Pontifícia Universidade Católica
do Rio de Janeiro

Bianca Moreira Cunha

Producing and Evaluating Visual Representations Toward Effective Explainable Artificial Intelligence

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Profª. Simone Diniz Junqueira Barbosa

Rio de Janeiro
April 2025

Pontifícia Universidade Católica
do Rio de Janeiro

Bianca Moreira Cunha

Producing and Evaluating Visual Representations Toward Effective Explainable Artificial Intelligence

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the Examination Committee:

Profª. Simone Diniz Junqueira Barbosa
Advisor
Departamento de Informática – PUC-Rio

Prof. Alberto Barbosa Raposo
Departamento de Informática – PUC-Rio

Profª. Greis Francy Mireya Silva Calpa
Departamento de Informática – PUC-Rio

Rio de Janeiro, April 28th, 2025

Bianca Moreira Cunha

Graduated in Industrial Engineering by the Pontifical Catholic University of Rio de Janeiro (PUC-Rio). Worked as a Research and Development Analyst at ExACTa PUC-Rio and as a Data Scientist at Aditum.

Bibliographic Data

Cunha, Bianca Moreira

Producing and Evaluating Visual Representations Toward Effective Explainable Artificial Intelligence / Bianca Moreira Cunha; advisor: Simone Diniz Junqueira Barbosa. – 2025.

268 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2025.

Inclui bibliografia

1. keywordpre – Teses. 2. keywordpre – Teses. 3. Aprendizado de Máquina. 4. Visualização. 5. Explicação. 6. Interpretabilidade. 7. Inteligência Artificial Explicável. 8. Valores SHAP. I. Barbosa, Simone Diniz Junqueira. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

To my friends and family, whose unwavering support, patience, and love sustained me throughout this journey

Acknowledgments

I would like to express my sincere gratitude to all those who, directly or indirectly, contributed to the completion of this dissertation and to my journey throughout this graduate program.

To my advisor, Simone Barbosa, for her invaluable guidance throughout the development of this work, for believing in the potential of this reasearch and for ensuring the process was as smooth and manageable as possible.

To my colleagues, for their valuable collaboration and mutual support throughout the various stages of the graduate program.

To my friends and family, whose unwavering support, patience, and love sustained me throughout this journey.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Abstract

Cunha, Bianca Moreira; Barbosa, Simone Diniz Junqueira (Advisor). **Producing and Evaluating Visual Representations Toward Effective Explainable Artificial Intelligence**. Rio de Janeiro, 2025. 268p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The employment of Machine Learning (ML) models across diverse domains has grown exponentially in recent years. These models undertake critical tasks spanning medical diagnoses, criminal sentencing, and loan approvals. To enable users to grasp the rationale behind predictions and engender trust, these models should be interpretable. Equally vital is the capability of developers to pinpoint and rectify any erroneous behaviors. In this context emerges the field of Explainable Artificial Intelligence (XAI), which aims to develop methods to make ML models more interpretable while maintaining their performance level. Various methods have been proposed, many leveraging visual explanations to elucidate model behavior. However, a notable gap remains: a lack of rigorous assessment regarding the effectiveness of these explanations in enhancing interpretability. Previous findings showed that the visualizations presented by these methods can be confusing even for users who have a mathematical background and that there is a need for XAI researchers to work collaboratively with Information Visualization experts to develop these visualizations, as well as test the visualizations with users of various backgrounds. One of the most used XAI methods recently is the SHAP method, whose visual representations have not had their efficacy assessed before. Therefore, we developed a study where we worked together with visualization researchers and developed visualizations based on the information that the SHAP method provides, having in mind factors that are considered in literature to engender effectiveness to an explanation. We evaluated these visualizations with people from various backgrounds in order to assess if the visualizations are efficient in improving their understanding of the model. With the results of this study we propose an approach to produce and evaluate visual representations of explanations targeting their effectiveness.

Keywords

Machine Learning; Visualization; Explanation; Interpretability; Explainable Artificial Intelligence; SHAP values.

Resumo

Cunha, Bianca Moreira; Barbosa, Simone Diniz Junqueira. **Produzindo e Avaliando Representações Visuais para uma Eficaz Inteligência Artificial Explicável**. Rio de Janeiro, 2025. 268p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O uso de modelos de Aprendizado de Máquina (ML) em diversos domínios tem crescido exponcialmente nos últimos anos. Esses modelos realizam tarefas críticas que abrangem por exemplo diagnósticos médicos, sentenças criminais e aprovações de empréstimo. Para permitir que usuários compreendam a lógica por trás das predições e gerar confiança, esses modelos deveriam ser interpretáveis. Igualmente vital é a capacidade de desenvolvedores de localizar e corrigir quaisquer comportamentos errôneos. Neste contexto surge o campo de Inteligência Artificial Explicável (XAI), que visa desenvolver métodos para tornar modelos de ML mais interpretáveis, enquanto mantém seu nível de performance. Diversos métodos foram propostos, muitos aproveitando-se de explicações visuais para elucidar o comportamento do modelo. Porém, uma lacuna notável permanece: a ausência de uma avaliação rigorosa em relação à eficácia dessas explicações em melhorar a interpretabilidade. Resultados anteriores mostraram que visualizações apresentadas por estes métodos podem ser confusos mesmo para usuários que têm um histórico matemático e que há a necessidade para pesquisadores de XAI trabalharem colaborativamente com especialistas de Visualização da Informação, além de testar as visualizac' oes com usuários com bases diversas. Um dos métodos de XAI mais utilizados recentemente é o método SHAP, cujas representações visuais não tiveram a sua eficácia avaliada anteriomente. Por conta disso, nós desenvolvemos um estudo onde trabalhamos em conjunto com pesquisadores de visualização e desenvolvemos viusalizações baseadas nas informações que o método SHAP fornece, tendo em mente fatores considerados na literatura como características que geram eficácia a uma explicação. Avaliamos estas visualizações com pessoas com diversos históricos com o objetivo de avaliar se as visualizações são eficazes em melhorar o seu entendimento do modelo. Com os resultados deste estudo, promomos uma abordagem para produzir e avaliar representações visuais de explicações tendo como objetivo a sua eficácia.

Palavras-chave

Aprendizado de Máquina; Visualização; Explicação; Interpretabilidade; Inteligência Artificial Explicável; Valores SHAP.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI – Artificial Intelligence

ML – Machine Learning

XAI – Explainable Artificial Intelligence

1
Introduction

Systems that use Artificial Intelligence (AI) and Machine Learning (ML) have been rapidly becoming increasingly available and adopted in various fields in the past years. Many of the tasks they intend to support are tasks of high risk or crucial importance to people, companies, or society as a whole. For instance, systems used in the medical, legal, or security fields are some in which the user demands a high level of trust in the system and, therefore, needs to be able to understand why a certain decision was made by the model and to know when it is not behaving as expected. These models are algorithms that learn patterns from historical data and then respond to a specific problem when they receive unseen input. They allow for performance improvement on various tasks.

The field of AI that focuses on developing methods to improve the models' interpretability is called Explainable Artificial Intelligence, or XAI. Explanation methods such as SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017a) and LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) have been proposed to generate explanations for the behavior of a model or for specific predictions that will help users understand how the model makes its decisions. However, many of these methods claim to improve the model's interpretability without actually evaluating whether that assertion is true. Furthermore, it has not been long since researchers have proposed a clear definition of interpretability, and few works propose a way to assess the interpretability of a model or an explanation generated by the previously cited methods.

Doshi-Velez and Kim (2017) provide a definition for interpretability and propose a taxonomy of interpretability evaluation. Based on the concept of *simulatability* they proposed, which says that a model is *simulatable* if one can predict its behavior for new inputs, Hase and Bansal (2020) developed an experiment that assesses how users' understanding of a model is affected when they have access to explanations generated by the methods LIME and Anchor (Ribeiro et al., 2016, 2018). There is an implementation of the SHAP method (Lundberg and Lee, 2017a) that generates visual representations of the explanations and has been widely used in the XAI field; however, we have not found any study that evaluates the efficacy of these visual representations

of the explanation.

Even though many explanation methods have been proposed in the past few years, there has been little analysis of the reliability and robustness of such methods, making their utility for critical applications uncertain. Slack et al. (2020) raised critical questions about these methods that must be repeatedly asked and answered, such as: will the methods make the models more interpretable, trustworthy, and accountable? To whom will the explanations be accessible, comprehensible, or useful? Some researchers have acknowledged that explanation assessment should concentrate on aspects such as trust, transparency, understandability, usability, and fairness (Brdnik et al., 2023). The evaluation of XAI explanations is not standardized, which makes it still an open challenge (Brdnik and Šumak, 2024; Aechtner et al., 2022; Kim et al., 2024). This issue makes it hard to compare studies in the field and possibly generate insights and patterns apart from specific use cases. Furthermore, it makes it challenging to get to evidence-based XAI guidelines of what makes a significant explanation for users (Kim et al., 2024).

We performed a preliminary study, described in chapter 4, where we concluded that, in order to develop efficient visualizations of explanations, there is a need for XAI researchers to work with Information Visualization researchers. Therefore, in our final work, we followed a process to develop visual representations of SHAP explanations with the support of Information Visualization researchers and then evaluated these new visual explanations with users from various backgrounds.

Our work focused on how to efficiently visually represent explanations to improve the explainability of the explanations. Our objective is to design a process for generating visualizations and then to be able to evaluate them in terms of what makes a good explanation. We bring up the following research questions:

Q1 Are visual representations of explanations generated by the SHAP library[1] effective in improving the models' interpretability?

Q2 Are the visual representations of explanations generated with the collaboration visualization researchers more effective in improving the models' interpretability?

Q3 How appropriate is the concept of "simulatability" to evaluate the explanations generated by the SHAP method?

[1] https://shap.readthedocs.io/en/latest/

Additionally, a subquestion arose regarding aspects that have been considered indicators for explanation goodness and evaluation approaches previously proposed. These concepts will be described in chapter 3.

SQ1 Is perceived confidence or trust a good measure for explanation goodness?

Given that context, we conducted a preliminary study, inspired by Hase and Bansal (2020), in order to assess how visualizations of the explanations generated by the SHAP method (Lundberg and Lee, 2017a) affect people's understanding of the model. We also took a step further and included a textual explanation of the visualization, intending to evaluate the differences in understanding between having only a visualization, which can be hard to understand for people unfamiliar with data charts, and having an additional verbal explanation. Through the study, we found that the visualizations of the explanations can be confusing even for users with a background in statistics, mathematics, or a similar field. We concluded that XAI researchers who are developing model explanation methods need to work in conjunction with visualization researchers so that they can construct visualizations that best communicate the information provided by the method. Furthermore, there is a need to evaluate the effectiveness of the explanation's visualization by testing it with users of various backgrounds.

Having that in mind, we conducted a second study, in which we worked together with visualization researchers and asked them to design how they would represent the information provided by the SHAP method. Then, we tested these visualizations with users from various backgrounds.

Our contributions are: (i) successful visualizations resulting from the second study (if any); (ii) a process for generating effective visualizations by working in conjunction with visualization researchers; and (iii) a process for evaluating the explanations with end users.

The remainder of this document is divided into seven chapters: 2 Theoretical Foundations, 3 Related work, 4 Preliminary Study, 5 Proposal, 6 Results, 7 Discussion, and 8 Conclusion.

2
Theoretical Foundations

In this chapter, we present discussions on model interpretability, focusing on the meaning of interpretability and why it is desired; the differences between interpretability and explainability; and works on explainable AI (XAI).

2.1
What does interpretability mean?

To assess the interpretability of a model or whether a certain method can make a model more interpretable, we first need to understand what *interpretability* means. Doshi-Velez and Kim (2017) and Lipton (2017) came to the conclusion that there is no consensus in the literature as to the meaning of interpretability or how to evaluate it. Therefore, the claim that a model is interpretable has a quasi-scientific character (Lipton, 2017). Doshi-Velez and Kim (2017) proposed a definition for the term: "In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human."

Other researchers have proposed some meanings for interpretability in machine learning. Kim (2015) stated that interpretable models enable communication from machines to humans without changing the internal states of the models. Hase and Bansal (2020) suggested that a model is interpretable if it is "simulatable", which they described as a property that makes it possible for a person to predict the model's behavior for unseen inputs. Kim (2015); Chatzimparmpas et al. (2020), and Ridgeway et al. (1998) asserted that interpretability is a means to engender trust. Doshi-Velez and Kim (2017) said that interpretability can be defined as the ability to explain something or to present it in understandable terms to a human. Mittelstadt et al. (2019); Biran and Cotton (2017), and Lou et al. (2013) understand an interpretable model to be one whose inner mechanisms a person can understand. The latter also argued that a model can be interpretable if a person can understand a specific prediction, even if they do not understand how the model functions. These two ways of understanding interpretability proposed by Lou et al. (2013) derived the concepts of *ad-hoc* and *post-hoc* interpretability, which describe an understanding of the inner working of a model and the comprehension of

what led a model to a specific output, respectively. These concepts are further explored in subsection 2.4.1.

Kim et al. (2018) gave a formal definition to "interpretation" of a ML model as a function $g : E_m \rightarrow E_h$, where E_m and E_h are vector spaces. The first is composed of basis vectors e_m corresponding to data such as input features and neural activations, and the second is composed of basis vectors e_h corresponding to an unknown set of human-interpretable concepts.

This lack of consensus in the definition of interpretability makes it challenging to determine which approaches are most appropriate for evaluating machine learning models in terms of their interpretability. In this work, we consider interpretability in the context of machine learning to refer to being able to understand how a certain model generated a specific output and, therefore, trusting its results.

2.2 Why is interpretability desired?

Having the proposed definition of interpretability in mind, we need to understand why there is a need for interpretable models. Why should we not settle for accurate models, knowing that we can measure their correctness, even if we do not understand why they reached a specific result? The desire for model interpretations indicates that model predictions and accuracy metrics are insufficient to characterize the model in a way that the user can know that the model correctly represents reality (Lipton, 2017).

Machine Learning models are applied in many high-stakes scenarios, so the need for interpretable models will only grow. In these situations, domain experts need to be able to understand and trust the model they are working with so that they can make important decisions (Lakkaraju et al., 2019). Some real-world problems are hard to formalize, and that is when interpretations are needed (Lipton, 2017). Hoffman et al. (2018) argued that "XAI systems should enable the user to know whether, when, and why to trust and rely upon the XAI system and know whether, when, or why to mistrust the XAI and either not rely upon it, or rely on it with caution."

Another reason brought up by Tamagnini et al. (2017) for the need for model interpretability is that a model that analysts can inspect and whose decisions they can observe can support them in understanding better the data and the phenomenon it describes. Furthermore, understanding the predictions given by a model can allow the user to identify possible causes of errors, misclassifications, or unfair outcomes (Poulin et al., 2006; Correll, 2019).

Doshi-Velez and Kim (2017) argued that interpretability is a way to

qualitatively assess whether objectives such as fairness, privacy, reliability, robustness, causality, usability, and trust are met. They also debate that the need for interpretability comes from an incompleteness in the problem formalization, which causes an obstacle to optimization and evaluation. Therefore, explanations are a way to make gaps in problem formalization visible so they can be addressed.

Apart from the previously cited practical reasons, there are also strategic motivations for having explanations for AI systems' decisions. There is a higher propensity for users to trust and rely on a system if they can understand its outputs, and therefore, a higher probability for them to adopt the system (Woodcock et al., 2021). Additionally, the European Union law for data protection and privacy, the General Data Protection Regulation, released in 2016, states that AI systems must provide "meaningful information about the logic involved" in the decision-making process and also should provide "an explanation of the decision reached" (Radley-Gardner et al., 2016). Correll (2019) also brings up the moral duty in legal cases to inform the impacted people of the decision-making process.

2.3 Interpretability vs Explainability

In the literature, we could encounter two concepts that are similar and often used interchangeably but have a subtle difference in meaning: "interpretability" and "explainability." These concepts have been widely used, but not always consistently (Lopes et al., 2025). Lipton (2017) considers interpretability as a measure of the transparency of a model, *i.e.*, the understanding of how the model works internally, how it gets to certain decisions. Transparency can be viewed as the opposite of opacity or "black box-ness." In contrast, he defines post-hoc explanations as further information that we can get from a model, such as the significance of various parameters. Miller (2019) uses the previously cited definition of interpretability given by Lipton (2017) and considers explanations as a means by which an observer may obtain an understanding of a model. In our work, we understand interpretable models as models in which an observer can understand their inner workings, *e.g.*, a simple linear model with few coefficients can be interpretable since it can be easily understood. As for explainability, we consider an explanation to be an instrument that assists the observer in understanding a model by giving them additional information. That is what Explainable AI (XAI) aims to do: to propose methods that will provide explanations for models that are not naturally interpretable. This is addressed in the next section.

2.4 Explainable AI

Simple models, such as linear regression or simple tree models, can be deemed interpretable, but they can also be less accurate than more complex models (Lundberg and Lee, 2017a). Some argue that even these less complex models cannot be interpretable when they are sufficiently high-dimensional (Lipton, 2017). Interpretability is desired; however, model performance is also needed. Therefore, several methods have been proposed in the past years to generate explanations for machine learning models' behavior or for the outputs they produce, thus improving the models' interpretability without having performance loss. The field of AI encompassing this kind of work is called Explainable AI (XAI) (Ali et al., 2023). The Defense Advanced Research Project Agency (DARPA) technical report (Gunning and Aha, 2019) defined XAI as "a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners". The field focuses on improving the transparency of how the models get to their outputs, aiming to provide clarity to computer programmers and expert users (Woodcock et al., 2021; Alicioglu and Sun, 2022).

A characteristic of current AI explanations presented by Woodcock et al. (2021) is that they give answers to *how* questions and not to *why* questions, even though the latter are people's preferences. These methods will then answer the question "*How* did you get to that decision?" Their work presents two ways in which the *how* questions are answered:

- **Input influence** presents the list of input variables with an impact value for each. It is also called a *local explanation* (Tamagnini et al., 2017; Lundberg and Lee, 2017a; Ribeiro et al., 2018; Hase and Bansal, 2020).
- **Case-based reasoning** displays an instance of the training set similar to the one being classified (Hase and Bansal, 2020).

