

LYE, D. C. et al. The impact of multidrug resistance in healthcare-associated and nosocomial Gram-negative bacteraemia on mortality and length of stay: cohort study. **Clinical Microbiology and Infection**, v. 18, n. 5, p. 502–508, maio 2012.

MACVANE, S. H. Antimicrobial Resistance in the Intensive Care Unit: A Focus on Gram-Negative Bacterial Infections. **Journal of Intensive Care Medicine**, v. 32, n. 1, p. 25–37, jan. 2017.

MAGIORAKOS, A.-P. et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. **Clinical Microbiology and Infection**, v. 18, n. 3, p. 268–281, mar. 2012.

MARCHENAY, P. et al. Acquisition of carbapenem-resistant Gram-negative bacilli in intensive care unit: Predictors and molecular epidemiology | Elsevier Enhanced Reader. **Médecine et maladies infectieuses**, n. 45, p. 7, 2015.

MAULDIN, P. D. et al. Attributable Hospital Cost and Length of Stay Associated with Health Care-Associated Infections Caused by Antibiotic-Resistant Gram-Negative Bacteria. **Antimicrobial Agents and Chemotherapy**, v. 54, n. 1, p. 109–115, 1 jan. 2010.

MOHD SAZLLY LIM, S. et al. Clinical prediction models for ESBL-Enterobacteriaceae colonization or infection: a systematic review. **Journal of Hospital Infection**, v. 102, n. 1, p. 8–16, maio 2019.

MOONS, K. G. M. et al. Using the outcome for imputation of missing predictor values was preferred. **Journal of Clinical Epidemiology**, v. 59, n. 10, p. 1092–1101, out. 2006.

NATTINO, G.; FINAZZI, S.; BERTOLINI, G. A new test and graphical tool to assess the goodness of fit of logistic regression models. **Statistics in Medicine**, v. 35, n. 5, p. 709–720, 2016.

NEMENYI, P. **Distribution-free Multiple Comparisons**. [s.l.] Princeton University, 1963.

NHSN. 2020 NHSN Patient Safety Component Manual. p. 434, 2020.

NIELS. **MICE – ahoi data**, 2020. Disponível em: <https://statistics.ohlsen-web.de/category/mice/>. Acesso em: 31 jan. 2020

OECD. **Delivering Quality Health Services: A Global Imperative**. [s.l.] OECD Publishing, 2018.

OSBORNE, J. W.; OVERBAY, A. The power of outliers (and why researchers should always check for them). **Practical Assessment, Research, and Evaluation**, v. 9, n. 1, p. 6, 2004.

PAPP-WALLACE, K. M. et al. Carbapenems: Past, Present, and Future. **Antimicrobial Agents and Chemotherapy**, v. 55, n. 11, p. 4943–4960, nov. 2011.

PARK, K. et al. Robust predictive model for evaluating breast cancer survivability. **Engineering Applications of Artificial Intelligence**, v. 26, n. 9, p. 2194–2205, out. 2013.

PARK, Y. S. et al. Acquisition of extensive drug-resistant Pseudomonas aeruginosa among hospitalized patients: risk factors and resistance mechanisms to carbapenems. **The Journal of Hospital Infection**, v. 79, n. 1, p. 54–58, set. 2011.

PATEL, J. B.; COCKERILL, F. R. M100 Performance standards for antimicrobial susceptibility testing. **United State: Clinical and Laboratory Standards Institute**, v. 240, 2017.

PATEL, N.; UPADHYAY, S. Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA. **International Journal of Computer Applications**, v. 60, n. 12, p. 6, 2012.

PATEL, S. J. et al. Risk factors and Outcomes of Infections Caused by Extremely Drug-Resistant Gram-Negative Bacilli in Patients Hospitalized in Intensive Care Units. **American journal of infection control**, v. 42, n. 6, p. 626–631, jun. 2014.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **MACHINE LEARNING IN PYTHON**, p. 6, 2011.

PELEG, A. Y.; HOOPER, D. C. Hospital-Acquired Infections Due to Gram-Negative Bacteria. **New England Journal of Medicine**, v. 362, n. 19, p. 1804–1813, 13 maio 2010.

PERIWAL, V. et al. Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. **BMC Research Notes**, v. 4, n. 1, p. 504, dez. 2011.

PERKINS, N. J.; SCHISTERMAN, E. F. The Inconsistency of “Optimal” Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve. **American Journal of Epidemiology**, v. 163, n. 7, p. 670–675, 1 abr. 2006.

PLAYFORD, E. G.; CRAIG, J. C.; IREDELL, J. R. Carbapenem-resistant Acinetobacter baumannii in intensive care unit patients: risk factors for acquisition, infection and their consequences. **Journal of Hospital Infection**, v. 65, n. 3, p. 204–211, mar. 2007.

PRADE, S. et al. Estudo brasileiro da magnitude das infecções hospitalates em hospitais terciários. **Controle de Infecção Hospitalar**, 1995.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior. In: MONROY, R. et al. (Eds.). . **MICAI 2004: Advances in Artificial Intelligence**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. v. 2972p. 312–321.

PROVOST, F.; FAWCETT, T. Robust Classification for Imprecise Environments. **Machine Learning**, v. 42, n. 3, p. 203–231, 2001.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. [s.l.] Elsevier, 2014.

RABHI, S.; JAKUBOWICZ, J.; METZGER, M.-H. Deep Learning versus Conventional Machine Learning for Detection of Healthcare-Associated Infections in French Clinical Narratives. **Methods of Information in Medicine**, v. 58, n. 01, p. 031–041, jun. 2019.

RAFTER, J. A. et al. **Statistics with Maple**. [s.l.] Academic Press, 2003.

RAMAN, G. et al. Risk factors for hospitalized patients with resistant or multidrug-resistant Pseudomonas aeruginosa infections: a systematic review

and meta-analysis. **Antimicrobial Resistance & Infection Control**, v. 7, n. 1, p. 79, dez. 2018.

ROMANELLI, R. et al. Outbreak of Resistant Acinetobacter baumannii – Measures and Proposal for Prevention and Control. **The Brazilian Journal of Infectious Diseases**, v. 13, n. 5, p. 7, 2009.

ROULSTON, M. S. Performance targets and the Brier score. **Meteorological Applications**, v. 14, n. 2, p. 185–194, 2007.

ROUTSI, C. et al. Risk factors for carbapenem-resistant Gram-negative bacteremia in intensive care unit patients. **Intensive Care Medicine**, v. 39, n. 7, p. 1253–1261, jul. 2013.

RUBIN, D. B. Inference and Missing Data. **Biometrika**, v. 63, n. 3, p. 581–592, 1976.

RUBIO, F. et al. Trends in bacterial resistance in a tertiary university hospital over one decade | Elsevier Enhanced Reader. **The Brazilian Journal of INFECTIOUS DISEASES**, v. 17, n. 4, 2013.

SAARELA, M.; RYYNÄNEN, O.-P.; ÄYRÄMÖ, S. Predicting hospital associated disability from imbalanced data using supervised learning. **Artificial Intelligence in Medicine**, v. 95, p. 88–95, abr. 2019.

SAEYS, Y.; INZA, I.; LARRANAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, v. 23, n. 19, p. 2507–2517, 1 out. 2007.

SALKIND, N. J. **Encyclopedia of Research Design**. [s.l.] SAGE Publications, 2010.

SCHMITT, P.; MANDEL, J.; GUEDJ, M. A Comparison of Six Methods for Missing Data Imputation. **Journal of Biometrics & Biostatistics**, v. 06, n. 01, 2015.

SCHWABER, M. J. et al. Predictors of Carbapenem-Resistant Klebsiella pneumoniae Acquisition among Hospitalized Adults and Effect of Acquisition on Mortality. **Antimicrobial Agents and Chemotherapy**, v. 52, n. 3, p. 1028–1033, 1 mar. 2008.

SEIFFERT, C. et al. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. **IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans**, v. 40, n. 1, p. 185–197, 2009.

SHELKE, M. M. S.; DESHMUKH, D. P. R.; SHANDILYA, V. K. A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. **Int J Recent Trends in Eng & Res**, v. 3, p. 444–449, 2017.

SIDEY-GIBBONS, J. A. M.; SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. **BMC Medical Research Methodology**, v. 19, n. 1, p. 64, dez. 2019.

SIPS, M. E.; BONTEN, M. J. M.; VAN MOURIK, M. S. M. Automated surveillance of healthcare-associated infections: state of the art. **Current Opinion in Infectious Diseases**, v. 30, n. 4, p. 425–431, ago. 2017.

SONG, J. Y.; JEONG, I. S. Development of a risk prediction model of carbapenem-resistant Enterobacteriaceae colonization among patients in intensive care units. **American Journal of Infection Control**, v. 46, n. 11, p. 1240–1244, nov. 2018.

SPIEGELHALTER, D. J. Probabilistic prediction in patient management and clinical trials. **Statistics in Medicine**, v. 5, n. 5, p. 421–433, set. 1986.

STERNE, J. A. C. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. **BMJ**, v. 338, n. jun29 1, p. b2393–b2393, 1 set. 2009.

SURASARANG, K. et al. Risk Factors for Multi-Drug Resistant Acinetobacter Baumannii Nosocomial Infection. **J Med Assoc Thai**, v. 90, n. 8, p. 7, 2007.

TACCONELLI, E. et al. Prediction models to identify hospitalized patients at risk of being colonized or infected with multidrug-resistant Acinetobacter baumannii calcoaceticus complex. **Journal of Antimicrobial Chemotherapy**, v. 62, n. 5, p. 1130–1137, 18 jul. 2008.

TACCONELLI, E. et al. Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics. **World Health Organization**, v. 27, p. 318–327, 2017.

TAN, D. et al. Identification of Risk Factors of Multidrug-Resistant Tuberculosis by using Classification Tree Method. **The American Journal of Tropical Medicine and Hygiene**, v. 97, n. 6, p. 1720–1725, 6 dez. 2017.

TAPLITZ, R.; RITTER, M.; TORRIANI, F. Infection Prevention and Control, and Antimicrobial Stewardship. **Infectious Diseases**, p. 54, 2017.

THAI-NGHE, N.; GANTNER, Z.; SCHMIDT-THIEME, L. **Cost-sensitive learning methods for imbalanced data**. The 2010 International Joint Conference on Neural Networks (IJCNN). **Anais**...Barcelona, Spain: IEEE, jul. 2010Disponível em: <http://ieeexplore.ieee.org/document/5596486/>. Acesso em: 31 jan. 2020

TOMEK, I. Two Modifications of CNN. **IEEE Transactions on Systems, Man, and Cybernetics**, v. SMC-6, n. 11, p. 769–772, nov. 1976.

TSENG, W.-P. et al. Predicting Multidrug-Resistant Gram-Negative Bacterial Colonization and Associated Infection on Hospital Admission. **Infection Control & Hospital Epidemiology**, v. 38, n. 10, p. 1216–1225, out. 2017.

TUMBARELLO, M. et al. Identifying Patients Harboring Extended-Spectrum-β-Lactamase-Producing Enterobacteriaceae on Hospital Admission: Derivation and Validation of a Scoring System. **Antimicrobial Agents and Chemotherapy**, v. 55, n. 7, p. 3485–3490, jul. 2011a.

TUMBARELLO, M. et al. Multidrug-resistant *Pseudomonas aeruginosa* bloodstream infections: risk factors and mortality. **Epidemiology and Infection**, v. 139, n. 11, p. 1740–1749, nov. 2011b.

VAN DUIN, D.; PATERSON, D. L. Multidrug-Resistant Bacteria in the Community. **Infectious Disease Clinics of North America**, v. 30, n. 2, p. 377–390, jun. 2016.

VANHOEYVELD, J.; MARTENS, D. Imbalanced classification in sparse and large behaviour datasets. **Data Mining and Knowledge Discovery**, v. 32, n. 1, p. 25–82, jan. 2018.

VARDAKAS, K. Z. et al. Predictors of mortality in patients with infections due to multi-drug resistant Gram negative bacteria: The study, the patient, the bug or the drug? **Journal of Infection**, v. 66, n. 5, p. 401–414, maio 2013.

VARDAKAS, K. Z. et al. Characteristics, risk factors and outcomes of carbapenem-resistant Klebsiella pneumoniae infections in the intensive care unit. **The Journal of Infection**, v. 70, n. 6, p. 592–599, jun. 2015.

VASUDEVAN, A. et al. A prediction tool for nosocomial multi-drug resistant gram-negative bacilli infections in critically ill patients - prospective observational study. **BMC Infectious Diseases**, v. 14, n. 1, p. 615, dez. 2014.

VAUS, D. D.; VAUS, D. DE. **Surveys In Social Research**. [s.l.] Routledge, 2013.

VESIN, A. et al. Reporting and handling missing values in clinical studies in intensive care units. **Intensive Care Medicine**, v. 39, n. 8, p. 1396–1404, ago. 2013.

WANG, S.; YAO, X. **Diversity analysis on imbalanced data sets by using ensemble models**. 2009 IEEE Symposium on Computational Intelligence and Data Mining. **Anais**...Nashville, TN, USA: IEEE, mar. 2009Disponível em: <http://ieeexplore.ieee.org/document/4938667/>. Acesso em: 31 jan. 2020

WHO. **Health care-associated infections. Fact Sheet 2011**. Disponível em: <https://www.who.int/gpsc/country_work/gpsc_ccisc_fact_sheet_en.pdf>. Acesso em: 28 nov. 2020.

WHO. **Delivering quality health services: a global imperative for universal health coverage**. [s.l.] OECD Publishing, 2018.

WILLMANN, M. et al. Clinical and treatment-related risk factors for nosocomial colonisation with extensively drug-resistant Pseudomonas aeruginosa in a haematological patient population: a matched case control study. **BMC infectious diseases**, v. 14, p. 650, 10 dez. 2014.

WILSON, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. **IEEE Transactions on Systems, Man, and Cybernetics**, v. SMC-2, n. 3, p. 408–421, jul. 1972.

WU, H. et al. **An Improved Apriori-based Algorithm for Association Rules Mining**. 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. **Anais**... In: 2009 SIXTH INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY. ago. 2009

YANG, D. et al. A model for predicting nosocomial carbapenem-resistant Klebsiella pneumoniae infection. **Biomedical Reports**, v. 5, n. 4, p. 501–505, out. 2016.

YANG, P. et al. **Ensemble-based wrapper methods for feature selection and class imbalance learning**. Pacific-Asia Conference on Knowledge Discovery and Data Mining. **Anais**...Berlin, Heidelberg: Springer, 2013

YANG, S. **An Introduction to Naïve Bayes Classifier**. Disponível em: <https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf>. Acesso em: 31 jan. 2020.

YANG, Z.; GAO, D. Classification for Imbalanced and Overlapping Classes Using Outlier Detection and Sampling Techniques. **Applied Mathematics & Information Sciences**, v. 7, n. 1L, p. 375–381, 1 fev. 2013.

YIN, L. et al. Feature selection for high-dimensional imbalanced data. **Neurocomputing**, v. 105, p. 3–11, abr. 2013.

YONG SUN; FENG LIU. **SMOTE-NCL: A re-sampling method with filter for network intrusion detection**. 2016 2nd IEEE International Conference on Computer and Communications (ICCC). **Anais**...Chengdu, China: IEEE, out. 2016Disponível em: <http://ieeexplore.ieee.org/document/7924886/>. Acesso em: 9 fev. 2020

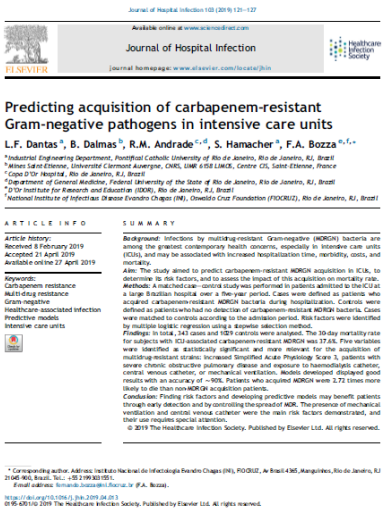
YOUDEN, W. J. Index for rating diagnostic tests. **Cancer**, v. 3, n. 1, p. 32–35, 1950.

ZHANG, Z. Missing data exploration: highlighting graphical presentation of missing pattern. **Annals of Translational Medicine**, v. 3, n. 22, dez. 2015.

Appendices

Appendix A

https://www.journalofhospitalinfection.com/article/S0195-6701(19)30182-3/pdf



Journal of Hospital Infection 103 (2019) 121–127

Available online at www.sciencedirect.com

Journal of Hospital Infection

Journal homepage: www.elsevier.com/locate/jhin

Predicting acquisition of carbapenem-resistant Gram-negative pathogens in intensive care units

L.F. Dantas [a], B. Dalmas [b], R.M. Andrade [c, d], S. Hamacher [a], F.A. Bozza [e, f, *]

[a] Industrial Engineering Department, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
[b] Mines Saint-Etienne, Université Clermont Auvergne, CNRS, UMR 6158 LIMOS, Centre CIS, Saint-Etienne, France
[c] Copa D'Or Hospital, Rio de Janeiro, RJ, Brazil
[d] Department of General Medicine, Federal University of the State of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
[e] D'Or Institute for Research and Education (IDOR), Rio de Janeiro, RJ, Brazil
[f] National Institute of Infectious Disease Evandro Chagas (INI), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, RJ, Brazil

ARTICLE INFO

Article history:
Received 8 February 2019
Accepted 21 April 2019
Available online 27 April 2019

Keywords:
Carbapenem resistance
Multi-drug resistance
Gram-negative
Healthcare-associated infection
Predictive models
Intensive care units

SUMMARY

Background: Infections by multidrug-resistant Gram-negative (MDRGN) bacteria are among the greatest contemporary health concerns, especially in intensive care units (ICUs), and may be associated with increased hospitalization time, morbidity, costs, and mortality.

Aim: The study aimed to predict carbapenem-resistant MDRGN acquisition in ICUs, to determine its risk factors, and to assess the impact of this acquisition on mortality rate.

Methods: A matched case–control study was performed in patients admitted to the ICU at a large Brazilian hospital over a five-year period. Cases were defined as patients who acquired carbapenem-resistant MDRGN bacteria during hospitalization. Controls were defined as patients who had no detection of carbapenem-resistant MDRGN bacteria. Cases were matched to controls according to the admission period. Risk factors were identified by multiple logistic regression using a stepwise selection method.

Findings: In total, 143 cases and 1089 controls were analysed. The 30-day mortality rate for subjects with ICU-associated carbapenem-resistant MDRGN was 37.6%. Five variables were identified as statistically significant and more relevant for the acquisition of multidrug-resistant strains: increased Simplified Acute Physiology Score 3, patients with severe chronic obstructive pulmonary disease and exposure to haemodialysis catheter, central venous catheter, or mechanical ventilation. Models developed displayed good results with an accuracy of ~90%. Patients who acquired MDRGN were 2.72 times more likely to die than non-MDRGN acquisition patients.

Conclusion: Finding risk factors and developing predictive models may benefit patients through early detection and by controlling the spread of MDR. The presence of mechanical ventilation and central venous catheter were the main risk factors demonstrated, and their use requires special attention.

* Corresponding author. Address: Instituto Nacional de Infectologia Evandro Chagas (INI), FIOCRUZ, Av Brasil 4365, Manguinhos, Rio de Janeiro, RJ 21045-900, Brazil. Tel.: +55 21 993031951.
E-mail address: fernando.bozza@ini.fiocruz.br (F.A. Bozza).

https://doi.org/10.1016/j.jhin.2019.04.013
0195-6701/© 2019 The Healthcare Infection Society. Published by Elsevier Ltd. All rights reserved.

Appendix B

Old Protocol for surveillance cultures

All patients with risk factors for multidrug-resistant (MDR) bacteria at the time of admission or during hospitalization were screened with pharyngeal, nasal, and rectal swab cultures.

The risk factors considered at the time of admission were:

- any hospital admission for more than 24 h in the last six months;
- medical or surgical procedures in the previous six months;
- use of antibiotics during the last six months;
- home care; or
- known colonization by any multidrug-resistant micro-organism in the last year.

If any of these risk factors were detected, the patient was submitted to swab cultures and remained in contact isolation until the test result was known.

At any time of hospitalization:

Surveillance cultures were collected from patients who had possible contact with patients colonized by MDR bacteria or weekly from all patients hospitalized in units in outbreak situations.

Our study aimed to analyze only the patients and cultures (surveillance or not) collected after 72 h of admission.

Appendix C

Table C.1 - All variables available in our dataset grouped by the description, also considering the new variables created.

Databases	
Code	**Description**
1	Laboratory tests
2	Information on ICU admission: Identification, Demographic Data, Diagnostics, and ICD-10.
3	Comorbidities and Functional Capacity
4	Invasive Device Use
5	Reasons for ICU admission
6	Antibiotic use
7	**New variables - variables created from variables extracted**

Laboratory Tests	**Patient Information**	**ICU Information**	**Hospital Information**	**Index**	**Comorbidities**	**Invasive Device during Hospitalization**	**Reasons for ICU admission**	**Antibiotic use**
TestDate	EpimedCode/PatientId	UnitCode	HospitalCode	Charlson Comorbidity Index	ChronicHealthStatusName	InvasiveDeviceGroup	ICDCode	AntibioticUseDate
ExamType	MedicalRecord	UnitAdmissionDate	HospitalAdmissionDate	MFI points	IsChfNyhaClass23	PlacementDate	ICDName	DrugName
ExamDescription	Age	UnitDischargeName	HospitalDischargeName	Frail Patient MFI	IsChfNyhaClass4	RemovalDate	AdmissionSource	Amount
Antibiogram	Gender	Unit Discharge Date	HospitalDischargeDate	Saps3Points	IsCrfNoDialysis	CatheterDuration	AdmissionType	ValuePayment
tests_before	BMI	UnitDestinationName	HospitalDestinationName	Sofa Score	IsCrfDialysis	InvasiveDeviceType	AdmissionReason	**ATC Classification: J01C, J01D, J01E, J01F, J01G, J01M, J01X, and J04A**
RESULT	IsHospitalReadmission	LengthHospitalStayPriorUnitAdmission	HospitalLengthStay		IsCirrhosisChildAB	InvasiveDeviceSite	AdmissionMainDiagnosis	**Antibiotic**
	IsReadmission24h	UnitLengthStay	**LOS_ICU_before_test**		IsCirrhosisChildC	**VesDURTOTAL**	PriorityType	
	IsReadmission48h		**LOS_hospital_before_test**		IsHepaticFailure	**VesDURMORE**	IsNeurologicalComaStuporObtundedDelirium	
					IsSolidTumorLocoregional	**VesTIMESTOTAL**	IsNeurologicalSeizures	
					IsSolidTumorMetastatic	**VesTIMESMORE**	IsNeurologicalFocalNeurologicDeficit	
					Anatomic Tumor Site Name	**VESICAL**	IsNeurologicalIntracranialMassEffect	
					IsHematologicalMalignancy	**CVCDURTOTAL**	IsCardiovascularHypovolemicHemorrhagicShock	
					HematologicalMalignancyTypeCode	**CVCDURMORE**	IsCardiovascularSepticShock	
					HematologicalMalignancyTypeName	**CVCTIMESTOTAL**	IsCardiovascularRhythmDisturbances	
					IsImmunossupression	**CVCTIMESMORE**	IsCardiovascularAnaphylacticMixedUndefinedShock	

Laboratory tests	Patient Information	ICU Information	Hospital Information	Index	Comorbidities	Invasive Device during Hospitalization	Reasons for ICU admission	Antibiotic use
					IsSevereCopd	CVC	IsDigestiveAcuteAbdomen	
					IsSteroidsUse	DiaDURTOTAL	IsDigestiveSeverePancreatitis	
					IsAids	DiaDURMORE	IsLiverFailure	
					IsArterialHypertension	DiaTIMESTOTAL	IsTransplantSolidOrgan	
					IsAsthma	DiaTIMESMORE	IsTraumaMultipleTrauma	
					IsDiabetesUncomplicated	DIALYSIS	IsCardiacSurgery	
					IsDiabetesComplicated	MVDURTOTAL	IsNeurosurgery	
					IsAngina	MVDURMORE		
					IsPreviousMI	MVTIMESTOTAL		
					IsCardiacArrhythmia	MVTIMESMORE		
					IsDeepVenousThrombosis	MV		
					IsPeripheralArteryDisease	PerDURTOTAL		
					IsChronicAtrialFibrilation	PerDURMORE		
					IsRheumaticDisease	PerTIMESTOTAL		
					IsStrokeSequelae	PerTIMESMORE		
					IsStrokeNoSequelae	PERIPHERAL		
					IsDementia	ArtDURTOTAL		
					IsTobaccoConsumption	ArtDURMORE		
					IsAlcoholism	ArtTIMESTOTAL		
					IsPsychiatricDisease	ArtTIMESMORE		
					IsMorbidObesity	ARTERIAL		
					IsMalnourishment			
					Is Peptic Disease			
					Is Solid Organ Transplant			
					Is Autologous BMT			
					Is Allogeneic BMT			
					Is Other Solid Organ Transplant			
					Is Cardiac Transplant			
					Is Combined Liverkidney Transplant			
					Is Combined Pancreaskidney Transplant			
					Is Liver Transplant			
					Is Intestinal Transplant			
					Is Pancreas Transplant			
					Is Lung Transplant			
					Is Kidney Transplant			
					Is Hypothyroidism			
					Is Hyperthyroidism			
					Is Dyslipidemias			
					Transplant			

Appendix D

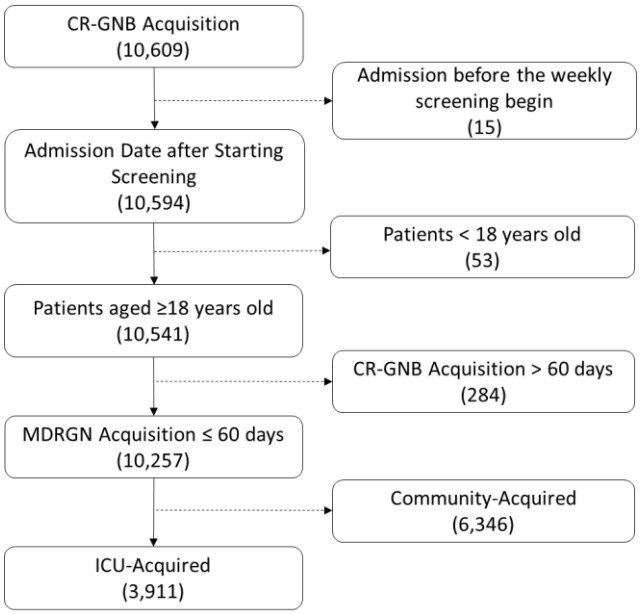


Figure D.1 - Selection of cases.

Appendix E

Table E.1 - Descriptive statistical analysis comparing between Negative and Positive groups.

Variables	Negative (N=3517)	Positive (N=394)	P-value
Laboratory tests			
tests_before			
Mean (SD)	1.38 (1.47)	1.49 (1.41)	0.032
Median [Min, Max]	1.00 [0, 10.0]	1.00 [0, 7.00]	
Hospital Information			
Hospital			
A	250 (7.1%)	60 (15.2%)	<0.001
B	749 (21.3%)	57 (14.5%)	
C	1000 (28.4%)	81 (20.6%)	
D	1518 (43.2%)	196 (49.7%)	
LOS_hospital_before_test			
Mean (SD)	14.8 (12.3)	19.2 (13.7)	<0.001
Median [Min, Max]	10.0 [3.00, 60.0]	15.0 [3.00, 60.0]	
Patient Information			
Age			
Mean (SD)	75.3 (15.4)	75.1 (14.9)	0.541
Median [Min, Max]	79.0 [18.0, 105]	78.0 [18.0, 99.0]	
Gender			
F	1902 (54.1%)	210 (53.3%)	0.809
M	1615 (45.9%)	184 (46.7%)	
BMI			
Mean (SD)	26.7 (13.2)	25.8 (5.48)	0.305
Median [Min, Max]	25.2 [10.6, 283]	24.7 [13.9, 57.4]	
Missing	901 (25.6%)	84 (21.3%)	
ICU Information			
LOS_ICU_before_test			
Mean (SD)	13.0 (11.9)	16.4 (12.7)	<0.001
Median [Min, Max]	9.00 [0, 60.0]	13.0 [0, 60.0]	
Index			
CharlsonIndex			
Mean (SD)	1.77 (1.96)	2.02 (2.06)	0.007
Median [Min, Max]	1.00 [0, 12.0]	2.00 [0, 12.0]	
Missing	2 (0.1%)	0 (0%)	
MFIpoints			
Mean (SD)	2.24 (1.40)	2.39 (1.51)	0.144
Median [Min, Max]	2.00 [0, 8.00]	2.00 [0, 7.00]	
Missing	60 (1.7%)	14 (3.6%)	
FrailPatientMFI			
NO	2876 (81.8%)	308 (78.2%)	0.094
YES	641 (18.2%)	86 (21.8%)	
Saps3Points			
Mean (SD)	52.8 (12.9)	57.0 (13.8)	<0.001
Median [Min, Max]	52.0 [8.00, 104]	56.0 [19.0, 104]	
SofaScore			
Mean (SD)	1.75 (2.91)	2.97 (3.81)	<0.001
Median [Min, Max]	1.00 [0, 17.0]	1.00 [0, 17.0]	
Missing	1108 (31.5%)	124 (31.5%)	
Priority			
Priority 1	419 (11.9%)	81 (20.6%)	0.001
Priority 2	1073 (30.5%)	109 (27.7%)	
Priority 3	2 (0.1%)	0 (0%)	
Priority 4	4 (0.1%)	0 (0%)	
Priority 5	12 (0.3%)	1 (0.3%)	
Missing	2007 (57.1%)	203 (51.5%)	
Comorbidities			
ChronicHealthStatus			
Independent	1872 (53.2%)	179 (45.4%)	0.019
Need for assistance	812 (23.1%)	106 (26.9%)	
Restricted / bedridden	824 (23.4%)	105 (26.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsChfNyhaClass23			
FALSE	3256 (92.6%)	359 (91.1%)	0.653
TRUE	252 (7.2%)	31 (7.9%)	
Missing	9 (0.3%)	4 (1.0%)	
IsChfNyhaClass4			
FALSE	3486 (99.1%)	387 (98.2%)	1
TRUE	22 (0.6%)	3 (0.8%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCrfNoDialysis			
FALSE	3132 (89.1%)	347 (88.1%)	0.921
TRUE	376 (10.7%)	43 (10.9%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCrfDialysis			
FALSE	3424 (97.4%)	379 (96.2%)	0.73
TRUE	84 (2.4%)	11 (2.8%)	
Missing	9 (0.3%)	4 (1.0%)	

Variables	Negative (N=3517)	Positive (N=394)	P-value
IsCirrhosisChildAB			
FALSE	3497 (99.4%)	388 (98.5%)	0.854
TRUE	11 (0.3%)	2 (0.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCirrhosisChildC			
FALSE	3505 (99.7%)	389 (98.7%)	0.868
TRUE	3 (0.1%)	1 (0.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsHepaticFailure			
FALSE	3507 (99.7%)	389 (98.7%)	0.48
TRUE	1 (0.0%)	1 (0.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsSolidTumorLocoregiol			
FALSE	2911 (82.8%)	313 (79.4%)	0.201
TRUE	597 (17.0%)	77 (19.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsSolidTumorMetastatic			
FALSE	3375 (96.0%)	376 (95.4%)	0.954
TRUE	133 (3.8%)	14 (3.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsHematologicalMalignancy			
FALSE	3446 (98.0%)	380 (96.4%)	0.363
TRUE	62 (1.8%)	10 (2.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsImmunossupression			
FALSE	3197 (90.9%)	351 (89.1%)	0.516
TRUE	311 (8.8%)	39 (9.9%)	
Missing	9 (0.3%)	4 (1.0%)	
IsSevereCopd			
FALSE	3122 (88.8%)	335 (85.0%)	0.08
TRUE	386 (11.0%)	55 (14.0%)	
Missing	9 (0.3%)	4 (1.0%)	
IsSteroidsUse			
FALSE	3404 (96.8%)	381 (96.7%)	0.566
TRUE	104 (3.0%)	9 (2.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsAids			
FALSE	3473 (98.7%)	389 (98.7%)	0.241
TRUE	35 (1.0%)	1 (0.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsArterialHypertension			
FALSE	1197 (34.0%)	138 (35.0%)	0.658
TRUE	2311 (65.7%)	252 (64.0%)	
Missing	9 (0.3%)	4 (1.0%)	
IsAsthma			
FALSE	3402 (96.7%)	371 (94.2%)	0.069
TRUE	106 (3.0%)	19 (4.8%)	
Missing	9 (0.3%)	4 (1.0%)	
IsDiabetesUncomplicated			
FALSE	2624 (74.6%)	289 (73.4%)	0.811
TRUE	884 (25.1%)	101 (25.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsDiabetesComplicated			
FALSE	3292 (93.6%)	363 (92.1%)	0.629
TRUE	216 (6.1%)	27 (6.9%)	
Missing	9 (0.3%)	4 (1.0%)	
IsAngina			
FALSE	3269 (92.9%)	376 (95.4%)	0.019
TRUE	239 (6.8%)	14 (3.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsPreviousMI			
FALSE	3093 (87.9%)	344 (87.3%)	1
TRUE	415 (11.8%)	46 (11.7%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCardiacArrhythmia			
FALSE	3160 (89.8%)	353 (89.6%)	0.855
TRUE	348 (9.9%)	37 (9.4%)	
Missing	9 (0.3%)	4 (1.0%)	
IsDeepVenousThrombosis			
FALSE	3346 (95.1%)	359 (91.1%)	0.006
TRUE	162 (4.6%)	31 (7.9%)	
Missing	9 (0.3%)	4 (1.0%)	
IsPeripheralArteryDisease			
FALSE	3400 (96.7%)	384 (97.5%)	0.12
TRUE	108 (3.1%)	6 (1.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsChronicAtrialFibrilation			
FALSE	2994 (85.1%)	328 (83.2%)	0.56
TRUE	514 (14.6%)	62 (15.7%)	
Missing	9 (0.3%)	4 (1.0%)	
IsRheumaticDisease			
FALSE	3489 (99.2%)	386 (98.0%)	0.403
TRUE	19 (0.5%)	4 (1.0%)	
Missing	9 (0.3%)	4 (1.0%)	