Both methods require the human recipient to have prior domain knowledge so that they can understand and evaluate the explanation. The method of input influence also has the downside that long lists of input feature contributions might be hard to interpret (Correll, 2019).

Woodcock et al. (2021) stated that *why* questions are preferred by humans and are implicitly contrastive, which means that the answer to them offers an explanation relative to some event that did not happen. They call that event a *foil*. They consider the core finding of their work to be that "in

order to close a user's information gap, the AI explanation must be generated with an understanding of that user's unique foil."

Explanations have been explored only recently in AI but have been largely studied in other fields. These studies should be considered when developing explainability methods (Hoffman et al., 2018). An important factor to be considered in Explainable AI is curiosity, since it is one of the core reasons people seek explanations. Therefore, XAI systems must take advantage of the power of curiosity (Hoffman et al., 2018). Furthermore, Kim (2015) raised the concern that traditional machine learning systems are not designed with the possibility to leverage the knowledge of domain experts through direct interaction with the system. For these systems to make a real impact in important domains, machine learning and humans should take advantage of their complementary skills and work collaboratively.

Li et al. (2020) created a visual analytics system that helps users compare ML models regarding their performance, consistency, and reliability. They present the system by utilizing a medical dataset as a case study and explain how their visualizations assist the user in understanding how different ML models differ in their prediction process, how each feature impacts each model, and how consistent each model is. They claim that they were able to demonstrate the usefulness of their system. However, they have not conducted any evaluation with actual end users to support their claim.

2.4.1 Types of interpretability methods

Two ways of working with interpretability have been widely recognized: ad-hoc and post-hoc interpretation. **Ad-hoc interpretation** makes explicit the inner workings of a model, used in models such as linear and logistic regression, decision tree, and kNN (Brdnik and Šumak, 2024), while **post hoc interpretation** concerns how the model reached a certain prediction by generating an approximation of the model's reasoning, and is used in black-box models such as neural networks and random forest (Mittelstadt et al., 2019) (Brdnik and Šumak, 2024). The explanation methods cited before focus mainly on post-hoc explanations. The advantage of post-hoc interpretations is that black-box models can be interpreted without having their predictive performance undermined (Lipton, 2017).

Explanations generated by these methods can be textual explanations or visualizations. A language model, for instance, can be trained to generate textual explanations for another model. In contrast, visualizations can be rendered to qualitatively show what a model has learned (Lipton, 2017),

which means that they can provide more information on how the model makes decisions than traditional metrics.

Another dimension of explanations divides them into local and global explanations. A **local explanation** will clarify how the model reached a specific prediction, while a **global explanation** explains the model's behavior as a whole. An example of local explanation is an *explanation by example*, a post-hoc mechanism that provides other instances the model considers similar (Lipton, 2017). Kim et al. (2024) found that local explanations are the most used in XAI studies.

Post-hoc interpretations, however, present some challenges. First, many models operate on features such as pixel values, which are difficult for humans to understand. Furthermore, the explanations must correctly reflect the model's complex internals (Kim et al., 2018).

2.5 SHAP method

SHAP is a method proposed by Lundberg and Lee (2017a) based on Shapley values as a unified measure of feature importance. It is a concept from game theory that assigns a contribution value to each game participant to get to a certain outcome. In the context of model explanation, Shapley values represent the impact that each feature has on the predictions. It is a model-agnostic method, which means that it can be used to generate explanations for various prediction models.

Considering a binary prediction task, the impact of the features can be "positive" (contribute to a classification of the instance as belonging to the positive class), zero (neutral), or negative (contribute to a classification as the negative class). By having the contribution of each feature on each observation of the dataset, it is possible to have local explanations, *i.e.*, to visualize how each feature contributed to a specific prediction; and global explanations, *i.e.*, to have a notion of how each feature contributes to the predictions generally.

2.6 Understanding models through visualization

Visualization is an important tool for conveying trust in ML solutions, which is not a trivial task (Chatzimparmpas et al., 2020; Poulin et al., 2006). Spinner et al. (2019) argued that visualizations are a natural way of obtaining human-interpretable explanations. Graphical explanations help users visualize the evidence for a classification decision efficiently and provide an audit of the model, thus making it possible for the user to identify when a decision is

unexpected or erroneous (Poulin et al., 2006). Kim et al. (2023) point out that visualization is preferred by users among attributes related to interpretability. They argue that it is a more effective means for humans to absorb context than text. Various approaches have been proposed to clarify how certain classes of models function (Lakkaraju et al., 2019).

Chatzimparmpas et al. (2020) found that visualization of feature importance, the impact of different characteristics of the data instances, the investigation of hyperparameters, the pre-processing steps, and the evaluation of the model are the most popular among users who wish to understand the ML process. Brdnik et al. (2023) conducted a study that assessed users' perceived trust and satisfaction with explanations generated by multiple XAI methods. They gathered the participants in a semi-controlled environment, presented the predictions and explanations to them, and asked them to answer a questionnaire to assess the explanations according to their perception. Having STEM college students as study participants, they found that local explanations represented by bar charts were the ones in which the participants reported the highest degrees of trust and satisfaction.

Poulin et al. (2006) proposed a graphical explanation framework that aims to increase the user's capability to understand and audit the classification process based on evidence. It has five representation capabilities: (1) representing the classification decision; (2) representing the decision evidence by showing the impact of each feature on each possible decision; (3) decision speculation, where they provide an interactive interface where users can do "what-if" analyses by changing the values of features and see how it would affect the decision; (4) representing ranks of evidence, by ranking the evidence of each feature for the overall behavior of the model; and (5) representing the source of evidence, *i.e.*, the data supporting evidence contributions.

Recent approaches focus on both visual design and interactive, mixed-initiative workflows, as provided by Visual Analytics (VA). The diverse backgrounds of the different user groups affected by AI bring various requirements for XAI tools (Spinner et al., 2019). Domain experts may use XAI tools to scrutinize models and change hyperparameters to facilitate the AI system's analysis, while AI specialists and developers may use them to discover flaws in the architecture of their models (Ali et al., 2023). Lopes et al. (2025) proposed an interactive and visual tool to support users in utilizing, interpreting, and refining ML models. Its development was guided by a study on users' needs. They presented an interface that provides complementary visualizations and supports text databases in order to offer more information completeness.

Alicioglu and Sun (2022) named a sub-field of XAI that focuses on VA

research, Visual-Based XAI (vXAI). Their review found that it is still an under-explored field with a limited number of papers published. They found that local explanations help the user to understand the model's behavior for similar instances. Additionally, they discovered that the most popular representations for local explanations are bar charts, breakdown charts (both can be applied in text, tabular, or image data), and heatmaps (mostly for image data). Global explanations are more challenging for black-box models due to their complex structure and computational process. Therefore, only a few papers have proposed methods that provide global explanations. Histograms are currently the most popular visualization for global explanations. Since vXAI is still a novel field, there is no common visual approach to represent XAI methods for different types of data and models, and there is no standardized way to depict local and global explanations. Researchers tend to develop visualizations customized to their data domain and application area.

3
Related work

This chapter presents the works in the literature that strongly relate to our research. The methods proposed by these studies aim to evaluate XAI methods and their outputs, focusing on visual representations of explanations generated by these models.

3.1
Evaluating explanations

Current explainability evaluation approaches rely on some notion of "you'll know it when you see it." There is a need for more rigor in that evaluation since not all models, even in the same class, may be comparable. Likewise, different applications may have different interpretability needs. To evolve the field and to be able to compare explanation methods and to understand when they may generalize, there is a demand for a formalization of the notions of what makes a model interpretable and for them to be evidence-based (Doshi-Velez and Kim, 2017). Hase and Bansal (2020) showed some issues we might encounter in explanation methods. First, many explanation methods may not actually help users understand how a model behaves. Also, a method that works successfully in one domain might not work as well in another. Ali et al. (2023) agreed with that and also debated that most of the existing approaches are built with explainability aims that are too general and lack well-defined context-specific use cases and, as a consequence, miss the unique needs of a certain domain, resulting in poor adoption and sub-optimal outcomes. Lastly, Hase and Bansal (2020) stated that combining information from explanations does not necessarily result in explicit improvements in simulatability.

Spinner et al. (2019) argued that the proposed tools are often implemented as standalone prototype solutions and lack integration with the ML developing and debugging process. They claim, therefore, that there is a gap between theory and practice, and confirmed in their study that the people developing these XAI methods mostly have no hands-on experience using their proposed tools.

There is also a need to consider the intended audience of the proposed

XAI approaches. Most existing methods do not state their intended audience and, therefore, do not assess whether their needs are met (Ali et al., 2023). We can see a latent necessity for the explanations to be customized to their audience (Aechtner et al., 2022). Considering current XAI techniques, little to nothing is known about how they are perceived by end-users when they are embedded in an AI system (Bernardo and Seva, 2023). Hoffman et al. (2023) argue that the quality of an explanation depends on the users' needs, knowledge, and goals. It is important to evaluate XAI user experience interface from a human-computer interaction (HCI) perspective, in order to ensure that the solutions are appropriately designed for end-users (Kim et al., 2023). Miller (2019) and Ali et al. (2023) argue that current explanation methods are extremely static, and ideal explanations should contain explainer–explainee interaction. The needs of the users, either ML experts or end-users, have hardly been considered when designing XAI solutions. As a result, few works lie on the intersection between user needs and ML systems (Lopes et al., 2025).

Lopes et al. (2025) developed a new XAI approach considering the problems previously cited throughout the development. Their target users are end users familiar with ML models through the utilization of these models in daily tasks, such as data scientists or domain experts who use ML systems for analysis, decision-making, or research. They developed their method by having users participate in the development process, validating each step of it with them, and taking their needs into account, which most of the other methods did not do. They also provide complementary visualizations that let users obtain insights about the models. The participants considered the explanations generated by their tool complementary and informative. The solution allowed them to understand the outcomes of the model and how it works. Through their tool, they could also identify strategies they could take to improve the model.

Some of the existing explanation methods generate approximations to the system's behavior. These might be helpful for pedagogical purposes, but can also be misleading when presented to lay users (Mittelstadt et al., 2019). Slack et al. (2020) performed some tests that showed that malicious actors can build discriminatory models that can fool post-hoc explanation techniques, which shows that these methods can be insufficiently robust to ensure that the model is trustworthy, reliable, and fair in sensitive applications. Moreover, design can manipulate emotions, which can wrongly affect trust and reliance. In this way, erroneous and manipulative XAI can be displayed in a system and be effective if its design can positively affect the user (Bernardo and Seva, 2023). A local explanation given to an individual who does not understand

the model's limitations can be incomprehensible or misleading. This kind of explanation can be useful in a specific context and help reach a specific decision, but it does not give insight into how a model functions as a whole. Therefore, Mittelstadt et al. (2019) argued that these explanations might be useful for understanding relationships between variables relevant to a particular decision, but they do not prove the model's trustworthiness overall.

3.2 Methods to measure and assess interpretability

Although the field of XAI emerged in 2018, assessing explanations through human-centered evaluations is still a recent topic (Kim et al., 2024). Ali et al. (2023) stated that the value of an explanation is significantly influenced by how valuable it is proven to be for an end user responsible for decision-making. Consequently, end users need to be involved in assessing explainability methods, preferably in a context with real tasks and data. They also argued that user performance, *e.g.*, accuracy or speed of the decision-making process, should be the measurement for the method evaluation. Additionally, it is important to consider the user's understanding and satisfaction with the given explanation. Therefore, a significant portion of the evaluation relies on qualitative assessments using surveys or interviews. A user's comprehension of the functioning of an AI system may be examined by questioning the associated decision-making process. Moreover, when designing an explainability solution for an AI system, user expectations should be considered (Ali et al., 2023) (Stumpf et al., 2018).

Kim et al. (2024) concluded that one of the challenges in the XAI field is a lack of consensus concerning what makes an explanation good and meaningful to users. They also found an absence of standardization in the evaluation methods. There is no agreement on whether human-centered evaluation is essential for XAI evaluation, and if so, what factors need to be evaluated and how.

Lakkaraju et al. (2016) proposed a way to capture the interpretability of models based on decision sets. Decision sets are sets of classification rules, where each rule is an independent classifier. They defined four metrics for measuring interpretability: size, length, cover, and overlap. "Minimizing size encourages decision sets with a small number of rules. Minimizing length captures the notion that interpretable rules are short and concise. We use cover to denote how many data points satisfy the itemset of a rule, which is necessary for defining subsequent metrics. Finally, minimizing overlap encourages each rule to cover an independent part of the feature space" (Lakkaraju et al.,

2016)). They stated that a set of rules is considered interpretable if (1) the rules in the set describe non-overlapping feature spaces; (2) most data points are covered by the rules in the rule set; (3) the set is composed of a small number of rules, and these rules are concise; and (4) the rules in the rule set describe most of the classes present in the data. That is a useful and intelligible way to assess interpretability; nevertheless, it only encompasses models based on decision sets.

Brdnik and Šumak (2024) found in their review that qualitative evaluations are still developing and are mostly based on user studies. Additionally, tasks where users are asked to predict the model's decisions and rate their confidence in the prediction have also been used. They also mentioned a work by Chou et al. (2021) that proposed three-view evaluations, one on an objective level, which does not have the participation of users; another on a functional level, which is focused on functions; and another on a user level, which is a human-centric approach.

Adebayo et al. (2018) also proposed a method to guide researchers in assessing the scope of model explanation methods. Their method is based on randomization tests to evaluate the adequacy of explanation methods and serve as sanity checks in constructing new explanation methods.

Hoffman et al. (2018) developed a framework to assess explanations by considering the following measurements, and provided a checklist for each of them with a series of questions to characterize that quality in an explanation:

1. Explanation goodness
2. Explanation satisfaction
3. Users' mental models
4. Curiosity
5. Trust
6. Performance

Aechtner et al. (2022) simplified these measurements as **understandability, usefulness, trustworthiness, informativeness and satisfaction** and gave one question to characterize each of them. These questions are listed in chapter 5 and were used in our study.

Kim et al. (2024) proposed a categorization of evaluation measures, which is divided into three aspects:

1. The in-context quality of the explanation;

2. The contribution of the explanation to human-AI interaction;

3. The contribution of the explanation to human-AI performance.

Doshi-Velez and Kim (2017) proposed a taxonomy of evaluation approaches for interpretability: application-grounded, human-grounded, and functionally grounded.

- **Application-grounded evaluation** is the conduct of human experiments using real applications, *i.e.*, the explanations are evaluated based on their end-task, such as identifying errors. The method is tested according to the application's objective, thus giving concrete evidence of success.
- **Human-grounded evaluation** involves conducting simpler human-subject experiments while maintaining the nature of the target application. This kind of assessment is useful when doing experiments with the target community is challenging.
- **Functionally grounded evaluation** requires no human experiment; it considers interpretability as a proxy for explanation quality. This type of experiment is easier to perform since it does not demand the time and cost that human experiments need. It can be used when a method has already been validated by a human-grounded experiment, when a method is not mature, or when human subject experiments are unethical.

For the purposes of this study, we decided to perform a human-grounded evaluation, as described in Section 4.1.

3.2.1 SHAP method evaluations

In their literature review, Brdnik and Šumak (2024) mentioned that the SHAP method was the most cited among the XAI methods, which evidences the relevance of this method in the field. However, Aechtner et al. (2022) showed lower trust in SHAP explanations when compared to other widely used explanation methods (LIME and PDP). One of our hypotheses is that this is because the visualizations of the explanations generated by the SHAP library are ineffective in engendering trust. Aechtner et al. (2022) agree with that hypothesis: " Probable causes for this variance may be attributed to the complexity of visualization that leads to confusion for both AI novices as well as experts."

Brdnik et al. (2023) concluded that local explanations generated by SHAP represented as bar charts were the ones that engendered the highest

degrees of satisfaction and trust when compared to global explanations and other types of charts. Lopes et al. (2025) presented an overview of various black box explanation approaches and showed that users were not involved in the development of the SHAP method. They also mentioned a formal evaluation of the method with users; however, we did not find in the literature an approach with quantitative and qualitative evaluations, which we have done in our research. Furthermore, we did not find a work that evaluates the SHAP explanations using the factors proposed by Aechtner et al. (2022).

We did not find in the literature any XAI method that represents the model's success or failure to make the correct prediction, along with its explanations. This is critical information for the user so they may know if the model is biased, for instance. The study conducted by Brdnik et al. (2023), where they assessed multiple explanation techniques, showed that the factor that had the worst evaluation for all assessed methods was completeness. We believe that this lack of information about the model's performance can influence this result.

3.3 Considerations

As we saw in this chapter, visual representations of explanations can be extremely useful for users to understand more easily these explanations and, consequently, the model behavior. Therefore, they are an important tool in the XAI field. Various solutions have been proposed that take advantage of visualizations to convey information about the model to users; however, there are still no standardized approaches to represent explanations.

Furthermore, a formalization of the evaluation approaches of these explanation methods is needed, taking into account the needs, expectations, and perceptions of the final users. This formalization would possibly broaden the discussion about these methods, making it possible to generate more insight and more efficient output.

In our work, we take into account the arguments about visual representations of explanations brought up in this chapter in order to develop visualizations for the SHAP method with the support of InfoVis experts. In addition, we aim to construct a process of evaluation of these representations, having as directives factors that have repeatedly been mentioned as crucial for a good explanation, such as trustworthiness, satisfaction, and transparency. This evaluation process was conducted with users of various backgrounds and was built based on the taxonomy of Doshi-Velez and Kim (2017), more specifically using the human-grounded evaluation approach. We contributed to formalizing the

evaluation of the explanation methods by examining how efficient this evaluation process is and how it can be improved further. Additionally, we have a set of visualizations that were evaluated as more effective in conveying the explanation information than the ones provided by the widely used SHAP library, and that can be the base for generating more optimal explanation visualizations.

4
Preliminary Study

In this chapter, we present the preliminary study developed to evaluate the effectiveness of the visual representations of a widely known and applied ML explanation method, SHAP (Lundberg and Lee, 2017a). We present the study methodology (section 4.1), followed by an analysis (section 4.2) and a discussion (section 4.3) of the results.

4.1
Methodology

In this section, we present the study we developed to evaluate whether visual representations of explanations generated by the SHAP method are efficient in helping users interpret machine learning models. We describe the dataset used (subsection 4.1.1) and the experimental design (subsection 4.1.2). We used an online questionnaire for the study, in which we explained to the participant the goals of the study and its ethical considerations, including the steps we took to respect their privacy and maintain their anonymity in all study reports. We also made clear that they could interrupt or abandon the study at any moment without justification or penalty. The questionnaire included an item for them to declare their informed consent.

4.1.1
Dataset

The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases (Smith et al., 1988). It has 1,548 records of diagnostic measurements of female patients who are at least 21 years old and of Pima Indian heritage. The target variable is a boolean that indicates whether the patient was diagnosed with diabetes. With the purpose of not letting the respondents' preconceptions or prior knowledge of diabetes influence their responses, we changed the description of the dataset and said that the data was collected from extraterrestrial beings to diagnose a disease that is unknown to humans. The features are named Symptom 1, Symptom 2, etc.

4.1.2 Experimental Design

This study was inspired by the work done by Hase and Bansal (2020), who conducted simulation tests. Their experiment was based on the property of "simulatability," described by Doshi-Velez and Kim (2017) as a crucial property for interpretable models. The idea is that a model is *simulatable* if one can predict its behavior for new inputs. Thus, if a person can simulate the behavior of a model, it means that they understand how it works. We replicated the experiment that the researchers did and adapted a few steps for our purposes.

We used a Random Forest Classifier to generate the predictions and the SHAP method to generate the explanations for the model. As seen in section , SHAP (Lundberg and Lee, 2017a) is a method based on Shapley values, which represent the contribution of each participant of a game to get to a result. Considering our prediction task, the impact of the features can be positive (contribute to the positive class), zero (neutral), or negative (contribute to the negative class).

We conducted a study using a questionnaire and following this procedure:

1. We present the respondents with four instances of the dataset, the real classification of each, and the classification given by the model of each so that they can observe them and try to understand how the model works, what it is getting right, and what it is getting wrong.

2. We give them four new instances, without the classifications, and ask them to classify each.

3. We give them two of the instances given in step (1) and two new ones, along with graphical explanations generated by the SHAP method.

4. We give two of the instances given in step (2) and two new ones, along with their graphical explanations, and ask them to classify them.

5. We repeat step (3), adding a textual explanation to the visualization for each observation.

6. We repeat step (4), adding a textual explanation to the visualization for each observation.

After each classification step (steps 2, 4, and 6), we ask the respondents how confident they are about their response using a 5-point Likert Scale (where 1 indicates "not confident at all" and 5 indicates "extremely confident") and also what made them choose the selected class as an open-ended question. For

each step, we gave four instances: one True Positive, one True Negative, one False Positive, and one False Negative, according to the model's predictions. That way, the respondents could observe what the model is getting right and what it is getting wrong, and then show whether they can give the correct classification. Furthermore, that was a strategy so that, in the classification steps, we could prevent the respondents from just guessing all the classifications and succeeding even so. This design aims to mimic real situations in which the model can be erroneous, and the explanations can be used to identify these mistakes.

In step 1, we give the respondents the observation without any explanation so that we can assess whether their ability to interpret the model improves when they are given the same or similar observations along with the explanation to classify. Along the steps, we always repeat two previously seen explanations – to compare the responses and observe their progress – and give two new ones so as not to suffer from the effect of time in the respondent's learning process. Table 4.1 presents the classification steps each instance is used in and, consequently, how many times they were asked to be classified.

Table 4.1: Steps in which each instance is used and how many times they were classified

Instance	Step	Count
1	2 and 6	22
2	2 and 4	22
3	2 and 4	22
4	2 and 6	22
5	4	11
6	4	11
7	6	11
8	6	11

In this study, we aimed to test two hypotheses:

H1 The visual explanations help people better understand the model's behavior. (Lipton, 2017)

H2 The textual explanations help people better understand the visual explanations and improve the interpretability of the model compared to visual explanations alone. (Lipton, 2017)

In the experiment, we divided the respondents into three groups. Each group received a visualization type in steps 3-6. They could analyze the visual explanations before giving the classifications. The chosen visualizations were

a beeswarm chart, a bar chart, and a waterfall chart, which are generated by the Python SHAP library (Lundberg and Lee, 2017b). An example of each visualization can be seen in Figures 4.1, 4.2, and 4.3.

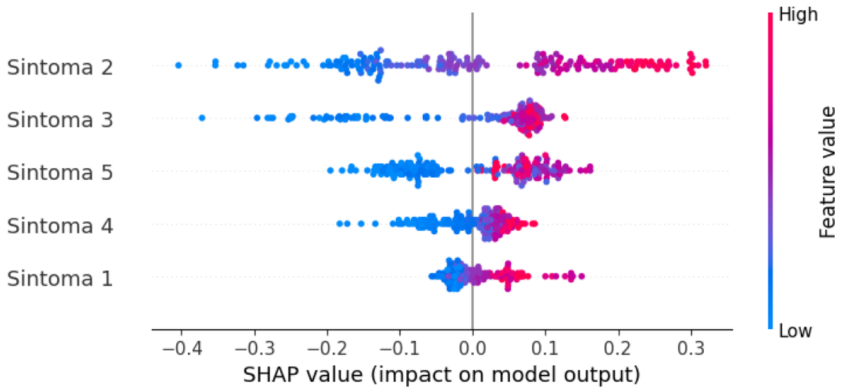


Figure 4.1: Beeswarm plot generated by Python SHAP library

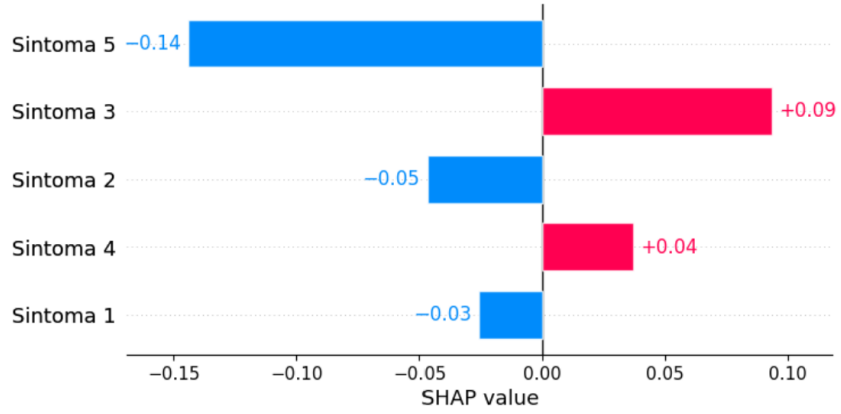


Figure 4.2: Bar plot generated by Python SHAP library

The purpose of giving three types of visualizations for the respondents to observe was to understand whether one would be easier to interpret than the others.

The textual explanations given to the participants were manually generated by us with the purpose of describing the visualization and each of its elements in order to help the participants understand their functioning. An example of a textual explanation is presented below. This explanation is for the graph in Figure 4.1:

"This plot shows the features sorted by magnitude of impact on the model in general, considering their absolute mean SHAP values. Each feature has a beeswarm, and they are composed of dots that represent each observation and are distributed along the x-axis according to the observations' SHAP values. On the right side of the plot, there is a vertical bar that gives the color shades

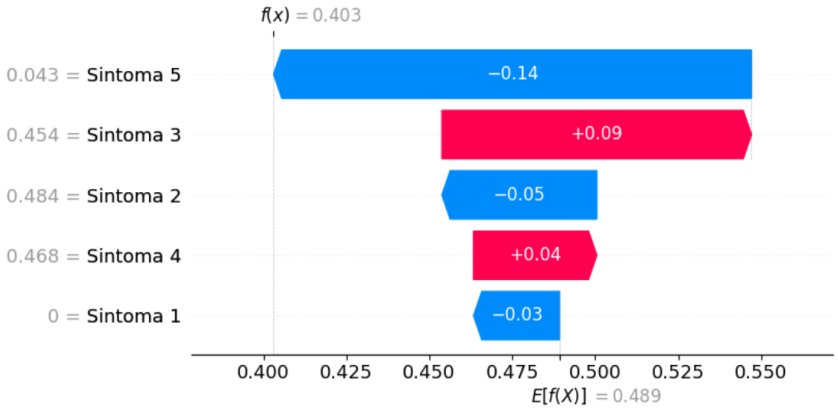


Figure 4.3: Waterfall plot generated by Python SHAP library

of the dots that represent, from high to low, values of the features. Red dots represent high feature values, shades of purple represent medium values, and blue dots represent low values. That means that, along with the information on how much impact each feature has on the model output, it is possible to know what ranges of feature values have positive, negative, or no impact. In this particular case, the 5 features that presented the highest impact in the model were Sintoma 2, Sintoma 3, Sintoma 5, Sintoma 4, and Sintoma 1, in that order."

4.2 Analysis and Results

There were 12 respondents to the questionnaire. One was disregarded since they did not have a mathematical or statistical background, resulting in 11 valid responses. 6 of the respondents had a graduate degree, 4 were attending graduate courses, and 1 had completed their undergraduate studies. To evaluate their responses, considering their understanding of machine learning and artificial intelligence, we asked them about their knowledge on the matter. 7 said that they had advanced knowledge, and 4 said that they had intermediate knowledge. In contrast, when asked about their knowledge of the interpretability of machine learning models, 5 stated they had intermediate knowledge of the subject, 3 basic knowledge, and 3 advanced knowledge. Considering that the textual explanations presented to the respondents in steps 5 and 6 of the experiment were written in English, we also asked them what they considered their English level. 8 were advanced, and 3 were fluent.

The analyses were conducted from several different perspectives. First, we analyzed the progress of the participants' understanding of the classification problem throughout the experiment as they received more information with the explanations. We also compared how each type of visualization impacted the

participants' confidence and correctness. Additionally, we assessed whether the instances given for classification had differences in difficulty for classification. Finally, we evaluated how the textual explanation affected the participants' understanding of the problem.

When observing the progress of each participant throughout the questionnaire, in most cases, there was no improvement from the step where the respondents did not have any explanation for the instances they were asked to classify to the step where they had the visual explanation, regardless of the type of visualization. Table 4.2 presents the percentage of correct classifications in the first part, when the participants had to classify the instances without having any explanation; in the second part, when they had only the visual explanation; and in the third part, when they had the visual and the textual explanation. In the last part, instead of presenting some improvement, as expected, since they also had the textual explanation to help understand the visualization, there was a significant drop in the percentage of correctly classified instances. We hypothesize this occurred (i) because the textual explanation generated some confusion for the respondents, (ii) because the instances presented in this part were harder to classify, or (iii) because of fatigue.

Table 4.2: Percentage of correct classifications in each step of the experiment

No explanation	**Visual explanation**	**Visual & textual explanation**
50%	59%	9%

Table 4.3 shows the overall percentages of correct and incorrect classifications per instance. This table was built to evaluate which instances are possibly easier or more difficult to classify. We found this analysis necessary when we observed that some instances tended to be more wrongly classified than others, regardless of whether they had an explanation along with them or not. That varying difficulty can happen because instances that present extreme values or significant differences between variable values can be more "obvious" to classify than others with mild values or less variance between variables. The analysis of the impact of the explanations can be affected by this difference in difficulty, since this condition can wrongly indicate that explanations are not helpful when the instances that do have explanations can simply be more difficult, for example.

The data suggests that instances 3, 7, and 8 are possibly more difficult to classify since they have been wrongly classified in over 90% of the cases.

When comparing the visualizations, the one that presented the highest overall success rate was the bar chart, with 40.6% of the questions correctly

Table 4.3: Overall percentages of correct and incorrect classifications per instance

Instance	Correct	Incorrect
1	50%	50%
2	**59%**	41%
3	9%	**91%**
4	**68%**	32%
5	**91%**	9%
6	**64%**	36%
7	0%	**100%**
8	0%	**100%**

answered. The second one was the beeswarm plot, with 37.5% of correct answers. Last was the waterfall plot, with only 25% of questions answered correctly (see Table 4.4). We note, however, that no visualization had a performance above 50%.

Table 4.4: Overall performance of each visualization type

Visualization type	Correct	Incorrect
Beeswarm	37.5%	62.5%
Bar	40.6%	59.4%
Waterfall	25%	75%

Table 4.5 presents the percentage of correct answers for each instance and each visualization type. With this view, we can now see that instance 1, along with the previously seen instances 3, 7, and 8, is probably more difficult to predict than other instances, regardless of the visualization type.

In Table 4.6, we present the percentages of confidence levels chosen by the respondents for each visualization type. The beeswarm plot generated the lowest confidence out of all the visualizations. The bar and the waterfall plot had the same percentage of high-confidence choices. However, the bar plot also had the highest percentage of low confidence. Thus, it is possible to conclude that out of these three visualizations, the one that generated more confidence overall was the waterfall plot (considering the confidence level ≥ 3), followed by the bar plot, and last, the beeswarm plot. We believe this is because bar and waterfall plots are much more common than beeswarm plots. Furthermore, they represent explanations for each predicted instance, while the beeswarm plot represents a global behavior of the model. For these reasons, the beeswarm plot can be harder to interpret than the others, even for people with a strong mathematical and statistical background.

Table 4.5: Percentage of **correct** answers for each instance and each visualization type

Instance	**Beeswarm**	**Bar**	**Waterfall**
1	0%	0%	0%
2	66%	75%	75%
3	33%	0%	0%
4	66%	50%	0%
5	100%	100%	75%
6	33%	100%	50%
7	0%	0%	0%
8	0%	0%	0%

Table 4.6: Distribution of respondents' confidence levels for each visualization type

Visualization type	**1-2**	**3**	**4-5**
Beeswarm	48%	26%	26%
Bar	36%	22%	42%
Waterfall	17%	42%	42%

However, it is interesting to see that the waterfall plot, which generated the highest confidence, was the one with the least correctly answered classifications. That can be an attention point since it may indicate that this kind of visualization can be misleading in some cases.

After the previously described observations, we performed statistical tests to verify whether there were significant differences between the answers to the question set without the explanations, the one with the graphical explanations, and the one with the graphical and textual explanations. The test gathered the number of correctly answered questions in each set and compared them pairwise using the Mann-Whitney U Test (Hart, 2001). Table 4.7 presents the p-values generated by the tests that compare the question sets that had the instances without explanations with the question sets that had the visual explanations and the one that had the visual and the textual explanations. The null hypothesis of each test is that the two compared groups are drawn from the same population. Both tests presented a p-value greater than or equal to 0.05; therefore, we could not reject the null hypothesis, meaning there is no significant difference between the conditions. That indicates that there is no significant difference in the number of correctly answered classification questions between the groups that do not have explanations and those that do have.

Similarly, we performed the same statistical test, but this time, we

Table 4.7: Results of the Mann Whitney U Test of responses without and with explanations

Pair of question set	**p-value**
Without explanation vs With visual explanations	0.77
Without explanation vs With visual and textual explanations	0.05

compared the groups of questions with each of the three visualization types. The results are presented in Table 4.8. Once more, the p-values were greater than 0.05; therefore, there was no significant difference between the conditions.

Table 4.8: Results of the Mann-Whitney U Test of responses using different visualizations as explanations

Pair of visualization type group	**p-value**
Beeswarm plot vs Bar plot	0.74
Beeswarm plot vs Waterfall plot	0.73
Bar plot vs Waterfall plot	0.49

The statistical tests thus showed that the availability of explanations did not make a difference in the number of instances correctly classified, nor did the type of visualization used in the graphic explanation. As we had a small number of responses and instances, the power of the tests was low, so there is a need to repeat the experiment with more instances to classify and try to get a larger number of responses to achieve more robust results.

The step with the textual explanation was the hardest for most respondents, considering they made more mistakes and presented less confidence in these questions. In these cases, we hypothesize that (1) the presented instances may have been too hard to interpret, or (2) the textual explanation may have made the respondents more confused instead of helping them to interpret the visualization. From examining their answers to the open-ended questions, we believe some participants who claimed advanced knowledge of Machine Learning and model interpretability understood the visual explanation; however, they still got most of the answers wrong. Additionally, one of the participants said that they did not understand the visualization when they did not have the textual explanation, and thought they had understood it better when they had it in hand. However, they got more wrong answers in the latter case. All of those observations, along with the analysis presented in Table 4.3, which showed two of the instances used in this step as some of the hardest ones to predict, made us believe that the instances selected for that part of the experiment were indeed too hard to predict, and this difficulty could not be overcome by the provided explanations.

We observed another issue with the understanding of what was being asked in the questionnaire, since some respondents said they were unsure if they should answer what they thought to be the right classification or the one that they thought the model would choose. The number of instances provided for observation seemed to have been an issue as well; some of the respondents said that they needed more instances to understand the problem better.

With the results obtained, we understood that the visual explanations given by the Python SHAP library can be hard to understand, even for people with a mathematical and/or statistical background who claim to have a good understanding of machine learning and model interpretability.

4.3 Discussion

As we observed, explanations presented by existing explanation methods can be hard to understand even for people in the mathematical and statistical fields; consequently, it can be even more difficult to achieve their objective when given to lay users. In scenarios where machine learning models are being used for critical tasks that require a high level of trust and transparency, those methods can still be insufficient for clarifying how a model reached a certain decision and ensuring it can be trusted. These methods have started paving a path to improving machine learning models' interpretability, but there is still a long way to go in order to actually achieve full transparency, trust, and accountability.