Variables	Negative (N=3517)	Positive (N=394)	P-value
IsStrokeSequelae			
FALSE	3374 (95.9%)	357 (90.6%)	<0.001
TRUE	134 (3.8%)	33 (8.4%)	
Missing	9 (0.3%)	4 (1.0%)	
IsStrokeNoSequelae			
FALSE	3225 (91.7%)	368 (93.4%)	0.111
TRUE	283 (8.0%)	22 (5.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsDementia			
FALSE	2782 (79.1%)	304 (77.2%)	0.576
TRUE	726 (20.6%)	86 (21.8%)	
Missing	9 (0.3%)	4 (1.0%)	
IsTobaccoConsumption			
FALSE	3247 (92.3%)	355 (90.1%)	0.325
TRUE	261 (7.4%)	35 (8.9%)	
Missing	9 (0.3%)	4 (1.0%)	
IsAlcoholism			
FALSE	3395 (96.5%)	374 (94.9%)	0.439
TRUE	113 (3.2%)	16 (4.1%)	
Missing	9 (0.3%)	4 (1.0%)	
IsPsychiatricDisease			
FALSE	3244 (92.2%)	360 (91.4%)	0.986
TRUE	264 (7.5%)	30 (7.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsMorbidObesity			
FALSE	3396 (96.6%)	381 (96.7%)	0.422
TRUE	112 (3.2%)	9 (2.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsMalnourishment			
FALSE	3493 (99.3%)	387 (98.2%)	0.582
TRUE	15 (0.4%)	3 (0.8%)	
Missing	9 (0.3%)	4 (1.0%)	
IsPepticDisease			
FALSE	3500 (99.5%)	388 (98.5%)	0.598
TRUE	8 (0.2%)	2 (0.5%)	
Missing	9 (0.3%)	4 (1.0%)	
Transplant			
FALSE	3157 (89.8%)	349 (88.6%)	0.82
TRUE	351 (10.0%)	41 (10.4%)	
Missing	9 (0.3%)	4 (1.0%)	
IsHypothyroidism			
FALSE	2926 (83.2%)	317 (80.5%)	0.32
TRUE	582 (16.5%)	73 (18.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsHyperthyroidism			
FALSE	3504 (99.6%)	388 (98.5%)	0.221
TRUE	4 (0.1%)	2 (0.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsDyslipidemias			
FALSE	2928 (83.3%)	327 (83.0%)	0.905
TRUE	580 (16.5%)	63 (16.0%)	
Missing	9 (0.3%)	4 (1.0%)	
IsChemotherapy			
FALSE	3367 (95.7%)	366 (92.9%)	0.064
TRUE	141 (4.0%)	24 (6.1%)	
Missing	9 (0.3%)	4 (1.0%)	
IsRadiationTherapy			
FALSE	3416 (97.1%)	376 (95.4%)	0.342
TRUE	92 (2.6%)	14 (3.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsHistoryOfPneumonia			
FALSE	3317 (94.3%)	360 (91.4%)	0.088
TRUE	191 (5.4%)	30 (7.6%)	
Missing	9 (0.3%)	4 (1.0%)	
Invasive Device during Hospitalization			
VesDURTOTAL			
Mean (SD)	6.72 (8.90)	10.9 (9.79)	<0.001
Median [Min, Max]	3.00 [0, 57.0]	9.00 [0, 52.0]	
VesDURMORE			
Mean (SD)	1.61 (3.41)	2.67 (3.98)	<0.001
Median [Min, Max]	0 [0, 56.0]	0 [0, 24.0]	
VesTIMESTOTAL			
Mean (SD)	0.928 (0.940)	1.30 (0.900)	<0.001
Median [Min, Max]	1.00 [0, 7.00]	1.00 [0, 5.00]	
VesTIMESMORE			
Mean (SD)	0.0836 (0.311)	0.124 (0.345)	0.003
Median [Min, Max]	0 [0, 5.00]	0 [0, 2.00]	
VESICAL			
NO	1231 (35.0%)	59 (15.0%)	<0.001
YES	2286 (65.0%)	335 (85.0%)	
ArtDURTOTAL			
Mean (SD)	3.88 (6.75)	7.83 (9.25)	<0.001
Median [Min, Max]	0 [0, 59.0]	5.00 [0, 53.0]	
ArtDURMORE			

Variables	Negative (N=3517)	Positive (N=394)	P-value
Mean (SD)	0.803 (2.40)	1.88 (3.66)	<0.001
Median [Min, Max]	0 [0, 39.0]	0 [0, 22.0]	
ArtTIMESTOTAL			
Mean (SD)	0.606 (0.868)	1.08 (1.06)	<0.001
Median [Min, Max]	0 [0, 6.00]	1.00 [0, 5.00]	
ArtTIMESMORE			
Mean (SD)	0.0427 (0.227)	0.109 (0.336)	<0.001
Median [Min, Max]	0 [0, 3.00]	0 [0, 2.00]	
ARTERIAL			
NO	2073 (58.9%)	143 (36.3%)	<0.001
YES	1444 (41.1%)	251 (63.7%)	
DiaDURTOTAL			
Mean (SD)	1.07 (4.31)	2.84 (6.91)	<0.001
Median [Min, Max]	0 [0, 41.0]	0 [0, 42.0]	
DiaDURMORE			
Mean (SD)	0.273 (1.54)	0.665 (2.48)	<0.001
Median [Min, Max]	0 [0, 31.0]	0 [0, 21.0]	
DiaTIMESTOTAL			
Mean (SD)	0.150 (0.542)	0.378 (0.827)	<0.001
Median [Min, Max]	0 [0, 5.00]	0 [0, 6.00]	
DiaTIMESMORE			
Mean (SD)	0.0199 (0.159)	0.0431 (0.203)	<0.001
Median [Min, Max]	0 [0, 2.00]	0 [0, 1.00]	
DIALYSIS			
NO	3187 (90.6%)	304 (77.2%)	<0.001
YES	330 (9.4%)	90 (22.8%)	
CVCDURTOTAL			
Mean (SD)	6.35 (8.73)	11.1 (10.2)	<0.001
Median [Min, Max]	3.00 [0, 60.0]	9.00 [0, 51.0]	
CVCDURMORE			
Mean (SD)	1.47 (3.34)	2.95 (4.59)	<0.001
Median [Min, Max]	0 [0, 56.0]	0 [0, 32.0]	
CVCTIMESTOTAL			
Mean (SD)	0.849 (0.975)	1.38 (1.09)	<0.001
Median [Min, Max]	1.00 [0, 6.00]	1.00 [0, 6.00]	
CVCTIMESMORE			
Mean (SD)	0.0893 (0.320)	0.193 (0.455)	<0.001
Median [Min, Max]	0 [0, 5.00]	0 [0, 3.00]	
CVC			
NO	1560 (44.4%)	81 (20.6%)	<0.001
YES	1957 (55.6%)	313 (79.4%)	
MVDURTOTAL			
Mean (SD)	4.19 (8.75)	8.51 (11.0)	<0.001
Median [Min, Max]	0 [0, 57.0]	5.00 [0, 49.0]	
MVDURMORE			
Mean (SD)	0.978 (2.79)	2.35 (4.61)	<0.001
Median [Min, Max]	0 [0, 56.0]	0 [0, 33.0]	
MVTIMESTOTAL			
Mean (SD)	0.400 (0.649)	0.766 (0.782)	<0.001
Median [Min, Max]	0 [0, 4.00]	1.00 [0, 5.00]	
MVTIMESMORE			
Mean (SD)	0.0205 (0.151)	0.0381 (0.192)	0.013
Median [Min, Max]	0 [0, 2.00]	0 [0, 1.00]	
MV			
NO	2379 (67.6%)	159 (40.4%)	<0.001
YES	1138 (32.4%)	235 (59.6%)	
PerDURTOTAL			
Mean (SD)	1.58 (3.84)	1.16 (2.76)	0.273
Median [Min, Max]	0 [0, 56.0]	0 [0, 26.0]	
PerDURMORE			
Mean (SD)	0.323 (1.33)	0.226 (0.966)	0.479
Median [Min, Max]	0 [0, 17.0]	0 [0, 7.00]	
PerTIMESTOTAL			
Mean (SD)	0.579 (1.27)	0.447 (0.980)	0.272
Median [Min, Max]	0 [0, 15.0]	0 [0, 8.00]	
PerTIMESMORE			
Mean (SD)	0.0887 (0.405)	0.0609 (0.321)	0.199
Median [Min, Max]	0 [0, 5.00]	0 [0, 3.00]	
PERIPHERAL			
NO	2543 (72.3%)	291 (73.9%)	0.552
YES	974 (27.7%)	103 (26.1%)	
Reasons for ICU admission			
AdmissionSource			
Emergency	2020 (57.4%)	188 (47.7%)	<0.001
Hemodynamic Room	55 (1.6%)	3 (0.8%)	
Operation Room	364 (10.3%)	45 (11.4%)	
Other ICU from hospital	419 (11.9%)	71 (18.0%)	
Others	24 (0.7%)	7 (1.8%)	
Semi Intensive Unit	201 (5.7%)	28 (7.1%)	
Transfer from another hospital	33 (0.9%)	10 (2.5%)	
Ward/Room	392 (11.1%)	38 (9.6%)	
Missing	9 (0.3%)	4 (1.0%)	
AdmissionReason			
Cardiovascular / Shock	846 (24.1%)	48 (12.2%)	<0.001

Variables	Negative (N=3517)	Positive (N=394)	P-value
Elective Surgery	253 (7.2%)	29 (7.4%)	
Emergency surgery	182 (5.2%)	18 (4.6%)	
Endocrine / Metabolic / Renal	85 (2.4%)	10 (2.5%)	
Infection / Sepsis	1200 (34.1%)	170 (43.1%)	
Liver and Pancreas / Gastrointestinal	193 (5.5%)	16 (4.1%)	
Neurological	303 (8.6%)	43 (10.9%)	
Non-surgical trauma	80 (2.3%)	10 (2.5%)	
Oncological / Hematological	67 (1.9%)	8 (2.0%)	
Others	60 (1.7%)	8 (2.0%)	
Respiratory	239 (6.8%)	30 (7.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsNeurologicalComaStuporObtundedDelirium			
FALSE	2968 (84.4%)	301 (76.4%)	<0.001
TRUE	540 (15.4%)	89 (22.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsNeurologicalSeizures			
FALSE	3350 (95.3%)	364 (92.4%)	0.074
TRUE	158 (4.5%)	26 (6.6%)	
Missing	9 (0.3%)	4 (1.0%)	
IsNeurologicalFocalNeurologicDeficit			
FALSE	3435 (97.7%)	373 (94.7%)	0.008
TRUE	73 (2.1%)	17 (4.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsNeurologicalIntracranialMassEffect			
FALSE	3467 (98.6%)	384 (97.5%)	0.696
TRUE	41 (1.2%)	6 (1.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCardiovascularHypovolemicHemorrhagicShock			
FALSE	3470 (98.7%)	381 (96.7%)	0.063
TRUE	38 (1.1%)	9 (2.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCardiovascularSepticShock			
FALSE	3335 (94.8%)	344 (87.3%)	<0.001
TRUE	173 (4.9%)	46 (11.7%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCardiovascularRhythmDisturbances			
FALSE	3058 (86.9%)	346 (87.8%)	0.429
TRUE	450 (12.8%)	44 (11.2%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCardiovascularAphylacticMixedUndefinedShock			
FALSE	3503 (99.6%)	389 (98.7%)	1
TRUE	5 (0.1%)	1 (0.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsDigestiveAcuteAbdomen			
FALSE	3418 (97.2%)	378 (95.9%)	0.665
TRUE	90 (2.6%)	12 (3.0%)	
Missing	9 (0.3%)	4 (1.0%)	
IsDigestiveSeverePancreatitis			
FALSE	3496 (99.4%)	389 (98.7%)	1
TRUE	12 (0.3%)	1 (0.3%)	
Missing	9 (0.3%)	4 (1.0%)	
IsLiverFailure			
FALSE	3492 (99.3%)	388 (98.5%)	1
TRUE	16 (0.5%)	2 (0.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsTransplantSolidOrgan			
FALSE	3501 (99.5%)	388 (98.5%)	0.505
TRUE	7 (0.2%)	2 (0.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsTraumaMultipleTrauma			
FALSE	3422 (97.3%)	380 (96.4%)	1
TRUE	86 (2.4%)	10 (2.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsCardiacSurgery			
FALSE	3471 (98.7%)	388 (98.5%)	0.452
TRUE	37 (1.1%)	2 (0.5%)	
Missing	9 (0.3%)	4 (1.0%)	
IsNeurosurgery			
FALSE	3496 (99.4%)	389 (98.7%)	1
TRUE	12 (0.3%)	1 (0.3%)	
Missing	9 (0.3%)	4 (1.0%)	
Antibiotic use			
J01A			
FALSE	3406 (96.8%)	350 (88.8%)	<0.001
TRUE	111 (3.2%)	44 (11.2%)	
J01C			
FALSE	1337 (38.0%)	114 (28.9%)	<0.001
TRUE	2180 (62.0%)	280 (71.1%)	
J01D			
FALSE	1625 (46.2%)	88 (22.3%)	<0.001
TRUE	1892 (53.8%)	306 (77.7%)	
J01E			
FALSE	3395 (96.5%)	368 (93.4%)	0.003
TRUE	122 (3.5%)	26 (6.6%)	

Variables	Negative (N=3517)	Positive (N=394)	P-value
J01F			
FALSE	2378 (67.6%)	233 (59.1%)	<0.001
TRUE	1139 (32.4%)	161 (40.9%)	
J01G			
FALSE	3300 (93.8%)	337 (85.5%)	<0.001
TRUE	217 (6.2%)	57 (14.5%)	
J01M			
FALSE	2975 (84.6%)	333 (84.5%)	1
TRUE	542 (15.4%)	61 (15.5%)	
J01X			
FALSE	2442 (69.4%)	163 (41.4%)	<0.001
TRUE	1075 (30.6%)	231 (58.6%)	
J04A			
FALSE	3494 (99.3%)	391 (99.2%)	1
TRUE	23 (0.7%)	3 (0.8%)	
Antibiotic			
FALSE	495 (14.1%)	10 (2.5%)	<0.001
TRUE	3022 (85.9%)	384 (97.5%)	

Appendix F

Visualization of the missingness pattern

We have in Table F.1 four patterns observed from possible patterns. The output table, “1” indicates the non-missing value, and “0” indicates a missing value. The first column shows the number of cases with that pattern, and the second one counts the number of missing variables. We can see, for example, that there are 74 cases where the variable “MFIpoints” is missing, whereas all the other variables are observed. We can also see two instances where only the variable “MFIPoints” has values among the variables presented. 3,824 records are complete cases.

Since the number of variables is too large, it was impossible to show all the table variables. However, all variables that do not appear have a value of "1", which indicates no missing values.

Table F.1 - Missingness Pattern to variables with some missing value (1=observed, 0=missing)

Number of cases	Number of missing values	MFIpoints	CharlsonIndex	Comorbidities and Functional Capacity																																						Reasons for ICU admission																
3824	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
74	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	59	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	60	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Count	843	74	2	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13

Figure F.1 gives us the frequencies for different combinations of variables missing. The blue color refers to observed data and the yellow color to the missing data. For example, the situation in which none of the variables are missing is the most frequent (97%). On the other hand, MFIpoints are missing values in 1.9% of the cases.

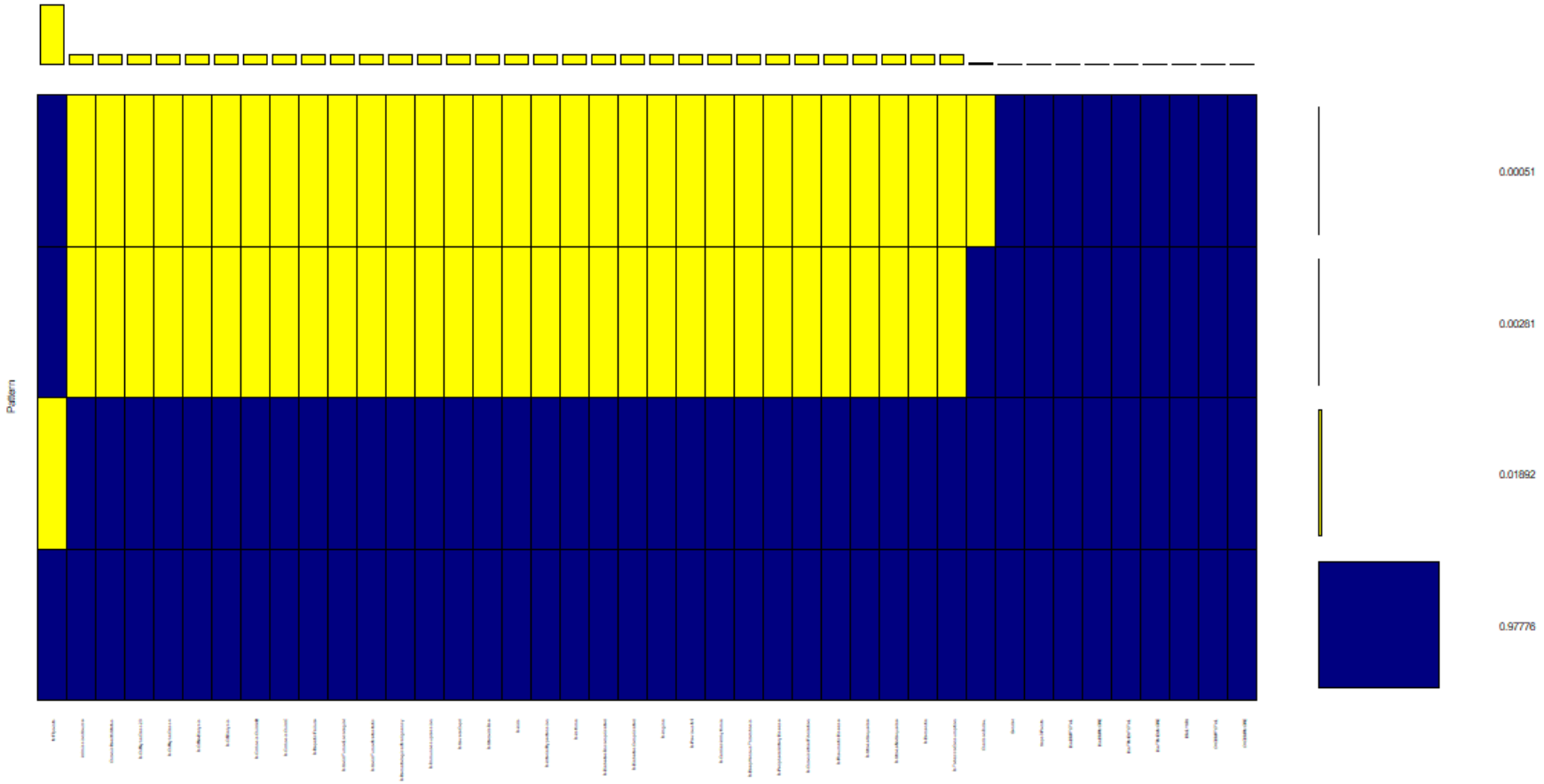


Figure F.1 - Aggregation Plot: combination among variables.

We can observe that when one comorbidity variable is non-available, the other comorbidities have no value either. The same thing happens with the reasons for admission. Thus, we decided to analyze any pattern between these missing values and the other database variables, explaining these missing.

We noted that one comorbidity does not depend on the missing value of another comorbidity. The same goes for the admission reason variables. The reason is that the information is not recorded at the time of ICU admission. Thus, we conclude that these variables may depend on the model's observed variables, not on the missing values. We consider that all these data are missing at random, and we use imputation to model them using the observed data.

Appendix G

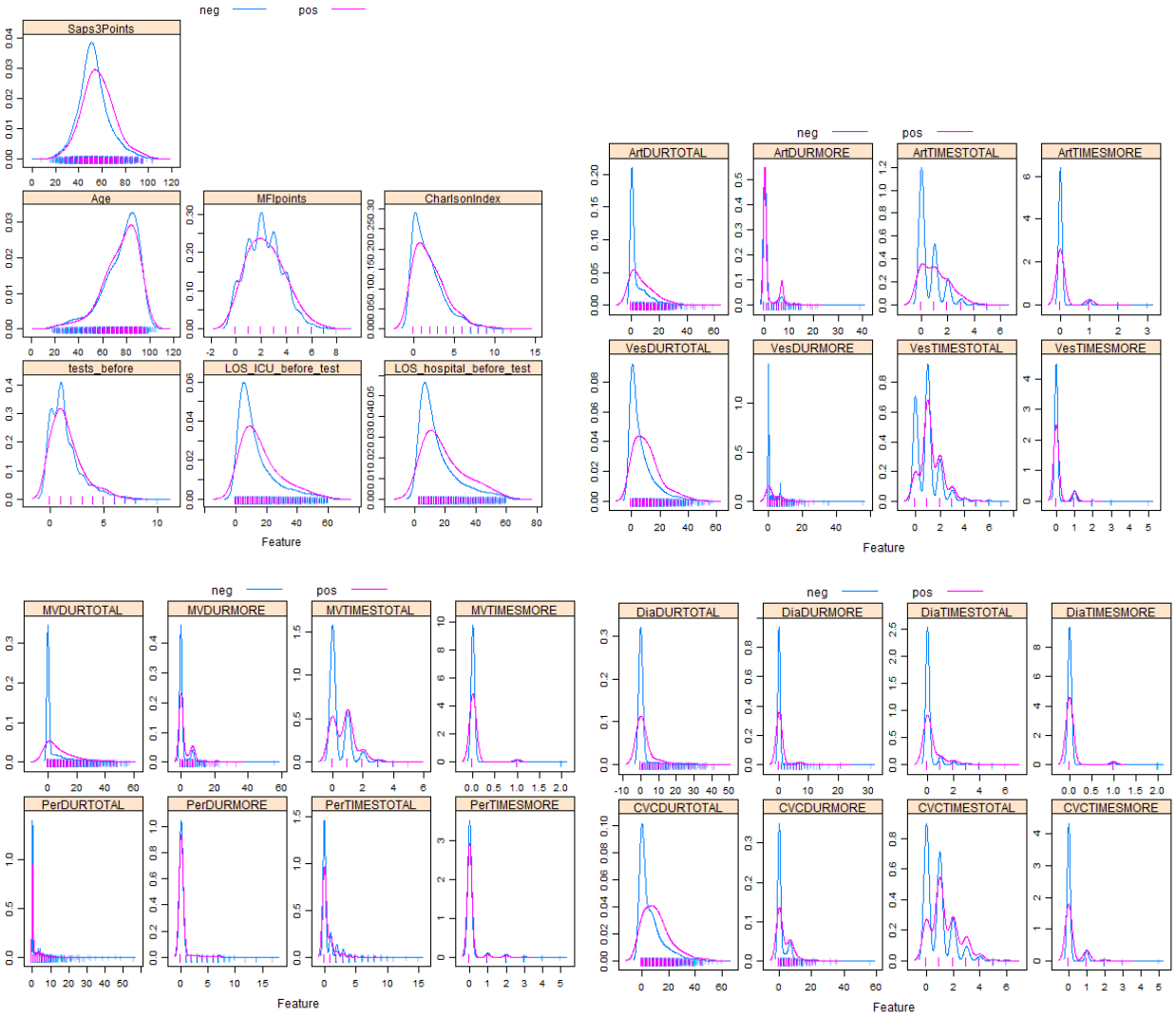


Figure G.1 - Overlaid Density Plots for each continuous variable.

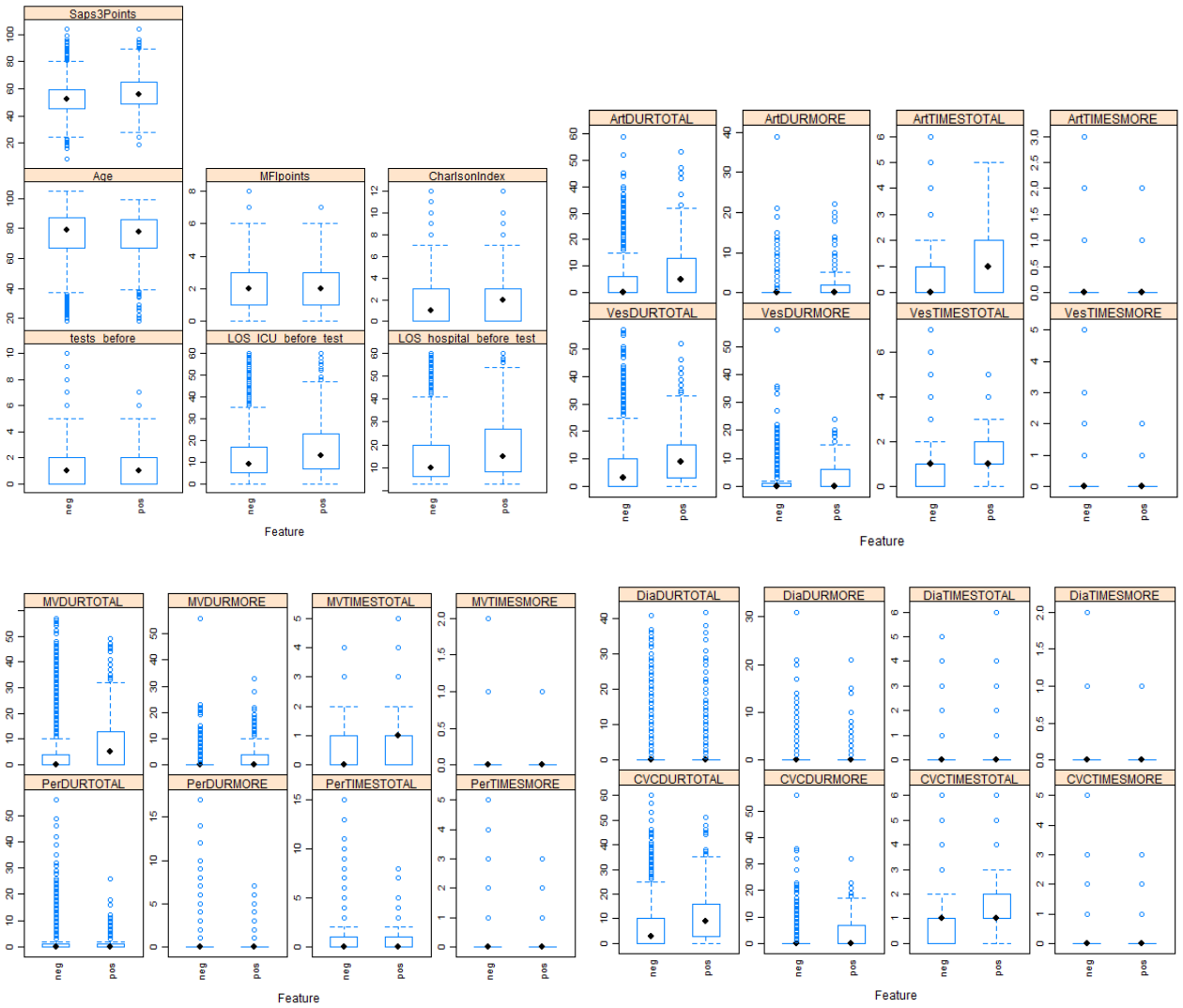


Figure G.2 - Boxplots for each continuous variable.

Table G.1 - Box-plot analysis for each continuous variable with the values of minimum, first and third quartile, IQR, maximum, and amount of outliers.

Variables	min	Q1	Q3	IQR	max	outlier_values
tests_before	0	0	2	2	5	87
LOS_ICU_before_test	0	5	18	13	38	247
LOS_hospital_before_test	0	6	20	14	41	225
Age	37	67	87	20	117	126
MFIpoints	0	1	3	2	6	13
CharlsonIndex	0	0	3	3	8	77
Saps3Points	22	45	60	15	82	141
VesDURTOTAL	0	0	11	11	28	178
VesDURMORE	0	0	2	2	5	636
VesTIMESTOTAL	0	0	1	1	2	230
VesTIMESMORE	0	0	0	0	0	316
ArtDURTOTAL	0	0	7	7	18	250
ArtDURMORE	0	0	0	0	0	588
ArtTIMESTOTAL	0	0	1	1	2	168
ArtTIMESMORE	0	0	0	0	0	173
PerDURTOTAL	0	0	1	1	2	825

Variables	min	Q1	Q3	IQR	max	outlier_values
PerDURMORE	0	0	0	0	0	300
PerTIMESTOTAL	0	0	1	1	2	283
PerTIMESMORE	0	0	0	0	0	213
MVDURTOTAL	0	0	5.5	6	14	515
MVDURMORE	0	0	0	0	0	651
MVTIMESTOTAL	0	0	1	1	2	51
MVTIMESMORE	0	0	0	0	0	82
CVCDURTOTAL	0	0	10	10	25	210
CVCDURMORE	0	0	1	1	2	845
CVCTIMESTOTAL	0	0	1	1	2	296
CVCTIMESMORE	0	0	0	0	0	352
DiaDURTOTAL	0	0	0	0	0	420
DiaDURMORE	0	0	0	0	0	180
DiaTIMESTOTAL	0	0	0	0	0	420
DiaTIMESMORE	0	0	0	0	0	77

Appendix H

Table H.1 - Results of the zero-variance analysis. Values considered "nzv" or zeroVar "are highlighted in red.

Variables	freqRatio	percentUnique	zeroVar	nzv
Hospital	1.60	0.13	FALSE	FALSE
LOS_ICU_before_test	1.30	1.95	FALSE	FALSE
LOS_hospital_before_test	1.27	1.85	FALSE	FALSE
tests_before	1.35	0.29	FALSE	FALSE
Age	1.03	2.72	FALSE	FALSE
Gender	1.17	0.06	FALSE	FALSE
AdmissionSource	4.48	0.26	FALSE	FALSE
AdmissionReason	1.55	0.35	FALSE	FALSE
CharlsonIndex	1.49	0.42	FALSE	FALSE
MFIpoints	1.24	0.29	FALSE	FALSE
FrailPatientMFI	4.29	0.06	FALSE	FALSE
Saps3Points	1.04	2.62	FALSE	FALSE
ChronicHealthStatus	2.15	0.10	FALSE	FALSE
IsChfNyhaClass23	13.05	0.06	FALSE	FALSE
IsChfNyhaClass4	154.95	0.06	FALSE	TRUE
IsCrfNoDialysis	8.26	0.06	FALSE	FALSE
IsCrfDialysis	42.32	0.06	FALSE	FALSE
IsCirrhosisChildAB	310.90	0.06	FALSE	TRUE
IsCirrhosisChildC	1038.67	0.06	FALSE	TRUE
IsHepaticFailure	1558.50	0.06	FALSE	TRUE
IsSolidTumorLocoregiol	4.77	0.06	FALSE	FALSE
IsSolidTumorMetastatic	24.36	0.06	FALSE	FALSE
IsHematologicalMalignancy	57.85	0.06	FALSE	TRUE
IsImmunossupression	10.55	0.06	FALSE	FALSE
IsSevereCopd	7.96	0.06	FALSE	FALSE
IsSteroidsUse	37.99	0.06	FALSE	FALSE
IsAids	123.76	0.06	FALSE	TRUE
IsArterialHypertension	1.94	0.06	FALSE	FALSE
IsAsthma	30.83	0.06	FALSE	FALSE
IsDiabetesUncomplicated	2.98	0.06	FALSE	FALSE
IsDiabetesComplicated	14.14	0.06	FALSE	FALSE
IsAngina	14.07	0.06	FALSE	FALSE
IsPreviousMI	7.41	0.06	FALSE	FALSE
IsCardiacArrhythmia	8.90	0.06	FALSE	FALSE
IsDeepVenousThrombosis	20.81	0.06	FALSE	FALSE
IsPeripheralArteryDisease	30.19	0.06	FALSE	FALSE
IsChronicAtrialFibrilation	5.68	0.06	FALSE	FALSE
IsRheumaticDisease	140.77	0.06	FALSE	TRUE
IsStrokeSequelae	20.81	0.06	FALSE	FALSE
IsStrokeNoSequelae	11.78	0.06	FALSE	FALSE
IsDementia	3.73	0.06	FALSE	FALSE
IsTobaccoConsumption	12.56	0.06	FALSE	FALSE
IsAlcoholism	31.83	0.06	FALSE	FALSE
IsPsychiatricDisease	12.68	0.06	FALSE	FALSE
IsMorbidObesity	32.54	0.06	FALSE	FALSE
IsMalnourishment	206.93	0.06	FALSE	TRUE
IsPepticDisease	444.57	0.06	FALSE	TRUE
Transplant	8.69	0.06	FALSE	FALSE
IsHypothyroidism	4.85	0.06	FALSE	FALSE
IsHyperthyroidism	622.80	0.06	FALSE	TRUE
IsDyslipidemias	4.93	0.06	FALSE	FALSE
IsChemotherapy	21.60	0.06	FALSE	FALSE
IsRadiationTherapy	35.27	0.06	FALSE	FALSE
IsHistoryOfPneumonia	17.24	0.06	FALSE	FALSE
VesDURTOTAL	4.74	1.63	FALSE	FALSE
VesDURMORE	7.30	0.86	FALSE	FALSE
VesTIMESTOTAL	1.38	0.26	FALSE	FALSE
VesTIMESMORE	11.87	0.13	FALSE	FALSE
VESICAL	2.02	0.06	FALSE	FALSE
ArtDURTOTAL	13.70	1.44	FALSE	FALSE
ArtDURMORE	14.83	0.67	FALSE	FALSE
ArtTIMESTOTAL	2.11	0.22	FALSE	FALSE
ArtTIMESMORE	24.37	0.13	FALSE	FALSE
ARTERIAL	1.30	0.06	FALSE	FALSE

Variables	freqRatio	percentUnique	zeroVar	nzv
DiaDURTOTAL	99.86	1.21	FALSE	TRUE
DiaDURMORE	49.68	0.61	FALSE	FALSE
DiaTIMESTOTAL	13.71	0.22	FALSE	FALSE
DiaTIMESMORE	62.73	0.10	FALSE	TRUE
DIALYSIS	8.37	0.06	FALSE	FALSE
CVCDURTOTAL	8.80	1.60	FALSE	FALSE
CVCDURMORE	7.74	0.83	FALSE	FALSE
CVCTIMESTOTAL	1.15	0.22	FALSE	FALSE
CVCTIMESMORE	11.09	0.13	FALSE	FALSE
CVC	1.40	0.06	FALSE	FALSE
MVDURTOTAL	27.27	1.69	FALSE	FALSE
MVDURMORE	11.33	0.73	FALSE	FALSE
MVTIMESTOTAL	2.25	0.19	FALSE	FALSE
MVTIMESMORE	52.88	0.10	FALSE	TRUE
MV	1.81	0.06	FALSE	FALSE
PerDURTOTAL	15.81	1.18	FALSE	FALSE
PerDURMORE	57.56	0.42	FALSE	TRUE
PerTIMESTOTAL	5.36	0.45	FALSE	FALSE
PerTIMESMORE	29.83	0.19	FALSE	FALSE
PERIPHERAL	2.67	0.06	FALSE	FALSE
IsNeurologicalComaStuporObtundedDelirium	5.19	0.06	FALSE	FALSE
IsNeurologicalSeizures	20.07	0.06	FALSE	FALSE
IsNeurologicalFocalNeurologicDeficit	41.73	0.06	FALSE	FALSE
IsNeurologicalIntracranialMassEffect	71.53	0.06	FALSE	TRUE
IsCardiovascularHypovolemicHemorrhagicShock	75.07	0.06	FALSE	TRUE
IsCardiovascularSepticShock	16.62	0.06	FALSE	FALSE
IsCardiovascularRhythmDisturbances	7.08	0.06	FALSE	FALSE
IsCardiovascularAphylacticMixedUndefinedShock	518.83	0.06	FALSE	TRUE
IsDigestiveAcuteAbdomen	37.51	0.06	FALSE	FALSE
IsDigestiveSeverePancreatitis	282.55	0.06	FALSE	TRUE
IsLiverFailure	221.79	0.06	FALSE	TRUE
IsTransplantSolidOrgan	518.83	0.06	FALSE	TRUE
IsTraumaMultipleTrauma	37.99	0.06	FALSE	FALSE
IsCardiacSurgery	81.08	0.06	FALSE	TRUE
IsNeurosurgery	282.55	0.06	FALSE	TRUE
J01A	25.30	0.06	FALSE	FALSE
J01C	1.71	0.06	FALSE	FALSE
J01D	1.30	0.06	FALSE	FALSE
J01E	26.70	0.06	FALSE	FALSE
J01F	2.03	0.06	FALSE	FALSE
J01G	13.49	0.06	FALSE	FALSE
J01M	5.59	0.06	FALSE	FALSE
J01X	1.98	0.06	FALSE	FALSE
J04A	148.05	0.06	FALSE	TRUE
Antibiotic	6.65	0.06	FALSE	FALSE

Legend:

zeroVar - if the predictor has only one distinct value;

nzv - if the predictor is a near zero variance predictor;

freqCut = the cutoff for the ratio of the most common value to the second most common value;

uniqueCut = the percentage of distinct values out of the number of total samples;

Cut - freqCut = 100/2, uniqueCut = 5

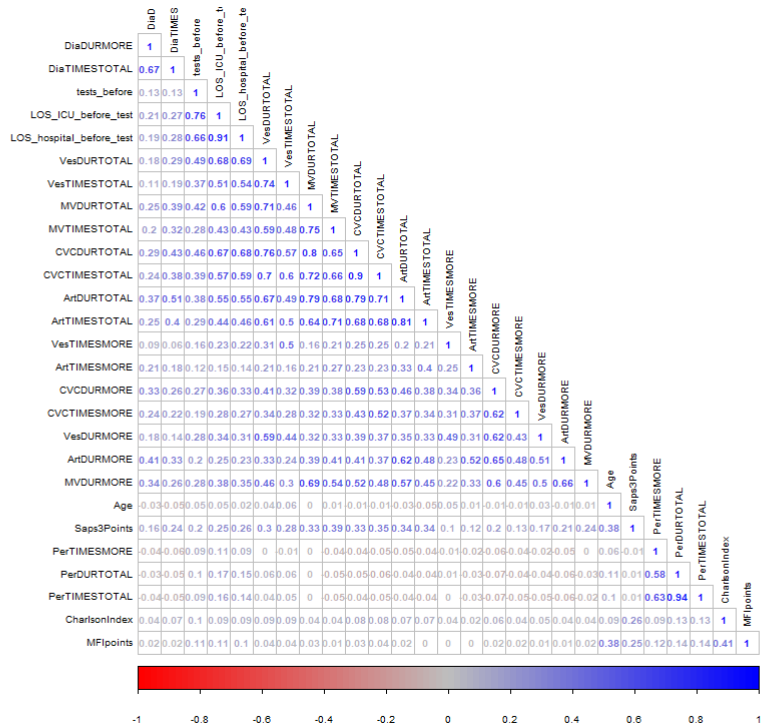


Figure H.1 - Correlation between continuous variables using the Pearson method.

Table H.2 - Results of the strength of the relationship between the categorical variables by the Goodman and Kruskal's tau (or lambda) measure. Pairs with an association higher than 0.40 are highlighted in red.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	4.00	0.01	0.05	0.01	0.01	0.10	0.02	0.03	0.00	0.03	0.01	0.03	0.02	0.02	0.02	0.01	0.00	0.02	0.03	0.01	0.06	0.02	0.00	0.01	0.03	0.00	0.01	0.01	0.01	0.01
2	0.00	2.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
3	0.07	0.01	8.00	0.10	0.01	0.03	0.01	0.02	0.00	0.01	0.00	0.02	0.02	0.02	0.01	0.00	0.00	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.00	0.03	0.01	0.01	0.01
4	0.01	0.01	0.21	11.00	0.02	0.03	0.06	0.04	0.01	0.02	0.01	0.03	0.08	0.00	0.03	0.02	0.01	0.01	0.06	0.02	0.02	0.01	0.01	0.02	0.02	0.02	0.07	0.02	0.01	0.01
5	0.01	0.00	0.00	0.00	2.00	0.06	0.03	0.02	0.00	0.00	0.00	0.00	0.03	0.00	0.09	0.00	0.05	0.11	0.10	0.14	0.00	0.01	0.07	0.02	0.05	0.03	0.06	0.00	0.00	0.00
6	0.08	0.02	0.01	0.01	0.10	3.00	0.01	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.02	0.00	0.00	0.01	0.02	0.00	0.09	0.00	0.00	0.00
7	0.01	0.00	0.00	0.02	0.03	0.00	2.00	0.03	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.01	0.01	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
8	0.01	0.01	0.00	0.01	0.02	0.01	0.03	2.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.02	0.01	0.02	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	2.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.03	2.00	0.00	0.27	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
13	0.01	0.00	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.05	0.01	0.01
14	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.27	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.01	0.00	0.00	0.01	0.09	0.00	0.01	0.02	0.00	0.00	0.00	0.01	0.00	0.00	2.00	0.00	0.05	0.01	0.02	0.04	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	2.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.01	0.00	0.00	0.00	0.11	0.01	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	2.00	0.02	0.04	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	0.01	0.01	0.00	0.02	0.10	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	2.00	0.11	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	0.00	0.02	0.00	0.00	0.14	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.04	0.11	2.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00
21	0.03	0.00	0.00	0.00	0.00	0.01	0.03	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
22	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
23	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.04	0.01	0.01	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	0.01	0.00	0.00	0.00	0.02	0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	2.00	0.01	0.00	0.01	0.00	0.00	0.00
25	0.00	0.00	0.00	0.01	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	2.00	0.00	0.02	0.00	0.00	0.00
26	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00
27	0.00	0.00	0.00	0.02	0.06	0.05	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.00	2.00	0.01	0.00	0.00
28	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	2.00	0.11	0.00
29	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	2.00	0.00
30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00
31	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
32	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.15	0.08	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
33	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
34	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.01	0.01	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
35	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.07	0.05	**0.49**	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00
36	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.01	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
37	0.00	0.00	0.00	0.00	0.04	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
38	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
39	0.02	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
40	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
41	0.03	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
42	0.02	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
43	0.15	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.03	0.00	0.04	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
44	0.02	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.00	0.00	0.00
45	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00
46	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.01	0.00	0.00	0.00
47	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
48	0.04	0.00	0.01	0.04	0.01	0.01	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.01	0.05	0.01	0.07	0.00	0.00	0.08	0.00	0.00	0.01	0.00	0.00	0.00
49	0.00	0.00	0.02	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
51	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
52	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
53	0.00	0.00	0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
55	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
57	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00
58	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
59	0.01	0.00	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
1	0.01	0.02	0.02	0.03	0.03	0.02	0.01	0.03	0.05	0.01	0.05	0.04	0.35	0.05	0.01	0.01	0.07	0.08	0.01	0.01	0.01	0.03	0.01	0.00	0.02	0.00	0.02	0.02	0.02
2	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.02	0.01	0.03	0.02	0.02	0.01	0.02	0.05	0.02	0.03	0.04	0.03	0.01	0.00	0.01	0.02	0.02	0.07	0.01	0.02	0.01	0.04	0.01	0.01	0.01	0.04	0.04	0.02
4	0.01	0.03	0.01	0.03	0.05	0.03	0.01	0.03	0.04	0.01	0.05	0.07	0.05	0.06	0.12	0.07	0.05	0.13	0.19	0.23	0.01	0.05	0.04	0.02	0.08	0.01	0.05	0.05	0.11
5	0.00	0.00	0.00	0.01	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.01	0.02	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.03	0.00	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.01
7	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.22	0.00	0.00	0.07	0.08	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.15	0.00	0.00	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.08	0.00	0.00	**0.49**	0.30	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01
14	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
15	0.00	0.01	0.00	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
17	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
19	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01
20	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
27	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.04	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01
28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
29	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
31	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00
32	0.00	2.00	0.00	0.00	0.10	0.04	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00
33	0.00	0.00	2.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
34	0.00	0.00	0.02	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
35	0.00	0.10	0.00	0.00	2.00	0.39	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
36	0.00	0.04	0.00	0.00	0.39	2.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
37	0.00	0.01	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.27	0.04	0.33	0.21	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.07	0.07	0.01	0.03	0.01	0.01	0.07	0.11
39	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.27	2.00	0.10	**0.43**	**0.54**	0.00	0.02	0.01	0.00	0.04	0.00	0.01	0.00	0.02	0.07	0.09	0.01	0.03	0.02	0.01	0.11	0.08
40	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.10	2.00	0.06	0.11	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.01	0.04	0.02	0.02	0.02	0.01	0.09	0.02
41	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.33	**0.43**	0.06	2.00	0.33	0.00	0.01	0.01	0.00	0.02	0.00	0.01	0.00	0.02	0.09	0.10	0.01	0.04	0.02	0.01	0.10	0.14
42	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.21	**0.54**	0.11	0.33	2.00	0.00	0.04	0.02	0.00	0.04	0.00	0.01	0.00	0.03	0.08	0.07	0.02	0.06	0.02	0.00	0.11	0.08
43	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
44	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.01	0.04	0.00	2.00	0.11	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
45	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.02	0.00	0.11	2.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00
46	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.02	0.02	0.04	0.00	0.01	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.01
48	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.01	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
49	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00
50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
51	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.03	0.02	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.03	0.02	0.00	0.03	0.01	0.02	0.01
52	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.07	0.01	0.09	0.08	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.08	0.00	0.00	0.01	0.26
53	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.07	0.09	0.04	0.10	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	2.00	0.01	0.00	0.03	0.00	0.13	0.20
54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	2.00	0.00	0.00	0.01	0.02	0.01
55	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.03	0.02	0.04	0.06	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.08	0.00	0.00	2.00	0.00	0.00	0.00	0.07
56	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.00	2.00	0.00	0.04	0.01
57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.01	0.00	0.00	2.00	0.07	0.03
58	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.07	0.11	0.09	0.10	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.01	0.13	0.02	0.00	0.04	0.07	2.00	0.08
59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.08	0.02	0.14	0.08	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.26	0.20	0.01	0.07	0.01	0.03	0.08	2.00

Code	Variables	Code	Variables
1	**Hospital**	**31**	**IsMorbidObesity**
2	**Gender**	**32**	**Transplant**
3	**AdmissionSource**	**33**	**IsHypothyroidism**
4	**AdmissionReason**	**34**	**IsDyslipidemias**
5	**FrailPatientMFI**	**35**	**IsChemotherapy**
6	**ChronicHealthStatus**	**36**	**IsRadiationTherapy**
7	**IsChfNyhaClass23**	**37**	**IsHistoryOfPneumonia**
8	**IsCrfNoDialysis**	**38**	**VESICAL**
9	**IsCrfDialysis**	**39**	**ARTERIAL**
10	**IsSolidTumorLocoregiol**	**40**	**DIALYSIS**
11	**IsSolidTumorMetastatic**	**41**	**CVC**
12	**IsImmunossupression**	**42**	**MV**

Code	Variables	Code	Variables
13	IsSevereCopd	43	PERIPHERAL
14	IsSteroidsUse	44	IsNeurologicalComaStuporObtundedDelirium
15	IsArterialHypertension	45	IsNeurologicalSeizures
16	IsAsthma	46	IsNeurologicalFocalNeurologicDeficit
17	IsDiabetesUncomplicated	47	IsCardiovascularSepticShock
18	IsDiabetesComplicated	48	IsCardiovascularRhythmDisturbances
19	IsAngina	49	IsDigestiveAcuteAbdomen
20	IsPreviousMI	50	IsTraumaMultipleTrauma
21	IsCardiacArrhythmia	51	J01A
22	IsDeepVenousThrombosis	52	J01C
23	IsPeripheralArteryDisease	53	J01D
24	IsChronicAtrialFibrilation	54	J01E
25	IsStrokeSequelae	55	J01F
26	IsStrokeNoSequelae	56	J01G
27	IsDementia	57	J01M
28	IsTobaccoConsumption	58	J01X
29	IsAlcoholism	59	Antibiotic
30	IsPsychiatricDisease		

Appendix I

Table I.1 - Variables selected by Recurse Filter Elimination (RFE) – 35 variables.

Recurse Filter Elimination (RFE)
Hospital
J01X
VesDURTOTAL
AdmissionSource
Saps3Points
AdmissionReason
VesTIMESTOTAL
tests_before
MFIpoints
Age
PerDURTOTAL
MVDURTOTAL
LOS_hospital_before_test
CVCTIMESTOTAL
J01D
J01G
ChronicHealthStatus
IsNeurologicalComaStuporObtundedDelirium
CVCDURMORE
DiaTIMESTOTAL
VesDURMORE
IsHistoryOfPneumonia
CharlsonIndex
J01C
ArtDURMORE
ArtTIMESTOTAL
IsAlcoholism
DIALYSIS
MVTIMESTOTAL
IsChronicAtrialFibrilation
IsStrokeSequelae
IsDiabetesUncomplicated
PERIPHERAL
IsDementia
MV

Table I.2 - Variables selected by Selection by Filter (SBF) – 42 variables.

Selection by Filter (SBF)	p.value
MVDURTOTAL	0.000
J01X	0.000
MVTIMESTOTAL	0.000
VesDURTOTAL	0.000
MV	0.000
CVCTIMESTOTAL	0.000
J01D	0.000
CVC	0.000
ArtTIMESTOTAL	0.000
VESICAL	0.000
VesTIMESTOTAL	0.000
DiaTIMESTOTAL	0.000
MVDURMORE	0.000
DIALYSIS	0.000
ArtDURMORE	0.000
Antibiotic	0.000
CVCDURMORE	0.000
LOS_hospital_before_test	0.000
VesDURMORE	0.000
Saps3Points	0.000
CVCTIMESMORE	0.000
J01A	0.000
ArtTIMESMORE	0.000
J01G	0.000
IsCardiovascularSepticShock	0.000
IsNeurologicalComaStuporObtundedDelirium	0.000

Selection by Filter (SBF)	p.value
IsStrokeSequelae	0.000
DiaDURMORE	0.000
Hospital	0.000
AdmissionSource	0.000
AdmissionReason	0.000
VesTIMESMORE	0.001
ChronicHealthStatus	0.002
J01F	0.002
FrailPatientMFI	0.005
J01C	0.012
CharlsonIndex	0.020
IsDeepVenousThrombosis	0.023
J01E	0.024
IsHistoryOfPneumonia	0.027
IsNeurologicalFocalNeurologicDeficit	0.045
IsSevereCopd	0.047

Table I.3 - Variables selected by Class Decomposition with random forest (D.RF) – 24 variables.

Random Forest after decomposition
ChronicHealthStatus
Hospital
AdmissionReason
J01F
IsArterialHypertension
MFIpoints
J01D
CVCTIMESTOTAL
Antibiotic
Gender
CVC
VesDURTOTAL
IsDementia
J01C
MVDURTOTAL
VesTIMESTOTAL
IsDiabetesUncomplicated
J01X
Age
MVTIMESTOTAL
PerDURTOTAL
VESICAL
MV
AdmissionSource

Table I.4 - Variables selected by Class Decomposition with filter (D.SBF) – 76 variables.

Filter after decomposition	p.value
Hospital	0.000
Gender	0.000
AdmissionSource	0.000
AdmissionReason	0.000
FrailPatientMFI	0.000
ChronicHealthStatus	0.000
IsChfNyhaClass23	0.000
IsCrfNoDialysis	0.000
IsCrfDialysis	0.001
IsSolidTumorLocoregiol	0.000
IsImmunossupression	0.000
IsSevereCopd	0.000
IsSteroidsUse	0.015
IsArterialHypertension	0.000
IsAsthma	0.000
IsDiabetesUncomplicated	0.000
IsDiabetesComplicated	0.000
IsAngina	0.000
IsPreviousMI	0.000
IsCardiacArrhythmia	0.000
IsDeepVenousThrombosis	0.001
IsPeripheralArteryDisease	0.011

Filter after decomposition	p.value
IsChronicAtrialFibrilation	0.000
IsStrokeSequelae	0.000
IsDementia	0.000
IsTobaccoConsumption	0.000
IsAlcoholism	0.000
IsPsychiatricDisease	0.042
IsMorbidObesity	0.000
Transplant	0.000
IsHypothyroidism	0.000
IsDyslipidemias	0.000
IsRadiationTherapy	0.000
IsHistoryOfPneumonia	0.000
VESICAL	0.000
DIALYSIS	0.000
CVC	0.000
MV	0.000
PERIPHERAL	0.000
IsNeurologicalComaStuporObtundedDelirium	0.000
IsNeurologicalSeizures	0.000
IsCardiovascularSepticShock	0.000
IsCardiovascularRhythmDisturbances	0.000
IsDigestiveAcuteAbdomen	0.000
J01A	0.000
J01C	0.000
J01D	0.000
J01E	0.000
J01F	0.000
J01G	0.000
J01M	0.000
J01X	0.000
Antibiotic	0.000
LOS_hospital_before_test	0.000
tests_before	0.000
Age	0.000
CharlsonIndex	0.000
MFIpoints	0.000
Saps3Points	0.000
VesDURTOTAL	0.000
VesDURMORE	0.000
VesTIMESTOTAL	0.000
VesTIMESMORE	0.000
ArtDURMORE	0.000
ArtTIMESTOTAL	0.000
ArtTIMESMORE	0.000
DiaDURMORE	0.000
DiaTIMESTOTAL	0.000
CVCDURMORE	0.000
CVCTIMESTOTAL	0.000
CVCTIMESMORE	0.000
MVDURTOTAL	0.000
MVDURMORE	0.000
MVTIMESTOTAL	0.000
PerDURTOTAL	0.000
PerTIMESMORE	0.000

Appendix J

Table J.1 - Results for each method from 10-fold cross-validation - without a strategy (NONE model).

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.63	0.68	0.71	0.71	0.74	0.79	0.00	0.01	0.03	0.04	0.06	0.13	0.98	0.99	0.99	0.99	1.00	1.00	0.00	0.17	0.25	0.42	0.57	1.00	0.90	0.90	0.90	0.90	0.91	0.91
LR_Regularization	0.65	0.68	0.72	0.72	0.75	0.80	0.00	0.00	0.00	0.01	0.00	0.06	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.17	0.33	0.44	0.67	1.00	0.90	0.90	0.90	0.90	0.90	0.90
LDA	0.64	0.67	0.72	0.71	0.75	0.78	0.03	0.06	0.09	0.09	0.10	0.19	0.96	0.98	0.98	0.98	0.99	1.00	0.15	0.23	0.32	0.38	0.49	0.75	0.90	0.90	0.90	0.91	0.91	0.91
NEAREST_SHRUNKEN_CENTROIDS	0.64	0.69	0.70	0.70	0.71	0.78	0.00	0.06	0.09	0.09	0.13	0.16	0.95	0.96	0.97	0.97	0.97	0.99	0.00	0.17	0.25	0.23	0.32	0.36	0.90	0.90	0.90	0.90	0.91	0.91
SVM_LINEAR	0.46	0.57	0.61	0.59	0.64	0.70	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
NEURAL_NETWORK	0.65	0.67	0.71	0.71	0.74	0.79	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
SVM_RADIAL	0.65	0.66	0.71	0.70	0.73	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.90	0.90	0.90	0.90	0.90
K_NEAREST_NEIGHBORS	0.56	0.61	0.65	0.64	0.67	0.74	0.00	0.00	0.00	0.01	0.02	0.06	0.99	0.99	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.21	0.50	0.67	0.90	0.90	0.90	0.90	0.90	0.91
NAIVE_BAYES	0.63	0.67	0.70	0.70	0.72	0.79	0.13	0.17	0.22	0.22	0.28	0.35	0.88	0.91	0.92	0.92	0.93	0.94	0.14	0.18	0.24	0.23	0.25	0.34	0.90	0.91	0.91	0.91	0.92	0.92
C45	0.54	0.57	0.62	0.61	0.64	0.73	0.00	0.01	0.03	0.04	0.06	0.10	0.97	0.98	0.99	0.99	0.99	1.00	0.00	0.03	0.17	0.22	0.35	0.60	0.90	0.90	0.90	0.90	0.90	0.91
CART	0.50	0.56	0.60	0.61	0.66	0.75	0.03	0.04	0.06	0.09	0.12	0.22	0.94	0.95	0.95	0.95	0.96	0.96	0.06	0.08	0.13	0.16	0.25	0.33	0.90	0.90	0.90	0.90	0.91	0.91
C50	0.50	0.50	0.50	0.55	0.50	0.81	0.00	0.00	0.00	0.02	0.00	0.09	0.99	1.00	1.00	1.00	1.00	1.00	0.50	0.63	0.75	0.75	0.88	1.00	0.90	0.90	0.90	0.90	0.90	0.91
RANDOM_FOREST	0.62	0.63	0.68	0.68	0.72	0.80	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
GBM	0.65	0.68	0.72	0.72	0.73	0.81	0.00	0.00	0.00	0.01	0.00	0.06	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.30	0.50	1.00	0.90	0.90	0.90	0.90	0.90	0.90
BAGGING	0.59	0.63	0.65	0.65	0.67	0.70	0.03	0.04	0.08	0.08	0.10	0.13	0.97	0.97	0.98	0.98	0.99	0.99	0.11	0.25	0.32	0.31	0.36	0.50	0.90	0.90	0.90	0.90	0.91	0.91
ADABOOST	0.56	0.65	0.66	0.68	0.72	0.77	0.00	0.00	0.03	0.03	0.05	0.06	0.98	0.98	0.99	0.98	0.99	1.00	0.00	0.00	0.13	0.17	0.36	0.40	0.90	0.90	0.90	0.90	0.90	0.90

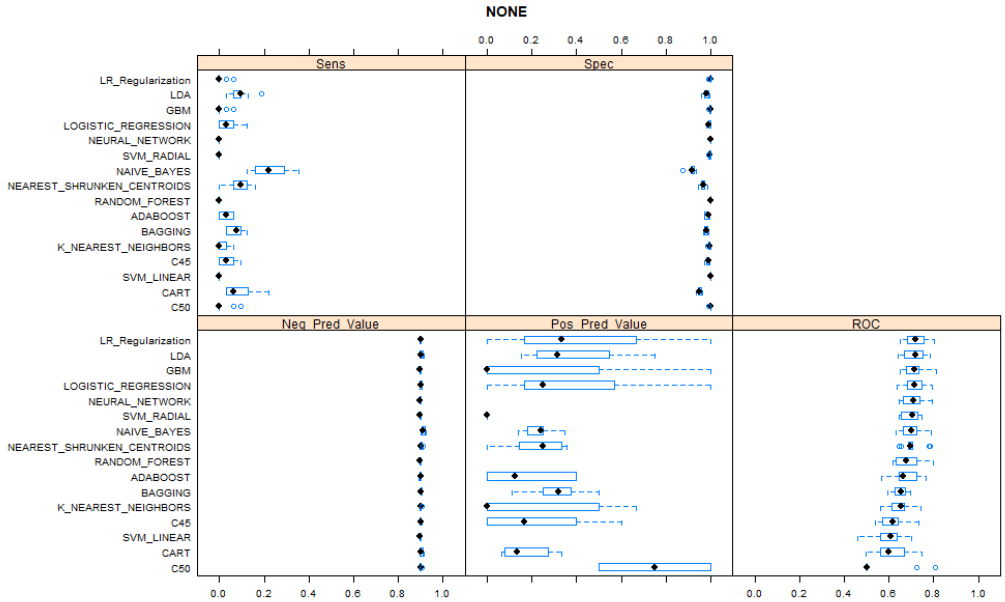


Figure J.1 - Cross-validation box-plot using the best hyperparameter - without a strategy (NONE model).

Table J.2 - Results for each method from 10-fold cross-validation - with the downsampling strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.63	0.67	0.68	0.69	0.74	0.75	0.50	0.59	0.65	0.64	0.69	0.77	0.60	0.63	0.65	0.65	0.66	0.72	0.13	0.15	0.17	0.17	0.18	0.22	0.91	0.93	0.94	0.94	0.95	0.96
LR_Regularization	0.64	0.69	0.69	0.71	0.74	0.77	0.53	0.58	0.62	0.63	0.65	0.74	0.64	0.67	0.67	0.68	0.69	0.74	0.16	0.16	0.17	0.18	0.18	0.24	0.93	0.94	0.94	0.94	0.95	0.96
LDA	0.63	0.65	0.67	0.69	0.74	0.75	0.53	0.57	0.65	0.63	0.67	0.72	0.61	0.63	0.65	0.66	0.69	0.71	0.13	0.16	0.17	0.17	0.18	0.21	0.92	0.93	0.94	0.94	0.95	0.96
NEAREST_SHRUNKEN_CENTROIDS	0.64	0.68	0.70	0.70	0.71	0.78	0.53	0.57	0.59	0.61	0.62	0.75	0.64	0.67	0.70	0.70	0.72	0.77	0.15	0.17	0.18	0.19	0.19	0.27	0.93	0.93	0.94	0.94	0.94	0.96
SVM_LINEAR	0.64	0.66	0.69	0.71	0.75	0.81	0.52	0.54	0.57	0.63	0.73	0.81	0.62	0.66	0.68	0.68	0.71	0.73	0.14	0.16	0.18	0.18	0.19	0.23	0.92	0.93	0.94	0.94	0.96	0.97
NEURAL_NETWORK	0.65	0.67	0.71	0.71	0.73	0.79	0.59	0.63	0.67	0.68	0.73	0.78	0.60	0.63	0.64	0.65	0.66	0.74	0.16	0.17	0.17	0.18	0.18	0.24	0.93	0.94	0.95	0.95	0.95	0.97
SVM_RADIAL	0.58	0.65	0.67	0.67	0.71	0.75	0.81	0.89	0.94	0.93	0.96	1.00	0.12	0.18	0.19	0.19	0.21	0.26	0.10	0.11	0.12	0.11	0.12	0.12	0.90	0.95	0.97	0.96	0.97	1.00
K_NEAREST_NEIGHBORS	0.58	0.66	0.69	0.68	0.69	0.78	0.41	0.57	0.60	0.59	0.64	0.71	0.63	0.68	0.71	0.70	0.72	0.77	0.13	0.16	0.19	0.18	0.19	0.25	0.91	0.93	0.94	0.94	0.94	0.96
NAIVE_BAYES	0.66	0.68	0.70	0.71	0.72	0.80	0.34	0.49	0.55	0.54	0.59	0.72	0.66	0.70	0.75	0.74	0.77	0.79	0.14	0.16	0.19	0.19	0.20	0.28	0.91	0.93	0.94	0.93	0.94	0.96
C45	0.59	0.63	0.66	0.67	0.70	0.76	0.53	0.58	0.66	0.65	0.70	0.78	0.53	0.61	0.63	0.64	0.67	0.72	0.14	0.16	0.16	0.17	0.18	0.22	0.93	0.93	0.94	0.94	0.95	0.96
CART	0.55	0.60	0.62	0.63	0.66	0.74	0.44	0.50	0.59	0.59	0.70	0.75	0.45	0.66	0.68	0.67	0.72	0.76	0.13	0.14	0.17	0.17	0.19	0.25	0.91	0.93	0.94	0.94	0.94	0.96
C50	0.59	0.64	0.67	0.68	0.72	0.82	0.50	0.59	0.64	0.64	0.68	0.84	0.58	0.63	0.65	0.65	0.66	0.70	0.14	0.15	0.16	0.17	0.18	0.24	0.92	0.93	0.94	0.94	0.94	0.98
RANDOM_FOREST	0.65	0.67	0.70	0.71	0.71	0.80	0.56	0.61	0.66	0.68	0.70	0.88	0.59	0.61	0.63	0.64	0.66	0.70	0.15	0.16	0.17	0.17	0.18	0.22	0.93	0.94	0.94	0.95	0.95	0.98
GBM	0.64	0.68	0.70	0.71	0.71	0.82	0.56	0.60	0.64	0.66	0.71	0.84	0.60	0.64	0.66	0.65	0.67	0.69	0.14	0.16	0.18	0.18	0.19	0.23	0.92	0.94	0.94	0.95	0.95	0.97
BAGGING	0.59	0.62	0.62	0.64	0.66	0.71	0.48	0.52	0.57	0.56	0.59	0.63	0.59	0.59	0.61	0.61	0.63	0.63	0.12	0.13	0.14	0.14	0.14	0.16	0.91	0.92	0.93	0.93	0.93	0.94
ADABOOST	0.60	0.66	0.68	0.69	0.70	0.80	0.58	0.61	0.69	0.67	0.71	0.78	0.56	0.60	0.61	0.63	0.66	0.69	0.14	0.16	0.17	0.17	0.18	0.22	0.93	0.94	0.94	0.94	0.95	0.97

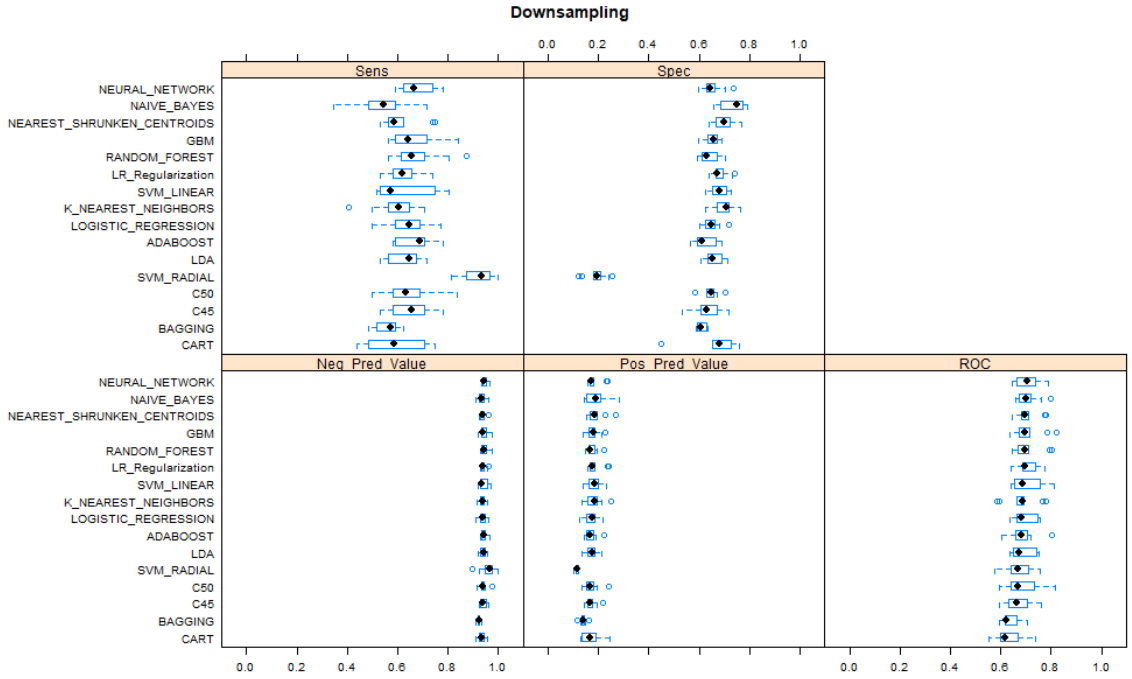


Figure J.2 - Cross-validation boxplot using the best hyperparameter - with the downsampling strategy.