The experiment performed in this study suggests that the explanations generated by the SHAP library do not help improve the model's interpretability in some cases. That can be because the investigated visualizations are hard to interpret, thus inefficient in helping users understand the model's behavior. Furthermore, we found that some visualizations can be misleading since they confer more confidence to the user, even when they are mistaken in their classification.

A few issues with the experiment made it not fully conclusive. Future work should reproduce the experiment, giving more instances as examples to the respondents so that they can get a wider view of the problem, and also more instances to classify so that we get more data points to analyze. Furthermore, there should be a way to guarantee that the selected instances have similar difficulty levels so that the complexity factor does not impact the experiment and that the results can be more reliable. We also felt that a qualitative evaluation of the visualizations was lacking and should be performed in further studies.

Despite the issues observed in the experiment, it made us confident that XAI and VA researchers should work collaboratively to develop more efficient visualizations to help the user interpret ML models, regardless of whether they are end users, ML specialists, or developers. Additionally, the proposed visualizations should always be evaluated with various audiences since each could understand the explanations differently, depending on their background and domain knowledge.

5
Proposal

We previously concluded that XAI researchers need to work together with VA experts to propose efficient visualizations for ML model explanations and suggest the need to evaluate the proposed visualizations by testing them with users from varied backgrounds. Most of the existing methods were developed targeting the needs of developers or data scientists and did not consider the needs of the final user. Having this in mind, in this work, we developed a process to create new visualizations for the SHAP explanations by working cooperatively with Information Visualization researchers and evaluating these new visualizations in order to assess their effectiveness. Based on this study, our aim is to propose a process that will help XAI researchers develop more effective explanation visualizations to be used by users from various backgrounds.

In this chapter, we present our study methodology, describing each step of the study, presenting the visualizations that were generated, and how we evaluated them.

5.1
Methodology

Our study was structured into four stages: (i) individual study sessions with Information Visualization (InfoVis) researchers, to develop the first versions of the visualizations; (ii) co-design sessions with all InfoVis researchers, to produce the final versions of the visualization; (iii) questionnaire with users to evaluate the new visualizations' effectiveness; and (iv) analysis of the evaluation's results. The development of visualizations for a specific goal should consist of two or three iterations, each being composed of a series of design meetings.

It would be ideal to have final users involved in the development also, since they would have deeper domain knowledge and experience with the issues encountered in the model utilization (Morelli et al., 2021). However, in the context of our study, we did not focus on a specific domain. We intended to develop visualizations that could be used in many domains. In our approach, we did not work with final users to develop the visual representations of the explanations; however, we worked with InfoVis researchers, who specialize in

developing visual representations that will be better comprehended by a target audience. Since co-design is a flexible approach, which should adapt to the goals of each project (Morelli et al., 2021), we found that having a group of InfoVis researchers for our purposes would be appropriate. Our evaluation approach includes a qualitative and a quantitative assessment. We consider the aspects gathered by Aechtner et al. (2022) to evaluate the explanations conveyed by the visualizations.

5.1.1 Study sessions

The first step of our study was to invite InfoVis researchers and develop new visualizations for SHAP explanations with their help. The participants were three researchers: two doctors and a doctoral candidate. A crucial aspect of co-design sessions is to have a clear understanding of the objective of the visualization (Morelli et al., 2021). Therefore, to start up the study session, the participants were presented with the SHAP method and a use case as an example to illustrate how the method works. We used as an example an instance from the same dataset that we used in the preliminary study (Smith et al., 1988). We explained how the method works and exemplified it with an instance from the Diabetes dataset, showing them the values of the instance and the SHAP values of each feature for that instance. The instance with the shap values that were presented to the participants is in Table 5.1. We also gave them the real diagnosis, which was negative, and the model prediction, which was 0, which means that it was correct.

Table 5.1: Instance used as example in the study sessions

Feature name	**Value**	**Shap value for the positive class**
Number of pregnancies	2	-0.035
Glucose	115	-0.043
Blood pressure	64	-0.033
Skin thickness	22	-0.010
Insulin	0	0.001
Body Mass Index (BMI)	30.8	0.033
Diabetes Pedigree Function	0.421	0.019
Age	21	-0.018

Then, we listed all the information that the method provides, which are:

- The value of each feature for each observation of the training set
- The shap values of each feature for each observation of the training set
- The real class for each observation of the training set

- The predicted class for each observation of the training set
- The expected value of each class for the model

Finally, we asked them a general question: *How would you create one (or more) visualization that helps the user to have a better understanding of the model's behavior?* Along with that question, we listed a few support questions to help them construct the visualizations. These questions are in Appendix B.

After the three study sessions, we had a few drafts of the visualizations developed by the researchers.

5.1.2 Co-design session

The next stage of the study was the co-design session, where we gathered the three researchers who developed the first versions of the visualizations, intending to evolve them into final versions that meet the requirements for the explanations. It is imperative that the objective of the visualization is clear and put in simple sentences before the co-design session begins, and that the crucial information to be represented in the visualizations is prioritized (Morelli et al., 2021). For this reason, we decided to use the aspects that Hoffman et al. (2018) listed as those that would demonstrate explanation goodness and that were simplified by Aechtner et al. (2022) into five questions that can be used to evaluate the explanations in terms of these aspects. These aspects and their respective assessment questions are as follows:

- **Understandability**: From the explanation, does the user understand how the model makes a decision?
- **Usefulness**: Is the explanation useful for the user to make better decisions or to perform an action?
- **Trustworthiness**: Does the explanation increase the user's trust in the model?
- **Informativeness**: Does the explanation provide sufficient information to explain how the model makes decisions?
- **Satisfaction**: Does the explanation of the model satisfy the user?

We should note that, during those sessions, the InfoVis researchers (and not real users) answered and discussed these questions. This is a common procedure in inspection methods, where the inspectors assess the quality of an artifact with certain users in mind (in this case, users with some knowledge of machine learning and artificial intelligence).

Additionally, as we analyzed the visualizations provided by the SHAP Python library and visualizations provided by other libraries that utilize other explanation methods, such as the LIME library[1], the ELI5 library[2] and the Skater library,[3] we noticed that they did not include in those explanations elements that showed when the model succeeds or when it misses in a prediction. Apart from understanding how a model works, a user will only be able to fully and appropriately rely on a model if they can explore when and why it fails (Hoffman et al., 2023). Additionally, a good explanation should deliver the appropriate amount of information, must contain only relevant information, and needs to be context-oriented (Ali et al., 2023). For this reason, another factor that we previously decided to include in the visual explanation representations was where the model succeeds and when it fails.

Having these aspects in mind, the group criticized the first versions of the visualizations in terms of the aspects described above and suggested improvements to them. Some of the first versions of the visualizations are in Figure 5.1. After all participants had the opportunity to provide suggestions for each visualization, the group designed the final versions together. The final versions sketched, together with their critiques, can be seen in Figure 5.2.

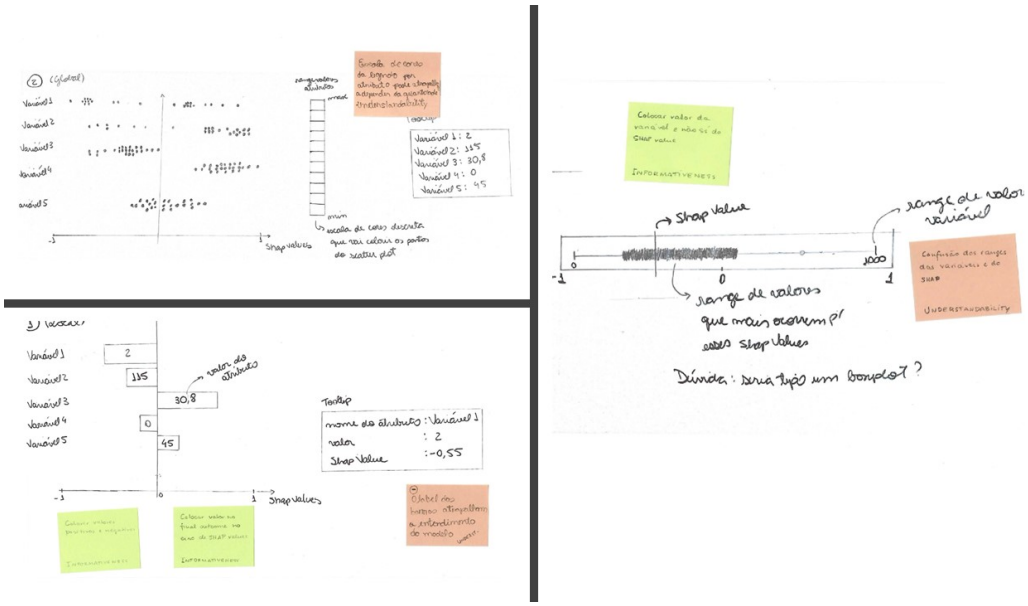


Figure 5.1: First versions of visualizations with reviews

Following the co-design session, we gathered the visualizations and

[1]https://github.com/marcotcr/lime
[2]https://eli5.readthedocs.io/en/latest/
[3]https://pypi.org/project/skater/

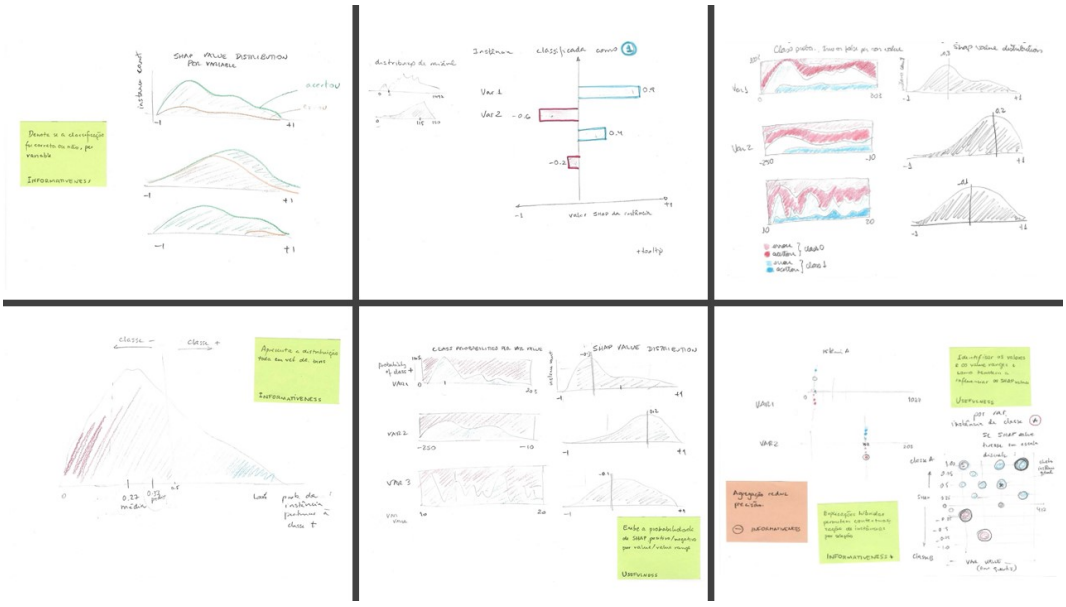


Figure 5.2: Final versions of visualizations with reviews

generated them digitally in a *Jupyter Notebook*[4] using the *Plotly*[5] library in Python. They were generated first by using synthetic instances to ensure that all the elements produced by the researchers were included in the visualizations. For the next step, we chose a dataset that would be used to train a model, generate the explanations, and generate the visualizations.

5.1.3 Evaluation questionnaire

For the evaluation phase, we chose a dataset to be used to train a machine learning model and then generate the SHAP explanations for it.

5.1.3.1 Dataset

We chose a Loan Approval Classification[6] dataset on *Kaggle*[7]. It is a dataset with an unbalanced target variable, which we thought could fairly represent reality, and is a problem of common knowledge. Table 5.2 lists the features present in the dataset. The target variable was the loan status, a binary variable, which would be 0 for *Rejected* and 1 for *Approved*.

This dataset was pre-processed and then utilized to train a machine learning model, for which we then generated the explanations using the *SHAP*

[4]https://jupyter.org/
[5]https://plotly.com/
[6]https://www.kaggle.com/code/nikola6453/loan-approval-classification-accuracy-91-3
[7]https://www.kaggle.com/

Table 5.2: Features of the dataset chosen for the study

Feature name	**Type**
Age	float
Highest education level	categorical
Annual income	float
Years of employment experience	integer
Home ownership status (e.g., rent, own, mortgage)	categorical
Loan amount requested	float
Purpose of the loan	categorical
Loan interest rate	float
Loan amount as a percentage of annual income	float
Length of credit history in years	float
Credit score	integer
Indicator of previous loan defaults	categorical

Python library.[8] During the pre-processing, we performed a scaling process in order to standardize the input variables.

5.1.3.2 Model

For the prediction model, we decided to use the *Random Forest Classifier*[9] implementation of the Scikit Learn library. The data was split into train and test sets, with 70% of the data in the training set and 30% in the test set. The data was then scaled, using the *MinMaxScaler*[10] implementation of the Scikit Learn library, which scales each feature to a given range, for example *(0, 1)*.

Having the scaled dataset, we trained the model with the training set and generated predictions for the test set. The model presented 91% accuracy. The dataset presents a class imbalance, having 78% of the loan status as *Rejected* and 22% as *Approved*. Considering that for this specific problem we need to be sure that no client has a loan rejected when they should have it approved in order to avoid losing these clients, the most critical metrics would be the precision of the negative class (class 0) and the recall of the positive class (class 1), which were 93% and 77%, respectively. The expressions of precision and recall[11] are:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

[8] https://shap.readthedocs.io/en/latest/
[9] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[10] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
[11] https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

We consider that to be a reasonable model since it has good metrics, and is good for our study because it still misses some predictions, which allows us to explore the understanding of the participants of the model's limitations.

5.1.3.3 Selecting the instances for the questionnaire

One of the limitations of our preliminary study was that the instances used in the questionnaire were randomly chosen and assigned to the steps, which created an imbalance that may have biased the results. It is expected that the participants' classification accuracy would be lower for more difficult instances than for easier ones. In order to prevent this from happening this time, we selected instances with similar difficulties. For that, we used the probabilities predicted by the model, which indicate the level of "confidence" in its prediction. Instances with probabilities close to 0.5 are generally more difficult to classify, while the ones with more extreme probabilities, *i.e.*, close to 0 or 1, are easier to classify.

Having that in mind, we considered the instance difficulty as:

$$Difficulty = |P(PositiveClass) - 0.5|$$

Then, we randomly selected instances that presented a difficulty value ranging from 0.4 to 0.5, which means these are instances that should not be hard to predict, and used them to generate the explanation visualizations to be used in the questionnaire. We randomly selected one True Positive, one False Positive, one True Negative, and one False Negative for each part of the questionnaire.

5.1.3.4 Visualizations

After selecting the instances, we then had to generate the visualizations according to the previous stages of the study, now using the data from these instances. The final versions of the visualizations for one of the instances that were used in the questionnaire can be seen in Figures 5.3, 5.4, 5.5, and 5.6.

The first visualization (Figure 5.3) is a bar graph where the bars represent the Shap values of each feature of the model in a prediction. The features are sorted by the Shap value module, which means that the features that have a stronger impact on the prediction, independently of the class they are contributing to, are at the top of the chart, and those that have a lower

impact are at the bottom. Features that have positive Shap values, *i.e.*, that contribute to the positive class, have blue bars, and the ones that have negative Shap values have red bars. The Shap value of each variable is displayed next to its respective bar. Under the title of the chart is highlighted the classification given by the model for that instance.

In the second visualization (Figure 5.4), we have histograms that present the distributions of the values of the model's variables. The values that compose the distributions come from the training set. The values of each feature in the instance being predicted are highlighted with a black vertical dashed line. The colors of the histograms represent the positive (blue) or negative (red) Shap values for the predicted instance, just as in the previous visualization. This visualization allows the user to understand whether the value of the variable in that instance is a common value or if it is an outlier, for example, and thus try to understand how this value may impact the prediction.

In the third visualization (Figure 5.5), there are two sets of graphs. In the set on the left column, we have the distributions of the features' values once again, but this time we have emphasized in different colors the count of each predicted class, and for each class, the count of cases where the model succeeded and the cases where it missed. The cases in which the model classified as *Approved* and was successful are colored in dark green, the cases in which it classified as *Approved* and failed are colored in light green, the cases in which it classified as *Rejected* and was successful are colored in dark orange, and the ones it classified as *Rejected* and failed are colored in light orange. The value of each feature for the instance being predicted is highlighted in a black vertical dashed line. The charts to the right of the visualization are distributions of the Shap values of each feature, and the Shap value of each feature for the instance being predicted is also highlighted in a black vertical dashed line. The minimum and maximum values of the feature values and of the Shap values are emphasized. That allows the user to instantly visualize which features have a higher variation of the Shap values, which ones have higher impact, and which have lower impact in general. The values for the distributions for this visualization also come from the training set.

The fourth and last visualization (Figure 5.6) shows the concentration of instances for each Shap value range and feature value quartile. The interval where the feature lies in that instance is highlighted with an X, and the background color indicates if the Shap value in that interval contributes to the positive (blue) or negative (red) class. This visualization can assist the user in understanding which value intervals of each variable produce a higher impact in the prediction, as well as which intervals contribute to the positive

class and which contribute to the negative class, since a single feature may contribute to both classes depending on its value.