Table J.3 - Results for each method from 10-fold cross-validation - with the upsampling strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.64	0.67	0.70	0.71	0.74	0.78	0.50	0.53	0.60	0.60	0.65	0.71	0.67	0.68	0.69	0.70	0.71	0.75	0.15	0.16	0.18	0.18	0.20	0.24	0.92	0.93	0.94	0.94	0.95	0.96
LR_Regularization	0.66	0.68	0.71	0.72	0.76	0.78	0.52	0.54	0.60	0.61	0.69	0.71	0.67	0.69	0.70	0.70	0.71	0.74	0.16	0.17	0.18	0.19	0.20	0.23	0.93	0.93	0.94	0.94	0.95	0.96
LDA	0.65	0.67	0.70	0.71	0.74	0.78	0.50	0.54	0.60	0.60	0.65	0.71	0.68	0.68	0.69	0.70	0.71	0.76	0.15	0.16	0.18	0.18	0.19	0.24	0.92	0.93	0.94	0.94	0.95	0.96
NEAREST_SHRUNKEN_CENTROIDS	0.65	0.68	0.70	0.70	0.71	0.79	0.53	0.56	0.59	0.61	0.62	0.78	0.65	0.68	0.70	0.70	0.72	0.75	0.16	0.17	0.18	0.19	0.19	0.26	0.93	0.93	0.94	0.94	0.94	0.97
SVM_LINEAR	0.63	0.66	0.69	0.71	0.75	0.79	0.47	0.55	0.60	0.61	0.70	0.75	0.67	0.69	0.69	0.70	0.71	0.74	0.15	0.16	0.18	0.19	0.21	0.24	0.92	0.93	0.94	0.94	0.95	0.96
NEURAL_NETWORK	0.65	0.68	0.71	0.72	0.74	0.80	0.53	0.58	0.62	0.63	0.67	0.72	0.68	0.69	0.71	0.71	0.72	0.73	0.16	0.18	0.19	0.19	0.20	0.23	0.93	0.94	0.94	0.94	0.95	0.96
SVM_RADIAL	0.60	0.65	0.69	0.69	0.73	0.76	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.98	0.98	0.98	0.99	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.90	0.90	0.90	0.90	0.90
K_NEAREST_NEIGHBORS	0.54	0.61	0.63	0.63	0.66	0.69	0.47	0.63	0.65	0.62	0.67	0.74	0.54	0.55	0.57	0.58	0.60	0.62	0.11	0.13	0.14	0.14	0.15	0.18	0.91	0.93	0.93	0.93	0.94	0.96
NAIVE_BAYES	0.64	0.67	0.70	0.71	0.72	0.79	0.34	0.42	0.50	0.50	0.53	0.69	0.72	0.75	0.78	0.77	0.79	0.80	0.13	0.15	0.21	0.20	0.21	0.28	0.91	0.92	0.93	0.93	0.94	0.96
C45	0.56	0.62	0.67	0.66	0.71	0.76	0.38	0.48	0.55	0.54	0.64	0.68	0.65	0.70	0.71	0.71	0.75	0.75	0.15	0.16	0.17	0.18	0.19	0.23	0.91	0.92	0.93	0.93	0.95	0.95
CART	0.56	0.60	0.63	0.63	0.64	0.73	0.50	0.52	0.58	0.60	0.64	0.78	0.52	0.60	0.68	0.65	0.68	0.71	0.12	0.15	0.16	0.16	0.16	0.21	0.92	0.93	0.93	0.93	0.94	0.96
C50	0.62	0.65	0.67	0.68	0.70	0.78	0.03	0.07	0.11	0.12	0.17	0.22	0.94	0.96	0.97	0.96	0.97	0.99	0.08	0.21	0.29	0.27	0.36	0.41	0.90	0.90	0.91	0.91	0.91	0.92
RANDOM_FOREST	0.66	0.67	0.69	0.71	0.71	0.81	0.55	0.59	0.63	0.64	0.66	0.78	0.61	0.62	0.66	0.65	0.67	0.68	0.14	0.16	0.17	0.17	0.17	0.22	0.93	0.93	0.94	0.94	0.94	0.97
GBM	0.66	0.67	0.70	0.71	0.74	0.80	0.50	0.52	0.56	0.60	0.64	0.78	0.69	0.70	0.71	0.72	0.73	0.74	0.16	0.17	0.18	0.19	0.21	0.25	0.92	0.93	0.94	0.94	0.95	0.97
BAGGING	0.49	0.59	0.62	0.62	0.67	0.73	0.03	0.07	0.10	0.10	0.13	0.19	0.93	0.94	0.95	0.95	0.96	0.98	0.09	0.14	0.18	0.19	0.20	0.33	0.90	0.90	0.90	0.90	0.91	0.91
ADABOOST	0.60	0.63	0.65	0.68	0.74	0.81	0.00	0.01	0.03	0.03	0.03	0.09	0.96	0.98	0.98	0.98	0.99	0.99	0.00	0.03	0.15	0.20	0.24	0.60	0.89	0.90	0.90	0.90	0.90	0.91

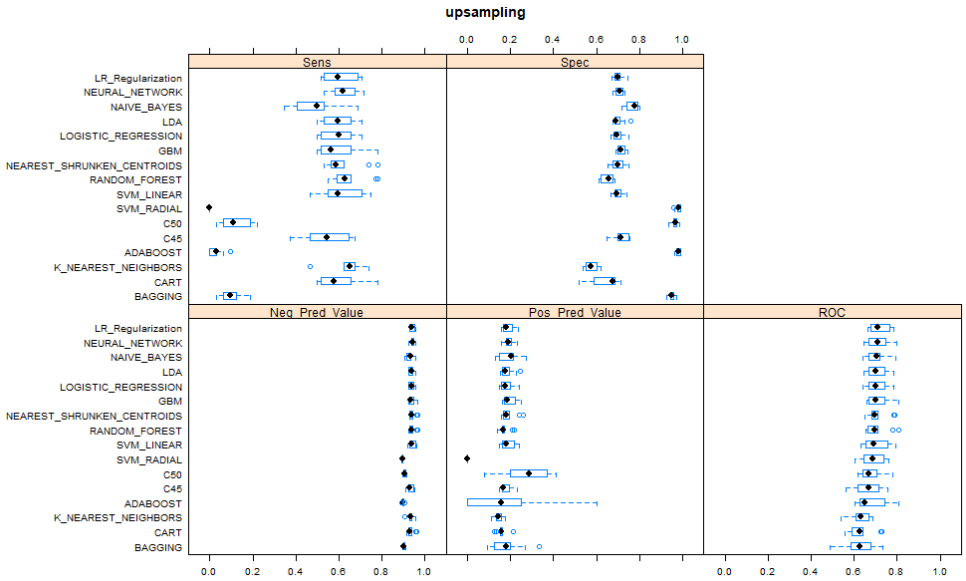


Figure J.3 - Cross-validation boxplot using the best hyperparameter - with the upsampling strategy.

Table J.4 - Results for each method from 10-fold cross-validation - with the SMOTE sampling strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.62	0.66	0.69	0.69	0.74	0.77	0.44	0.45	0.48	0.52	0.55	0.72	0.72	0.73	0.75	0.75	0.78	0.79	0.15	0.17	0.18	0.19	0.20	0.28	0.92	0.93	0.93	0.93	0.94	0.96
LR_Regularization	0.65	0.68	0.72	0.72	0.76	0.77	0.38	0.47	0.49	0.50	0.56	0.63	0.74	0.77	0.77	0.78	0.80	0.83	0.16	0.18	0.19	0.21	0.23	0.29	0.92	0.93	0.93	0.93	0.94	0.95
LDA	0.63	0.67	0.69	0.70	0.74	0.77	0.45	0.47	0.48	0.52	0.56	0.69	0.71	0.74	0.76	0.76	0.78	0.79	0.16	0.17	0.18	0.19	0.22	0.26	0.92	0.93	0.93	0.93	0.94	0.96
NEAREST_SHRUNKEN_CENTROIDS	0.65	0.68	0.70	0.70	0.71	0.78	0.47	0.49	0.52	0.54	0.57	0.69	0.70	0.73	0.74	0.75	0.77	0.82	0.16	0.17	0.18	0.20	0.21	0.31	0.92	0.93	0.93	0.94	0.94	0.96
SVM_LINEAR	0.64	0.66	0.71	0.71	0.75	0.77	0.42	0.45	0.47	0.50	0.54	0.63	0.72	0.77	0.78	0.78	0.80	0.81	0.16	0.19	0.20	0.20	0.21	0.27	0.92	0.93	0.93	0.93	0.94	0.95
NEURAL_NETWORK	0.65	0.69	0.70	0.71	0.73	0.78	0.41	0.45	0.51	0.51	0.55	0.63	0.72	0.75	0.77	0.77	0.79	0.80	0.16	0.17	0.20	0.20	0.22	0.26	0.92	0.92	0.94	0.93	0.94	0.95
SVM_RADIAL	0.61	0.62	0.69	0.66	0.69	0.71	0.00	0.03	0.05	0.04	0.06	0.06	0.91	0.93	0.94	0.94	0.96	0.97	0.00	0.05	0.09	0.09	0.12	0.20	0.89	0.90	0.90	0.90	0.90	0.90
K_NEAREST_NEIGHBORS	0.59	0.63	0.64	0.65	0.67	0.74	0.41	0.44	0.56	0.53	0.60	0.66	0.64	0.65	0.66	0.67	0.68	0.70	0.11	0.13	0.16	0.15	0.17	0.20	0.90	0.92	0.93	0.93	0.94	0.95
NAIVE_BAYES	0.65	0.68	0.71	0.71	0.72	0.79	0.41	0.50	0.55	0.55	0.59	0.69	0.67	0.69	0.73	0.73	0.75	0.82	0.13	0.17	0.18	0.19	0.19	0.30	0.91	0.93	0.93	0.93	0.94	0.96
C45	0.61	0.64	0.65	0.66	0.67	0.74	0.16	0.19	0.28	0.27	0.34	0.35	0.84	0.87	0.90	0.90	0.92	0.94	0.12	0.18	0.24	0.23	0.28	0.31	0.90	0.91	0.92	0.92	0.92	0.93
CART	0.57	0.60	0.65	0.64	0.67	0.72	0.03	0.06	0.13	0.13	0.18	0.25	0.85	0.95	0.96	0.95	0.97	0.99	0.08	0.14	0.30	0.27	0.39	0.44	0.90	0.90	0.91	0.91	0.91	0.92
C50	0.61	0.65	0.67	0.69	0.71	0.83	0.16	0.21	0.25	0.25	0.30	0.32	0.88	0.89	0.91	0.91	0.92	0.94	0.17	0.20	0.21	0.23	0.26	0.38	0.90	0.91	0.92	0.91	0.92	0.93
RANDOM_FOREST	0.58	0.67	0.69	0.70	0.75	0.81	0.00	0.03	0.03	0.07	0.09	0.25	0.87	0.96	0.97	0.96	0.98	0.99	0.00	0.10	0.14	0.20	0.31	0.50	0.89	0.90	0.90	0.90	0.90	0.91
GBM	0.63	0.66	0.70	0.70	0.73	0.81	0.09	0.20	0.23	0.24	0.33	0.38	0.89	0.90	0.91	0.92	0.93	0.94	0.13	0.20	0.21	0.24	0.26	0.39	0.90	0.91	0.91	0.92	0.92	0.93
BAGGING	0.56	0.59	0.65	0.65	0.69	0.76	0.16	0.20	0.25	0.26	0.31	0.44	0.84	0.87	0.88	0.88	0.89	0.92	0.12	0.14	0.19	0.20	0.26	0.28	0.90	0.91	0.91	0.91	0.92	0.93
ADABOOST	0.60	0.66	0.69	0.69	0.72	0.76	0.19	0.23	0.25	0.28	0.31	0.41	0.85	0.88	0.88	0.89	0.90	0.92	0.15	0.21	0.21	0.21	0.24	0.27	0.90	0.91	0.92	0.92	0.92	0.93

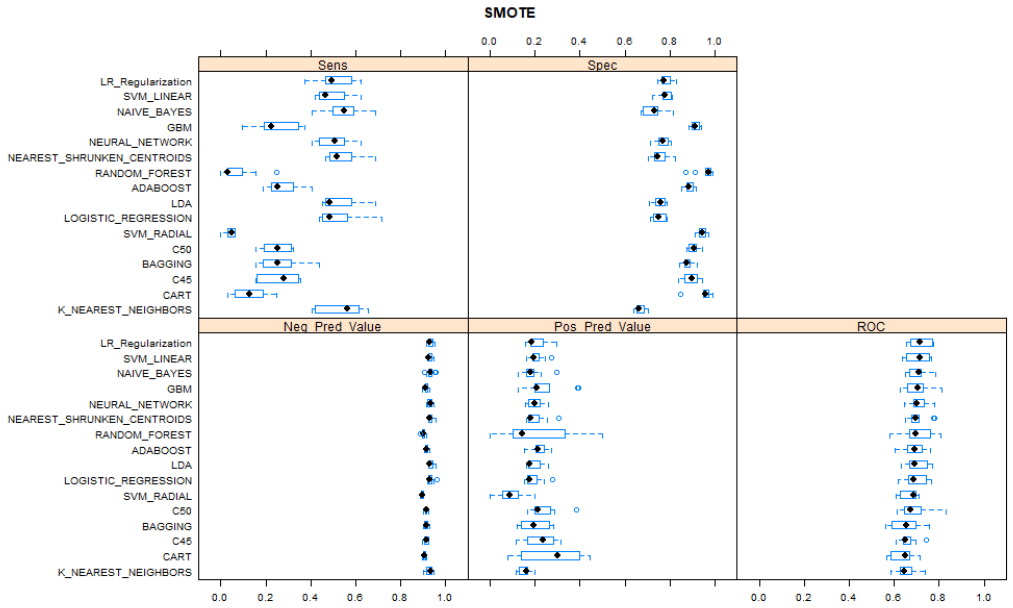


Figure J.4 - Cross-validation boxplot using the best hyperparameter - with the SMOTE sampling strategy.

Table J.5 - Results for each method from 10-fold cross-validation - with the Tomek Link strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.64	0.68	0.71	0.72	0.75	0.80	0.00	0.04	0.06	0.07	0.09	0.16	0.97	0.98	0.98	0.98	1.00	1.00	0.00	0.19	0.35	0.34	0.48	0.75	0.90	0.90	0.90	0.90	0.91	0.91
LR_Regularization	0.65	0.68	0.72	0.72	0.76	0.80	0.00	0.00	0.00	0.02	0.03	0.06	0.99	0.99	1.00	1.00	1.00	1.00	0.00	0.05	0.35	0.31	0.50	0.67	0.90	0.90	0.90	0.90	0.90	0.90
LDA	0.65	0.67	0.71	0.71	0.75	0.78	0.06	0.10	0.13	0.13	0.16	0.19	0.94	0.97	0.97	0.97	0.98	1.00	0.15	0.24	0.33	0.38	0.42	0.80	0.90	0.90	0.91	0.91	0.91	0.92
NEAREST_SHRUNKEN_CENTROIDS	0.64	0.69	0.70	0.70	0.71	0.78	0.06	0.09	0.11	0.12	0.15	0.23	0.94	0.95	0.96	0.95	0.96	0.97	0.14	0.18	0.21	0.23	0.25	0.35	0.90	0.90	0.90	0.91	0.91	0.92
SVM_LINEAR	0.47	0.59	0.62	0.61	0.65	0.72	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
NEURAL_NETWORK	0.65	0.67	0.71	0.71	0.74	0.79	0.00	0.00	0.00	0.01	0.00	0.06	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.23	0.50	0.67	0.90	0.90	0.90	0.90	0.90	0.90
SVM_RADIAL	0.64	0.66	0.70	0.70	0.73	0.74	0.00	0.00	0.00	0.01	0.00	0.03	0.93	0.96	0.97	0.97	0.97	0.98	0.00	0.00	0.00	0.02	0.00	0.11	0.89	0.89	0.90	0.90	0.90	0.90
K_NEAREST_NEIGHBORS	0.55	0.63	0.64	0.65	0.69	0.75	0.00	0.03	0.03	0.04	0.06	0.13	0.98	0.99	0.99	0.99	1.00	1.00	0.00	0.15	0.29	0.28	0.48	0.50	0.90	0.90	0.90	0.90	0.91	0.91
NAIVE_BAYES	0.62	0.66	0.70	0.70	0.72	0.77	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
C45	0.53	0.56	0.61	0.60	0.63	0.68	0.03	0.09	0.10	0.10	0.13	0.19	0.95	0.95	0.96	0.96	0.97	0.98	0.07	0.20	0.22	0.22	0.29	0.40	0.90	0.90	0.90	0.90	0.91	0.91
CART	0.50	0.58	0.60	0.60	0.61	0.73	0.00	0.03	0.03	0.05	0.08	0.13	0.96	0.97	0.98	0.98	0.99	1.00	0.08	0.11	0.20	0.21	0.27	0.44	0.90	0.90	0.90	0.90	0.90	0.91
C50	0.50	0.50	0.50	0.50	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
RANDOM_FOREST	0.60	0.62	0.66	0.67	0.70	0.77	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
GBM	0.66	0.67	0.69	0.71	0.72	0.82	0.00	0.00	0.00	0.02	0.03	0.06	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.39	1.00	1.00	0.90	0.90	0.90	0.90	0.90	0.90
BAGGING	0.59	0.63	0.67	0.67	0.69	0.74	0.03	0.03	0.06	0.08	0.09	0.16	0.95	0.96	0.97	0.97	0.98	0.99	0.08	0.11	0.23	0.21	0.30	0.33	0.90	0.90	0.90	0.90	0.90	0.91
ADABOOST	0.61	0.63	0.68	0.68	0.70	0.77	0.00	0.03	0.08	0.07	0.10	0.13	0.95	0.96	0.98	0.97	0.98	1.00	0.00	0.08	0.26	0.24	0.36	0.67	0.89	0.90	0.90	0.90	0.91	0.91

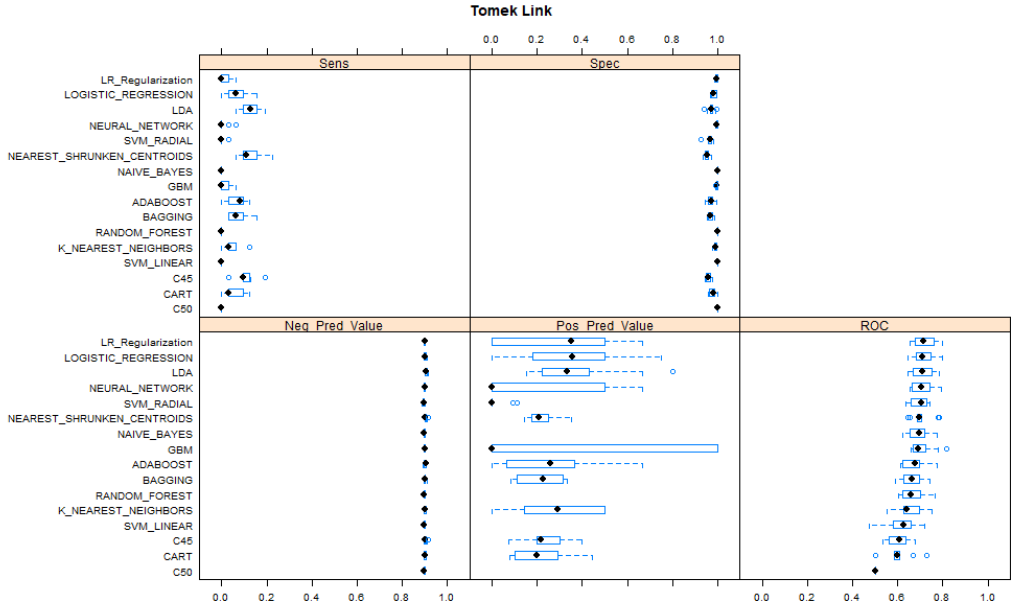


Figure J.5 - Cross-validation boxplot using the best hyperparameter - with the Tomek Link strategy.

Table J.6 - Results for each method from 10-fold cross-validation - with the Neighborhood Cleaning Rule (NCL) strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.64	0.68	0.71	0.72	0.75	0.80	0.03	0.04	0.09	0.08	0.10	0.16	0.96	0.98	0.98	0.98	0.99	1.00	0.13	0.25	0.32	0.38	0.49	0.75	0.90	0.90	0.91	0.90	0.91	0.91
LR_Regularization	0.66	0.68	0.71	0.72	0.76	0.79	0.00	0.01	0.03	0.04	0.05	0.13	0.98	0.98	0.99	0.99	1.00	1.00	0.00	0.14	0.25	0.26	0.44	0.50	0.90	0.90	0.90	0.90	0.90	0.91
LDA	0.65	0.68	0.70	0.71	0.75	0.78	0.06	0.10	0.14	0.15	0.20	0.25	0.93	0.95	0.96	0.96	0.97	0.98	0.15	0.21	0.33	0.31	0.42	0.44	0.90	0.91	0.91	0.91	0.91	0.92
NEAREST_SHRUNKEN_CENTROIDS	0.64	0.69	0.70	0.70	0.71	0.78	0.13	0.13	0.15	0.17	0.21	0.26	0.92	0.93	0.94	0.94	0.95	0.96	0.16	0.20	0.24	0.24	0.26	0.35	0.90	0.90	0.91	0.91	0.91	0.92
SVM_LINEAR	0.56	0.61	0.63	0.64	0.66	0.72	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
NEURAL_NETWORK	0.66	0.68	0.70	0.71	0.74	0.79	0.00	0.00	0.00	0.01	0.00	0.03	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.17	0.33	0.28	0.42	0.50	0.90	0.90	0.90	0.90	0.90	0.90
SVM_RADIAL	0.60	0.64	0.68	0.68	0.71	0.74	0.00	0.00	0.03	0.03	0.06	0.10	0.89	0.94	0.94	0.94	0.96	0.96	0.00	0.00	0.05	0.06	0.12	0.15	0.89	0.89	0.90	0.90	0.90	0.90
K_NEAREST_NEIGHBORS	0.57	0.62	0.64	0.65	0.69	0.75	0.03	0.04	0.06	0.08	0.10	0.16	0.96	0.97	0.98	0.98	0.99	0.99	0.20	0.21	0.32	0.30	0.33	0.40	0.90	0.90	0.90	0.90	0.91	0.91
NAIVE_BAYES	0.62	0.66	0.70	0.70	0.71	0.78	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.90	0.90	0.90	0.90	0.90
C45	0.56	0.57	0.59	0.61	0.63	0.68	0.03	0.08	0.13	0.12	0.16	0.19	0.92	0.95	0.95	0.96	0.96	0.98	0.15	0.21	0.24	0.24	0.27	0.33	0.90	0.90	0.91	0.91	0.91	0.91
CART	0.57	0.59	0.61	0.62	0.65	0.69	0.09	0.14	0.17	0.20	0.24	0.34	0.88	0.92	0.93	0.93	0.93	0.94	0.14	0.18	0.23	0.22	0.26	0.29	0.90	0.91	0.91	0.91	0.91	0.92
C50	0.50	0.50	0.50	0.57	0.66	0.76	0.00	0.00	0.00	0.01	0.00	0.03	0.98	0.99	1.00	1.00	1.00	1.00	0.00	0.10	0.20	0.18	0.27	0.33	0.90	0.90	0.90	0.90	0.90	0.90
RANDOM_FOREST	0.60	0.65	0.67	0.68	0.72	0.78	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
GBM	0.67	0.68	0.69	0.71	0.74	0.79	0.00	0.01	0.05	0.04	0.06	0.09	0.98	0.99	0.99	0.99	1.00	1.00	0.00	0.07	0.35	0.41	0.58	1.00	0.90	0.90	0.90	0.90	0.90	0.91
BAGGING	0.56	0.62	0.63	0.64	0.66	0.71	0.03	0.04	0.08	0.08	0.10	0.16	0.94	0.95	0.96	0.96	0.97	0.98	0.06	0.10	0.17	0.19	0.25	0.38	0.89	0.90	0.90	0.90	0.90	0.91
ADABOOST	0.61	0.62	0.67	0.68	0.72	0.75	0.00	0.03	0.05	0.06	0.09	0.13	0.95	0.96	0.97	0.97	0.98	0.99	0.00	0.11	0.17	0.18	0.24	0.33	0.90	0.90	0.90	0.90	0.90	0.91

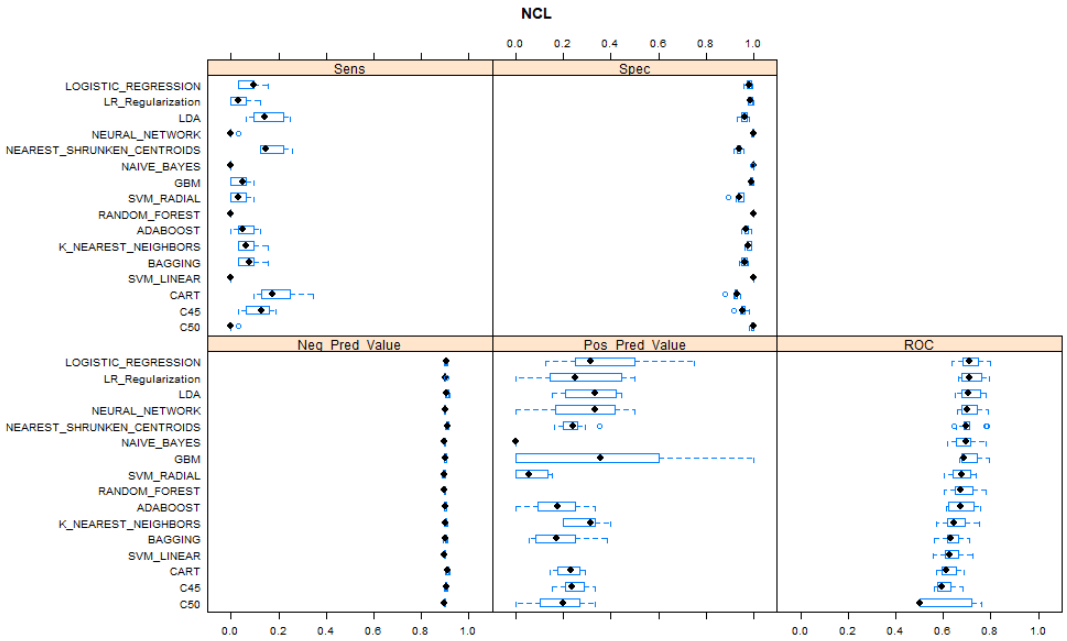


Figure J.6 - Cross-validation boxplot using the best hyperparameter - with the Neighborhood Cleaning Rule (NCL) strategy.

Table J.7 - Results for each method from 10-fold cross-validation - with the One Side Selection (OSS) strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.64	0.68	0.71	0.72	0.74	0.80	0.00	0.04	0.06	0.07	0.09	0.16	0.97	0.98	0.98	0.99	1.00	1.00	0.00	0.23	0.40	0.42	0.50	1.00	0.90	0.90	0.90	0.90	0.91	0.91
LR_Regularization	0.65	0.68	0.72	0.72	0.75	0.80	0.00	0.00	0.00	0.02	0.03	0.06	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.33	0.50	0.57	1.00	1.00	0.90	0.90	0.90	0.90	0.90	0.90
LDA	0.65	0.67	0.71	0.71	0.75	0.78	0.06	0.09	0.13	0.12	0.15	0.19	0.94	0.97	0.98	0.97	0.98	1.00	0.15	0.24	0.32	0.38	0.44	0.80	0.90	0.90	0.91	0.91	0.91	0.92
NEAREST_SHRUNKEN_CENTROIDS	0.64	0.69	0.70	0.70	0.71	0.78	0.06	0.09	0.10	0.12	0.15	0.23	0.94	0.95	0.96	0.96	0.96	0.98	0.14	0.20	0.22	0.23	0.25	0.35	0.90	0.90	0.90	0.91	0.91	0.92
SVM_LINEAR	0.50	0.53	0.58	0.58	0.62	0.69	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
NEURAL_NETWORK	0.65	0.67	0.71	0.72	0.74	0.79	0.00	0.00	0.00	0.01	0.00	0.03	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.30	0.50	1.00	0.90	0.90	0.90	0.90	0.90	0.90
SVM_RADIAL	0.64	0.66	0.70	0.70	0.72	0.75	0.00	0.00	0.00	0.01	0.00	0.03	0.93	0.95	0.97	0.96	0.98	0.98	0.00	0.00	0.00	0.02	0.00	0.10	0.89	0.89	0.90	0.90	0.90	0.90
K_NEAREST_NEIGHBORS	0.55	0.63	0.64	0.65	0.68	0.75	0.00	0.03	0.03	0.04	0.06	0.09	0.98	0.99	0.99	0.99	1.00	1.00	0.00	0.27	0.37	0.36	0.48	0.67	0.90	0.90	0.90	0.90	0.90	0.91
NAIVE_BAYES	0.62	0.66	0.70	0.70	0.71	0.77	0.00	0.00	0.00	0.00	0.00	0.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.90	0.90	0.90	0.90	0.90
C45	0.53	0.56	0.58	0.59	0.62	0.68	0.03	0.07	0.11	0.11	0.15	0.19	0.95	0.95	0.96	0.96	0.96	0.98	0.07	0.16	0.22	0.23	0.29	0.50	0.90	0.90	0.91	0.91	0.91	0.91
CART	0.50	0.58	0.60	0.60	0.61	0.73	0.00	0.03	0.03	0.05	0.08	0.13	0.96	0.97	0.98	0.98	0.99	1.00	0.08	0.11	0.20	0.21	0.27	0.44	0.90	0.90	0.90	0.90	0.90	0.91
C50	0.50	0.50	0.50	0.55	0.50	0.78	0.00	0.00	0.00	0.01	0.00	0.03	0.99	1.00	1.00	1.00	1.00	1.00	0.20	0.20	0.20	0.20	0.20	0.20	0.90	0.90	0.90	0.90	0.90	0.90
RANDOM_FOREST	0.62	0.63	0.66	0.68	0.71	0.79	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA	NA	NA	NA	NA	0.90	0.90	0.90	0.90	0.90	0.90
GBM	0.65	0.68	0.70	0.71	0.72	0.81	0.00	0.00	0.00	0.01	0.03	0.03	0.99	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.13	0.34	0.63	1.00	0.90	0.90	0.90	0.90	0.90	0.90
BAGGING	0.58	0.62	0.63	0.65	0.69	0.72	0.03	0.07	0.10	0.09	0.10	0.13	0.96	0.96	0.97	0.97	0.98	0.99	0.10	0.20	0.21	0.26	0.32	0.50	0.90	0.90	0.91	0.90	0.91	0.91
ADABOOST	0.60	0.66	0.68	0.67	0.70	0.75	0.00	0.03	0.03	0.04	0.05	0.10	0.97	0.97	0.98	0.98	0.98	1.00	0.00	0.10	0.17	0.18	0.20	0.50	0.90	0.90	0.90	0.90	0.90	0.91

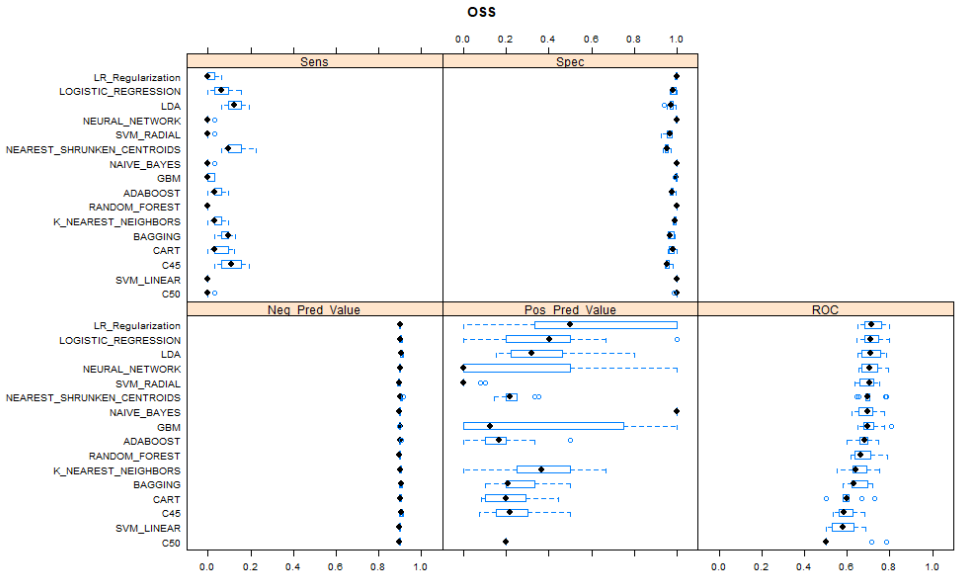


Figure J.7 - Cross-validation boxplot using the best hyperparameter - with the One Side Selection (OSS) strategy.

Table J.8 - Results for each method from 10-fold cross-validation - with the SMOTE + Tomek strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.61	0.66	0.68	0.69	0.73	0.75	0.44	0.47	0.54	0.54	0.59	0.72	0.67	0.71	0.72	0.72	0.74	0.77	0.14	0.15	0.17	0.18	0.20	0.24	0.92	0.92	0.93	0.93	0.94	0.96
LR_Regularization	0.65	0.68	0.72	0.72	0.76	0.77	0.41	0.47	0.51	0.53	0.61	0.66	0.73	0.75	0.77	0.77	0.79	0.81	0.16	0.17	0.19	0.21	0.24	0.28	0.92	0.92	0.93	0.94	0.95	0.95
LDA	0.63	0.67	0.69	0.70	0.74	0.77	0.45	0.47	0.48	0.53	0.58	0.72	0.70	0.73	0.75	0.75	0.77	0.79	0.16	0.16	0.17	0.19	0.21	0.27	0.92	0.92	0.93	0.93	0.94	0.96
NEAREST_SHRUNKEN_CENTROIDS	0.65	0.68	0.70	0.70	0.71	0.78	0.47	0.49	0.52	0.54	0.57	0.69	0.70	0.73	0.74	0.75	0.77	0.82	0.16	0.17	0.18	0.20	0.21	0.31	0.92	0.93	0.93	0.94	0.94	0.96
SVM_LINEAR	0.63	0.66	0.71	0.71	0.75	0.77	0.47	0.49	0.50	0.52	0.53	0.63	0.73	0.76	0.77	0.77	0.80	0.80	0.17	0.19	0.20	0.21	0.22	0.26	0.93	0.93	0.93	0.93	0.94	0.95
NEURAL_NETWORK	0.65	0.69	0.70	0.71	0.73	0.78	0.41	0.47	0.52	0.52	0.55	0.63	0.70	0.74	0.76	0.76	0.79	0.80	0.15	0.18	0.20	0.20	0.22	0.25	0.92	0.92	0.94	0.93	0.94	0.95
SVM_RADIAL	0.61	0.62	0.68	0.66	0.69	0.71	0.00	0.03	0.06	0.05	0.06	0.10	0.91	0.93	0.94	0.94	0.96	0.97	0.00	0.05	0.09	0.10	0.14	0.27	0.89	0.89	0.90	0.90	0.90	0.91
K_NEAREST_NEIGHBORS	0.59	0.63	0.64	0.65	0.67	0.74	0.41	0.44	0.57	0.53	0.59	0.66	0.63	0.65	0.66	0.66	0.67	0.70	0.11	0.13	0.16	0.15	0.17	0.20	0.90	0.91	0.93	0.93	0.94	0.95
NAIVE_BAYES	0.62	0.65	0.68	0.68	0.70	0.77	0.48	0.52	0.57	0.57	0.59	0.75	0.65	0.67	0.71	0.71	0.74	0.78	0.14	0.16	0.18	0.18	0.19	0.27	0.92	0.93	0.93	0.94	0.94	0.96
C45	0.62	0.64	0.67	0.68	0.69	0.79	0.31	0.35	0.39	0.43	0.48	0.74	0.74	0.77	0.80	0.80	0.83	0.85	0.14	0.16	0.19	0.19	0.21	0.29	0.91	0.92	0.92	0.93	0.93	0.97
CART	0.58	0.62	0.64	0.66	0.67	0.79	0.22	0.32	0.38	0.39	0.47	0.56	0.72	0.78	0.80	0.80	0.83	0.88	0.11	0.15	0.16	0.19	0.22	0.35	0.90	0.91	0.92	0.92	0.93	0.95
C50	0.63	0.66	0.69	0.69	0.70	0.79	0.22	0.30	0.33	0.34	0.38	0.47	0.85	0.87	0.88	0.88	0.89	0.91	0.20	0.21	0.23	0.24	0.26	0.33	0.91	0.92	0.92	0.92	0.92	0.94
RANDOM_FOREST	0.64	0.68	0.69	0.70	0.71	0.81	0.34	0.42	0.49	0.49	0.55	0.65	0.74	0.75	0.79	0.78	0.81	0.83	0.15	0.17	0.20	0.21	0.24	0.29	0.91	0.92	0.93	0.93	0.94	0.95
GBM	0.65	0.67	0.69	0.70	0.73	0.80	0.25	0.30	0.32	0.33	0.37	0.42	0.85	0.87	0.90	0.89	0.92	0.93	0.18	0.22	0.27	0.27	0.31	0.38	0.91	0.92	0.92	0.92	0.93	0.94
BAGGING	0.61	0.63	0.66	0.67	0.69	0.77	0.22	0.30	0.35	0.34	0.38	0.47	0.80	0.82	0.83	0.84	0.86	0.90	0.14	0.18	0.19	0.20	0.20	0.31	0.91	0.92	0.92	0.92	0.92	0.93
ADABOOST	0.59	0.64	0.67	0.68	0.72	0.80	0.16	0.27	0.34	0.34	0.41	0.56	0.84	0.85	0.87	0.87	0.89	0.91	0.14	0.19	0.23	0.23	0.27	0.33	0.90	0.92	0.92	0.92	0.93	0.95

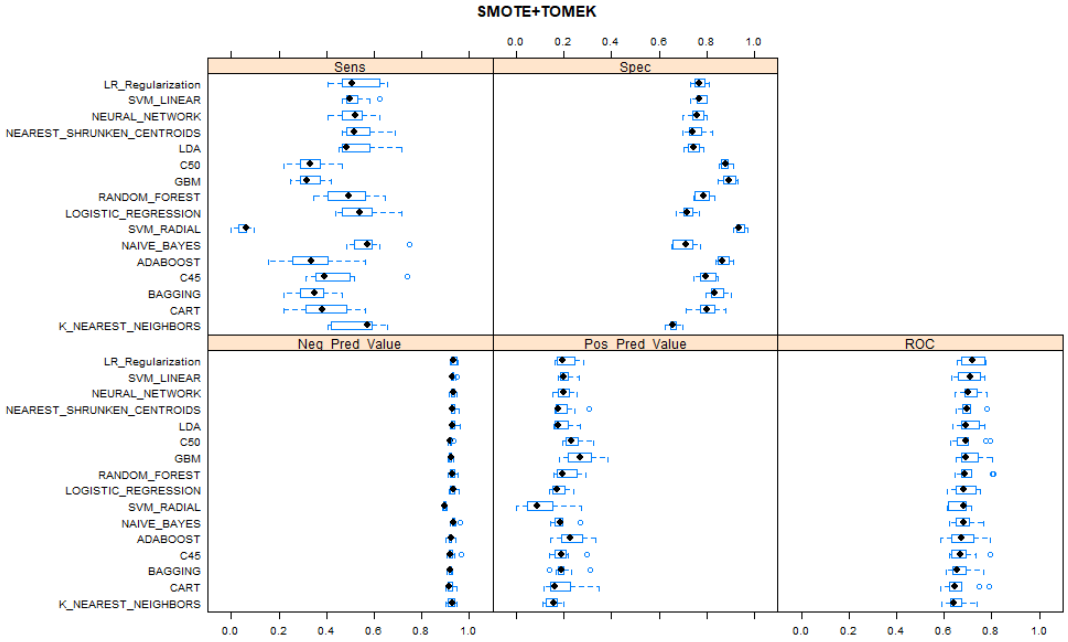


Figure J.8 - Cross-validation boxplot using the best hyperparameter - with the SMOTE + Tomek strategy.

Table J.9 - Results for each method from 10-fold cross-validation - with the SMOTE + NCL strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.62	0.65	0.68	0.69	0.72	0.76	0.47	0.51	0.59	0.59	0.65	0.72	0.64	0.67	0.69	0.69	0.70	0.73	0.14	0.15	0.17	0.18	0.20	0.23	0.92	0.92	0.94	0.94	0.95	0.96
LR_Regularization	0.66	0.68	0.71	0.71	0.76	0.77	0.44	0.53	0.56	0.56	0.63	0.66	0.67	0.71	0.74	0.73	0.75	0.79	0.16	0.17	0.18	0.19	0.22	0.26	0.92	0.93	0.93	0.94	0.95	0.95
LDA	0.63	0.67	0.69	0.70	0.74	0.77	0.45	0.50	0.55	0.56	0.62	0.75	0.68	0.70	0.71	0.71	0.73	0.75	0.14	0.16	0.17	0.18	0.20	0.25	0.92	0.92	0.93	0.94	0.94	0.96
NEAREST_SHRUNKEN_CENTROIDS	0.65	0.68	0.70	0.70	0.71	0.78	0.48	0.53	0.56	0.58	0.60	0.72	0.67	0.71	0.72	0.73	0.74	0.80	0.16	0.17	0.18	0.19	0.20	0.29	0.93	0.93	0.93	0.94	0.94	0.96
SVM_LINEAR	0.63	0.69	0.72	0.71	0.75	0.77	0.50	0.52	0.56	0.57	0.59	0.72	0.69	0.73	0.74	0.74	0.76	0.78	0.16	0.18	0.19	0.20	0.21	0.27	0.93	0.93	0.94	0.94	0.94	0.96
NEURAL_NETWORK	0.64	0.68	0.70	0.71	0.73	0.77	0.44	0.51	0.56	0.56	0.59	0.69	0.67	0.71	0.73	0.73	0.74	0.78	0.15	0.18	0.19	0.19	0.20	0.24	0.92	0.93	0.94	0.94	0.94	0.96
SVM_RADIAL	0.60	0.64	0.64	0.65	0.67	0.71	0.03	0.06	0.06	0.07	0.09	0.13	0.88	0.91	0.91	0.92	0.93	0.95	0.04	0.06	0.09	0.09	0.11	0.17	0.89	0.89	0.90	0.90	0.90	0.90
K_NEAREST_NEIGHBORS	0.59	0.62	0.65	0.66	0.68	0.74	0.47	0.53	0.60	0.59	0.62	0.72	0.61	0.62	0.63	0.64	0.65	0.67	0.12	0.14	0.15	0.15	0.16	0.20	0.91	0.92	0.94	0.93	0.94	0.95
NAIVE_BAYES	0.62	0.65	0.67	0.69	0.71	0.78	0.48	0.54	0.59	0.58	0.59	0.72	0.66	0.67	0.69	0.71	0.73	0.78	0.15	0.17	0.18	0.18	0.19	0.26	0.93	0.93	0.94	0.94	0.94	0.96
C45	0.61	0.65	0.66	0.68	0.68	0.78	0.35	0.41	0.47	0.48	0.51	0.68	0.72	0.78	0.79	0.79	0.81	0.85	0.13	0.18	0.20	0.21	0.23	0.31	0.91	0.92	0.93	0.93	0.94	0.96
CART	0.58	0.62	0.63	0.65	0.68	0.74	0.28	0.42	0.47	0.46	0.50	0.69	0.66	0.75	0.80	0.77	0.80	0.82	0.13	0.15	0.18	0.19	0.22	0.29	0.91	0.92	0.92	0.93	0.93	0.96
C50	0.63	0.66	0.68	0.70	0.70	0.82	0.22	0.35	0.38	0.38	0.40	0.56	0.82	0.84	0.85	0.85	0.86	0.89	0.16	0.20	0.22	0.22	0.23	0.32	0.91	0.92	0.92	0.92	0.93	0.94
RANDOM_FOREST	0.64	0.66	0.69	0.70	0.71	0.81	0.31	0.42	0.50	0.49	0.55	0.68	0.69	0.75	0.77	0.77	0.79	0.83	0.16	0.17	0.18	0.19	0.21	0.26	0.91	0.92	0.93	0.93	0.94	0.96
GBM	0.65	0.67	0.69	0.71	0.74	0.81	0.32	0.38	0.40	0.40	0.43	0.53	0.82	0.83	0.85	0.85	0.88	0.89	0.18	0.21	0.23	0.24	0.25	0.33	0.92	0.92	0.93	0.93	0.93	0.94
BAGGING	0.60	0.62	0.65	0.66	0.67	0.76	0.25	0.32	0.39	0.39	0.46	0.53	0.76	0.80	0.83	0.82	0.84	0.87	0.13	0.16	0.19	0.20	0.25	0.28	0.90	0.92	0.92	0.92	0.93	0.94
ADABOOST	0.60	0.63	0.68	0.68	0.72	0.79	0.19	0.27	0.32	0.32	0.36	0.44	0.83	0.85	0.86	0.86	0.87	0.89	0.13	0.18	0.19	0.20	0.23	0.28	0.90	0.91	0.92	0.92	0.92	0.93

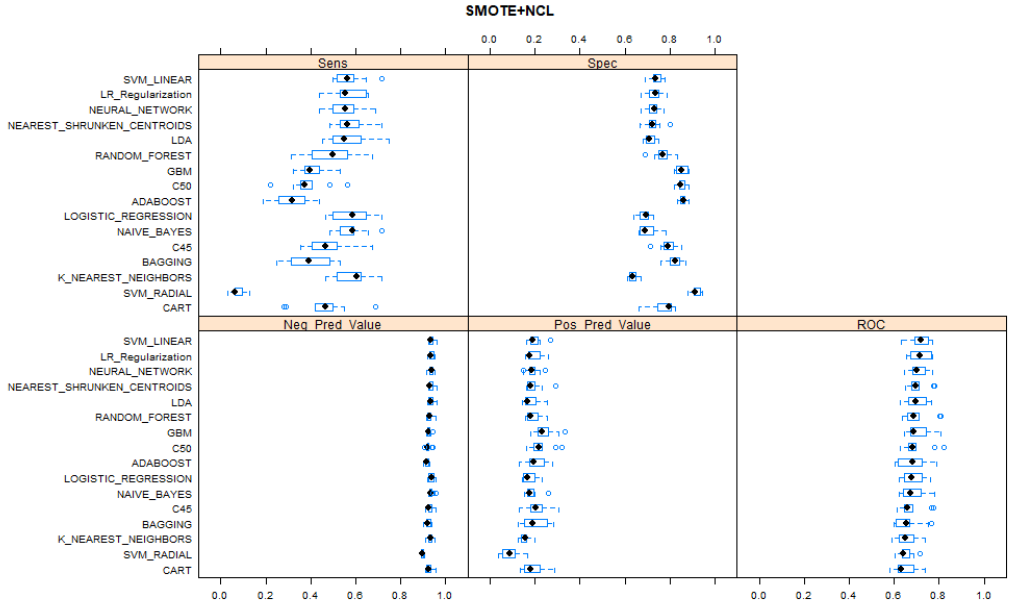


Figure J.9 - Cross-validation boxplot using the best hyperparameter - with the SMOTE + NCL strategy.

Table J.10 - Results for each method from 10-fold cross-validation - with the SMOTE + OSS strategy.

Methods	ROC						Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.61	0.66	0.68	0.69	0.73	0.75	0.44	0.47	0.54	0.55	0.59	0.72	0.67	0.71	0.72	0.72	0.74	0.77	0.14	0.15	0.17	0.18	0.20	0.24	0.92	0.92	0.93	0.93	0.94	0.96
LR_Regularization	0.65	0.68	0.72	0.72	0.76	0.78	0.38	0.47	0.51	0.53	0.61	0.66	0.73	0.75	0.77	0.77	0.79	0.81	0.16	0.17	0.19	0.21	0.24	0.28	0.92	0.93	0.93	0.94	0.95	0.95
LDA	0.63	0.67	0.69	0.70	0.74	0.77	0.45	0.47	0.48	0.53	0.58	0.72	0.70	0.73	0.75	0.75	0.77	0.78	0.16	0.16	0.17	0.19	0.21	0.27	0.92	0.92	0.93	0.93	0.94	0.96
NEAREST_SHRUNKEN_CENTROIDS	0.65	0.68	0.70	0.70	0.71	0.78	0.47	0.49	0.52	0.54	0.57	0.69	0.70	0.73	0.74	0.75	0.77	0.82	0.16	0.17	0.18	0.20	0.21	0.31	0.92	0.93	0.93	0.94	0.94	0.96
SVM_LINEAR	0.63	0.66	0.71	0.71	0.75	0.77	0.47	0.47	0.49	0.51	0.53	0.63	0.72	0.77	0.78	0.78	0.80	0.80	0.16	0.19	0.21	0.21	0.22	0.27	0.92	0.93	0.93	0.93	0.94	0.95
NEURAL_NETWORK	0.64	0.69	0.70	0.71	0.73	0.78	0.41	0.47	0.52	0.52	0.55	0.63	0.70	0.74	0.76	0.76	0.78	0.80	0.15	0.18	0.20	0.20	0.22	0.25	0.92	0.92	0.94	0.93	0.94	0.95
SVM_RADIAL	0.61	0.62	0.68	0.66	0.69	0.71	0.00	0.03	0.06	0.05	0.06	0.09	0.91	0.93	0.94	0.94	0.95	0.96	0.00	0.05	0.09	0.09	0.11	0.21	0.89	0.89	0.90	0.90	0.90	0.90
K_NEAREST_NEIGHBORS	0.59	0.63	0.64	0.65	0.67	0.74	0.41	0.44	0.57	0.53	0.61	0.66	0.63	0.65	0.66	0.66	0.67	0.70	0.11	0.12	0.16	0.15	0.17	0.20	0.90	0.91	0.93	0.93	0.94	0.95
NAIVE_BAYES	0.65	0.67	0.69	0.69	0.71	0.76	0.56	0.64	0.73	0.74	0.82	0.90	0.19	0.51	0.53	0.54	0.64	0.75	0.11	0.15	0.15	0.16	0.17	0.21	0.93	0.94	0.95	0.95	0.95	0.98
C45	0.62	0.64	0.67	0.68	0.69	0.79	0.31	0.35	0.39	0.43	0.48	0.74	0.74	0.77	0.80	0.80	0.83	0.85	0.14	0.16	0.19	0.19	0.21	0.29	0.91	0.92	0.92	0.93	0.93	0.97
CART	0.55	0.61	0.65	0.66	0.68	0.77	0.29	0.35	0.42	0.41	0.46	0.53	0.74	0.79	0.82	0.81	0.83	0.86	0.13	0.15	0.20	0.20	0.23	0.30	0.91	0.91	0.92	0.92	0.93	0.94
C50	0.63	0.67	0.68	0.70	0.71	0.80	0.22	0.28	0.32	0.34	0.38	0.50	0.85	0.87	0.88	0.88	0.90	0.92	0.17	0.20	0.24	0.25	0.28	0.37	0.91	0.92	0.92	0.92	0.93	0.94
RANDOM_FOREST	0.65	0.68	0.69	0.70	0.71	0.80	0.28	0.44	0.49	0.50	0.55	0.66	0.72	0.75	0.77	0.77	0.79	0.80	0.12	0.17	0.19	0.19	0.21	0.28	0.90	0.92	0.93	0.93	0.94	0.95
GBM	0.65	0.68	0.69	0.70	0.73	0.81	0.28	0.31	0.32	0.34	0.38	0.41	0.86	0.87	0.89	0.89	0.92	0.93	0.20	0.23	0.26	0.27	0.31	0.39	0.92	0.92	0.92	0.92	0.93	0.93
BAGGING	0.62	0.63	0.66	0.66	0.69	0.71	0.22	0.26	0.32	0.33	0.38	0.44	0.80	0.81	0.84	0.84	0.86	0.87	0.12	0.16	0.20	0.18	0.21	0.24	0.91	0.91	0.92	0.92	0.92	0.93
ADABOOST	0.60	0.64	0.68	0.69	0.73	0.79	0.16	0.23	0.29	0.28	0.33	0.41	0.84	0.87	0.87	0.88	0.89	0.91	0.14	0.18	0.20	0.21	0.23	0.30	0.90	0.91	0.92	0.92	0.92	0.93

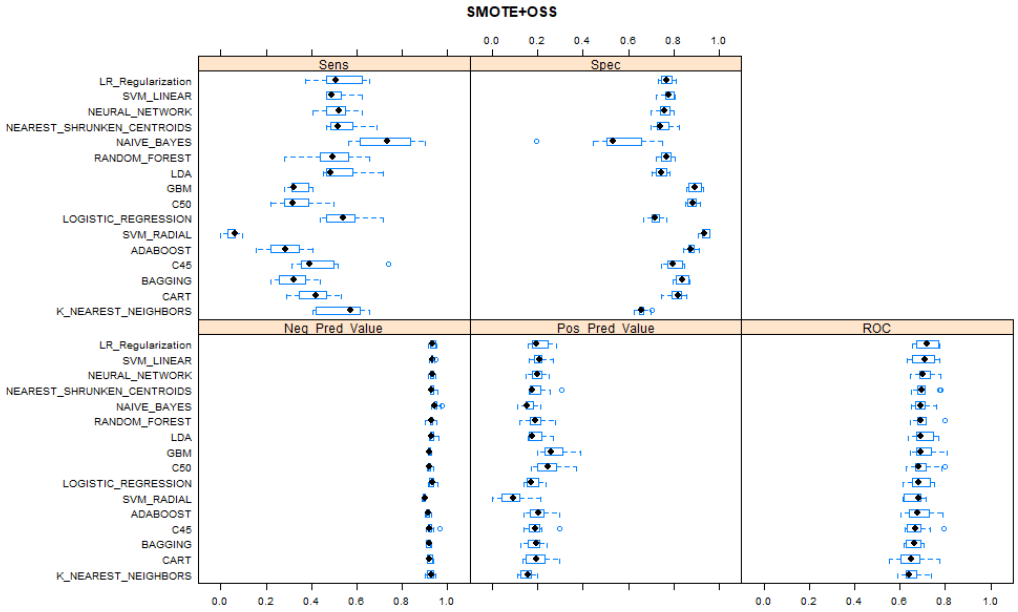


Figure J.10 - Cross-validation boxplot using the best hyperparameter - with the SMOTE + OSS strategy.

Table J.11 - The best hyperparameters values for each method from 10-fold cross-validation.

Strategy	LR parameter	LR_Regularization alpha	LR_Regularization lambda	LDA parameter	NSC threshold	SVM_LINEAR C	NEURAL_NETWORK size	NEURAL_NETWORK decay	SVM_RADIAL sigma	SVM_RADIAL C	KNN k	NAIVE_BAYES Laplace	NAIVE_BAYES usekernel	NAIVE_BAYES adjust
None	none	0.2	0.01	none	0	0.25	5	0.5	0.125	0.25	10	5.00	TRUE	5.00
Down	none	0	0.114210526	none	0	8	4	0.5	0.125	0.125	7	2.74	TRUE	2.74
Up	none	0.8	0.01	none	1	0.125	4	1.5	0.125	0.125	10	5.00	TRUE	5.00
SMOTE	none	0.6	0.01	none	0	4	3	2	0.125	8	9	3.97	TRUE	3.97
Tomek Link	none	0.2	0.01	none	0	0.25	5	0.5	0.125	0.125	10	3.77	TRUE	3.77
NCL	none	0.2	0.01	none	0	0.25	2	1	0.125	0.25	10	4.18	TRUE	4.18
OSS	none	0.4	0.01	none	0	0.125	2	0.5	0.125	0.125	10	3.87	TRUE	3.87
SMOTE + Tomek	none	0.6	0.01	none	0	4	3	2	0.125	8	9	2.44	FALSE	2.44
SMOTE + NCL	none	0.6	0.01	none	0	4	3	2	0.125	1	9	2.44	FALSE	2.44
SMOTE + OSS	none	0.6	0.01	none	0	4	3	2	0.125	8	9	2.95	TRUE	2.95

Strategy	C45 C	C45 M	CART cp	C50 trials	C50 model	C50 winnow	RF mtry	GBM n.trees	GBM interaction.depth	GBM shrinkage	GBM n.minobsinnode	BAGGING mfinal	BAGGING maxdepth	BAGGING coeflearn	ADABOOST mfinal	ADABOOST maxdepth	AdaBoost coeflearn
None	0.5	10	0.001	30	tree	FALSE	19	300	5	0.01	10	50	12	Breiman	50	12	Breiman
Down	0.3	25	0.1	25	tree	FALSE	5	150	5	0.01	10	20	12	Breiman	20	12	Breiman
Up	0.05	30	0.01	25	tree	FALSE	5	300	5	0.01	10	40	12	Breiman	40	12	Breiman
SMOTE	0.5	30	0.01	30	tree	FALSE	19	150	3	0.1	10	40	12	Breiman	40	12	Breiman
Tomek Link	0.5	10	0.005	1	tree	TRUE	11	300	5	0.01	20	50	8	Breiman	50	8	Breiman
NCL	0.5	10	0.001	15	tree	FALSE	17	300	5	0.01	20	50	12	Breiman	50	12	Breiman
OSS	0.5	10	0.005	25	tree	TRUE	29	300	5	0.01	10	50	12	Breiman	50	12	Breiman
SMOTE + Tomek	0.5	40	0.01	30	tree	FALSE	11	300	5	0.01	10	50	8	Breiman	50	8	Breiman
SMOTE + NCL	0.00 1	40	0.01	30	tree	FALSE	17	300	5	0.01	10	50	12	Breiman	50	12	Breiman
SMOTE + OSS	0.5	40	0.005	30	tree	FALSE	10	270	5	0.01	10	40	12	Breiman	40	12	Breiman

Appendix K

Table K.1 - Performance of the ML models computed from the independent test set for each combination considering a weight two times higher for false-negative records when compared with a false positive classification **with** changing the cut-off value.

METHODS	NONE								DOWNSAMPLING								UPSAMPLING								SMOTE							
	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC
LOGISTIC_REGRESSION	0.31	0.12	0.97	0.31	0.91	0.08	0.73	0.14	0.32	0.88	0.41	0.14	0.97	0.22	0.73	0.18	0.35	0.85	0.50	0.16	0.97	0.19	0.73	0.21	0.31	0.76	0.53	0.15	0.95	0.17	0.71	0.17
LR_Regularization	0.31	0.06	0.98	0.26	0.90	0.09	0.74	0.09	0.34	0.94	0.32	0.13	0.98	0.21	0.73	0.17	0.40	0.81	0.54	0.16	0.96	0.20	0.75	0.21	0.34	0.77	0.55	0.16	0.96	0.17	0.73	0.19
LDA	0.33	0.14	0.96	0.31	0.91	0.09	0.73	0.15	0.28	0.88	0.38	0.14	0.97	0.22	0.72	0.17	0.35	0.85	0.50	0.16	0.97	0.19	0.73	0.21	0.30	0.77	0.54	0.16	0.95	0.17	0.71	0.18
NEAREST_SHRUNKEN_CENTROIDS	0.66	0.04	0.99	0.30	0.90	0.09	0.73	0.08	0.34	0.81	0.52	0.16	0.96	0.21	0.73	0.20	0.33	0.81	0.49	0.15	0.96	0.21	0.74	0.18	0.27	0.77	0.53	0.15	0.95	0.18	0.73	0.18
SVM_LINEAR	0.13	0.00	1.00	NA	0.90	0.09	0.57	0.00	0.39	0.83	0.42	0.14	0.96	0.21	0.71	0.15	0.36	0.81	0.50	0.15	0.96	0.19	0.72	0.19	0.36	0.69	0.60	0.16	0.95	0.17	0.71	0.17
NEURAL_NETWORK	0.28	0.10	0.97	0.31	0.91	0.08	0.74	0.13	0.39	0.79	0.49	0.15	0.96	0.22	0.72	0.17	0.29	0.86	0.42	0.14	0.96	0.19	0.73	0.17	0.40	0.73	0.65	0.19	0.96	0.17	0.72	0.23
SVM_RADIAL	0.08	0.91	0.36	0.14	0.97	0.09	0.69	0.17	0.51	0.03	0.97	0.10	0.90	0.25	0.69	0.00	0.10	0.03	0.94	0.05	0.90	0.11	0.67	-0.04	0.42	0.01	0.94	0.02	0.90	0.12	0.65	-0.07
K_NEAREST_NEIGHBORS	0.29	0.18	0.94	0.25	0.91	0.09	0.63	0.14	0.33	0.71	0.53	0.14	0.94	0.20	0.68	0.14	0.70	0.42	0.75	0.16	0.92	0.26	0.60	0.12	0.35	0.72	0.55	0.15	0.95	0.22	0.69	0.16
NAIVE_BAYES	1.00	0.01	1.00	0.25	0.90	0.13	0.75	0.04	0.00	0.92	0.32	0.13	0.97	0.20	0.74	0.16	0.00	0.88	0.43	0.15	0.97	0.22	0.74	0.19	0.01	0.86	0.49	0.16	0.97	0.18	0.73	0.21
C45	0.32	0.08	0.97	0.25	0.90	0.09	0.64	0.09	0.31	0.83	0.34	0.12	0.95	0.24	0.62	0.11	0.32	0.49	0.71	0.16	0.93	0.20	0.57	0.13	0.32	0.41	0.86	0.25	0.93	0.11	0.68	0.22
CART	0.29	0.15	0.96	0.29	0.91	0.09	0.73	0.15	NA	1.00	0.00	0.10	NA	0.23	0.67	0.00	0.46	0.65	0.67	0.18	0.95	0.21	0.66	0.20	0.31	0.32	0.88	0.23	0.92	0.11	0.67	0.17
C50	Inf	0.00	1.00	NA	0.90	0.10	0.50	0.00	0.48	0.71	0.63	0.17	0.95	0.22	0.73	0.20	0.57	0.05	0.99	0.29	0.90	0.10	0.69	0.08	0.40	0.44	0.85	0.24	0.93	0.11	0.69	0.22
RANDOM_FOREST	0.05	0.17	0.95	0.25	0.91	0.10	0.69	0.14	0.19	0.92	0.39	0.14	0.98	0.21	0.75	0.20	0.13	0.91	0.37	0.14	0.97	0.22	0.75	0.18	0.13	0.67	0.67	0.18	0.95	0.10	0.71	0.21
GBM	0.22	0.15	0.95	0.27	0.91	0.08	0.77	0.14	0.41	0.85	0.45	0.15	0.96	0.22	0.73	0.18	0.40	0.82	0.53	0.16	0.96	0.19	0.75	0.21	0.30	0.45	0.79	0.19	0.93	0.11	0.71	0.17
BAGGING	0.35	0.19	0.94	0.27	0.91	0.09	0.69	0.16	0.55	0.62	0.70	0.18	0.94	0.23	0.71	0.20	0.80	0.00	1.00	0.00	0.90	0.10	0.65	-0.01	0.45	0.29	0.86	0.19	0.92	0.13	0.67	0.13
ADABOOST	0.49	0.00	0.98	0.00	0.90	0.10	0.65	-0.04	0.69	0.19	0.92	0.20	0.91	0.22	0.69	0.11	0.51	0.00	0.99	0.00	0.90	0.10	0.66	-0.04	0.57	0.10	0.96	0.22	0.91	0.14	0.66	0.09

METHODS	SMOTE + TOMEK LINK								SMOTE + NCL								SMOTE + OSS							
	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC
LOGISTIC_REGRESSION	0.37	0.72	0.58	0.16	0.95	0.19	0.71	0.18	0.39	0.72	0.56	0.15	0.95	0.20	0.71	0.17	0.34	0.76	0.54	0.15	0.95	0.19	0.71	0.18
LR_Regularization	0.35	0.77	0.56	0.16	0.96	0.17	0.73	0.20	0.34	0.81	0.51	0.15	0.96	0.18	0.73	0.19	0.35	0.77	0.56	0.16	0.96	0.17	0.73	0.20
LDA	0.32	0.76	0.57	0.16	0.95	0.17	0.71	0.19	0.37	0.69	0.60	0.16	0.95	0.19	0.72	0.18	0.34	0.74	0.58	0.17	0.95	0.17	0.71	0.20
NEAREST_SHRUNKEN_CENTROIDS	0.26	0.78	0.51	0.15	0.96	0.18	0.73	0.18	0.28	0.77	0.51	0.15	0.95	0.20	0.73	0.17	0.27	0.77	0.52	0.15	0.95	0.18	0.73	0.17
SVM_LINEAR	0.35	0.77	0.55	0.16	0.96	0.17	0.71	0.19	0.35	0.78	0.54	0.16	0.96	0.18	0.71	0.19	0.33	0.78	0.53	0.16	0.96	0.17	0.72	0.19
NEURAL_NETWORK	0.40	0.76	0.64	0.19	0.96	0.17	0.72	0.24	0.43	0.72	0.64	0.18	0.95	0.19	0.72	0.22	0.41	0.74	0.65	0.19	0.96	0.18	0.72	0.24
SVM_RADIAL	0.45	0.01	0.95	0.03	0.90	0.12	0.65	-0.06	0.54	0.01	0.93	0.02	0.89	0.13	0.64	-0.07	0.43	0.01	0.95	0.03	0.90	0.12	0.65	-0.06
K_NEAREST_NEIGHBORS	0.38	0.72	0.54	0.15	0.95	0.22	0.69	0.16	0.38	0.74	0.53	0.15	0.95	0.24	0.70	0.16	0.38	0.72	0.54	0.15	0.95	0.22	0.69	0.16
NAIVE_BAYES	0.06	0.68	0.55	0.14	0.94	0.29	0.66	0.14	0.04	0.73	0.51	0.14	0.94	0.30	0.66	0.15	0.02	0.83	0.50	0.16	0.96	0.26	0.72	0.20
C45	0.34	0.49	0.74	0.17	0.93	0.16	0.66	0.15	0.48	0.55	0.78	0.22	0.94	0.17	0.69	0.23	0.30	0.50	0.71	0.16	0.93	0.16	0.66	0.14
CART	0.47	0.45	0.71	0.15	0.92	0.19	0.68	0.10	0.28	0.71	0.60	0.16	0.95	0.20	0.67	0.19	0.28	0.49	0.73	0.17	0.93	0.18	0.60	0.14
C50	0.50	0.26	0.88	0.19	0.91	0.14	0.67	0.12	0.50	0.32	0.88	0.23	0.92	0.14	0.67	0.17	0.50	0.31	0.88	0.23	0.92	0.13	0.69	0.17
RANDOM_FOREST	0.18	0.79	0.52	0.16	0.96	0.16	0.73	0.19	0.34	0.65	0.65	0.17	0.94	0.18	0.73	0.19	0.23	0.77	0.58	0.17	0.96	0.16	0.73	0.21
GBM	0.36	0.58	0.75	0.21	0.94	0.13	0.73	0.22	0.40	0.59	0.76	0.21	0.94	0.14	0.74	0.23	0.36	0.58	0.73	0.19	0.94	0.13	0.73	0.20
BAGGING	0.55	0.23	0.90	0.20	0.91	0.14	0.67	0.12	0.55	0.31	0.86	0.20	0.92	0.16	0.67	0.14	0.55	0.32	0.89	0.25	0.92	0.14	0.68	0.19
ADABOOST	0.56	0.19	0.94	0.27	0.91	0.17	0.70	0.16	0.61	0.17	0.95	0.29	0.91	0.15	0.71	0.16	0.63	0.06	0.98	0.26	0.90	0.15	0.71	0.09

METHODS	SMOTEBoost								RUSBoost								SMOTEBagging								UnderBagging							
	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC	lrThresh	Sens	Spec	PPV	NPV	brierScore	AUC	MCC
LOGISTIC_REGRESSION																																
LR_Regularization																																
LDA																																
NEAREST_SHRUNKEN_CENTROIDS																																
SVM_LINEAR																																
NEURAL_NETWORK																																
SVM_RADIAL	0.27	0.71	0.64	0.18	0.95	0.12	0.70	0.21	0.27	0.62	0.75	0.22	0.95	0.10	0.72	0.24	0.28	0.72	0.65	0.18	0.95	0.15	0.70	0.22	0.09	0.55	0.79	0.22	0.94	0.09	0.68	0.24
K_NEAREST_NEIGHBORS																																
NAIVE_BAYES	0.24	0.81	0.49	0.15	0.96	0.14	0.70	0.18	0.23	0.78	0.42	0.13	0.95	0.12	0.70	0.13	0.00	0.94	0.30	0.13	0.98	0.23	0.73	0.16	0.00	0.88	0.45	0.15	0.97	0.19	0.73	0.20
C45																																
CART	0.24	0.88	0.39	0.14	0.97	0.12	0.70	0.17	0.25	0.77	0.59	0.17	0.96	0.11	0.74	0.22	0.24	0.90	0.45	0.15	0.98	0.15	0.74	0.21	NA	1.00	0.00	0.10	NA	0.09	0.50	0.00
C50	0.38	0.19	0.93	0.23	0.91	0.10	0.61	0.13	0.37	0.22	0.93	0.25	0.91	0.10	0.66	0.16	0.51	0.09	0.97	0.23	0.91	0.10	0.67	0.09	0.07	0.82	0.55	0.17	0.96	0.08	0.74	0.22
RANDOM_FOREST	0.51	0.01	0.99	0.10	0.90	0.11	0.73	0.00	0.57	0.00	0.99	0.00	0.90	0.11	0.72	-0.02	0.60	0.00	1.00	0.00	0.90	0.09	0.73	-0.02	0.39	0.06	0.98	0.26	0.90	0.08	0.75	0.09
GBM																																
BAGGING																																
AdaBoost																																

Appendix L

Table L.1 - The p-values of the Post hoc Nemenyi test comparing the NPV among the balancing strategies, where p-values marked with red color have results different statistically for a 95% confidence level.