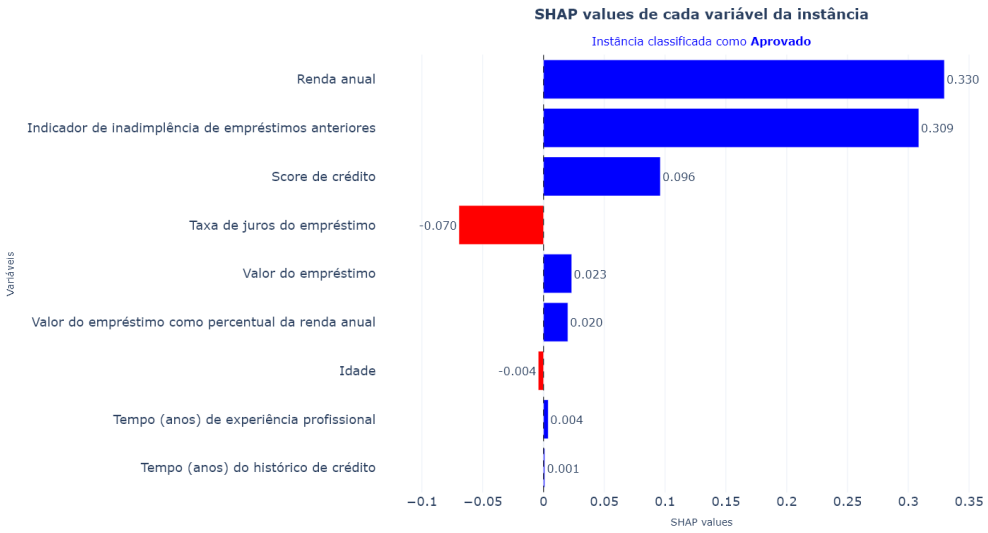


Figure 5.3: Bar plot of the SHAP values for an instance

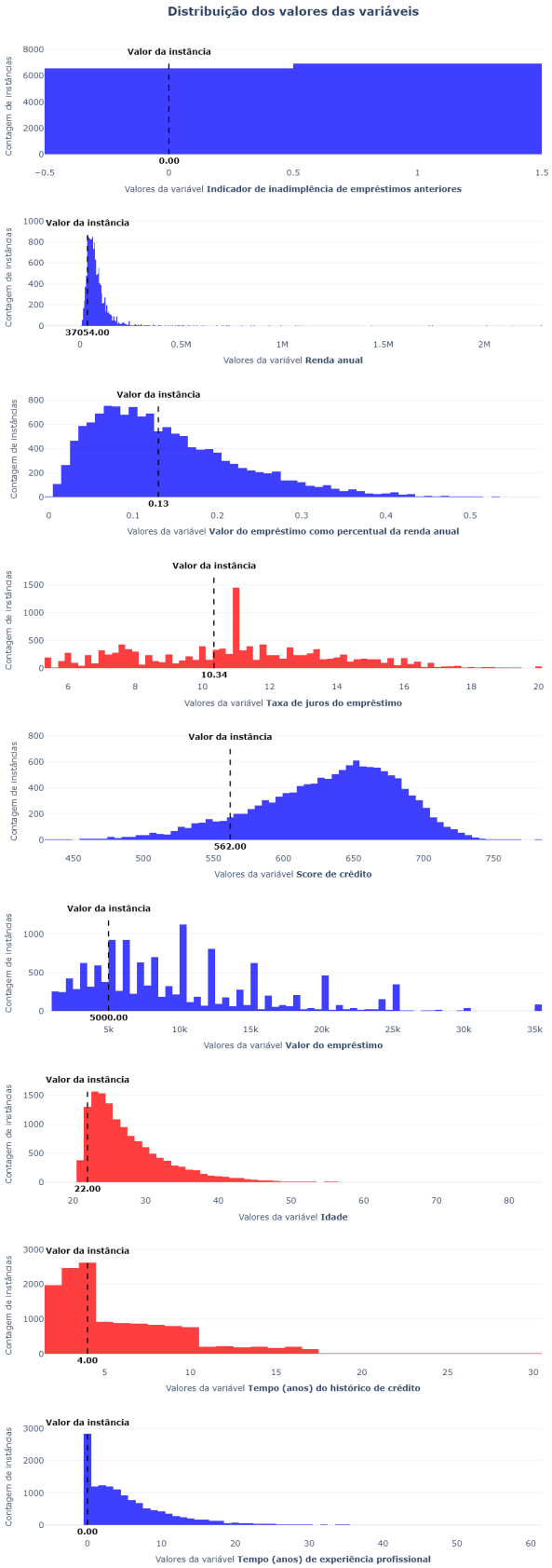


Figure 5.4: Distributions of the features values for an instance

Figure 5.5: Distributions of the feature values divided by class and model success or failure, and SHAP values distributions for an instance

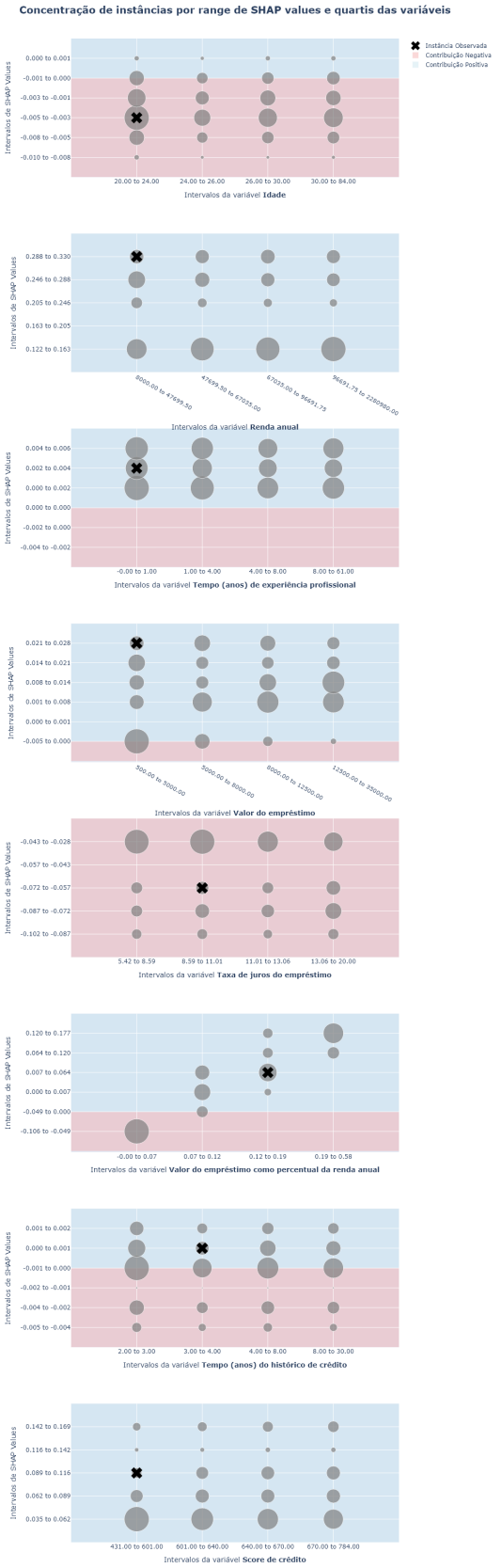


Figure 5.6: Density of instances for each SHAP value range and variable value range for an instance

5.1.3.5 Questionnaire

In this section, we describe how the questionnaire was structured. We chose the questionnaire with the purpose of having more answers than we would be able to get in interviews, since we would not have the time and resources to conduct numerous interviews. It was divided into five parts: (i) problem presentation and Informed Consent Form; (ii) profile questions; (iii) instance observation and prediction without explanations; (iv) instance observation and prediction with explanations and their evaluation; and (v) visualizations comparison. For the evaluation, we chose to have the classification along with the assessment questions because, as Hoffman et al. (2023) said, *"A prediction task can serve as a method for peering into users' mental models, especially if its application is be accompanied by a confidence rating and a free response elaboration in which the users explain or justify their predictions, or respond to a probe about counterfactuals."*

Initially, we presented the SHAP method, gave a simple explanation of its functioning, and described the concepts of *understandability, usefulness, trustworthiness, informativeness, and satisfaction*, which we intended to use to evaluate the explanations in the following parts of the questionnaire.

Then we gave them a setting in which they should imagine that they were placed so that their prediction evaluations could be more contextualized, as done in the work of Aechtner et al. (2022). We asked them to imagine that they work in a bank in the loan approval department. Each loan request is evaluated by a Machine Learning model, which decides whether the loan should be approved or not. We then present them with the features used in the model and introduce an issue. We say that frequently, especially when loans are rejected, clients request explanations for the refusal, and therefore, they need a way to understand how the model reached the decision, so that they can give their clients an explanation. Additionally, we highlighted how important it is to be aware of how the model makes its decisions so that they can know when it is mistaken.

After that, the following sections were described, so that they knew what to expect from the questionnaire, and the Informed Consent Form was displayed. The following sections were only presented to the participants who agreed to the terms.

The next section had the profile questions. We asked the participants what their instruction level was and whether their field of study was STEM (Science, Technology, Engineering, Math). We also asked what their level of knowledge was of Machine Learning/Artificial Intelligence, XAI (Explainable

Artificial Intelligence), and Information Visualization. To conclude that section, we asked what the participants considered to be the most important aspects of a visual representation of a prediction explanation and let them select all that they think apply from the five aspects previously described (understandability, usefulness, trustworthiness, informativeness, and satisfaction). Finally, we asked an open question for them to describe what they expected from a visual representation of a prediction explanation.

In the third part of the questionnaire, we started by presenting them with four instances, giving them the features and their values, and the model classification along with the right classification. We asked them to examine this information and try to gain some knowledge of how the model makes decisions. After that, we gave them four new instances, now only presenting the features and their values, and asked them to classify those instances according to how they believed the model would classify them. After each classification, we asked them how confident they were in their prediction by using a 5-point Scale, where 1 means *"Low confidence"* and 5 means *"High confidence"*.

The next part was divided into an observation task and a prediction task. For the observation task, we presented four new instances, but instead of their feature values, we presented only the model classification, the correct classification, and the explanation visualizations. The purpose of the observation task was for participants to learn about the explanation visualizations and which explanation visualizations were associated with correct or incorrect classifications. For the prediction task, we presented another four instances and the corresponding explanation visualizations and asked them to classify each one according to the class they believed the model would choose. We asked again a question of level of confidence for each predicted instance, but this time we also posed ten more 5-point scale questions. Five were assertions about the visualizations that described them as being understandable, useful, trustworthy, informative, and satisfying, and they had to score their degree of agreement with that assertion in a 5-point Likert scale (1 meaning *"Disagree"* and 5 meaning *"Agree"*). Since the visualization set was composed of four visualizations, for each one we added an assertion for he participant to rate how much they believed that visualization contributed to the previous questions, also in a 5-point scale (from 1, meaning *"Did not contribute at all"*, to 5, meaning *"Contributed fully"*). Finally, we added an open question asking them to give their opinion on the understandability, usefulness, trustworthiness, informativeness, and satisfaction engendered by each visualization or combination of visualizations.

In the last section, we presented the two visualizations that the SHAP

library provides for local explanations as a set of visualizations for an explanation of an instance, along with the set of visualizations generated in our study as a second visualization set, so that they could compare the efficacy of each set. To avoid biasing the results, we did not tell the participants that one of the visualization sets was from an existing library and that the other was designed for this study. We gave them the real classification and the model classification, and posed six questions. The first five were about the five target aspects of explanations, and each asked them which of the two sets better met the requirements of that aspect. They could choose between *"Set 1"*, *"Set 2"*, *Both* or *Neither*. To conclude, we asked again the same open question that we asked after each prediction in section four, but now concerning each set of visualizations.

We conducted a pilot test of the questionnaire to ensure that everything we asked was clear and to avoid any confusion issues. After the pilot test, we made some adjustments. We made clear that the instances of each section were different and independent, we highlighted that the predictions we wanted from the participants were what they believed that the model would give as output, we adjusted the legend of one of the visualizations that was not clear and added the last open question in the comparison section.

The questionnaire was created in Google Forms[12] and was completed remotely. It was available for answers for two weeks. After that time, we closed the form and gathered the data for analysis, which will be described in the next chapter. But we also held two in-person study sessions, described in the next section.

5.1.4 Extra study sessions

During the analysis of the responses to the questionnaire, which will be detailed in the next chapter, we saw that many of the respondents were confused by the visualizations, and some said that they needed an explanation of the visualizations' functioning and how they relate to each other. We decided then to have two new study sessions with new participants in order to test whether, after having a previous explanation of the SHAP method and the visualizations that they would use to try to predict the instances, the participants would have a better performance, higher confidence, and rate the explanations higher.

The first study session gathered undergraduate students who were taking an Introduction to Information Visualization course. We gave them an

[12]https://docs.google.com/forms/

introductory presentation, where we briefly explained prediction models, the XAI field, and the SHAP method, and presented the visualizations they would encounter in the questionnaire. We described the visualizations in detail and answered any questions they had. After that, we gave them a printed version of the visualizations and asked them to write on the paper and highlight any element of the visualizations that they did not understand, that they thought helped understand the model, or that did not help at all.

For the second study session, we had the visualization researchers who helped us to develop the visualizations answer the questionnaire as well. The idea was to have people who understood the visualizations deeply and who were also familiar with the issue of the explanations evaluation, so that we could see whether their performance would differ from the other participants.

Both groups also received a printed version of the visualizations on A3-sized paper. We asked them to highlight any elements in the visualizations that they found helped them or that they did not understand and to write their impressions.

6
Results

In this chapter, we present the results of the study described in the previous chapter.

6.1
Participants' profiles

In total, we had 15 responses: 7 for the first part, where the participants only had the questionnaire, 5 for the second, and 3 for the third, both having either previous knowledge or an explanation of the visualizations before responding. 5 of the respondents are either doctors or doctoral students, 5 are either master's or master's students, 5 are undergraduate students, and 1 has finished their undergraduate studies. The participants of the first extra study session are all graduate students, while the ones in the second extra study session are all either doctors or doctoral students.

One of the participants claims not to have a STEM (Science, Technology, Engineering, Math) field of study. Most participants have intermediate knowledge of Machine Learning/AI, no to basic knowledge of XAI, and basic to intermediate knowledge of Information Visualization. The distribution of each of these knowledge questions is in Figure 6.1.

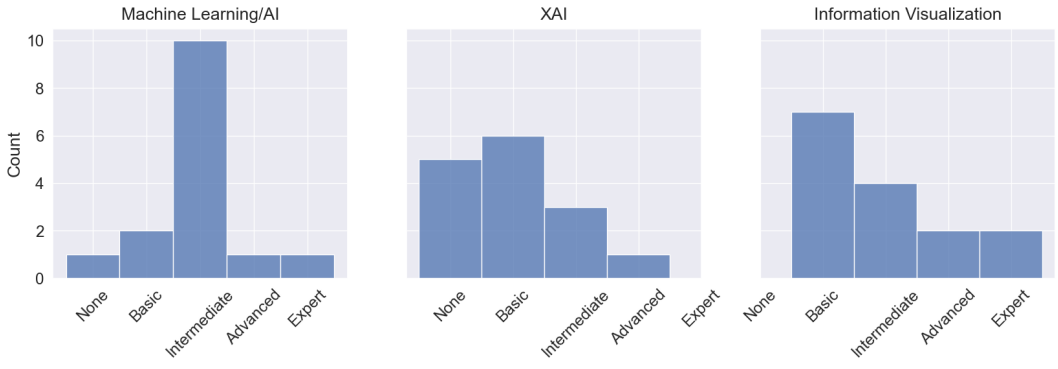


Figure 6.1: Distributions of the knowledge levels for each field

Finally, we asked respondents to say what they expected from the visual representations of the model explanations. The answers to this question indicate that they expect the visual representations of the explanations to be

clear and intuitive and to facilitate the comprehension of the model's decision process. They highlighted the requirement for transparency, allowing users to understand which information influenced the prediction and how the variables affected the result. Furthermore, they mentioned the need for support in the decision-making process by reducing uncertainty and improving trust in the model. To satisfy diverse users, including experts and non-technical users, they suggested that there should be a combination of graphical and textual elements so that the explanations could be accessible. Lastly, they emphasized the need to balance the amount of information given and the ease of interpretation, thus avoiding cognitive overload.

6.2 Confidence level

After each prediction, we asked participants about their confidence level in their response on a 5-point scale, where 1 meant *low confidence* and 5 meant *high confidence*. The general average confidence for the predictions where they did not have the explanations was **3.0**, and for the predictions where they did have the explanations, it was **3.11**.

When we calculate the mean confidence of the participants in each of the three groups (participants who only had the questionnaire, graduate students with previous explanation, and visualization researchers with previous knowledge), we can see a difference between the groups. First, we can see that the mean confidence in both stages (without explanation and with explanation) was much higher for the first group (3.43 without explanation, 3.27 with explanation), which did not have previous knowledge of the problem, the SHAP method, or the visualizations. The group with the lowest confidence for both stages was the third one (2.5 without explanation, 2.92 with explanation), which was composed of visualization researchers who had a deep knowledge of the visualizations used and previous knowledge of the SHAP method. The confidence distributions for each group at each experiment stage can be seen in Figure 6.2.

We divided the confidence scale into *Low* (1 and 2), *Medium* (3), and *High* (4 and 5). As seen in Figure 6.3, the percentages of low and medium confidence were lower when they had the explanations, and the percentage of high confidence was higher in that case.