	Downsampling	Upsampling	SMOTE	SMOTE_Tomek	SMOTE_NCL	SMOTE_OSS	SMOTEBoost	RUSBoost	SMOTEBagging
Upsampling	0.800	NA	NA	NA	NA	NA	NA	NA	NA
SMOTE	0.568	1.000	NA	NA	NA	NA	NA	NA	NA
SMOTE_Tomek	0.444	1.000	1.000	NA	NA	NA	NA	NA	NA
SMOTE_NCL	0.404	1.000	1.000	1.000	NA	NA	NA	NA	NA
SMOTE_OSS	0.589	1.000	1.000	1.000	1.000	NA	NA	NA	NA
SMOTEBoost	1.000	0.404	0.205	0.135	0.117	0.218	NA	NA	NA
RUSBoost	0.987	0.145	0.056	0.033	0.027	0.061	1.000	NA	NA
SMOTEBagging	0.424	0.003	0.001	0.000	0.000	0.001	0.816	0.977	NA
UnderBagging	0.056	0.000	0.000	0.000	0.000	0.000	0.247	0.568	0.997

Table L.2 - The p-values of the Post hoc Nemenyi test comparing the NPV among the machine learning techniques, where p-values marked with red color have results different statistically for a 95% confidence level.

	LR	LR_regularization	LDA	NSC	SVM_LINEAR	NN	SVM_RADIAL	KNN	NB	C45	CART	C50	RF	GBM	BAGGING
LR_regularization	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
LDA	1.000	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NSC	1.000	1.000	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
SVM_LINEAR	1.000	1.000	1.000	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NN	0.999	1.000	1.000	1.000	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
SVM_RADIAL	0.062	0.003	0.018	0.018	0.002	0.001	NA	NA	NA	NA	NA	NA	NA	NA	NA
KNN	1.000	1.000	1.000	1.000	1.000	1.000	0.024	NA	NA	NA	NA	NA	NA	NA	NA
NB	1.000	1.000	1.000	1.000	0.999	0.993	0.115	1.000	NA	NA	NA	NA	NA	NA	NA
C45	1.000	1.000	1.000	1.000	0.999	0.993	0.115	1.000	1.000	NA	NA	NA	NA	NA	NA
CART	1.000	0.973	1.000	1.000	0.955	0.896	0.361	1.000	1.000	1.000	NA	NA	NA	NA	NA
C50	0.315	0.029	0.130	0.130	0.021	0.010	1.000	0.166	0.462	0.462	0.815	NA	NA	NA	NA
RF	1.000	0.985	1.000	1.000	0.973	0.930	0.300	1.000	1.000	1.000	1.000	0.757	NA	NA	NA
GBM	1.000	1.000	1.000	1.000	1.000	1.000	0.011	1.000	1.000	1.000	0.998	0.094	0.999	NA	NA
BAGGING	1.000	0.991	1.000	1.000	0.982	0.950	0.258	1.000	1.000	1.000	1.000	0.709	1.000	1.000	NA
ADABOOST	1.000	0.980	1.000	1.000	0.965	0.914	0.330	1.000	1.000	1.000	1.000	0.787	1.000	0.999	1.000

Table L.3 - The p-values of the Post hoc Nemenyi test comparing the MCC among the balancing strategies, where p-values marked with red color have results different statistically for a 95% confidence level.

	Downsampling	Upsampling	SMOTE	SMOTE_Tomek	SMOTE_NCL	SMOTE_OSS	SMOTEBoost	RUSBoost	SMOTEBagging
Upsampling	1.000	NA	NA	NA	NA	NA	NA	NA	NA
SMOTE	1.000	1.000	NA	NA	NA	NA	NA	NA	NA
SMOTE_Tomek	1.000	1.000	1.000	NA	NA	NA	NA	NA	NA
SMOTE_NCL	1.000	1.000	1.000	1.000	NA	NA	NA	NA	NA
SMOTE_OSS	0.991	1.000	1.000	1.000	1.000	NA	NA	NA	NA
SMOTEBoost	0.205	0.385	0.610	0.484	0.505	0.846	NA	NA	NA
RUSBoost	0.027	0.073	0.167	0.108	0.117	0.366	0.999	NA	NA
SMOTEBagging	0.004	0.014	0.040	0.022	0.025	0.117	0.963	1.000	NA
UnderBagging	0.000	0.000	0.001	0.000	0.000	0.003	0.347	0.831	0.984

Table L.4 - The p-values of the Post hoc Nemenyi test comparing the MCC among the machine learning techniques, where p-values marked with red color have results different statistically for a 95% confidence level.

	LR	LR_regularization	LDA	NSC	SVM_LINEAR	NN	SVM_RADIAL	KNN	NB	C45	CART	C50	RF	GBM	BAGGING
LR_regularization	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
LDA	1.000	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NSC	1.000	1.000	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
SVM_LINEAR	1.000	1.000	1.000	1.000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NN	0.864	1.000	0.999	0.989	0.998	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
SVM_RADIAL	0.445	0.046	0.062	0.157	0.088	0.001	NA	NA	NA	NA	NA	NA	NA	NA	NA
KNN	1.000	1.000	1.000	1.000	1.000	0.815	0.515	NA	NA	NA	NA	NA	NA	NA	NA
NB	1.000	0.922	0.950	0.993	0.973	0.272	0.955	1.000	NA	NA	NA	NA	NA	NA	NA
C50	1.000	0.896	0.930	0.989	0.960	0.233	0.969	1.000	1.000	0.886	1.000	NA	NA	NA	NA
RF	1.000	1.000	1.000	1.000	1.000	0.815	0.515	1.000	1.000	1.000	0.991	1.000	NA	NA	NA
GBM	0.569	0.987	0.977	0.886	0.955	1.000	0.000	0.497	0.088	0.989	0.010	0.071	0.497	NA	NA
BAGGING	1.000	1.000	1.000	1.000	1.000	0.997	0.094	1.000	0.977	1.000	0.709	0.965	1.000	0.950	NA
ADABOOST	1.000	1.000	1.000	1.000	1.000	0.937	0.315	1.000	1.000	1.000	0.950	0.999	1.000	0.709	1.000

Appendix M

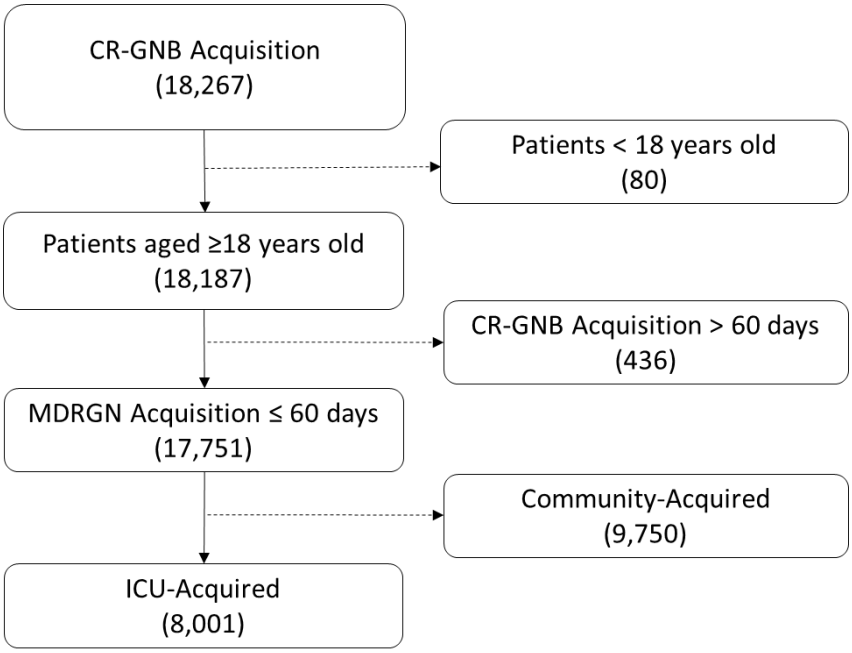


Figure M.1 - Selection of cases.

Appendix N

Table N.1 - Descriptive statistical analysis comparing between Negative and Positive patients.

Variables	Negative (N=1543)	Positive (N=527)	P-value
Hospital Information			
Hospital			
BANGU	151 (9.8%)	63 (12.0%)	0.737
BARRA	264 (17.1%)	88 (16.7%)	
COPA	699 (45.3%)	233 (44.2%)	
NITEROI	204 (13.2%)	68 (12.9%)	
QUINTA	225 (14.6%)	75 (14.2%)	
LOS_hospital_before_test			
Mean (SD)	11.7 (11.2)	17.8 (12.6)	<0.001
Median [Min, Max]	8.00 [3.00, 60.0]	14.0 [3.00, 60.0]	
Patient Information			
Age			
Mean (SD)	74.6 (16.2)	75.4 (14.4)	0.825
Median [Min, Max]	79.0 [18.0, 105]	78.0 [18.0, 102]	
Gender			
F	838 (54.3%)	284 (53.9%)	0.907
M	705 (45.7%)	243 (46.1%)	
BMI			
Mean (SD)	26.9 (12.1)	27.4 (17.8)	0.295
Median [Min, Max]	25.4 [14.1, 283]	25.0 [13.9, 260]	
Missing	272 (17.6%)	105 (19.9%)	
ICU Information			
LOS_ICU_before_test			
Mean (SD)	9.90 (10.5)	15.5 (12.0)	<0.001
Median [Min, Max]	6.00 [0, 60.0]	12.0 [0, 60.0]	
Index			
CharlsonIndex			
Mean (SD)	1.68 (1.85)	2.05 (2.01)	<0.001
Median [Min, Max]	1.00 [0, 11.0]	2.00 [0, 12.0]	
MFIpoints			
Mean (SD)	2.14 (1.38)	2.39 (1.47)	0.002
Median [Min, Max]	2.00 [0, 7.00]	2.00 [0, 7.00]	
Missing	50 (3.2%)	14 (2.7%)	
FrailPatientMFI			
NO	1303 (84.4%)	409 (77.6%)	<0.001
YES	240 (15.6%)	118 (22.4%)	
Saps3Points			
Mean (SD)	51.4 (12.8)	57.9 (13.7)	<0.001
Median [Min, Max]	51.0 [16.0, 104]	56.0 [23.0, 97.0]	
SofaScore			
Mean (SD)	1.71 (2.65)	2.79 (3.67)	<0.001
Median [Min, Max]	1.00 [0, 16.0]	1.00 [0, 17.0]	
Missing	390 (25.3%)	140 (26.6%)	
Priority			
Priority 1	167 (10.8%)	98 (18.6%)	<0.001
Priority 2	516 (33.4%)	127 (24.1%)	
Priority 3	1 (0.1%)	0 (0%)	
Priority 4	2 (0.1%)	0 (0%)	
Priority 5	10 (0.6%)	1 (0.2%)	
Missing	847 (54.9%)	301 (57.1%)	
Comorbidities			
ChronicHealthStatus			
Independent	905 (58.7%)	235 (44.6%)	<0.001
Need for assistance	360 (23.3%)	136 (25.8%)	
Restricted / bedridden	274 (17.8%)	152 (28.8%)	
Missing	4 (0.3%)	4 (0.8%)	
IsChfNyhaClass23			
FALSE	1440 (93.3%)	484 (91.8%)	0.479
TRUE	99 (6.4%)	39 (7.4%)	
Missing	4 (0.3%)	4 (0.8%)	
IsChfNyhaClass4			
FALSE	1532 (99.3%)	522 (99.1%)	0.667
TRUE	7 (0.5%)	1 (0.2%)	
Missing	4 (0.3%)	4 (0.8%)	
IsCrfNoDialysis			
FALSE	1394 (90.3%)	465 (88.2%)	0.307
TRUE	145 (9.4%)	58 (11.0%)	
Missing	4 (0.3%)	4 (0.8%)	
IsCrfDialysis			
FALSE	1505 (97.5%)	504 (95.6%)	0.106
TRUE	34 (2.2%)	19 (3.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsCirrhosisChildAB			
FALSE	1533 (99.4%)	520 (98.7%)	0.868
TRUE	6 (0.4%)	3 (0.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsCirrhosisChildC			

Variables	Negative (N=1543)	Positive (N=527)	P-value
FALSE	1531 (99.2%)	522 (99.1%)	0.548
TRUE	8 (0.5%)	1 (0.2%)	
Missing	4 (0.3%)	4 (0.8%)	
IsHepaticFailure			
FALSE	1532 (99.3%)	520 (98.7%)	1
TRUE	7 (0.5%)	3 (0.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsSolidTumorLocoregiol			
FALSE	1257 (81.5%)	416 (78.9%)	0.311
TRUE	282 (18.3%)	107 (20.3%)	
Missing	4 (0.3%)	4 (0.8%)	
IsSolidTumorMetastatic			
FALSE	1491 (96.6%)	504 (95.6%)	0.667
TRUE	48 (3.1%)	19 (3.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsHematologicalMalignancy			
FALSE	1517 (98.3%)	509 (96.6%)	0.091
TRUE	22 (1.4%)	14 (2.7%)	
Missing	4 (0.3%)	4 (0.8%)	
IsImmunossupression			
FALSE	1397 (90.5%)	463 (87.9%)	0.159
TRUE	142 (9.2%)	60 (11.4%)	
Missing	4 (0.3%)	4 (0.8%)	
IsSevereCopd			
FALSE	1405 (91.1%)	455 (86.3%)	0.006
TRUE	134 (8.7%)	68 (12.9%)	
Missing	4 (0.3%)	4 (0.8%)	
IsSteroidsUse			
FALSE	1515 (98.2%)	512 (97.2%)	0.525
TRUE	24 (1.6%)	11 (2.1%)	
Missing	4 (0.3%)	4 (0.8%)	
IsAids			
FALSE	1530 (99.2%)	522 (99.1%)	0.45
TRUE	9 (0.6%)	1 (0.2%)	
Missing	4 (0.3%)	4 (0.8%)	
IsArterialHypertension			
FALSE	482 (31.2%)	168 (31.9%)	0.774
TRUE	1057 (68.5%)	355 (67.4%)	
Missing	4 (0.3%)	4 (0.8%)	
IsAsthma			
FALSE	1497 (97.0%)	498 (94.5%)	0.032
TRUE	42 (2.7%)	25 (4.7%)	
Missing	4 (0.3%)	4 (0.8%)	
IsDiabetesUncomplicated			
FALSE	1122 (72.7%)	379 (71.9%)	0.891
TRUE	417 (27.0%)	144 (27.3%)	
Missing	4 (0.3%)	4 (0.8%)	
IsDiabetesComplicated			
FALSE	1464 (94.9%)	492 (93.4%)	0.407
TRUE	75 (4.9%)	31 (5.9%)	
Missing	4 (0.3%)	4 (0.8%)	
IsAngina			
FALSE	1455 (94.3%)	506 (96.0%)	0.057
TRUE	84 (5.4%)	17 (3.2%)	
Missing	4 (0.3%)	4 (0.8%)	
IsPreviousMI			
FALSE	1349 (87.4%)	460 (87.3%)	0.918
TRUE	190 (12.3%)	63 (12.0%)	
Missing	4 (0.3%)	4 (0.8%)	
IsCardiacArrhythmia			
FALSE	1375 (89.1%)	475 (90.1%)	0.38
TRUE	164 (10.6%)	48 (9.1%)	
Missing	4 (0.3%)	4 (0.8%)	
IsDeepVenousThrombosis			
FALSE	1479 (95.9%)	480 (91.1%)	<0.001
TRUE	60 (3.9%)	43 (8.2%)	
Missing	4 (0.3%)	4 (0.8%)	
IsPeripheralArteryDisease			
FALSE	1500 (97.2%)	512 (97.2%)	0.697
TRUE	39 (2.5%)	11 (2.1%)	
Missing	4 (0.3%)	4 (0.8%)	
IsChronicAtrialFibrilation			
FALSE	1335 (86.5%)	437 (82.9%)	0.082
TRUE	204 (13.2%)	86 (16.3%)	
Missing	4 (0.3%)	4 (0.8%)	
IsRheumaticDisease			
FALSE	1532 (99.3%)	519 (98.5%)	0.622
TRUE	7 (0.5%)	4 (0.8%)	
Missing	4 (0.3%)	4 (0.8%)	
IsStrokeSequelae			
FALSE	1494 (96.8%)	483 (91.7%)	<0.001
TRUE	45 (2.9%)	40 (7.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsStrokeNoSequelae			

Variables	Negative (N=1543)	Positive (N=527)	P-value
FALSE	1429 (92.6%)	490 (93.0%)	0.581
TRUE	110 (7.1%)	33 (6.3%)	
Missing	4 (0.3%)	4 (0.8%)	
IsDementia			
FALSE	1283 (83.1%)	416 (78.9%)	0.055
TRUE	256 (16.6%)	107 (20.3%)	
Missing	4 (0.3%)	4 (0.8%)	
IsTobaccoConsumption			
FALSE	1417 (91.8%)	477 (90.5%)	0.593
TRUE	122 (7.9%)	46 (8.7%)	
Missing	4 (0.3%)	4 (0.8%)	
IsAlcoholism			
FALSE	1486 (96.3%)	503 (95.4%)	0.787
TRUE	53 (3.4%)	20 (3.8%)	
Missing	4 (0.3%)	4 (0.8%)	
IsPsychiatricDisease			
FALSE	1444 (93.6%)	488 (92.6%)	0.75
TRUE	95 (6.2%)	35 (6.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsMorbidObesity			
FALSE	1483 (96.1%)	509 (96.6%)	0.363
TRUE	56 (3.6%)	14 (2.7%)	
Missing	4 (0.3%)	4 (0.8%)	
IsMalnourishment			
FALSE	1535 (99.5%)	520 (98.7%)	0.528
TRUE	4 (0.3%)	3 (0.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsPepticDisease			
FALSE	1538 (99.7%)	519 (98.5%)	0.022
TRUE	1 (0.1%)	4 (0.8%)	
Missing	4 (0.3%)	4 (0.8%)	
Transplant			
FALSE	1391 (90.1%)	466 (88.4%)	0.446
TRUE	148 (9.6%)	57 (10.8%)	
Missing	4 (0.3%)	4 (0.8%)	
IsHypothyroidism			
FALSE	1280 (83.0%)	428 (81.2%)	0.527
TRUE	259 (16.8%)	95 (18.0%)	
Missing	4 (0.3%)	4 (0.8%)	
IsHyperthyroidism			
FALSE	1534 (99.4%)	521 (98.9%)	1
TRUE	5 (0.3%)	2 (0.4%)	
Missing	4 (0.3%)	4 (0.8%)	
IsDyslipidemias			
FALSE	1292 (83.7%)	446 (84.6%)	0.515
TRUE	247 (16.0%)	77 (14.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsChemotherapy			
FALSE	1458 (94.5%)	493 (93.5%)	0.763
TRUE	81 (5.2%)	30 (5.7%)	
Missing	4 (0.3%)	4 (0.8%)	
IsRadiationTherapy			
FALSE	1493 (96.8%)	502 (95.3%)	0.317
TRUE	46 (3.0%)	21 (4.0%)	
Missing	4 (0.3%)	4 (0.8%)	
IsHistoryOfPneumonia			
FALSE	1477 (95.7%)	484 (91.8%)	0.003
TRUE	62 (4.0%)	39 (7.4%)	
Missing	4 (0.3%)	4 (0.8%)	
Invasive Device during Hospitalization			
VesDURTOTAL			
Mean (SD)	4.64 (7.35)	10.2 (9.28)	<0.001
Median [Min, Max]	2.00 [0, 58.0]	8.00 [0, 52.0]	
VesTIMESTOTAL			
Mean (SD)	0.735 (0.808)	1.23 (0.861)	<0.001
Median [Min, Max]	1.00 [0, 7.00]	1.00 [0, 5.00]	
VESICAL			
NO	665 (43.1%)	81 (15.4%)	<0.001
YES	878 (56.9%)	446 (84.6%)	
ArtDURTOTAL			
Mean (SD)	2.51 (5.50)	7.26 (8.37)	<0.001
Median [Min, Max]	0 [0, 59.0]	5.00 [0, 53.0]	
ArtTIMESTOTAL			
Mean (SD)	0.437 (0.743)	1.05 (1.01)	<0.001
Median [Min, Max]	0 [0, 6.00]	1.00 [0, 5.00]	
ARTERIAL			
NO	1046 (67.8%)	185 (35.1%)	<0.001
YES	497 (32.2%)	342 (64.9%)	
DiaDURTOTAL			
Mean (SD)	0.804 (3.63)	2.29 (6.13)	<0.001
Median [Min, Max]	0 [0, 33.0]	0 [0, 42.0]	
DiaTIMESTOTAL			
Mean (SD)	0.107 (0.432)	0.326 (0.770)	<0.001
Median [Min, Max]	0 [0, 4.00]	0 [0, 6.00]	

Variables	Negative (N=1543)	Positive (N=527)	P-value
DIALYSIS			
NO	1433 (92.9%)	420 (79.7%)	<0.001
YES	110 (7.1%)	107 (20.3%)	
CVCDURTOTAL			
Mean (SD)	4.55 (7.30)	10.6 (9.71)	<0.001
Median [Min, Max]	0 [0, 60.0]	9.00 [0, 54.0]	
CVCTIMESTOTAL			
Mean (SD)	0.655 (0.838)	1.36 (1.07)	<0.001
Median [Min, Max]	0 [0, 5.00]	1.00 [0, 6.00]	
CVC			
NO	815 (52.8%)	111 (21.1%)	<0.001
YES	728 (47.2%)	416 (78.9%)	
MVDURTOTAL			
Mean (SD)	2.30 (6.60)	7.95 (10.0)	<0.001
Median [Min, Max]	0 [0, 57.0]	5.00 [0, 50.0]	
MVTIMESTOTAL			
Mean (SD)	0.260 (0.525)	0.765 (0.749)	<0.001
Median [Min, Max]	0 [0, 4.00]	1.00 [0, 5.00]	
MV			
NO	1195 (77.4%)	204 (38.7%)	<0.001
YES	348 (22.6%)	323 (61.3%)	
PerDURTOTAL			
Mean (SD)	1.31 (2.92)	0.962 (2.59)	0.006
Median [Min, Max]	0 [0, 29.0]	0 [0, 26.0]	
PerTIMESTOTAL			
Mean (SD)	0.512 (1.06)	0.361 (0.871)	0.005
Median [Min, Max]	0 [0, 9.00]	0 [0, 8.00]	
PERIPHERAL			
NO	1118 (72.5%)	411 (78.0%)	0.015
YES	425 (27.5%)	116 (22.0%)	
Reasons for ICU admission			
AdmissionSource			
Emergency	922 (59.8%)	253 (48.0%)	<0.001
Hemodynamic Room	17 (1.1%)	4 (0.8%)	
Operation Room	183 (11.9%)	63 (12.0%)	
Other ICU from hospital	139 (9.0%)	85 (16.1%)	
Others	12 (0.8%)	11 (2.1%)	
Semi Intensive Unit	80 (5.2%)	35 (6.6%)	
Transfer from another hospital	17 (1.1%)	17 (3.2%)	
Ward/Room	169 (11.0%)	55 (10.4%)	
Missing	4 (0.3%)	4 (0.8%)	
AdmissionReason			
Cardiovascular / Shock	402 (26.1%)	74 (14.0%)	<0.001
Elective Surgery	147 (9.5%)	39 (7.4%)	
Emergency surgery	63 (4.1%)	27 (5.1%)	
Endocrine / Metabolic / Renal	47 (3.0%)	13 (2.5%)	
Infection / Sepsis	499 (32.3%)	216 (41.0%)	
Liver and Pancreas / Gastrointestinal	90 (5.8%)	22 (4.2%)	
Neurological	131 (8.5%)	55 (10.4%)	
Non-surgical trauma	31 (2.0%)	14 (2.7%)	
Oncological / Hematological	34 (2.2%)	13 (2.5%)	
Others	22 (1.4%)	10 (1.9%)	
Respiratory	73 (4.7%)	40 (7.6%)	
Missing	4 (0.3%)	4 (0.8%)	
IsNeurologicalComaStuporObtundedDelirium			
FALSE	1269 (82.2%)	380 (72.1%)	<0.001
TRUE	268 (17.4%)	143 (27.1%)	
Missing	6 (0.4%)	4 (0.8%)	
IsNeurologicalSeizures			
FALSE	1489 (96.5%)	492 (93.4%)	0.006
TRUE	48 (3.1%)	31 (5.9%)	
Missing	6 (0.4%)	4 (0.8%)	
IsNeurologicalFocalNeurologicDeficit			
FALSE	1512 (98.0%)	496 (94.1%)	<0.001
TRUE	25 (1.6%)	27 (5.1%)	
Missing	6 (0.4%)	4 (0.8%)	
IsNeurologicalIntracranialMassEffect			
FALSE	1527 (99.0%)	518 (98.3%)	0.68
TRUE	10 (0.6%)	5 (0.9%)	
Missing	6 (0.4%)	4 (0.8%)	
IsCardiovascularHypovolemicHemorrhagicShock			
FALSE	1514 (98.1%)	509 (96.6%)	0.118
TRUE	23 (1.5%)	14 (2.7%)	
Missing	6 (0.4%)	4 (0.8%)	
IsCardiovascularSepticShock			
FALSE	1434 (92.9%)	455 (86.3%)	<0.001
TRUE	103 (6.7%)	68 (12.9%)	
Missing	6 (0.4%)	4 (0.8%)	
IsCardiovascularRhythmDisturbances			
FALSE	1367 (88.6%)	468 (88.8%)	0.792
TRUE	170 (11.0%)	55 (10.4%)	
Missing	6 (0.4%)	4 (0.8%)	
IsCardiovascularAphylacticMixedUndefinedShock			
FALSE	1533 (99.4%)	522 (99.1%)	1

Variables	Negative (N=1543)	Positive (N=527)	P-value
TRUE	4 (0.3%)	1 (0.2%)	
Missing	6 (0.4%)	4 (0.8%)	
IsDigestiveAcuteAbdomen			
FALSE	1492 (96.7%)	501 (95.1%)	0.2
TRUE	45 (2.9%)	22 (4.2%)	
Missing	6 (0.4%)	4 (0.8%)	
IsDigestiveSeverePancreatitis			
FALSE	1532 (99.3%)	522 (99.1%)	0.983
TRUE	5 (0.3%)	1 (0.2%)	
Missing	6 (0.4%)	4 (0.8%)	
IsLiverFailure			
FALSE	1532 (99.3%)	519 (98.5%)	0.351
TRUE	5 (0.3%)	4 (0.8%)	
Missing	6 (0.4%)	4 (0.8%)	
IsTransplantSolidOrgan			
FALSE	1537 (99.6%)	521 (98.9%)	0.107
TRUE	0 (0%)	2 (0.4%)	
Missing	6 (0.4%)	4 (0.8%)	
IsTraumaMultipleTrauma			
FALSE	1508 (97.7%)	511 (97.0%)	0.693
TRUE	29 (1.9%)	12 (2.3%)	
Missing	6 (0.4%)	4 (0.8%)	
IsCardiacSurgery			
FALSE	1523 (98.7%)	520 (98.7%)	0.648
TRUE	14 (0.9%)	3 (0.6%)	
Missing	6 (0.4%)	4 (0.8%)	
IsNeurosurgery			
FALSE	1535 (99.5%)	522 (99.1%)	1
TRUE	2 (0.1%)	1 (0.2%)	
Missing	6 (0.4%)	4 (0.8%)	
Antibiotic use			
J01A			
FALSE	1471 (95.3%)	466 (88.4%)	<0.001
TRUE	72 (4.7%)	61 (11.6%)	
J01C			
FALSE	714 (46.3%)	143 (27.1%)	<0.001
TRUE	829 (53.7%)	384 (72.9%)	
J01D			
FALSE	813 (52.7%)	128 (24.3%)	<0.001
TRUE	730 (47.3%)	399 (75.7%)	
J01E			
FALSE	1503 (97.4%)	497 (94.3%)	0.001
TRUE	40 (2.6%)	30 (5.7%)	
J01F			
FALSE	1092 (70.8%)	305 (57.9%)	<0.001
TRUE	451 (29.2%)	222 (42.1%)	
J01G			
FALSE	1483 (96.1%)	466 (88.4%)	<0.001
TRUE	60 (3.9%)	61 (11.6%)	
J01M			
FALSE	1320 (85.5%)	446 (84.6%)	0.658
TRUE	223 (14.5%)	81 (15.4%)	
J01X			
FALSE	1138 (73.8%)	249 (47.2%)	<0.001
TRUE	405 (26.2%)	278 (52.8%)	
J04A			
FALSE	1540 (99.8%)	525 (99.6%)	0.815
TRUE	3 (0.2%)	2 (0.4%)	
Antibiotic			
FALSE	306 (19.8%)	15 (2.8%)	<0.001
TRUE	1237 (80.2%)	512 (97.2%)	

Appendix O

Table O.1 - Results for each method from 10-fold cross-validation.