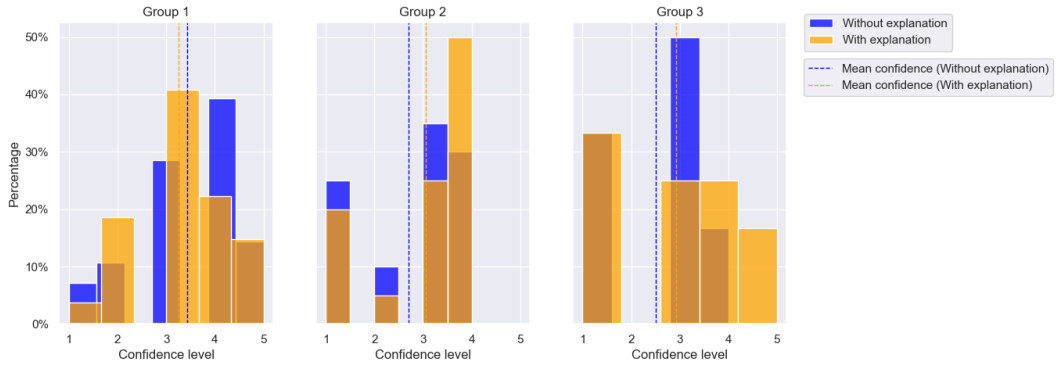


Figure 6.2: Distributions of the confidence levels for each experiment group

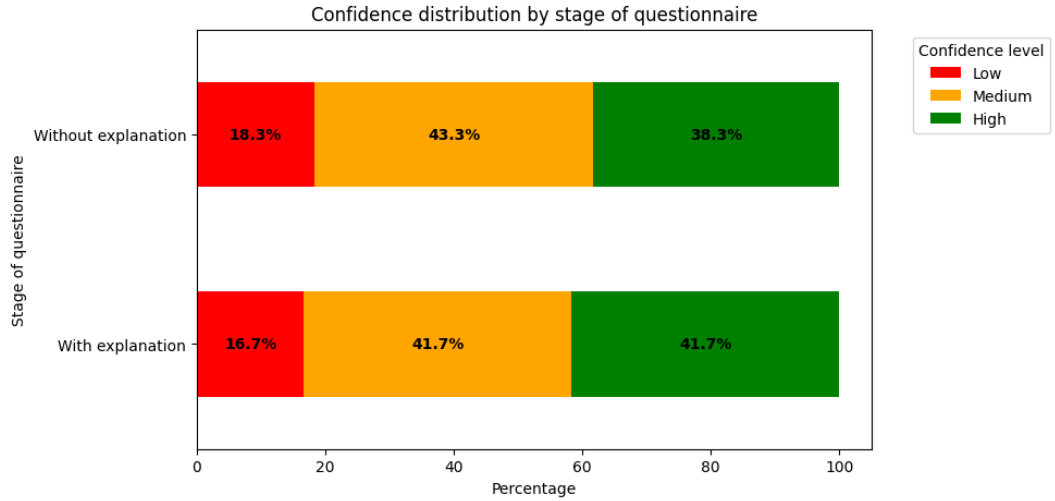


Figure 6.3: Distributions of the confidence levels

6.3 Performance

Ideally, when someone has higher confidence in their response, they should be successful more frequently than when they are not confident, or else their confidence is deceiving them. By observing the general results of the predictions, there was a slight improvement in the success rate when they had the explanations. Without the explanations, they got **23.3%** of the predictions right, while when they had the explanations, they succeeded in **28.3%** of the responses.

When we looked at the groups separately, the first and the third groups performed better in the second stage, when they had the explanations. The first group, which had higher confidence for the first stage of the experiment, presented a higher success rate for the second stage. The second group also had an inversion of confidence and performance: in the second state, they

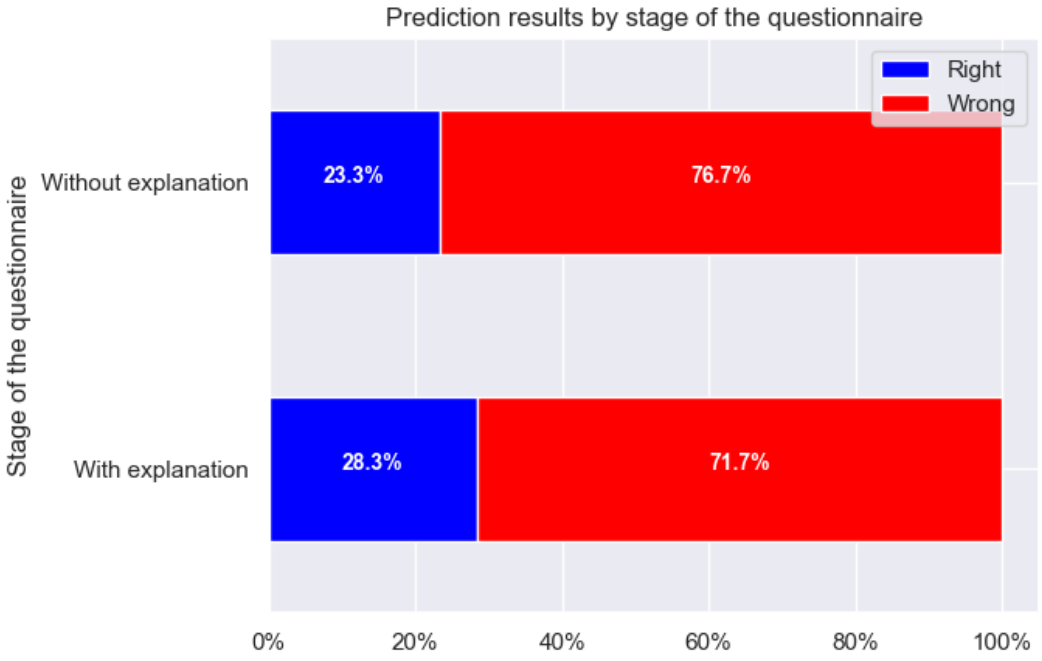


Figure 6.4: Results of the predictions without explanation and with explanation

presented higher confidence but a lower success rate. The only group where confidence and performance were coherent, that is, the group that presented a higher success rate when they had higher confidence, was the third group. Furthermore, the third group had the most expressive improvement when they had the explanations, which makes us believe that having a deeper knowledge of the visualization is of huge importance when we talk about improving the interpretability of a model through visual representations of explanations. All results can be seen in Figure 6.5.

As was done in the preliminary study, we also did a Mann-Whitney U Test (Hart, 2001) to verify whether there was a statistical difference between the number of correct answers for each instance when they did not have the explanations and when they had the explanations. We did the test considering all the answers and then divided them into three groups. In all cases, the p-value was 1.0, meaning there was no significant difference between the samples.

6.4 Correlation confidence vs performance

As said above, higher confidence should mean better performance, and the contrary could be risky in the context of a critical task. It could mean one would trust a model when it is inaccurate. Therefore, we investigated the indication that participants were wrongly confident in the first set of predictions.

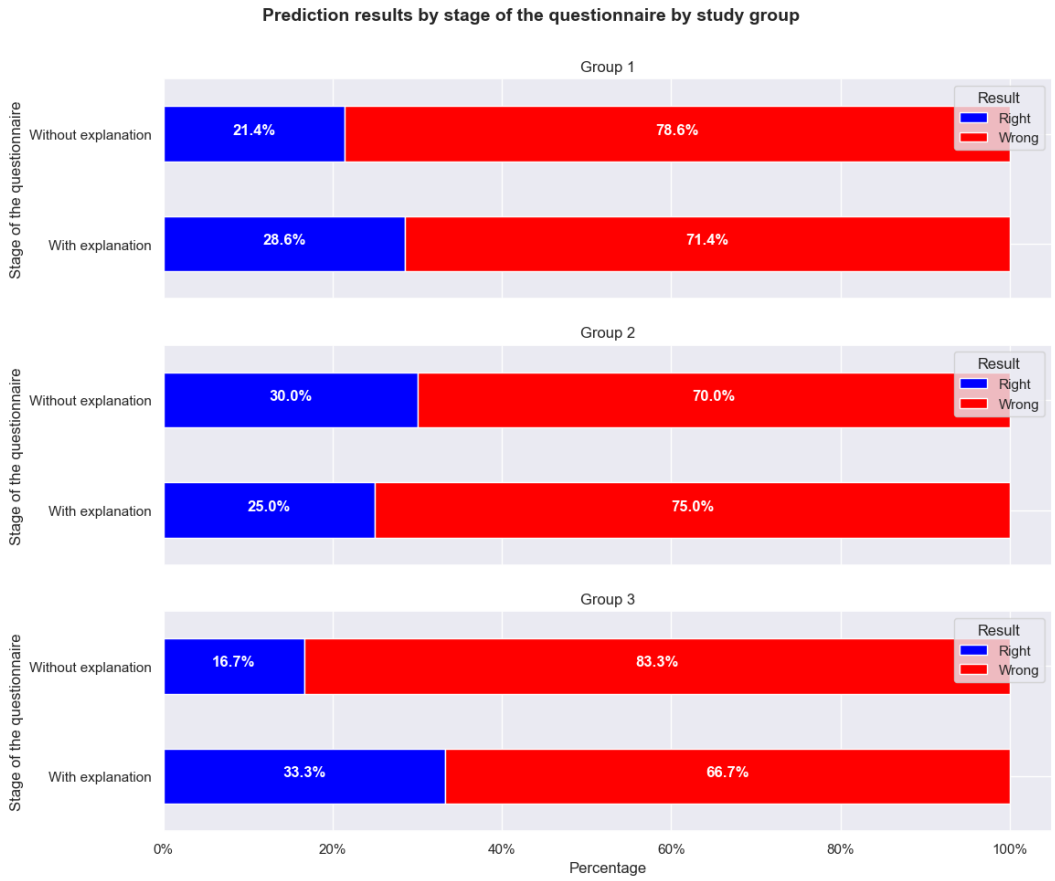


Figure 6.5: Results of the predictions without explanation and with explanation by study group

In order to do that, we used a Fisher's exact test to get the correlation between the result of their response (right or wrong) and the confidence level (low, medium, or high). The Fisher exact test is used to verify whether two categorical variables have a significant relationship when at least one value in the contingency table is lower than 5. Since the Python implementation for that test only accepts 2 x 2 contingency tables, and we have one of the variables having three categories, we considered medium and high as one category and low as the other category since we believe that some amount of confidence, even when it is not high, can be deceiving if it is misconceived. The method generates the p-value and the odds ratio. The test's null hypothesis is that there is no relationship between performance and confidence. Considering a significance level of 0.05, if the p-value is less than 0.05, we can reject the null hypothesis and say that there is some relationship between the two variables.

The Fisher exact test was conducted for the participants' responses in the condition without explanations and then for that with the explanations.

We used the Fisher exact method from the Scipy[1] library. For the first set of responses, the p-value was 1.0, and for the second, the p-value was 0.448. Neither was less than 0.05, so we could not reject the null hypothesis, meaning that the two variables are unrelated.

6.5 Explanation quality factors

In the second part of the experiment, when we gave the participants the visual representation of the explanations, for each of the instances that the participants had to classify according to what they believed would be the model's prediction, we asked them to rate in a 5-point Likert scale of 1 to 5 how much they agreed with some assertions. Each assertion was related to one of the previously determined explanation quality factors (understandability, usefulness, trustworthiness, informativeness, and satisfaction). The objective was for them to indicate how much they thought the explanations met these criteria for explanation quality.

The factor with the highest percentage of low agreement was satisfaction, with 49.1% of the responses being 1 or 2. It also had the lowest percentage of high agreement, 35.6%. Informativeness was the factor that presented the lowest difference between levels of agreement, having 35.6% of low agreement, 27.1% of medium agreement, and 37.3% of high agreement. Understandability, usefulness, and trustworthiness were rated with high agreement at most, with trustworthiness the factor with the highest percentage of high agreement, 50.8%.

Those results show once again what we saw in the confidence versus performance association, where the satisfaction with the visualizations was low since the participants felt some uncertainty and difficulty in interpreting the visualizations, but even so, they still feel that the explanations engender more trust in the model.

With the open question after each prediction, we could get a better sense of what the participants felt about the visualizations and how successful they were in assisting them in better understanding the model. In terms of *understandability*, there was a divergence of opinion since some of the visualizations were considered helpful for a better understanding of the impact of the variables in the predictions, but others were considered unclear.

As to *usefulness*, some participants recognized that the visualizations help identify the most influential variables. However, there were some criticisms of their effectiveness in predicting the model's output. Some answers indicate

[1] https://docs.scipy.org/doc/scipy-1.15.2/reference/generated/scipy.stats.fisher_exact.html

that even after analyzing the visualizations, the participants were still not confident in their predictions, which implies that the practical usefulness of the visualizations might be limited, or they need improvement.

The *trustworthiness* engendered by the visual representations differed substantially. While some visualizations improved the participants' trust in the model by providing a clear vision of the variables' value distributions, others had the inverse impact. Some reports said that some visualizations made the understanding of the variables' impact more difficult, decreasing the confidence of the participants when making decisions based on the explanations.

Informativeness was also a divergent factor: while some found the visualizations presented redundant or confusing information, others highlighted that these same visualizations contributed to the comprehension of the impact of the variable in the decision. The combination of different visualizations was mentioned as a strategy for having a more complete vision of the model's decision process.

Finally, the *satisfaction* with the visual representations was directly influenced by the clarity of the provided information and the possibility of understanding the model's reasoning. The participants who could extract relevant insights demonstrated higher satisfaction, while others who encountered difficulties interpreting the charts expressed the need to improve the explanations.

While the visual representations of the SHAP explanations offered valuable insights into the model's decision process, their efficacy still faces some challenges. Difficulty of comprehension, information redundancy, and visual ambiguities were frequent criticisms, indicating that some improvement is needed to make them more accessible and reliable to different users.

6.6 Contribution of each visualization

When asked how much each visualization contributed to their prediction, visualization D, in Figure 5.6, was the one they considered to contribute the most, where 48.3% of the responses were at levels 4 or 5 on the Likert Scale. That was also emphasized in the open question, where one of the respondents said that visualization D was the one they relied on to give their response. One of their responses is presented below:

"Through the color palette of visualization D, I looked for what the model understood about the loan status; the other visualizations did not provide enough information for me to give an answer" (freely translated from Portuguese).

Conversely, some respondents said that visualization D was hard to

interpret or ambiguous, mentioning that the colors generated some uncertainty in some cases.

Visualizations A (Figure 5.3) and B (Figure 5.4) had a similar percentage of high contribution ratings, 47.5% and 45.8%, respectively. In contrast, visualization B was said to be insufficient if used in an isolated manner, and A, B, and D were said to have redundant information. Figure 6.6 shows the distribution of the visualizations' contribution ratings.

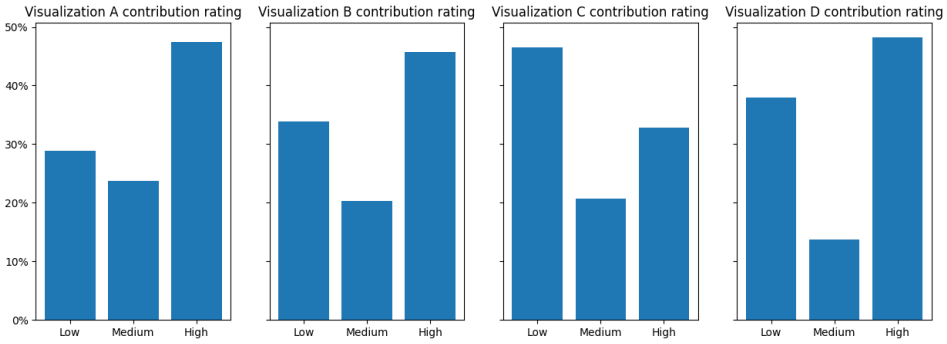


Figure 6.6: Contribution rating of each generated visualization

Visualization A (Figure 5.3) was well perceived regarding friendliness and general comprehension of the model's decision patterns. The feedback in the printed charts showed that this visualization helped participants understand which variables were more important to the model and which were not. Some participants used that knowledge when looking at the other visualizations and analyzing how each important variable impacted the decision. That can be seen in Figure 6.7. Nevertheless, there were some criticisms about its ability to provide detailed information on the impact of each variable on the decisions. When comparing different instances, some observations seemed identical when analyzed only with this visualization, but presented more apparent differences in other visualizations. That implies that visualization A can be useful for capturing global patterns but is insufficient for more detailed analyses of specific instances.

Visualization B (Figure 5.4) got mixed evaluations. Some participants thought it helped them understand which variables impacted the predictions most. By contrast, when presented separately, this visualization was considered insufficient for predicting the model's decisions. On the printed visualizations, some participants mentioned that the colors in this visualization were unclear, and there should be a legend to facilitate the interpretation. Others said that it was their favorite visualization and that it helped them understand when the instance is hard to predict, since they could see when the variable values were in a common range of the distribution or not. Its utility was mainly highlighted

when combined with other visualizations, especially visualization C.

Visualization C (Figure 5.5) helped the participants identify cases of false positives, which might indicate that it provides valuable information about the model's errors and uncertainties. Even so, there was some criticism about the colors used, with some participants saying there should be a more explicit color schema. One of them highlighted in the printed visualization that they did not understand why the SHAP distributions on this visualization did not have a color scheme like the other visualizations did. By contrast, they mentioned that the minimum-maximum range of the SHAP distributions presented in the visualization might be interesting to explore since it might indicate a possible volatility introduced by a variable. The visualization was more useful in providing an understanding of the model's global behavior than in predicting specific instances. One of the participants highlighted (freely translated from Portuguese): *"I believe that there is a utility in the analysis of this data for improving the model, but for understandability it might not be ideal, especially taking visualization C under consideration [...]".*

Visualization D (Figure 5.6) was the one that was most criticized regarding its clarity and informativeness. Although some participants used its colors to interpret the model's reasoning, many reported that this approach did not provide enough information to support the decision-making. One of the participants also said that visualization D seems repetitive when they have visualization B. On the printed chart, one of them said they found that visualization difficult to interpret, but they could see that it coincides with visualization A and makes clear which variables have a negative impact. Furthermore, in some cases, there was some ambiguity in the interpretation, which engendered uncertainty instead of clarity. That made some participants find this the least useful visualization compared to others.

Another issue that was pointed out was that the feature sorting was not consistent throughout the visualizations, which caused more difficulty in analyzing them together.

6.7 Previous expectation vs what was presented

After all the predictions, we asked them if their expectations of the explanations were met. Generally, they said that their expectations were not fully met. Most said that they felt that a textual explanation, together with the visualization, was needed. Two respondents said their expectations were met, but one said they would add a textual annotation to the plots. Some argued that they needed a previous explanation of the visualizations' interpretation

and how they relate to each other. Furthermore, one said they needed a better explanation of how the SHAP values are calculated. Some participants suggested the addition of tooltips or explanatory labels.