Methods	MCC						Brier score						ROC						prAUC					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.153	0.234	0.284	0.281	0.342	0.391	0.151	0.154	0.159	0.162	0.168	0.177	0.726	0.750	0.768	0.766	0.783	0.799	0.415	0.454	0.490	0.486	0.527	0.547
LR_Regularization	0.122	0.209	0.255	0.253	0.289	0.366	0.149	0.153	0.157	0.158	0.163	0.167	0.742	0.758	0.771	0.777	0.800	0.813	0.435	0.455	0.495	0.499	0.524	0.598
LDA	0.139	0.285	0.305	0.306	0.356	0.400	0.153	0.154	0.161	0.164	0.171	0.186	0.713	0.746	0.760	0.761	0.777	0.804	0.397	0.444	0.484	0.488	0.513	0.594
NEAREST_SHRUNKEN_CENTROIDS	0.266	0.300	0.319	0.325	0.363	0.377	0.145	0.150	0.161	0.160	0.169	0.172	0.721	0.740	0.763	0.770	0.802	0.825	0.421	0.450	0.477	0.484	0.510	0.576
SVM_LINEAR	0.107	0.158	0.170	0.183	0.188	0.313	0.163	0.169	0.170	0.172	0.176	0.181	0.698	0.743	0.760	0.755	0.770	0.805	0.425	0.455	0.476	0.480	0.503	0.545
NEURAL_NETWORK	0.198	0.248	0.300	0.297	0.352	0.405	0.147	0.150	0.155	0.156	0.161	0.172	0.728	0.759	0.779	0.776	0.794	0.816	0.414	0.477	0.508	0.498	0.530	0.547
SVM_RADIAL	-0.018	0.044	0.088	0.111	0.173	0.295	0.161	0.168	0.173	0.173	0.178	0.189	0.662	0.703	0.725	0.727	0.764	0.772	0.365	0.377	0.416	0.422	0.455	0.514
K_NEAREST_NEIGHBORS	0.037	0.161	0.226	0.218	0.279	0.349	0.149	0.164	0.169	0.169	0.180	0.183	0.698	0.708	0.731	0.737	0.755	0.796	0.380	0.405	0.444	0.450	0.480	0.539
NAIVE_BAYES	0.276	0.290	0.299	0.326	0.339	0.431	0.171	0.190	0.212	0.208	0.221	0.239	0.730	0.746	0.768	0.776	0.795	0.851	0.423	0.466	0.480	0.493	0.529	0.594
C45	0.306	0.320	0.355	0.355	0.381	0.438	0.148	0.156	0.157	0.159	0.161	0.172	0.656	0.695	0.709	0.711	0.741	0.756	0.425	0.465	0.495	0.499	0.543	0.585
CART	0.207	0.307	0.318	0.327	0.373	0.389	0.156	0.158	0.162	0.166	0.168	0.185	0.639	0.680	0.708	0.701	0.727	0.743	0.365	0.391	0.451	0.449	0.500	0.541
C50	0.044	0.331	0.354	0.310	0.366	0.376	0.138	0.151	0.156	0.157	0.162	0.172	0.739	0.744	0.774	0.775	0.793	0.845	0.412	0.471	0.525	0.507	0.541	0.586
RANDOM_FOREST	0.206	0.265	0.292	0.306	0.351	0.405	0.156	0.162	0.169	0.168	0.172	0.187	0.725	0.767	0.778	0.781	0.794	0.835	0.399	0.472	0.526	0.521	0.577	0.620
GBM	0.154	0.262	0.312	0.296	0.344	0.390	0.142	0.148	0.152	0.154	0.159	0.170	0.732	0.772	0.792	0.789	0.804	0.845	0.403	0.492	0.544	0.529	0.579	0.609
BAGGING	0.090	0.201	0.333	0.288	0.353	0.448	0.152	0.161	0.168	0.174	0.184	0.207	0.670	0.719	0.749	0.743	0.775	0.812	0.358	0.395	0.481	0.461	0.516	0.538
ADABOOST	0.090	0.206	0.247	0.250	0.282	0.485	0.156	0.165	0.169	0.171	0.178	0.187	0.687	0.717	0.743	0.736	0.752	0.782	0.366	0.401	0.455	0.450	0.472	0.551

Methods	Sensitivity						Specificity						PPV						NPV					
	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
LOGISTIC_REGRESSION	0.214	0.263	0.294	0.311	0.310	0.476	0.855	0.890	0.907	0.913	0.935	0.968	0.414	0.493	0.549	0.565	0.600	0.765	0.771	0.785	0.791	0.795	0.799	0.828
LR_Regularization	0.143	0.190	0.235	0.223	0.256	0.286	0.911	0.935	0.935	0.946	0.955	0.992	0.429	0.528	0.579	0.606	0.612	0.889	0.762	0.777	0.780	0.781	0.787	0.795
LDA	0.310	0.328	0.345	0.363	0.379	0.476	0.821	0.889	0.902	0.897	0.909	0.951	0.371	0.528	0.547	0.558	0.575	0.700	0.777	0.799	0.804	0.805	0.812	0.828
NEAREST_SHRUNKEN_CENTROIDS	0.326	0.333	0.365	0.382	0.429	0.476	0.831	0.865	0.903	0.897	0.927	0.944	0.462	0.515	0.558	0.572	0.632	0.682	0.797	0.802	0.808	0.810	0.815	0.831
SVM_LINEAR	0.070	0.143	0.155	0.152	0.167	0.214	0.927	0.951	0.951	0.956	0.966	0.984	0.438	0.500	0.519	0.548	0.553	0.800	0.750	0.765	0.767	0.767	0.770	0.780
NEURAL_NETWORK	0.214	0.281	0.329	0.337	0.375	0.476	0.837	0.889	0.915	0.908	0.927	0.959	0.429	0.505	0.546	0.564	0.613	0.706	0.777	0.790	0.795	0.801	0.810	0.835
SVM_RADIAL	0.048	0.077	0.129	0.123	0.143	0.214	0.902	0.923	0.947	0.943	0.966	0.968	0.222	0.333	0.402	0.427	0.520	0.692	0.744	0.751	0.757	0.759	0.767	0.784
K_NEAREST_NEIGHBORS	0.095	0.226	0.294	0.268	0.310	0.405	0.837	0.895	0.903	0.906	0.927	0.951	0.308	0.423	0.485	0.493	0.555	0.684	0.752	0.772	0.788	0.784	0.793	0.816
NAIVE_BAYES	0.381	0.395	0.429	0.434	0.476	0.500	0.806	0.833	0.870	0.866	0.895	0.943	0.467	0.482	0.500	0.537	0.568	0.708	0.805	0.811	0.817	0.818	0.824	0.836
C45	0.256	0.292	0.333	0.334	0.368	0.452	0.887	0.935	0.939	0.939	0.949	0.967	0.576	0.622	0.650	0.659	0.678	0.762	0.788	0.796	0.805	0.805	0.814	0.827
CART	0.310	0.328	0.345	0.358	0.381	0.429	0.855	0.897	0.915	0.912	0.931	0.952	0.438	0.553	0.601	0.591	0.627	0.700	0.791	0.801	0.804	0.806	0.809	0.825
C50	0.119	0.314	0.341	0.329	0.375	0.429	0.886	0.903	0.915	0.921	0.941	0.960	0.313	0.574	0.593	0.585	0.663	0.722	0.752	0.801	0.806	0.801	0.811	0.821
RANDOM_FOREST	0.238	0.266	0.298	0.308	0.331	0.476	0.886	0.905	0.931	0.928	0.949	0.967	0.462	0.561	0.588	0.604	0.637	0.750	0.784	0.786	0.795	0.797	0.801	0.835
GBM	0.233	0.244	0.286	0.294	0.326	0.405	0.886	0.913	0.923	0.930	0.955	0.967	0.417	0.557	0.595	0.600	0.658	0.765	0.773	0.784	0.796	0.794	0.802	0.819
BAGGING	0.262	0.385	0.417	0.403	0.450	0.524	0.789	0.816	0.874	0.861	0.879	0.935	0.333	0.430	0.516	0.508	0.555	0.704	0.765	0.792	0.817	0.808	0.824	0.835
ADABOOST	0.238	0.266	0.357	0.346	0.405	0.465	0.821	0.848	0.871	0.874	0.899	0.943	0.333	0.443	0.482	0.489	0.513	0.741	0.765	0.782	0.797	0.797	0.812	0.835

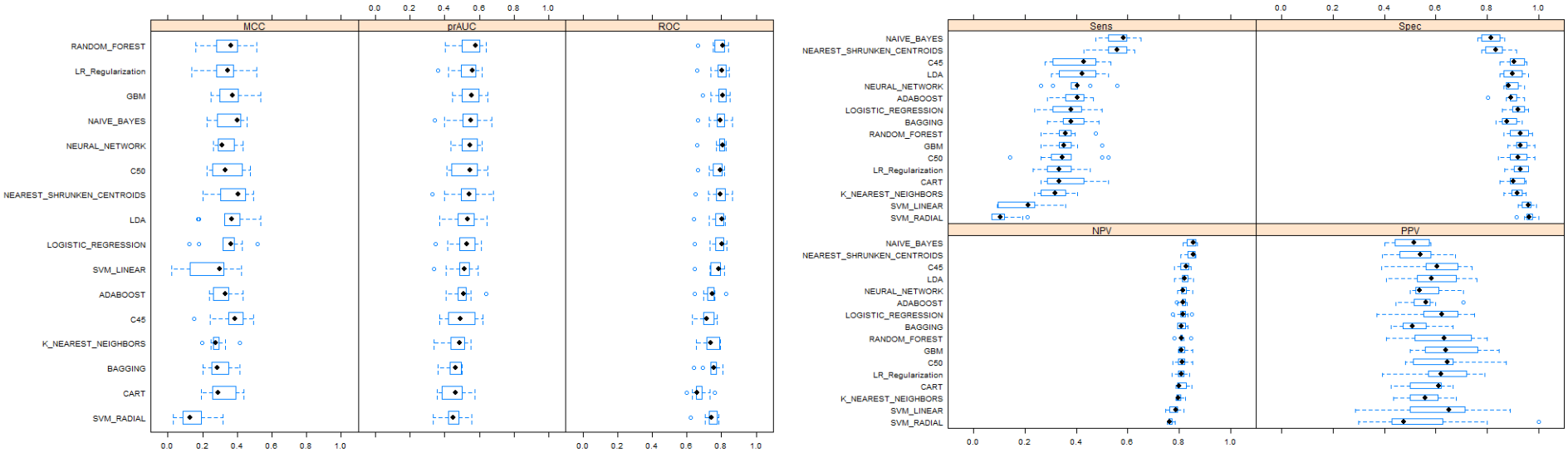


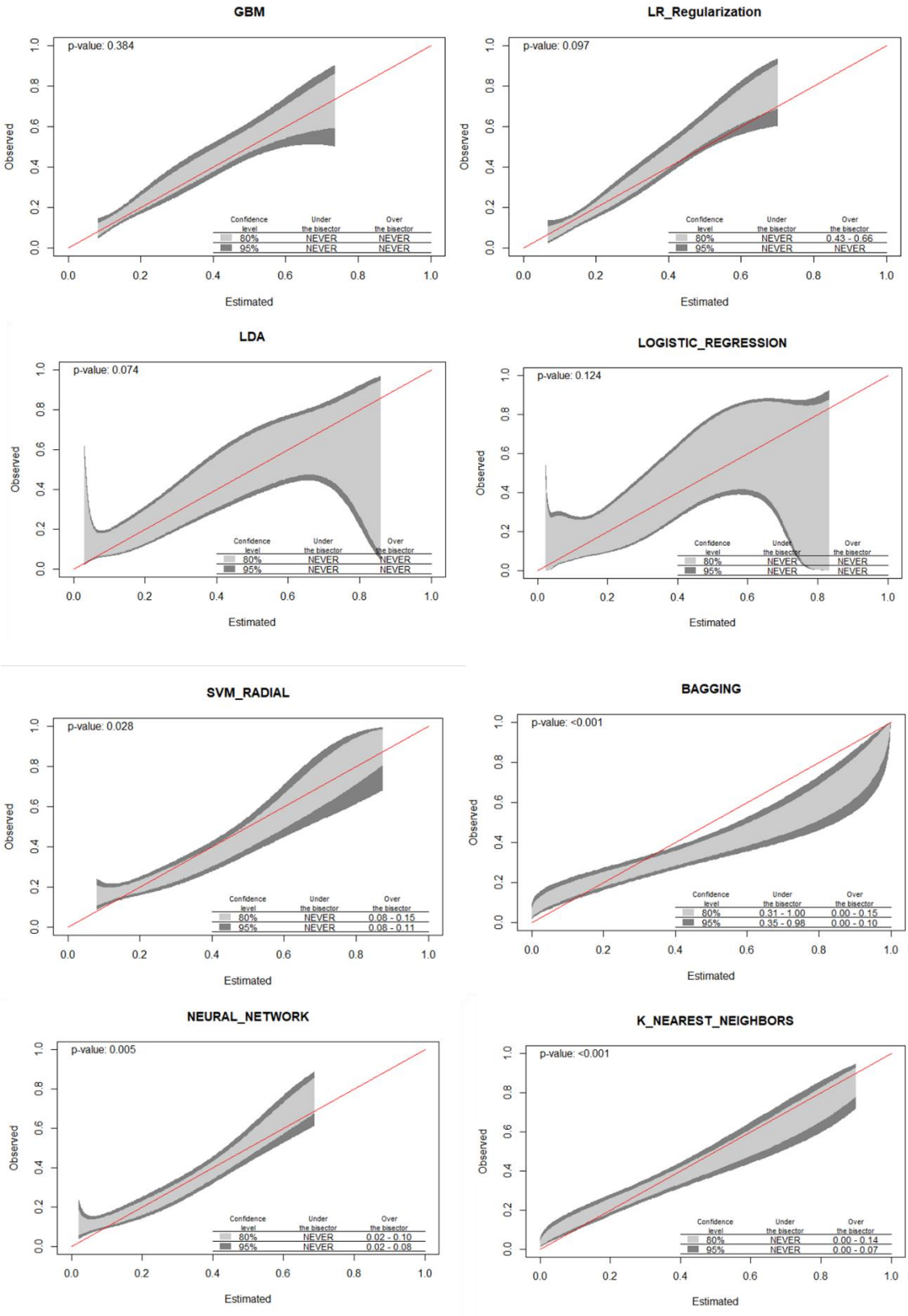
Figure O.1 - Cross-validation boxplot using the best hyperparameter.

Table O.2 - The best hyperparameters values for each method from 10-fold cross-validation.

Hyperparameters	LR parameter	LR_Regularization.alpha	LR_Regularization lambda	LDA parameter	NSC threshold	SVM_LINEAR C	NEURAL_NETWORK size	NEURAL_NETWORK decay	SVM_RADIAL sigma	SVM_RADIAL C	k	NAIVE_BAYES laplace	NAIVE_BAYES usekernel	NAIVE_BAYES adjust	C45 C	C45 M	CART cp
Best Tune	none	0.2	0.06	none	2	0.25	2	0.5	0.125	1	10	5	TRUE	5	0.01	10	0.01

Hyperparameters	C50 trials	C50 model	C50 winnow	RF mtry	GBM n.trees	GBM interaction.depth	GBM shrinkage	GBM n.minobsinnode	BAGGING mfinal	BAGGING maxdepth	BAGGING coeflearn	ADABOOST mfinal	AdaBoost maxdepth	AdaBoost coeflearn
Best Tune	30	tree	FALSE	14	270	5	0.01	10	30	12	Breiman	30	12	Breiman

Appendix P



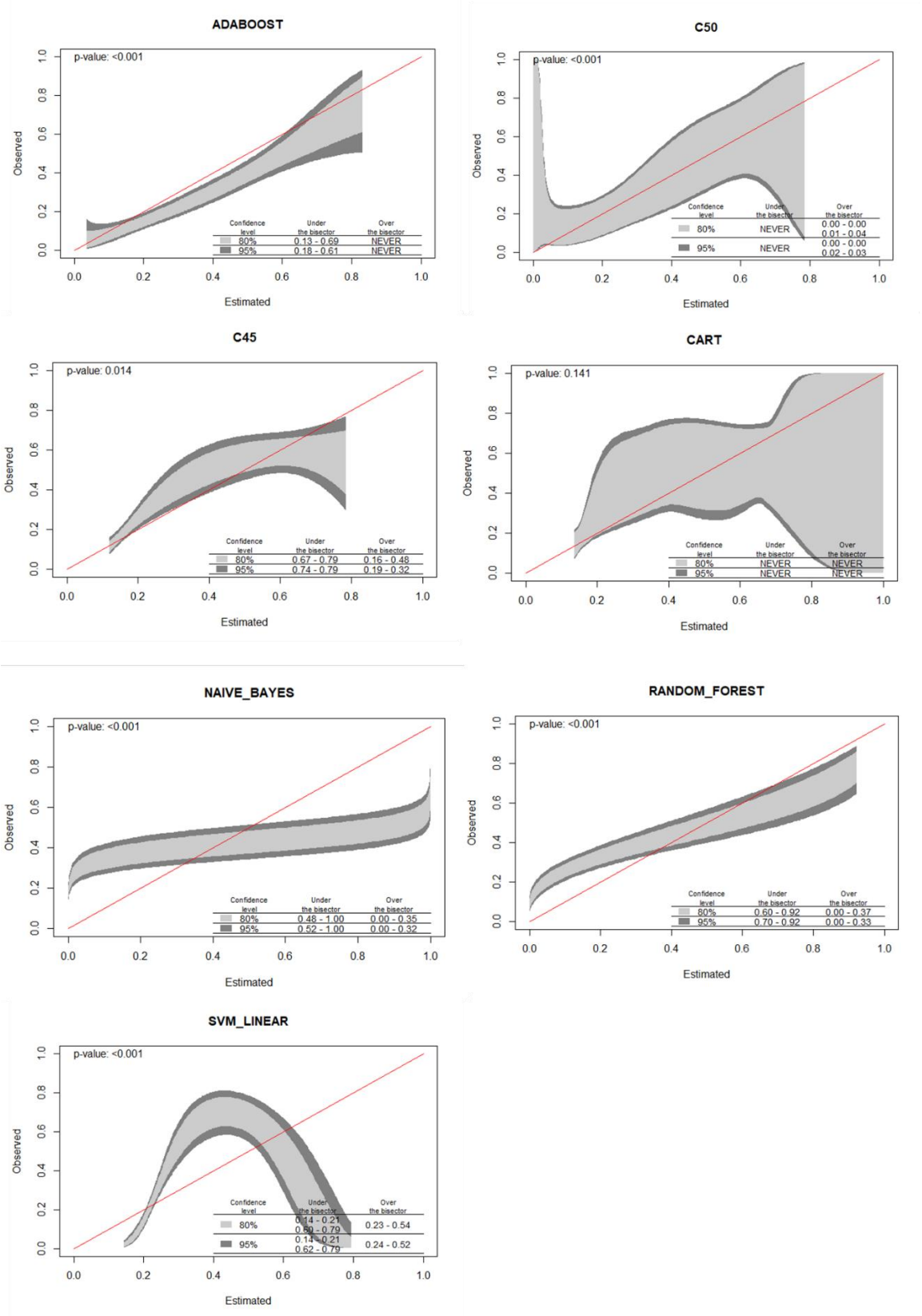


Figure P.1 - Calibration belts for the methods at two confidence levels. CI:0-80% (light shaded area) and CI:0-95% (dark shaded area).

Appendix Q

Table Q.1 - Clinical characteristics of patients considered for each hospital.

Variables	Hospital A (N=214)	Hospital B (N=272)	Hospital C (N=352)	Hospital D (N=932)	Hospital E (N=300)	TOTAL (N=2070)
RESULT						
Negative	151 (70.6%)	204 (75.0%)	264 (75.0%)	699 (75.0%)	225 (75.0%)	1543 (74.5%)
Positive	63 (29.4%)	68 (25.0%)	88 (25.0%)	233 (25.0%)	75 (25.0%)	527 (25.5%)
LOS_ICU_before_test						
Mean (SD)	8.67 (10.0)	11.8 (11.3)	11.1 (11.4)	12.3 (11.8)	9.76 (9.73)	11.3 (11.3)
Median [Min, Max]	5.00 [0, 59.0]	8.00 [0, 60.0]	7.00 [0, 54.0]	8.00 [0, 60.0]	6.00 [0, 56.0]	7.00 [0, 60.0]
LOS_hospital_before_test						
Mean (SD)	10.6 (11.3)	12.9 (11.9)	13.8 (12.5)	13.9 (12.2)	12.9 (10.7)	13.3 (12.0)
Median [Min, Max]	6.00 [3.00, 59.0]	9.00 [3.00, 60.0]	9.00 [3.00, 59.0]	10.0 [3.00, 60.0]	9.00 [3.00, 59.0]	9.00 [3.00, 60.0]
Age						
Mean (SD)	70.8 (15.6)	74.8 (15.5)	70.7 (17.7)	77.4 (14.5)	72.8 (17.1)	74.6 (15.9)
Median [Min, Max]	74.0 [20.0, 95.0]	80.0 [18.0, 98.0]	74.0 [20.0, 102]	80.0 [18.0, 105]	78.0 [19.0, 99.0]	78.0 [18.0, 105]
Gender						
F	136 (63.6%)	155 (57.0%)	170 (48.3%)	491 (52.7%)	161 (53.7%)	1113 (53.8%)
M	78 (36.4%)	117 (43.0%)	182 (51.7%)	441 (47.3%)	139 (46.3%)	957 (46.2%)
BMI						
Mean (SD)	27.6 (14.1)	NA (NA)	27.6 (15.4)	26.6 (13.7)	26.8 (9.37)	26.9 (13.5)
Median [Min, Max]	25.7 [15.6, 215]	NA [NA, NA]	26.0 [13.9, 283]	25.0 [14.1, 263]	25.4 [14.5, 139]	25.3 [13.9, 283]
Missing	4 (1.9%)	272 (100%)	30 (8.5%)	38 (4.1%)	34 (11.3%)	378 (18.3%)
AdmissionSource						
Emergency	142 (66.4%)	217 (79.8%)	179 (50.9%)	501 (53.8%)	141 (47.0%)	1180 (57.0%)
Hemodynamic Room	0 (0%)	1 (0.4%)	3 (0.9%)	16 (1.7%)	5 (1.7%)	25 (1.2%)
Operation Room	37 (17.3%)	9 (3.3%)	48 (13.6%)	105 (11.3%)	43 (14.3%)	242 (11.7%)
Other ICU from hospital	0 (0%)	10 (3.7%)	30 (8.5%)	154 (16.5%)	27 (9.0%)	221 (10.7%)
Others	0 (0%)	6 (2.2%)	6 (1.7%)	6 (0.6%)	5 (1.7%)	23 (1.1%)
Semi Intensive Unit	0 (0%)	0 (0%)	39 (11.1%)	55 (5.9%)	18 (6.0%)	112 (5.4%)
Transfer from another hospital	2 (0.9%)	1 (0.4%)	7 (2.0%)	9 (1.0%)	15 (5.0%)	34 (1.6%)
Ward/Room	25 (11.7%)	28 (10.3%)	39 (11.1%)	86 (9.2%)	46 (15.3%)	224 (10.8%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
AdmissionReason						
Cardiovascular / Shock	49 (22.9%)	76 (27.9%)	65 (18.5%)	238 (25.5%)	65 (21.7%)	493 (23.8%)
Elective Surgery	32 (15.0%)	8 (2.9%)	34 (9.7%)	72 (7.7%)	37 (12.3%)	183 (8.8%)
Emergency surgery	8 (3.7%)	5 (1.8%)	23 (6.5%)	43 (4.6%)	13 (4.3%)	92 (4.4%)
Endocrine / Metabolic / Renal	4 (1.9%)	16 (5.9%)	7 (2.0%)	24 (2.6%)	9 (3.0%)	60 (2.9%)
Infection / Sepsis	66 (30.8%)	84 (30.9%)	150 (42.6%)	319 (34.2%)	94 (31.3%)	713 (34.4%)
Liver and Pancreas / Gastrointestinal	18 (8.4%)	14 (5.1%)	18 (5.1%)	51 (5.5%)	8 (2.7%)	109 (5.3%)
Neurological	21 (9.8%)	24 (8.8%)	23 (6.5%)	69 (7.4%)	41 (13.7%)	178 (8.6%)
Non-surgical trauma	2 (0.9%)	13 (4.8%)	6 (1.7%)	16 (1.7%)	4 (1.3%)	41 (2.0%)
Oncological / Hematological	1 (0.5%)	5 (1.8%)	6 (1.7%)	25 (2.7%)	11 (3.7%)	48 (2.3%)
Others	3 (1.4%)	2 (0.7%)	7 (2.0%)	18 (1.9%)	3 (1.0%)	33 (1.6%)
Respiratory	2 (0.9%)	25 (9.2%)	12 (3.4%)	57 (6.1%)	15 (5.0%)	111 (5.4%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
CharlsonIndex						
Mean (SD)	1.51 (1.74)	1.33 (1.48)	1.26 (1.97)	2.08 (1.97)	2.04 (1.93)	1.78 (1.92)
Median [Min, Max]	1.00 [0, 10.0]	1.00 [0, 9.00]	0 [0, 11.0]	2.00 [0, 12.0]	2.00 [0, 10.0]	1.00 [0, 12.0]
Missing	0 (0%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	1 (0.0%)
MFIpoints						
Mean (SD)	2.06 (1.54)	1.96 (1.29)	2.01 (1.41)	2.39 (1.42)	2.18 (1.33)	2.20 (1.42)
Median [Min, Max]	2.00 [0, 7.00]	2.00 [0, 5.00]	2.00 [0, 6.00]	2.00 [0, 7.00]	2.00 [0, 6.00]	2.00 [0, 7.00]
Missing	0 (0%)	0 (0%)	15 (4.3%)	40 (4.3%)	13 (4.3%)	68 (3.3%)
FrailPatientMFI						
NO	176 (82.2%)	241 (88.6%)	295 (83.8%)	747 (80.2%)	252 (84.0%)	1711 (82.7%)
YES	38 (17.8%)	31 (11.4%)	57 (16.2%)	185 (19.8%)	48 (16.0%)	359 (17.3%)
Saps3Points						
Mean (SD)	50.0 (13.7)	47.9 (8.26)	50.1 (12.4)	55.2 (13.8)	56.7 (14.8)	53.1 (13.5)
Median [Min, Max]	51.0 [16.0, 87.0]	48.0 [21.0, 78.0]	50.0 [16.0, 96.0]	54.0 [19.0, 104]	55.5 [16.0, 101]	52.0 [16.0, 104]
SofaScore						
Mean (SD)	1.96 (2.99)	0.0691 (0.254)	0.727 (0.447)	2.97 (3.48)	0.757 (0.430)	1.97 (2.97)
Median [Min, Max]	1.00 [0, 12.0]	0 [0, 1.00]	1.00 [0, 1.00]	2.00 [0, 17.0]	1.00 [0, 1.00]	1.00 [0, 17.0]
Missing	8 (3.7%)	84 (30.9%)	231 (65.6%)	136 (14.6%)	70 (23.3%)	529 (25.6%)
Priority						
Priority 1	9 (4.2%)	0 (0%)	0 (0%)	249 (26.7%)	0 (0%)	258 (12.5%)
Priority 2	2 (0.9%)	0 (0%)	0 (0%)	643 (69.0%)	0 (0%)	645 (31.2%)
Priority 3	0 (0%)	0 (0%)	0 (0%)	1 (0.1%)	0 (0%)	1 (0.0%)
Priority 4	0 (0%)	0 (0%)	0 (0%)	3 (0.3%)	0 (0%)	3 (0.1%)
Priority 5	0 (0%)	0 (0%)	0 (0%)	13 (1.4%)	0 (0%)	13 (0.6%)
Missing	203 (94.9%)	272 (100%)	352 (100%)	23 (2.5%)	300 (100%)	1150 (55.6%)
ChronicHealthStatus						
Independent	132 (61.7%)	119 (43.8%)	234 (66.5%)	486 (52.1%)	167 (55.7%)	1138 (55.0%)
Need for assistance	57 (26.6%)	0 (0%)	76 (21.6%)	306 (32.8%)	46 (15.3%)	485 (23.4%)
Restricted / bedridden	17 (7.9%)	153 (56.2%)	41 (11.6%)	140 (15.0%)	87 (29.0%)	438 (21.2%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsChfNyhaClass23						
FALSE	187 (87.4%)	271 (99.6%)	320 (90.9%)	857 (92.0%)	287 (95.7%)	1922 (92.9%)
TRUE	19 (8.9%)	1 (0.4%)	31 (8.8%)	75 (8.0%)	13 (4.3%)	139 (6.7%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsChfNyhaClass4						
FALSE	203 (94.9%)	272 (100%)	349 (99.1%)	930 (99.8%)	299 (99.7%)	2053 (99.2%)
TRUE	3 (1.4%)	0 (0%)	2 (0.6%)	2 (0.2%)	1 (0.3%)	8 (0.4%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)

Variables	Hospital A (N=214)	Hospital B (N=272)	Hospital C (N=352)	Hospital D (N=932)	Hospital E (N=300)	TOTAL (N=2070)
IsCrfNoDialysis						
FALSE	190 (88.8%)	264 (97.1%)	319 (90.6%)	810 (86.9%)	268 (89.3%)	1851 (89.4%)
TRUE	16 (7.5%)	8 (2.9%)	32 (9.1%)	122 (13.1%)	32 (10.7%)	210 (10.1%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsCrfDialysis						
FALSE	201 (93.9%)	262 (96.3%)	347 (98.6%)	906 (97.2%)	291 (97.0%)	2007 (97.0%)
TRUE	5 (2.3%)	10 (3.7%)	4 (1.1%)	26 (2.8%)	9 (3.0%)	54 (2.6%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsCirrhosisChildAB						
FALSE	205 (95.8%)	272 (100%)	348 (98.9%)	926 (99.4%)	300 (100%)	2051 (99.1%)
TRUE	1 (0.5%)	0 (0%)	3 (0.9%)	6 (0.6%)	0 (0%)	10 (0.5%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsCirrhosisChildC						
FALSE	206 (96.3%)	272 (100%)	349 (99.1%)	929 (99.7%)	297 (99.0%)	2053 (99.2%)
TRUE	0 (0%)	0 (0%)	2 (0.6%)	3 (0.3%)	3 (1.0%)	8 (0.4%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsHepaticFailure						
FALSE	206 (96.3%)	272 (100%)	351 (99.7%)	931 (99.9%)	291 (97.0%)	2051 (99.1%)
TRUE	0 (0%)	0 (0%)	0 (0%)	1 (0.1%)	9 (3.0%)	10 (0.5%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsSolidTumorLocoregiol						
FALSE	202 (94.4%)	234 (86.0%)	292 (83.0%)	707 (75.9%)	238 (79.3%)	1673 (80.8%)
TRUE	4 (1.9%)	38 (14.0%)	59 (16.8%)	225 (24.1%)	62 (20.7%)	388 (18.7%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsSolidTumorMetastatic						
FALSE	202 (94.4%)	267 (98.2%)	332 (94.3%)	908 (97.4%)	286 (95.3%)	1995 (96.4%)
TRUE	4 (1.9%)	5 (1.8%)	19 (5.4%)	24 (2.6%)	14 (4.7%)	66 (3.2%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsHematologicalMalignancy						
FALSE	204 (95.3%)	272 (100%)	344 (97.7%)	912 (97.9%)	288 (96.0%)	2020 (97.6%)
TRUE	2 (0.9%)	0 (0%)	7 (2.0%)	20 (2.1%)	12 (4.0%)	41 (2.0%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsImmunossupression						
FALSE	203 (94.9%)	269 (98.9%)	321 (91.2%)	825 (88.5%)	243 (81.0%)	1861 (89.9%)
TRUE	3 (1.4%)	3 (1.1%)	30 (8.5%)	107 (11.5%)	57 (19.0%)	200 (9.7%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsSevereCopd						
FALSE	196 (91.6%)	254 (93.4%)	321 (91.2%)	814 (87.3%)	274 (91.3%)	1859 (89.8%)
TRUE	10 (4.7%)	18 (6.6%)	30 (8.5%)	118 (12.7%)	26 (8.7%)	202 (9.8%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsSteroidsUse						
FALSE	205 (95.8%)	272 (100%)	335 (95.2%)	916 (98.3%)	297 (99.0%)	2025 (97.8%)
TRUE	1 (0.5%)	0 (0%)	16 (4.5%)	16 (1.7%)	3 (1.0%)	36 (1.7%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsAids						
FALSE	205 (95.8%)	272 (100%)	349 (99.1%)	929 (99.7%)	298 (99.3%)	2053 (99.2%)
TRUE	1 (0.5%)	0 (0%)	2 (0.6%)	3 (0.3%)	2 (0.7%)	8 (0.4%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsArterialHypertension						
FALSE	54 (25.2%)	113 (41.5%)	138 (39.2%)	277 (29.7%)	78 (26.0%)	660 (31.9%)
TRUE	152 (71.0%)	159 (58.5%)	213 (60.5%)	655 (70.3%)	222 (74.0%)	1401 (67.7%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsAsthma						
FALSE	200 (93.5%)	272 (100%)	334 (94.9%)	899 (96.5%)	287 (95.7%)	1992 (96.2%)
TRUE	6 (2.8%)	0 (0%)	17 (4.8%)	33 (3.5%)	13 (4.3%)	69 (3.3%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsDiabetesUncomplicated						
FALSE	160 (74.8%)	192 (70.6%)	260 (73.9%)	699 (75.0%)	200 (66.7%)	1511 (73.0%)
TRUE	46 (21.5%)	80 (29.4%)	91 (25.9%)	233 (25.0%)	100 (33.3%)	550 (26.6%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsDiabetesComplicated						
FALSE	182 (85.0%)	269 (98.9%)	340 (96.6%)	861 (92.4%)	299 (99.7%)	1951 (94.3%)
TRUE	24 (11.2%)	3 (1.1%)	11 (3.1%)	71 (7.6%)	1 (0.3%)	110 (5.3%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsAngina						
FALSE	197 (92.1%)	271 (99.6%)	330 (93.8%)	854 (91.6%)	299 (99.7%)	1951 (94.3%)
TRUE	9 (4.2%)	1 (0.4%)	21 (6.0%)	78 (8.4%)	1 (0.3%)	110 (5.3%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsPreviousMI						
FALSE	189 (88.3%)	252 (92.6%)	310 (88.1%)	803 (86.2%)	266 (88.7%)	1820 (87.9%)
TRUE	17 (7.9%)	20 (7.4%)	41 (11.6%)	129 (13.8%)	34 (11.3%)	241 (11.6%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsCardiacArrhythmia						
FALSE	206 (96.3%)	263 (96.7%)	319 (90.6%)	809 (86.8%)	260 (86.7%)	1857 (89.7%)
TRUE	0 (0%)	9 (3.3%)	32 (9.1%)	123 (13.2%)	40 (13.3%)	204 (9.9%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsDeepVenousThrombosis						
FALSE	200 (93.5%)	272 (100%)	318 (90.3%)	887 (95.2%)	283 (94.3%)	1960 (94.7%)
TRUE	6 (2.8%)	0 (0%)	33 (9.4%)	45 (4.8%)	17 (5.7%)	101 (4.9%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsPeripheralArteryDisease						
FALSE	198 (92.5%)	264 (97.1%)	346 (98.3%)	902 (96.8%)	298 (99.3%)	2008 (97.0%)
TRUE	8 (3.7%)	8 (2.9%)	5 (1.4%)	30 (3.2%)	2 (0.7%)	53 (2.6%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsChronicAtrialFibrilation						
FALSE	191 (89.3%)	250 (91.9%)	298 (84.7%)	775 (83.2%)	262 (87.3%)	1776 (85.8%)
TRUE	15 (7.0%)	22 (8.1%)	53 (15.1%)	157 (16.8%)	38 (12.7%)	285 (13.8%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsRheumaticDisease						
FALSE	204 (95.3%)	272 (100%)	345 (98.0%)	930 (99.8%)	299 (99.7%)	2050 (99.0%)
TRUE	2 (0.9%)	0 (0%)	6 (1.7%)	2 (0.2%)	1 (0.3%)	11 (0.5%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)

Variables	Hospital A (N=214)	Hospital B (N=272)	Hospital C (N=352)	Hospital D (N=932)	Hospital E (N=300)	TOTAL (N=2070)
IsStrokeSequelae						
FALSE	186 (86.9%)	268 (98.5%)	338 (96.0%)	896 (96.1%)	286 (95.3%)	1974 (95.4%)
TRUE	20 (9.3%)	4 (1.5%)	13 (3.7%)	36 (3.9%)	14 (4.7%)	87 (4.2%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsStrokeNoSequelae						
FALSE	200 (93.5%)	253 (93.0%)	327 (92.9%)	861 (92.4%)	275 (91.7%)	1916 (92.6%)
TRUE	6 (2.8%)	19 (7.0%)	24 (6.8%)	71 (7.6%)	25 (8.3%)	145 (7.0%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsDementia						
FALSE	169 (79.0%)	209 (76.8%)	288 (81.8%)	782 (83.9%)	255 (85.0%)	1703 (82.3%)
TRUE	37 (17.3%)	63 (23.2%)	63 (17.9%)	150 (16.1%)	45 (15.0%)	358 (17.3%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsTobaccoConsumption						
FALSE	190 (88.8%)	255 (93.8%)	323 (91.8%)	850 (91.2%)	283 (94.3%)	1901 (91.8%)
TRUE	16 (7.5%)	17 (6.2%)	28 (8.0%)	82 (8.8%)	17 (5.7%)	160 (7.7%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsAlcoholism						
FALSE	198 (92.5%)	265 (97.4%)	338 (96.0%)	890 (95.5%)	297 (99.0%)	1988 (96.0%)
TRUE	8 (3.7%)	7 (2.6%)	13 (3.7%)	42 (4.5%)	3 (1.0%)	73 (3.5%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsPsychiatricDisease						
FALSE	198 (92.5%)	264 (97.1%)	322 (91.5%)	851 (91.3%)	290 (96.7%)	1925 (93.0%)
TRUE	8 (3.7%)	8 (2.9%)	29 (8.2%)	81 (8.7%)	10 (3.3%)	136 (6.6%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsMorbidObesity						
FALSE	202 (94.4%)	271 (99.6%)	334 (94.9%)	902 (96.8%)	284 (94.7%)	1993 (96.3%)
TRUE	4 (1.9%)	1 (0.4%)	17 (4.8%)	30 (3.2%)	16 (5.3%)	68 (3.3%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsMalnourishment						
FALSE	201 (93.9%)	272 (100%)	351 (99.7%)	931 (99.9%)	300 (100%)	2055 (99.3%)
TRUE	5 (2.3%)	0 (0%)	0 (0%)	1 (0.1%)	0 (0%)	6 (0.3%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsPepticDisease						
FALSE	203 (94.9%)	272 (100%)	349 (99.1%)	932 (100%)	300 (100%)	2056 (99.3%)
TRUE	3 (1.4%)	0 (0%)	2 (0.6%)	0 (0%)	0 (0%)	5 (0.2%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
Transplant						
FALSE	204 (95.3%)	260 (95.6%)	314 (89.2%)	818 (87.8%)	260 (86.7%)	1856 (89.7%)
TRUE	2 (0.9%)	12 (4.4%)	37 (10.5%)	114 (12.2%)	40 (13.3%)	205 (9.9%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsHypothyroidism						
FALSE	189 (88.3%)	234 (86.0%)	300 (85.2%)	727 (78.0%)	255 (85.0%)	1705 (82.4%)
TRUE	17 (7.9%)	38 (14.0%)	51 (14.5%)	205 (22.0%)	45 (15.0%)	356 (17.2%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsHyperthyroidism						
FALSE	205 (95.8%)	272 (100%)	351 (99.7%)	928 (99.6%)	298 (99.3%)	2054 (99.2%)
TRUE	1 (0.5%)	0 (0%)	0 (0%)	4 (0.4%)	2 (0.7%)	7 (0.3%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsDyslipidemias						
FALSE	190 (88.8%)	248 (91.2%)	291 (82.7%)	772 (82.8%)	239 (79.7%)	1740 (84.1%)
TRUE	16 (7.5%)	24 (8.8%)	60 (17.0%)	160 (17.2%)	61 (20.3%)	321 (15.5%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsChemotherapy						
FALSE	205 (95.8%)	271 (99.6%)	341 (96.9%)	856 (91.8%)	280 (93.3%)	1953 (94.3%)
TRUE	1 (0.5%)	1 (0.4%)	10 (2.8%)	76 (8.2%)	20 (6.7%)	108 (5.2%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsRadiationTherapy						
FALSE	205 (95.8%)	271 (99.6%)	348 (98.9%)	884 (94.8%)	289 (96.3%)	1997 (96.5%)
TRUE	1 (0.5%)	1 (0.4%)	3 (0.9%)	48 (5.2%)	11 (3.7%)	64 (3.1%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
IsHistoryOfPneumonia						
FALSE	190 (88.8%)	269 (98.9%)	327 (92.9%)	880 (94.4%)	290 (96.7%)	1956 (94.5%)
TRUE	16 (7.5%)	3 (1.1%)	24 (6.8%)	52 (5.6%)	10 (3.3%)	105 (5.1%)
Missing	8 (3.7%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	9 (0.4%)
VesDURTOTAL						
Mean (SD)	3.19 (5.15)	4.37 (7.74)	5.97 (8.81)	7.30 (8.90)	6.01 (7.49)	6.08 (8.33)
Median [Min, Max]	1.00 [0, 32.0]	1.00 [0, 44.0]	3.00 [0, 58.0]	5.00 [0, 55.0]	3.00 [0, 45.0]	3.00 [0, 58.0]
VesTIMESTOTAL						
Mean (SD)	0.631 (0.698)	0.695 (0.896)	0.761 (0.837)	0.981 (0.853)	0.843 (0.784)	0.850 (0.841)
Median [Min, Max]	1.00 [0, 3.00]	1.00 [0, 7.00]	1.00 [0, 5.00]	1.00 [0, 6.00]	1.00 [0, 4.00]	1.00 [0, 7.00]
VESICAL						
NO	103 (48.1%)	133 (48.9%)	150 (42.6%)	269 (28.9%)	100 (33.3%)	755 (36.5%)
YES	111 (51.9%)	139 (51.1%)	202 (57.4%)	663 (71.1%)	200 (66.7%)	1315 (63.5%)
ArtDURTOTAL						
Mean (SD)	2.33 (5.16)	1.85 (4.93)	3.41 (6.70)	4.53 (7.36)	3.89 (5.83)	3.67 (6.62)
Median [Min, Max]	0 [0, 33.0]	0 [0, 32.0]	0 [0, 43.0]	0 [0, 59.0]	0 [0, 27.0]	0 [0, 59.0]
ArtTIMESTOTAL						
Mean (SD)	0.369 (0.642)	0.309 (0.631)	0.486 (0.777)	0.735 (0.952)	0.623 (0.802)	0.583 (0.851)
Median [Min, Max]	0 [0, 4.00]	0 [0, 3.00]	0 [0, 4.00]	0 [0, 6.00]	0 [0, 4.00]	0 [0, 6.00]
ARTERIAL						
NO	150 (70.1%)	210 (77.2%)	233 (66.2%)	487 (52.3%)	160 (53.3%)	1240 (59.9%)
YES	64 (29.9%)	62 (22.8%)	119 (33.8%)	445 (47.7%)	140 (46.7%)	830 (40.1%)
DiaDURTOTAL						
Mean (SD)	0.421 (3.34)	0.838 (3.66)	1.33 (4.51)	1.28 (4.80)	1.27 (4.09)	1.14 (4.39)
Median [Min, Max]	0 [0, 36.0]	0 [0, 29.0]	0 [0, 33.0]	0 [0, 42.0]	0 [0, 30.0]	0 [0, 42.0]
DiaTIMESTOTAL						
Mean (SD)	0.0467 (0.318)	0.125 (0.493)	0.176 (0.520)	0.180 (0.599)	0.193 (0.563)	0.160 (0.545)
Median [Min, Max]	0 [0, 3.00]	0 [0, 4.00]	0 [0, 3.00]	0 [0, 6.00]	0 [0, 3.00]	0 [0, 6.00]
DIALYSIS						
NO	208 (97.2%)	251 (92.3%)	309 (87.8%)	829 (88.9%)	261 (87.0%)	1858 (89.8%)
YES	6 (2.8%)	21 (7.7%)	43 (12.2%)	103 (11.1%)	39 (13.0%)	212 (10.2%)
CVCDURTOTAL						

Variables	Hospital A (N=214)	Hospital B (N=272)	Hospital C (N=352)	Hospital D (N=932)	Hospital E (N=300)	TOTAL (N=2070)
Mean (SD)	3.51 (6.47)	3.95 (7.19)	5.28 (7.89)	7.25 (8.98)	7.03 (8.38)	6.06 (8.38)
Median [Min, Max]	0 [0, 44.0]	0 [0, 40.0]	0 [0, 45.0]	5.00 [0, 60.0]	5.00 [0, 42.0]	3.00 [0, 60.0]
CVCTIMESTOTAL						
Mean (SD)	0.486 (0.690)	0.540 (0.836)	0.682 (0.887)	1.01 (1.02)	0.963 (0.972)	0.831 (0.958)
Median [Min, Max]	0 [0, 4.00]	0 [0, 4.00]	0 [0, 4.00]	1.00 [0, 6.00]	1.00 [0, 5.00]	1.00 [0, 6.00]
CVC						
NO	127 (59.3%)	170 (62.5%)	190 (54.0%)	338 (36.3%)	109 (36.3%)	934 (45.1%)
YES	87 (40.7%)	102 (37.5%)	162 (46.0%)	594 (63.7%)	191 (63.7%)	1136 (54.9%)
MVDURTOTAL						
Mean (SD)	1.75 (5.86)	2.30 (6.77)	3.00 (7.08)	4.52 (8.70)	4.51 (7.80)	3.68 (7.88)
Median [Min, Max]	0 [0, 47.0]	0 [0, 54.0]	0 [0, 45.0]	0 [0, 57.0]	0 [0, 50.0]	0 [0, 57.0]
MVTIMESTOTAL						
Mean (SD)	0.192 (0.418)	0.224 (0.506)	0.301 (0.555)	0.472 (0.705)	0.487 (0.636)	0.384 (0.632)
Median [Min, Max]	0 [0, 2.00]	0 [0, 3.00]	0 [0, 3.00]	0 [0, 5.00]	0 [0, 4.00]	0 [0, 5.00]
MV						
NO	175 (81.8%)	221 (81.2%)	262 (74.4%)	580 (62.2%)	172 (57.3%)	1410 (68.1%)
YES	39 (18.2%)	51 (18.8%)	90 (25.6%)	352 (37.8%)	128 (42.7%)	660 (31.9%)
PerDURTOTAL						
Mean (SD)	2.75 (2.75)	0 (0)	0.0142 (0.192)	2.21 (4.17)	0 (0)	1.28 (3.15)
Median [Min, Max]	3.00 [0, 19.0]	0 [0, 0]	0 [0, 3.00]	0 [0, 56.0]	0 [0, 0]	0 [0, 56.0]
PerTIMESTOTAL						
Mean (SD)	1.14 (1.03)	0 (0)	0.00568 (0.0753)	0.822 (1.40)	0 (0)	0.489 (1.09)
Median [Min, Max]	1.00 [0, 5.00]	0 [0, 0]	0 [0, 1.00]	0 [0, 15.0]	0 [0, 0]	0 [0, 15.0]
PERIPHERAL						
NO	62 (29.0%)	272 (100%)	350 (99.4%)	541 (58.0%)	300 (100%)	1525 (73.7%)
YES	152 (71.0%)	0 (0%)	2 (0.6%)	391 (42.0%)	0 (0%)	545 (26.3%)
IsNeurologicalComaStuporObtundedDelirium						
FALSE	165 (77.1%)	259 (95.2%)	322 (91.5%)	709 (76.1%)	188 (62.7%)	1643 (79.4%)
TRUE	41 (19.2%)	13 (4.8%)	28 (8.0%)	222 (23.8%)	112 (37.3%)	416 (20.1%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsNeurologicalSeizures						
FALSE	201 (93.9%)	269 (98.9%)	341 (96.9%)	882 (94.6%)	294 (98.0%)	1987 (96.0%)
TRUE	5 (2.3%)	3 (1.1%)	9 (2.6%)	49 (5.3%)	6 (2.0%)	72 (3.5%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsNeurologicalFocalNeurologicDeficit						
FALSE	198 (92.5%)	272 (100%)	344 (97.7%)	915 (98.2%)	275 (91.7%)	2004 (96.8%)
TRUE	8 (3.7%)	0 (0%)	6 (1.7%)	16 (1.7%)	25 (8.3%)	55 (2.7%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsNeurologicalIntracranialMassEffect						
FALSE	200 (93.5%)	272 (100%)	344 (97.7%)	929 (99.7%)	300 (100%)	2045 (98.8%)
TRUE	6 (2.8%)	0 (0%)	6 (1.7%)	2 (0.2%)	0 (0%)	14 (0.7%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsCardiovascularHypovolemicHemorrhagicShock						
FALSE	185 (86.4%)	272 (100%)	347 (98.6%)	928 (99.6%)	290 (96.7%)	2022 (97.7%)
TRUE	21 (9.8%)	0 (0%)	3 (0.9%)	3 (0.3%)	10 (3.3%)	37 (1.8%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsCardiovascularSepticShock						
FALSE	160 (74.8%)	271 (99.6%)	337 (95.7%)	867 (93.0%)	249 (83.0%)	1884 (91.0%)
TRUE	46 (21.5%)	1 (0.4%)	13 (3.7%)	64 (6.9%)	51 (17.0%)	175 (8.5%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsCardiovascularRhythmDisturbances						
FALSE	199 (93.0%)	266 (97.8%)	326 (92.6%)	748 (80.3%)	285 (95.0%)	1824 (88.1%)
TRUE	7 (3.3%)	6 (2.2%)	24 (6.8%)	183 (19.6%)	15 (5.0%)	235 (11.4%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsCardiovascularAphylacticMixedUndefinedShock						
FALSE	204 (95.3%)	272 (100%)	350 (99.4%)	928 (99.6%)	300 (100%)	2054 (99.2%)
TRUE	2 (0.9%)	0 (0%)	0 (0%)	3 (0.3%)	0 (0%)	5 (0.2%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsDigestiveAcuteAbdomen						
FALSE	189 (88.3%)	272 (100%)	340 (96.6%)	912 (97.9%)	276 (92.0%)	1989 (96.1%)
TRUE	17 (7.9%)	0 (0%)	10 (2.8%)	19 (2.0%)	24 (8.0%)	70 (3.4%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsDigestiveSeverePancreatitis						
FALSE	204 (95.3%)	272 (100%)	349 (99.1%)	928 (99.6%)	300 (100%)	2053 (99.2%)
TRUE	2 (0.9%)	0 (0%)	1 (0.3%)	3 (0.3%)	0 (0%)	6 (0.3%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsLiverFailure						
FALSE	205 (95.8%)	272 (100%)	346 (98.3%)	929 (99.7%)	297 (99.0%)	2049 (99.0%)
TRUE	1 (0.5%)	0 (0%)	4 (1.1%)	2 (0.2%)	3 (1.0%)	10 (0.5%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsTransplantSolidOrgan						
FALSE	206 (96.3%)	272 (100%)	350 (99.4%)	928 (99.6%)	300 (100%)	2056 (99.3%)
TRUE	0 (0%)	0 (0%)	0 (0%)	3 (0.3%)	0 (0%)	3 (0.1%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsTraumaMultipleTrauma						
FALSE	205 (95.8%)	272 (100%)	339 (96.3%)	905 (97.1%)	299 (99.7%)	2020 (97.6%)
TRUE	1 (0.5%)	0 (0%)	11 (3.1%)	26 (2.8%)	1 (0.3%)	39 (1.9%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsCardiacSurgery						
FALSE	206 (96.3%)	272 (100%)	348 (98.9%)	918 (98.5%)	298 (99.3%)	2042 (98.6%)
TRUE	0 (0%)	0 (0%)	2 (0.6%)	13 (1.4%)	2 (0.7%)	17 (0.8%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
IsNeurosurgery						
FALSE	206 (96.3%)	272 (100%)	350 (99.4%)	929 (99.7%)	299 (99.7%)	2056 (99.3%)
TRUE	0 (0%)	0 (0%)	0 (0%)	2 (0.2%)	1 (0.3%)	3 (0.1%)
Missing	8 (3.7%)	0 (0%)	2 (0.6%)	1 (0.1%)	0 (0%)	11 (0.5%)
J01A						
FALSE	209 (97.7%)	266 (97.8%)	320 (90.9%)	907 (97.3%)	235 (78.3%)	1937 (93.6%)
TRUE	5 (2.3%)	6 (2.2%)	32 (9.1%)	25 (2.7%)	65 (21.7%)	133 (6.4%)

Variables	Hospital A (N=214)	Hospital B (N=272)	Hospital C (N=352)	Hospital D (N=932)	Hospital E (N=300)	TOTAL (N=2070)
J01C						
FALSE	121 (56.5%)	158 (58.1%)	120 (34.1%)	349 (37.4%)	109 (36.3%)	857 (41.4%)
TRUE	93 (43.5%)	114 (41.9%)	232 (65.9%)	583 (62.6%)	191 (63.7%)	1213 (58.6%)
J01D						
FALSE	135 (63.1%)	115 (42.3%)	158 (44.9%)	420 (45.1%)	131 (43.7%)	959 (46.3%)
TRUE	79 (36.9%)	157 (57.7%)	194 (55.1%)	512 (54.9%)	169 (56.3%)	1111 (53.7%)
J01E						
FALSE	209 (97.7%)	267 (98.2%)	334 (94.9%)	894 (95.9%)	293 (97.7%)	1997 (96.5%)
TRUE	5 (2.3%)	5 (1.8%)	18 (5.1%)	38 (4.1%)	7 (2.3%)	73 (3.5%)
J01F						
FALSE	166 (77.6%)	205 (75.4%)	230 (65.3%)	601 (64.5%)	182 (60.7%)	1384 (66.9%)
TRUE	48 (22.4%)	67 (24.6%)	122 (34.7%)	331 (35.5%)	118 (39.3%)	686 (33.1%)
J01G						
FALSE	199 (93.0%)	253 (93.0%)	327 (92.9%)	894 (95.9%)	280 (93.3%)	1953 (94.3%)
TRUE	15 (7.0%)	19 (7.0%)	25 (7.1%)	38 (4.1%)	20 (6.7%)	117 (5.7%)
J01M						
FALSE	160 (74.8%)	251 (92.3%)	301 (85.5%)	773 (82.9%)	270 (90.0%)	1755 (84.8%)
TRUE	54 (25.2%)	21 (7.7%)	51 (14.5%)	159 (17.1%)	30 (10.0%)	315 (15.2%)
J01X						
FALSE	154 (72.0%)	203 (74.6%)	246 (69.9%)	587 (63.0%)	211 (70.3%)	1401 (67.7%)
TRUE	60 (28.0%)	69 (25.4%)	106 (30.1%)	345 (37.0%)	89 (29.7%)	669 (32.3%)
J04A						
FALSE	213 (99.5%)	272 (100%)	349 (99.1%)	932 (100%)	300 (100%)	2066 (99.8%)
TRUE	1 (0.5%)	0 (0%)	3 (0.9%)	0 (0%)	0 (0%)	4 (0.2%)
Antibiotic						
FALSE	54 (25.2%)	68 (25.0%)	41 (11.6%)	128 (13.7%)	36 (12.0%)	327 (15.8%)
TRUE	160 (74.8%)	204 (75.0%)	311 (88.4%)	804 (86.3%)	264 (88.0%)	1743 (84.2%)

Appendix R

Table R.1 - Importance attributes after feature selection by Information Gain.

HOSPITAL A	
Attributes	**Importance**
MVDURTOTAL	0.088
Saps3Points	0.079
MV	0.072
MVTIMESTOTAL	0.072
Antibiotic	0.068
J01D	0.061
ChronicHealthStatus	0.041
IsDiabetesUncomplicated	0.019
AdmissionSource	0.018
IsMalnourishment	0.013
J01E	0.010

HOSPITAL B	
Attributes	**Importance**
MVDURTOTAL	0.071
VesDURTOTAL	0.070
MV	0.064
MVTIMESTOTAL	0.064
J01X	0.063
ArtTIMESTOTAL	0.050
J01D	0.048
VesTIMESTOTAL	0.046
VESICAL	0.046
ArtDURTOTAL	0.046
CVC	0.042
CVCTIMESTOTAL	0.042
DiaTIMESTOTAL	0.040
DIALYSIS	0.040
J01C	0.023
IsCrfNoDialysis	0.021
ChronicHealthStatus	0.018
AdmissionSource	0.018
J01E	0.016
J01G	0.012
Transplant	0.004
IsSolidTumorLocoregiol	0.001

HOSPITAL C	
Attributes	**Importance**
MVDURTOTAL	0.080
CVCTIMESTOTAL	0.069
CVC	0.069
MV	0.064
MVTIMESTOTAL	0.064
AdmissionSource	0.060
LOS_ICU_before_test	0.054
VESICAL	0.052
VesTIMESTOTAL	0.052
J01D	0.038
DiaDURTOTAL	0.037
DIALYSIS	0.036
Saps3Points	0.035
AdmissionReason	0.031
J01X	0.026
ChronicHealthStatus	0.026
IsCardiovascularSepticShock	0.025
J01G	0.020
J01A	0.019
Antibiotic	0.018
IsSevereCopd	0.012
IsAsthma	0.011
J01F	0.010
IsHypothyroidism	0.009
J01C	0.008
IsNeurologicalSeizures	0.007
IsPreviousMI	0.006
IsDementia	0.005
IsStrokeSequelae	0.005
IsSolidTumorMetastatic	0.005
J01E	0.004
J01M	0.002
IsSteroidsUse	0.002
IsAlcoholism	0.001
IsPsychiatricDisease	0.001

HOSPITAL E	
Attributes	**Importance**
MVTIMESTOTAL	0.058
MVDURTOTAL	0.058
MV	0.058
LOS_ICU_before_test	0.056
CVCTIMESTOTAL	0.047
J01G	0.032
AdmissionReason	0.026
IsNeurologicalFocalNeurologicDeficit	0.012
ChronicHealthStatus	0.010
IsDyslipidemias	0.009
IsNeurologicalSeizures	0.001

HOSPITAL D	
Attributes	**Importance**
MVDURTOTAL	0.070
MVTIMESTOTAL	0.064
MV	0.064
VesDURTOTAL	0.063
ArtTIMESTOTAL	0.049
CVCTIMESTOTAL	0.043
CVC	0.043
VesTIMESTOTAL	0.041
VESICAL	0.041
AdmissionReason	0.028
Antibiotic	0.026
J01D	0.025
J01X	0.023
LOS_hospital_before_test	0.022
Saps3Points	0.019
J01C	0.015
AdmissionSource	0.015
DiaTIMESTOTAL	0.014
DIALYSIS	0.014
IsNeurologicalComaStuporObtundedDelirium	0.008
IsNeurologicalSeizures	0.007
ChronicHealthStatus	0.006
IsHistoryOfPneumonia	0.005
J01E	0.005
IsAngina	0.004
IsStrokeSequelae	0.004
IsDeepVenousThrombosis	0.004
J01F	0.004
IsDementia	0.003
IsCardiovascularSepticShock	0.003
J01A	0.003
IsChronicAtrialFibrilation	0.003
PERIPHERAL	0.002
IsStrokeNoSequelae	0.002
IsCrfDialysis	0.002
IsSevereCopd	0.001
FrailPatientMFI	0.001
IsCrfNoDialysis	0.001
IsHypothyroidism	0.001
IsArterialHypertension	0.001
IsDiabetesUncomplicated	0.001
IsAlcoholism	0.001

ALL HOSPITALS	
Attributes	**Importance**
MVDURTOTAL	0.066
MV	0.060
MVTIMESTOTAL	0.060
VesDURTOTAL	0.053
CVCTIMESTOTAL	0.052
ArtTIMESTOTAL	0.043
CVC	0.041
LOS_hospital_before_test	0.039
VESICAL	0.037
VesTIMESTOTAL	0.037
J01D	0.032
J01X	0.029
Antibiotic	0.027
Saps3Points	0.018
DiaTIMESTOTAL	0.017
DIALYSIS	0.017
J01C	0.014
AdmissionSource	0.013
AdmissionReason	0.012
ChronicHealthStatus	0.012
J01G	0.009
J01A	0.008
J01F	0.006
IsNeurologicalComaStuporObtundedDelirium	0.006
IsStrokeSequelae	0.005
IsDeepVenousThrombosis	0.005
IsNeurologicalFocalNeurologicDeficit	0.005
IsCardiovascularSepticShock	0.004
IsHistoryOfPneumonia	0.004
FrailPatientMFI	0.003
IsNeurologicalSeizures	0.002
J01E	0.002
IsDementia	0.002
IsAngina	0.001
PERIPHERAL	0.001
IsSevereCopd	0.001
IsDigestiveAcuteAbdomen	0.001
IsCrfDialysis	0.001
IsAsthma	0.001
IsChronicAtrialFibrilation	0.001

Appendix S

Table S1 - List of the rules generated from the association rule mining and their respective measures (support, confidence, and lift).

#	Rules	Support	Confidence	Lift
1	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
2	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.100	0.575	2.257
3	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
4	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01D=TRUE} => {RESULT=pos}	0.100	0.575	2.257
5	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
6	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.100	0.575	2.257
7	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
8	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01D=TRUE} => {RESULT=pos}	0.100	0.575	2.257
9	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
10	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.100	0.575	2.257
11	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
12	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE} => {RESULT=pos}	0.100	0.575	2.257
13	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
14	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.100	0.575	2.257
15	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.100	0.575	2.257
16	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE} => {RESULT=pos}	0.100	0.575	2.257
17	{MVDURTOTAL=[4,57],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
18	{MVDURTOTAL=[4,57],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
19	{MVDURTOTAL=[4,57],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
20	{MVDURTOTAL=[4,57],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
21	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
22	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
23	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
24	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.108	0.569	2.234
25	{MVDURTOTAL=[4,57],J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.110	0.559	2.195
26	{MVDURTOTAL=[4,57],J01D=TRUE} => {RESULT=pos}	0.110	0.559	2.195
27	{MVDURTOTAL=[4,57],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.110	0.559	2.195
28	{MVDURTOTAL=[4,57],MV=YES,J01D=TRUE} => {RESULT=pos}	0.110	0.559	2.195
29	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.110	0.559	2.195
30	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],J01D=TRUE} => {RESULT=pos}	0.110	0.559	2.195
31	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.110	0.559	2.195
32	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE} => {RESULT=pos}	0.110	0.559	2.195
33	{VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.107	0.556	2.185
34	{VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.107	0.556	2.185
35	{VesDURTOTAL=[6,58],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.107	0.556	2.185
36	{VesDURTOTAL=[6,58],MV=YES,J01D=TRUE} => {RESULT=pos}	0.107	0.556	2.185
37	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.107	0.556	2.185
38	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.107	0.556	2.185
39	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.107	0.556	2.185
40	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01D=TRUE} => {RESULT=pos}	0.107	0.556	2.185
41	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.107	0.556	2.185
42	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.107	0.556	2.185
43	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.107	0.556	2.185
44	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE} => {RESULT=pos}	0.107	0.556	2.185
45	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
46	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
47	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
48	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
49	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
50	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
51	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
52	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
53	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
54	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
55	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
56	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
57	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
58	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
59	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
60	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.100	0.548	2.151
61	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.119	0.542	2.128
62	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],Antibiotic=TRUE} => {RESULT=pos}	0.119	0.542	2.128

#	Rules	Support	Confidence	Lift
63	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.119	0.542	2.128
64	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,Antibiotic=TRUE} => {RESULT=pos}	0.119	0.542	2.128
65	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES,VESICAL=YES} => {RESULT=pos}	0.119	0.542	2.128
66	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MV=YES} => {RESULT=pos}	0.119	0.542	2.128
67	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.119	0.542	2.128
68	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],Antibiotic=TRUE} => {RESULT=pos}	0.119	0.542	2.128
69	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,VESICAL =YES} => {RESULT=pos}	0.119	0.542	2.128
70	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE} => {RESULT=pos}	0.119	0.542	2.128
71	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,VESICAL=YES} => {RESULT=pos}	0.119	0.542	2.128
72	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES} => {RESULT=pos}	0.119	0.542	2.128
73	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],VESICAL=YES} => {RESULT=pos}	0.119	0.542	2.128
74	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5]} => {RESULT=pos}	0.119	0.542	2.128
75	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58],VESICAL=YES} => {RESULT=pos}	0.119	0.542	2.128
76	{MVDURTOTAL=[4,57],VesDURTOTAL=[6,58]} => {RESULT=pos}	0.119	0.542	2.128
77	{MVDURTOTAL=[4,57],Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.110	0.539	2.117
78	{MVDURTOTAL=[4,57],MV=YES,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.110	0.539	2.117
79	{MVDURTOTAL=[4,57],MV=YES,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.110	0.539	2.117
80	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.110	0.539	2.117
81	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.110	0.539	2.117
82	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.110	0.539	2.117
83	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.110	0.539	2.117
84	{MVDURTOTAL=[4,57],VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.110	0.539	2.117
85	{VesDURTOTAL=[6,58],MV=YES,Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
86	{VesDURTOTAL=[6,58],MV=YES,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
87	{VesDURTOTAL=[6,58],MV=YES,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
88	{VesDURTOTAL=[6,58],MV=YES,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
89	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
90	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
91	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
92	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
93	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
94	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
95	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
96	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.109	0.537	2.109
97	{MVDURTOTAL=[4,57],Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.131	0.534	2.099
98	{MVDURTOTAL=[4,57],MV=YES,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.131	0.534	2.099
99	{MVDURTOTAL=[4,57],MV=YES,VESICAL=YES} => {RESULT=pos}	0.131	0.534	2.099
100	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.131	0.534	2.099
101	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.131	0.534	2.099
102	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,VESICAL=YES} => {RESULT=pos}	0.131	0.534	2.099
103	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],VESICAL=YES} => {RESULT=pos}	0.131	0.534	2.099
104	{MVDURTOTAL=[4,57],VESICAL=YES} => {RESULT=pos}	0.131	0.534	2.099
105	{VesDURTOTAL=[6,58],MV=YES,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.129	0.528	2.073
106	{VesDURTOTAL=[6,58],MV=YES,Antibiotic=TRUE} => {RESULT=pos}	0.129	0.528	2.073
107	{VesDURTOTAL=[6,58],MV=YES,VESICAL=YES} => {RESULT=pos}	0.129	0.528	2.073
108	{VesDURTOTAL=[6,58],MV=YES} => {RESULT=pos}	0.129	0.528	2.073
109	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.129	0.528	2.073
110	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],Antibiotic=TRUE} => {RESULT=pos}	0.129	0.528	2.073
111	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.129	0.528	2.073
112	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE} => {RESULT=pos}	0.129	0.528	2.073
113	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES,VESICAL=YES} => {RESULT=pos}	0.129	0.528	2.073
114	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],MV=YES} => {RESULT=pos}	0.129	0.528	2.073
115	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5],VESICAL=YES} => {RESULT=pos}	0.129	0.528	2.073
116	{VesDURTOTAL=[6,58],MVTIMESTOTAL=[1,5]} => {RESULT=pos}	0.129	0.528	2.073
117	{MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.122	0.527	2.071
118	{MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.122	0.527	2.071
119	{MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.122	0.527	2.071
120	{MVTIMESTOTAL=[1,5],J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.122	0.527	2.071
121	{MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.122	0.527	2.071
122	{MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,VESICAL=YES} => {RESULT=pos}	0.122	0.527	2.071
123	{MVDURTOTAL=[4,57],Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.112	0.525	2.062
124	{MVDURTOTAL=[4,57],J01C=TRUE} => {RESULT=pos}	0.112	0.525	2.062
125	{MVDURTOTAL=[4,57],MV=YES,Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.112	0.525	2.062
126	{MVDURTOTAL=[4,57],MV=YES,J01C=TRUE} => {RESULT=pos}	0.112	0.525	2.062
127	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.112	0.525	2.062
128	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],J01C=TRUE} => {RESULT=pos}	0.112	0.525	2.062
129	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.112	0.525	2.062
130	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,J01C=TRUE} => {RESULT=pos}	0.112	0.525	2.062
131	{MVDURTOTAL=[4,57],Antibiotic=TRUE} => {RESULT=pos}	0.135	0.523	2.056
132	{MVDURTOTAL=[4,57],MV=YES,Antibiotic=TRUE} => {RESULT=pos}	0.135	0.523	2.056
133	{MVDURTOTAL=[4,57],MV=YES} => {RESULT=pos}	0.135	0.523	2.056
134	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],Antibiotic=TRUE} => {RESULT=pos}	0.135	0.523	2.056
135	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE} => {RESULT=pos}	0.135	0.523	2.056
136	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5],MV=YES} => {RESULT=pos}	0.135	0.523	2.056
137	{MVDURTOTAL=[4,57],MVTIMESTOTAL=[1,5]} => {RESULT=pos}	0.135	0.523	2.056
138	{MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.126	0.519	2.038
139	{MV=YES,J01D=TRUE} => {RESULT=pos}	0.126	0.519	2.038
140	{MVTIMESTOTAL=[1,5],J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.126	0.519	2.038
141	{MVTIMESTOTAL=[1,5],J01D=TRUE} => {RESULT=pos}	0.126	0.519	2.038
142	{MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.126	0.519	2.038
143	{MVTIMESTOTAL=[1,5],MV=YES,J01D=TRUE} => {RESULT=pos}	0.126	0.519	2.038
144	{VesDURTOTAL=[6,58],J01X=TRUE,Antibiotic=TRUE,VESICAL=YES} => {RESULT=pos}	0.103	0.511	2.006
145	{VesDURTOTAL=[6,58],J01X=TRUE,Antibiotic=TRUE} => {RESULT=pos}	0.103	0.511	2.006
146	{VesDURTOTAL=[6,58],J01X=TRUE,VESICAL=YES} => {RESULT=pos}	0.103	0.511	2.006
147	{VesDURTOTAL=[6,58],J01X=TRUE} => {RESULT=pos}	0.103	0.511	2.006
148	{MV=YES,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.126	0.502	1.972
149	{MV=YES,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.126	0.502	1.972
150	{MVTIMESTOTAL=[1,5],Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.126	0.502	1.972
151	{MVTIMESTOTAL=[1,5],MV=YES,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.126	0.502	1.972

#	Rules	Support	Confidence	Lift
152	{MVTIMESTOTAL=[1,5],MV=YES,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.126	0.502	1.972
153	{MVTIMESTOTAL=[1,5],VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.126	0.502	1.972
154	{VesDURTOTAL=[6,58],J01D=TRUE,Antibiotic=TRUE,J01C=TRUE} => {RESULT=pos}	0.103	0.501	1.969
155	{VesDURTOTAL=[6,58],J01D=TRUE,Antibiotic=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.103	0.501	1.969
156	{VesDURTOTAL=[6,58],J01D=TRUE,J01C=TRUE} => {RESULT=pos}	0.103	0.501	1.969
157	{VesDURTOTAL=[6,58],J01D=TRUE,VESICAL=YES,J01C=TRUE} => {RESULT=pos}	0.103	0.501	1.969