Although some visualizations were considered useful for perceiving the model's logic, many participants reported that they could not identify why the instances were classified as positive or negative. Visualizations A and B were considered clearer but insufficient to provide definitive confidence in the interpretation. In addition, visualization C caused some confusion, and some participants did not understand how it could help them interpret the model's decisions. In conclusion, the participants generally had significant difficulty in interpreting the visualizations.

6.8 Comparison with the SHAP library visualizations

In the last part of the questionnaire, we compared the set of visualizations provided by the SHAP library for local explanations with our set of visualizations. We asked them which set met the criteria for each one of the factors used in the evaluation (understandability, usefulness, informativeness, trustworthiness, and satisfaction), and they could choose one of them, both, or neither. The results can be seen in Table 6.1. For all of the factors, most of the participants chose our visualization set as the one to best meet the criteria of understandability, usefulness, informativeness, trustworthiness, and satisfaction.

When comparing the visualizations in the printed charts, one of the participants, *i.e.*, one of the visualization researchers, pointed out that the construction of the waterfall plot (from the shap library) does not add value to the interpretation of the prediction. Additionally, they felt that the only element that added value and was not present in our visualizations was the final value of the prediction, which is the probability of the instance being of the positive class.

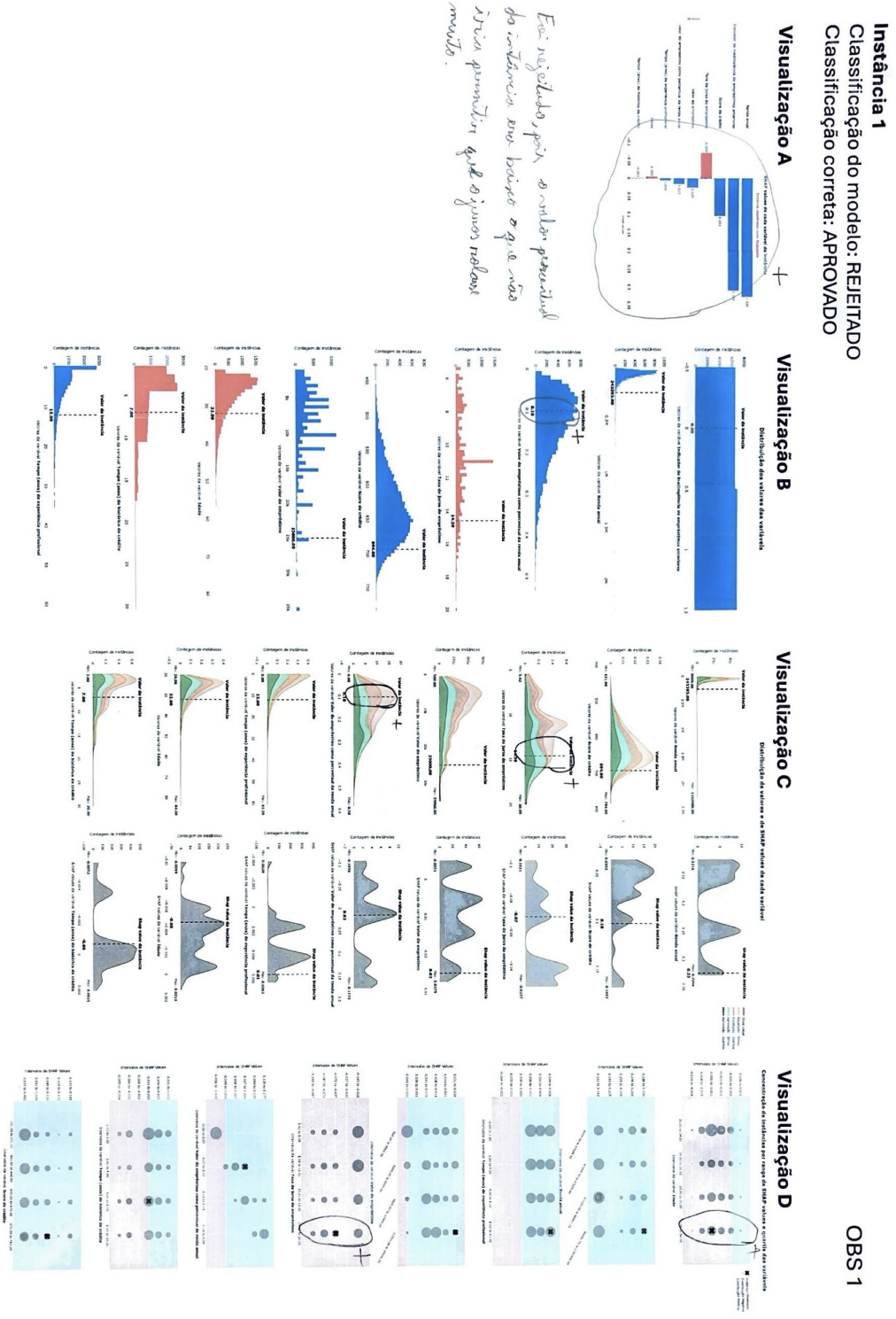


Figure 6.7: Printed version of the charts where a participant highlighted the parts that they considered relevant for the prediction

Table 6.1: Comparison between the visualizations generated in our study and the ones provided by the SHAP Python library

Visualization set	**Understandability**	**Usefulness**	**Trustworthiness**	**Informativeness**	**Satisfaction**
SHAP library visualizations	28.6%	26.7%	20%	20%	28.6%
Our visualizations	**57.1%**	**53.3%**	**60%**	**53.3%**	**57.1%**
Both	7.1%	20%	6.7%	13.3%	7.1%
Neihter	7.1%	0%	13.3%	13.3%	7.1%

7
Discussion

In this chapter, we discuss the results of the study described in the previous chapter and raise some analysis of the participants' confidence level and performance, the quality factors investigated, and the visualizations themselves. We close the chapter with a discussion about the limitations of our work.

7.1
Confidence level

As seen in the previous chapter, the participants presented higher confidence in their prediction when they did not have the explanations than when they had them. However, when we look at the success rate, *i.e.*, when they accurately predicted the model's response, it was slightly higher in the latter. That shows that higher confidence does not mean higher performance, which can be concerning when talking about critical tasks.

The overall confidence was average, close to 3 (on a 5-point scale from 1 to 5), for both stages of the experiment, but we can see that the participants who had a deeper knowledge of the visualizations and of the SHAP method presented higher confidence when they had the explanations. Conversely, the participants presented only with the information in the questionnaire displayed the highest confidence and had an inversion between confidence and performance. Participants with the least previous information presented the highest confidence, and those who had previous knowledge of the SHAP method and a deep understanding of the visualizations presented the lowest confidence. This inversion shows that confidence can be misleading, even when we have the explanations in hand. That can be risky and lead users to trust a model even when they understand it poorly.

7.2
Performance

Generally, the performance improvement engendered by the explanations was low. The group that presented the highest improvement was the one with the visualization researchers, who had a deep knowledge of the visualizations

and previous knowledge of the SHAP method. Nevertheless, their success rate was critically low. We can see that understanding how a model behaves and trying to predict how it would respond is not a trivial task. Even when the user has previous knowledge of the method that generated the explanation and has a deep knowledge of its visual representations, it is an extremely hard task to predict how the model would respond. In contrast, even if one could not predict the exact output of the model, the explanations can be of enormous importance in assisting the user to identify biases, for example.

Having that in mind, we question whether the simulatability property proposed by Doshi-Velez and Kim (2017) is an appropriate method to evaluate an explanation method in non application-grounded evaluation, *i.e.*, where the explanation method is not used in a real application. We believe that the most suitable way to use this property as an evaluation metric would be in a scenario of a real application where the users who participate in the evaluation are knowledgeable users who have a deep understanding of the domain and the problem being addressed by the model.

7.3 Quality factors

The results of the quality factors show low satisfaction, with the participants feeling some uncertainty and difficulty in interpreting the visualizations. By contrast, they show high trustworthiness. That is more evidence that users can to trust a model even when they do not understand it or when they are uncertain of its behavior. That can be extremely risky, especially in highly critical tasks, because one can end up trusting the outputs of a model that is biased or unfair without questioning its decisions.

7.4 Visualizations

We concluded that the explanations alone, without more complementary information such as the probabilities for each class, the distributions of the features, how the instance's feature values relate to the distribution, and the model's metrics, are not as efficient as desired, since this information engenders a more complete vision of the model and its behavior. The explanations alone are generally insufficient to provide this overall understanding of the model and, consequently, more interpretability.

On one hand, we saw that visualizations with complementary information can be very helpful in delivering sufficient data for the user to understand the model's behavior. On the other hand, some visualizations can be hard to

interpret, and different visualizations may engender conflicting insights. The task of generating visualizations to represent model explanations needs to be an iterative task, where users give their feedback and the visualizations get continuously improved, adding missing information or removing information that is redundant or that generates a cognitive overload.

Compared to the already existing and used visualizations generated by the SHAP Python library, our visualization set was preferred in all quality factors. That shows that, even with the results not being highly satisfactory, we managed to have an improvement when compared to an existing set of explanation visualizations. Our visualization set engendered a broader view of the model's behavior by adding complementary information other than only the SHAP values. We also have evidence that we have proposed an efficient process of generating effective explanation visualizations, keeping in mind that this process should be iterative and have the participation of end users in order to improve the visualizations' effectiveness even more.

8
Conclusion

This chapter concludes our work, highlighting its contributions, discussing its limitations, and pointing to future work.

8.0.1
Contributions

After our studies, we reached some interesting conclusions that corroborate or refine previous findings in the literature. First, there is a clear need to dedicate some attention to the development of visual representations of explanations, keeping in mind what makes a good explanation and users' expectations towards these explanations (Alicioglu and Sun, 2022). The support of Information Visualization experts in this development is crucial in order to create effective visualizations, *i.e.*, visualizations that convey explanations that meet the quality criteria considered in this work. Apart from effectiveness, the people who participate in this development should be able to create visualizations that will not be misleading or wrongly influence the users.

The development process also needs to take the users' needs and expectations into consideration. It would be ideal to include end users in the development process so that they could give feedback, and the visualizations could be iteratively improved to fulfill their needs.

The results of our study showed that the support of InfoVis experts in the development of the visual representations of the explanations is indeed crucial. They could construct a visualization set that was considered more appropriate by all the study participants compared to the one provided by the existing SHAP library. Their support allowed us to create a more informative visualization set, which contained complementary information, and also one that engendered more trust in the users. It is not an optimal visualization set, but we believe that it is a step further in the development of more efficient explanation representations.

Regarding the methodology of evaluation of the explanations, it is crucial to also engage end users in this activity. The explanations should be evaluated in terms of how they meet the users' expectations and how effective they are regarding understandability, usefulness, informativeness, trustworthiness,

and satisfaction. In our work, we proposed a way of using these traits in the evaluation of the explanations. We also generated some discussion on how the balance between informativeness and understandability should be considered in this evaluation, which opens a path for further studies.

In terms of human-AI performance, *i.e.*, contribution to the performance on the task at hand, we brought up the question whether simulatability is the most suitable property to be evaluated in an explanation when you have a human-grounded or functionally-grounded evaluation. These are scenarios where the individual performing the evaluation does not have a deep understanding of the problem being addressed by the model. We reasoned that trying to predict the output of a model, even having an explanation in hand, is not a trivial task. Even respondents who had a deep understanding of the visual explanations and previous knowledge of the SHAP method found significant difficulty in making predictions. Therefore, having previous knowledge of the context of the explanation is crucial for a proper "simulatability-based" evaluation. We believe that with this contribution, future studies will be able to have a direction on how to use "simulatability-based" evaluation, bearing in mind that it should be applied in an application-grounded approach.

8.1 Limitations and future work

Due to time constraints, we could not have improvement rounds in order to develop our visualizations further. It would be interesting to gather a group of end users who had contact with the model and the problem for a while and then tried to apply the explanations in their work in order to try to understand the model and take action on its results. Then, we would be able to enhance the visualizations and conduct another evaluation round with the users in order to get a more efficient set of visualizations. In future work, we could use the results of our evaluation process with users to improve the visualizations that we have created or even create new ones. Morelli et al. (2021) claim that it is important to have content experts and users in the co-design process, which was not possible in our case. They also pointed out that the co-design process should have iterative evaluations and user tests of the drafts in order to continuously improve it.

The evaluation of understandability, usefulness, informativeness, trustworthiness, and satisfaction was done in a one-time experiment where the participants had their first contact with the explanations. Trust levels should be measured over a period of time of use of the explanations (Hoffman et al., 2023), and that is probably true for the other factors as well.

We evaluated the visualizations generated in our study by choosing a domain of common knowledge so that the study participants would have a minimum understanding of the problem being addressed by the model. However, we could not affirm that these visualizations will function in the same way for other domains. Each domain or problem being addressed might demand a different visualization set that will depend on which information is more relevant for each. Future work might test these or new visualizations in other domains. Additionally, interactivity and personalization might be interesting to test for explanations, since each user or domain will have a different objective with the explanations. Another study that can also be done is on how the domain influences bias in the model, and how that could be translated into the explanations.

The sample size we had for our study is also a limitation since we had a reduced sample size and were also restricted to a specific context. In order to achieve some generalization, the study must be replicated on a larger scale and in diverse domains. Additionally, the evaluation done in this study was done specifically for the explanations generated by the SHAP method and may not work the same way for other XAI methods.

One hypothesis we had during our study was that understandability and informativeness could be opposites, meaning that if an explanation has too much information, it could be highly informative but also be hardly understandable because of the large amount of information; and, conversely, if it has too little information it could be extremely understandable, but missing relevant information. It is also a study we would like to see in future work, and examine how one could get to a balance between understandability and informativeness. Furthermore, after our findings, we believe that there could be a difference between perceived informativeness and actual informativeness, which could be an important factor to be evaluated not only in the context of model explanations but also in other contexts.

During our analysis, we also felt that a deeper understanding of the problem is crucial for the users to better exploit the explanations. When we think of the objectives of having an interpretable model, it is mainly for its users to have an understanding of how it reasons and, therefore, be able to have a critical vision of its decisions, thus avoiding bias, unfairness, and bad performance, and being able to trust the model's decisions. Having this in mind, we question if the human-grounded and functionally grounded evaluations proposed by Doshi-Velez and Kim (2017) are appropriate evaluation approaches. They are not based on a real application where the users have a previous understanding of the problem, which we consider crucial for one to properly interpret and use

the explanations.

Bibliography

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Aechtner, J., Cabrera, L., Katwal, D., Onghena, P., Valenzuela, D. P., and Wilbik, A. (2022). Comparing User Perception of Explanations Developed with XAI Methods. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, Padua, Italy. IEEE.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805.

Alicioglu, G. and Sun, B. (2022). A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics*, 102:502–520.

Bernardo, E. and Seva, R. (2023). Affective Design Analysis of Explainable Artificial Intelligence (XAI): A User-Centric Perspective. *Informatics*, 10(1):32.

Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13.

Brdnik, S., Podgorelec, V., and Šumak, B. (2023). Assessing Perceived Trust and Satisfaction with Multiple Explanation Techniques in XAI-Enhanced Learning Analytics. *Electronics*, 12(12):2594.

Brdnik, S. and Šumak, B. (2024). Current Trends, Challenges and Techniques in XAI Field; A Tertiary Study of XAI Research. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2032–2038, Opatija, Croatia. IEEE.

Chatzimparmpas, A., Martins, R. M., Jusufi, I., Kucher, K., Rossi, F., and Kerren, A. (2020). The State of the Art in Enhanc-

ing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum*, 39(3):713–756. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14034.

Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., and Jorge, J. (2021). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81.

Correll, M. (2019). Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. arXiv:1811.07271 [cs].

Doshi-Velez, F. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [cs, stat].

Gunning, D. and Aha, D. W. (2019). Darpa's explainable artificial intelligence program. *AI Magazine*, 40(2):44–58.

Hart, A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ*, 323(7309):391–393.

Hase, P. and Bansal, M. (2020). Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? arXiv:2005.01831 [cs].

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5:1096257.

Kim, B. (2015). Interactive and Interpretable Machine Learning Models for Human Machine Collaboration.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). arXiv:1711.11279 [stat].

Kim, D., Song, Y., Kim, S., Lee, S., Wu, Y., Shin, J., and Lee, D. (2023). How should the results of artificial intelligence be explained to users? - Research on consumer preferences in user-centered explainable artificial intelligence. *Technological Forecasting and Social Change*, 188:122343.

Kim, J., Maathuis, H., and Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence*, 7:1456486.

Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1675–1684, San Francisco California USA. ACM.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2019). Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, Honolulu HI USA. ACM.

Li, Y., Fujiwara, T., Choi, Y. K., Kim, K. K., and Ma, K.-L. (2020). A visual analytics system for multi-model comparison on clinical data predictions. *Visual Informatics*, 4(2):122–131.

Lipton, Z. C. (2017). The Mythos of Model Interpretability. arXiv:1606.03490 [cs, stat].

Lopes, B., Soares, L. S., Gonçalves, M. A., and Prates, R. O. (2025). A Human-Centered Multiperspective and Interactive Visual Tool For Explainable Machine Learning. *Journal of the Brazilian Computer Society*, 31(1):11–35.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631, Chicago Illinois USA. ACM.

Lundberg, S. M. and Lee, S.-I. (2017a). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lundberg, S. M. and Lee, S.-I. (2017b). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Mittelstadt, B., Russell, C., and Wachter, S. (2019). Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 279–288, New York, NY, USA. Association for Computing Machinery.

Morelli, A., Johansen, T. G., Pidcock, R., Harold, J., Pirani, A., Gomis, M., Lorenzoni, I., Haughey, E., and Coventry, K. (2021). Co-designing engaging and accessible data visualisations: a case study of the IPCC reports. *Climatic Change*, 168(3-4):26.

Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D. S., Fyshe, A., Pearcy, B., MacDonell, C., and Anvik, J. (2006). Visual Explanation of Evidence in Additive Classifiers.

Radley-Gardner, O., Beale, H., and Zimmermann, R., editors (2016). *Fundamental Texts On European Private Law*. Hart Publishing.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs, stat].

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Ridgeway, G., Madigan, D., Richardson, T., and O'Kane, J. (1998). Interpretable Boosted Naïve Bayes Classification.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. arXiv:1911.02508 [cs, stat].

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proc Annu Symp Comput Appl Med Care*, pages 261–265.

Spinner, T., Schlegel, U., Schafer, H., and El-Assady, M. (2019). explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1.

Stumpf, S., Skrebe, S., Aymer, G., and Hobson, J. (2018). Explaining smart heating systems to discourage fiddling with optimized behavior.

Tamagnini, P., Krause, J., Dasgupta, A., and Bertini, E. (2017). Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, pages 1–6, Chicago IL USA. ACM.

Woodcock, C., Mittelstadt, B., Busbridge, D., and Blank, G. (2021). The Impact of Explanations on Layperson Trust in Artificial Intelligence–Driven Symptom Checker Apps: Experimental Study. *Journal of Medical Internet Research*, 23(11):e29386.

A
Preliminary Study Material

The study material consisted of an informed consent form (Termo de Consentimento Livre e Esclarecido TCLE), a respondent profile questionnaire, and a simulation questionnaire, all presented insection A.1).

A.1
Study Form

Questionário trabalho INF2424 2023.2 - Bianca Cunha

Este questionário faz parte de um estudo para a disciplina de Visualização da Informação do curso de Mestrado em Informática na PUC-Rio.
Olá! Este questionário faz parte de um estudo para a disciplina de Visualização da Informação do curso de Mestrado em Informática na PUC-Rio. O estudo tem como objetivo entender se explicações geradas por métodos de explicação de modelos de machine learning realmente atingem o seu objetivo de melhorar a interpretabilidade dos modelos. Iremos fazer uma breve introdução sobre o assunto antes de começar o questionário.

Os modelos de machine learning vêm evoluindo ao longo dos anos e se tornando cada vez mais acurados e precisos. Porém, grande parte deles acaba não sendo interpretável, por conta de sua complexidade. A interpretabilidade é um fator importante para que se possa entender o comportamento do modelo e assim poder identificar o que gerou uma saída específica, como possivelmente mudar essa saída, ou até um comportamento incorreto que indique que o modelo precisa ser retreinado. Alguns métodos foram propostos com o objetivo de explicar o comportamento e/ou as saídas do modelo.

Neste questionário iremos apresentar um conjunto de dados que foi utilizado para treinar um modelo de machine learning e fazer algumas perguntas em relação à classificação.

Existe(m) 69 questão(ões) neste questionário.

Apresentação e Termo de Consentimento

1
Apresentação e Termo de Consentimento

Você está sendo convidado(a) a participar de uma pesquisa que investiga a eficácia do método de explicação de modelos de aprendizado de máquina SHAP. A sua participação nesta pesquisa é totalmente voluntária.

Título da Pesquisa: Avaliação da interpretabilidade de explicações geradas pelo método SHAP em classificações.

Os pesquisadores responsáveis pelo estudo poderão fornecer qualquer esclarecimento sobre o mesmo, assim como tirar dúvidas.
Os dados para contato são os seguintes:
Pesquisador Responsável: Bianca Moreira Cunha
Endereço: mcunhabianca@gmail.com

1) Objetivo Geral: Os objetivos primários desta pesquisa são: (i) Avaliar método de explicação de modelos de machine learning

2) Objetivos Específicos: (a) Avaliar se as saídas geradas por método de explicação de modelos de machine learning melhoram a interpretabilidade de modelos (b) Avaliar quais visualizações de geradas pelo método tem maior impacto na interpretabilidade dos modelos (c) Avaliar se ter uma explicação textual junto com a visualização gerada pelo método ajuda no entendimento do comportamento do modelo/da predição

3) Riscos: (a) Toda pesquisa realizada com seres humanos apresenta riscos. No entanto, os riscos apresentados nesta pesquisa são mínimos, pois as tecnologias utilizadas não alteram aspectos fisiológicos, psicológicos ou sociais dos participantes. Os questionários aplicados não tratam de aspectos relacionados à pessoa em si, mas sim às técnicas e processos utilizados. (b) Para evitar constrangimentos, sua identidade será mantida em sigilo. Além disso, você poderá ausentar-se do local do estudo a qualquer momento, caso sinta-se desconfortável. (c) Para assegurar a natureza voluntária da participação, o termo de consentimento livre e esclarecido (TCLE) associado a cada atividade será destacado

do material gerado e o professor da disciplina não terá acesso até que todas as suas notas da disciplina tenham sido lançadas. Dessa forma, os professores não saberão se o estudante consentiu ou não ter seus dados utilizados na pesquisa durante o período letivo, evitando constrangimentos entre estudantes universitários e professores ou prejuízos nas notas escolares. (d) Caso você tenha qualquer despesa decorrente da participação na pesquisa, haverá ressarcimento em espécie, logo após o término do experimento. (e) Todos os materiais necessários para realização da pesquisa serão fornecidos pelos próprios pesquisadores, de forma que não acarrete custos para os participantes. Além disso, está assegurado o direito a indenizações e qualquer tipo de assistência necessária para reparação a qualquer prejuízo causado pela pesquisa.

4) Benefícios: Através desta pesquisa espera-se identificar problemas e oportunidades de melhorias no uso de técnicas e processos de Interação Humano-Computador. Além disso, espera-se contribuir para a melhoria da aprendizagem destas técnicas por estudantes do ensino superior de computação, contribuindo desta forma para sua formação. Os benefícios gerados serão: (a) Materiais didáticos aprimorados e enriquecidos com base nas experiências práticas dos estudantes universitários em sala de aula, disponibilizados gratuitamente para todos os participantes do estudo e para estudantes de computação em geral; (b) Melhorias no uso das técnicas propostas, para que estas técnicas, posteriormente, possam ser melhor adaptadas para serem utilizadas na indústria de software; (c) Melhor apoio ao desenvolvimento de softwares inovadores, com maior qualidade, mais fáceis de utilizar para os usuários e, em última análise, que melhorem a sociedade como um todo.

5) Procedimentos: Você responderá a um questionário com perguntas relacionadas à avaliação de método de explicação de modelos de machine learning. Todos os questionários respondidos serão descartados após a sua transcrição. Quando os dados forem transcritos, seu nome será removido das transcrições e não será utilizado em nenhum momento durante a análise ou apresentação dos resultados. Algumas informações podem ser gravadas durante o estudo, no

entanto, após análise os áudios e vídeos serão descartados. Sua imagem não será utilizada e nem divulgada, de forma a manter sua identidade em sigilo.

6) Tratamento de possíveis riscos e desconfortos: A sua participação consiste tão somente em [[uma entrevista, que terá duração de 60 a 120 minutos, e o preenchimento de um questionário]]. Ainda assim, serão tomadas todas as providências durante a coleta desses dados de forma a garantir a sua privacidade e seu anonimato. Os dados coletados durante o estudo destinam-se estritamente a atividades da pesquisa. Desta forma, não serão utilizados para qualquer forma de avaliação profissional ou pessoal.

7) Custos: Você não terá nenhum gasto ou ônus com a sua participação no estudo e, também, não receberá qualquer espécie de reembolso ou gratificação devido à participação nesta pesquisa. No entanto, caso você tenha qualquer despesa decorrente da participação na pesquisa, haverá ressarcimento em espécie.

8) Confidencialidade da Pesquisa: Toda informação coletada neste estudo é confidencial, e seu nome e o da organização não serão identificados de modo algum, a não ser em caso de autorização explícita para esse fim.

9) Participação: Sua participação neste estudo é muito importante e voluntária. Você tem o direito de não querer participar ou de sair deste estudo a qualquer momento, sem penalidades. Você também tem o direito de se recusar a responder a qualquer pergunta do questionário. Para participar deste estudo você deverá ser maior de idade (ter 18 anos ou mais).

10) Declaração de Consentimento: Li ou alguém leu para mim as informações contidas neste documento antes de assinar este termo de consentimento. Declaro que toda a linguagem técnica utilizada na descrição deste estudo de pesquisa foi explicada satisfatoriamente e que recebi respostas para todas as minhas dúvidas. Confirmo também que recebi uma cópia deste Termo de Consentimento Livre e Esclarecido. Compreendo que sou livre para me retirar do estudo em qualquer momento, sem qualquer penalidade. Declaro ter mais de 18 anos e dou meu consentimento de livre e espontânea vontade para participar deste estudo.

Toda a dúvida a respeito desta pesquisa, poderá ser perguntada diretamente para Bianca Cunha, através do e-mail: mcunhabianca@gmail.com. Enquanto, as objeções a respeito da conduta ética poderão ser questionadas ao Comitê de Ética em Pesquisa da PUC-Rio.

* Required

Email address *

Por favor, concorde com o Termo de Consentimento Livre e Esclarecido (TCLE) abaixo para continuar. *

Se você discordar dos termos apresentados neste questionário sua participação será excluída deste estudo e seus dados não serão utilizados assim como não será necessário preencher as questões apresentadas neste questionário. A sua participação é de extrema importância para os autores desta pesquisa e todos os mesmos obedecem e seguem restritivamente os termos postados no TCLE devidamente assinado por você.

*

Escolha uma das seguintes respostas:
Favor escolher apenas uma das opções a seguir:

◯ Concordo com o Termo de Consentimento Livre e Esclarecido (TCLE).

◯ Não concordo com o Termo (TCLE) e não desejo participar deste estudo.

Perguntas de perfil

Grupo de perguntas para entender o perfil do participante.

2 Qual é o seu nível de escolaridade? *

Favor escolher apenas uma das opções a seguir:

◯ Ensino Médio completo

◯ Cursando Ensino Superior

◯ Ensino Superior completo

◯ Cursando Pós Graduação

◯ Pós Graduação completa

3 Você tem formação ou está cursando curso de base matemática ou estatística? *

Favor escolher apenas uma das opções a seguir:

- ◯ Sim
- ◯ Não

4 Qual você considera que seja o seu nível de conhecimento sobre Aprendizado de Máquina e/ou Inteligência Artificial? *

Favor escolher apenas uma das opções a seguir:

- ◯ Nenhum conhecimento
- ◯ Conhecimento básico
- ◯ Conhecimento intermediário
- ◯ Conhecimento avançado

5 Qual você considera que seja o seu nível de conhecimento sobre interpretabilidade de modelos de Aprendizado de Máquina? *

Favor escolher apenas uma das opções a seguir:

- ◯ Nenhum conhecimento
- ◯ Conhecimento básico
- ◯ Conhecimento intermediário
- ◯ Conhecimento avançado

6 Qual é o seu nível de inglês? *

Favor escolher apenas uma das opções a seguir:

- ◯ Básico
- ◯ Intermediário
- ◯ Avançado
- ◯ Fluente

Perguntas predição sem explicação

Grupo de perguntas feitas para o participante responder após observar registros com suas classes reais e classes preditas pelo modelo, sem fornecer explicações para ele.

7

O conjunto de dados utilizado para este estudo contém valores de sintomas apresentados por extraterrestes que geram um diagnóstico se eles têm ou não uma doença específica do planeta onde eles vivem. O valor de cada sintoma pode variar de 0 a 1, e o resultado do diagnóstico é 1 se for positivo e 0 se for negativo.

Ao longo do questionário serão apresentadas explicações para o comportamento do modelo que gera predições para o diagnóstico de cada extraterrestre a partir de 5 sintomas. Essas explicações são costruídas a partir de valores chamados SHAP values. SHAP value é uma medida de impacto de uma variável na saída de um modelo de machine learning. Portanto, um SHAP value alto indica que uma variável específica teve alta influência na produção da saída do modelo. Além disso, o SHAP value pode ser positivo ou negativo, sendo assim, em uma classificação, como no caso deste estudo, um SHAP value positivo indica influência na direção de uma classe e um SHAP value negativo indica influência na direção da outra classe. Olhando para o caso do nosso questionário então, estaremos avaliando o impacto de cada sintoma para o diagonóstico positivo ou negativo da doença extraterrestre, e o SHAP value de cada uma das variáveis indicará a magnitude desse impacto e a direção dele.

Abaixo estão alguns registros do conjunto de dados contendo os valores dos 5 sintomas apresentados pelo extraterrestre junto com a sua classificação real e a classificação dada pelo modelo. Observe estas instâncias com atenção:

Instância 1:

Sintoma 1	0.23
Sintoma 2	0.76
Sintoma 3	0.52
Sintoma 4	0.41
Sintoma 5	0.06

Classificação real	1
Classificação do modelo	1

Instância 2:

Sintoma 1	0,46
Sintoma 2	0,45
Sintoma 3	0,26
Sintoma 4	0,12
Sintoma 5	0,96

Classificação real	0
Classificação do modelo	0

Instância 3:

Sintoma 1	0,38
Sintoma 2	0,73
Sintoma 3	0,31
Sintoma 4	0,09
Sintoma 5	0,89

Classificação real	0
Classificação do modelo	1

Instância 4:

Sintoma 1	0,15
Sintoma 2	0,53
Sintoma 3	0,39
Sintoma 4	0,34
Sintoma 5	0,13

Classificação real	1
Classificação do modelo	0

8

A partir da observação dos registros exibidos anteriormente, das suas classes reais e suas classes preditas pelo modelo, classifique o registro a seguir:

Sintoma 1	0,08
Sintoma 2	0,89
Sintoma 3	0,59
Sintoma 4	0,85
Sintoma 5	0,02

*

Favor escolher apenas uma das opções a seguir:

◯ Positivo

◯ Negativo

9

A partir da observação dos registros exibidos anteriormente, das suas classes reais e suas classes preditas pelo modelo, classifique o registro a seguir:

Sintoma 1	0,08
Sintoma 2	0,41
Sintoma 3	0,47
Sintoma 4	0,24
Sintoma 5	0,06

*

Favor escolher apenas uma das opções a seguir:

◯ Positivo

◯ Negativo

10

A partir da observação dos registros exibidos anteriormente, das suas classes reais e suas classes preditas pelo modelo, classifique o registro a seguir:

Sintoma 1	0,08
Sintoma 2	0,64
Sintoma 3	0,32
Sintoma 4	0,59
Sintoma 5	0,04

*

Favor escolher apenas uma das opções a seguir:

◯ Positivo

◯ Negativo

11
A partir da observação dos registros exibidos anteriormente, das suas classes reais e suas classes preditas pelo modelo, classifique o registro a seguir:

Sintoma 1	0,15
Sintoma 2	0,66
Sintoma 3	0,27
Sintoma 4	0,19
Sintoma 5	0,17

*

Favor escolher apenas uma das opções a seguir:

◯ Positivo
◯ Negativo

12 Quão confiante você está da sua classificação? *

Favor escolher apenas uma das opções a seguir:

◯ 1 - nada confiante
◯ 2 - não tão confiante
◯ 3 - relativamente confiante
◯ 4 - bem confiante
◯ 5 - extremamente confiante

13 O que te fez escolher essa classificação? (Explique que elementos dos dados te levaram a fazer essa escolha) *

Por favor, coloque sua resposta aqui:

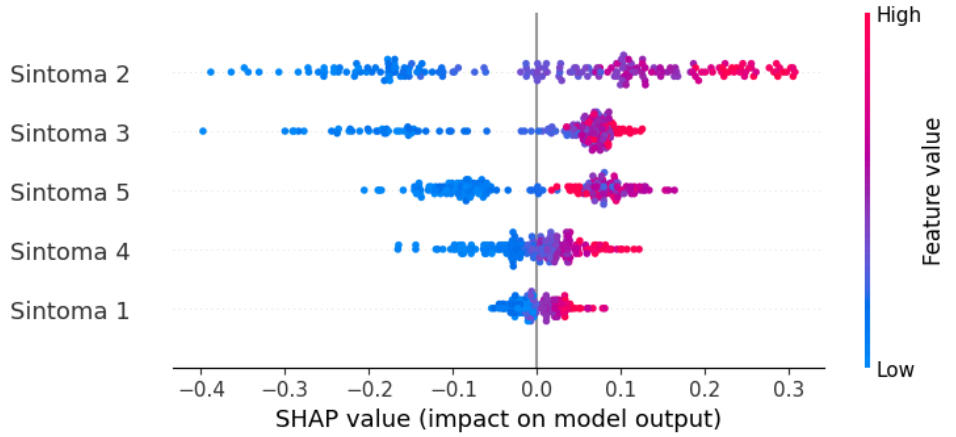
Observações com visualização da explicação

Observações com visualização da explicação

14
{if(is_empty(randnumber1.NAOK),rand(1,3),randnumber1.NAOK)}

15

Abaixo está a visualização da explicação do comportamento do modelo gerada pelo método SHAP, que deve ser observada juntamente com algumas instâncias, suas classificações reais e as classificações dadas pelo modelo:



Instância 1:

Sintoma 1	0
Sintoma 2	0,48
Sintoma 3	0,45
Sintoma 4	0,47
Sintoma 5	0,04

Classificação real	0
Classificação do modelo	0

Instância 2:

Sintoma 1	0.23
Sintoma 2	0.76
Sintoma 3	0.52
Sintoma 4	0.41
Sintoma 5	0.06

Classificação real	1
Classificação do modelo	1

Instância 3:

Sintoma 1	0,15
Sintoma 2	0,53
Sintoma 3	0,39
Sintoma 4	0,34
Sintoma 5	0,13

Classificação real	1
Classificação do modelo	0

Instância 4:

Sintoma 1	0,15
Sintoma 2	0,40
Sintoma 3	0,33
Sintoma 4	0,96
Sintoma 5	0,28

Classificação real	0
Classificação do modelo	1

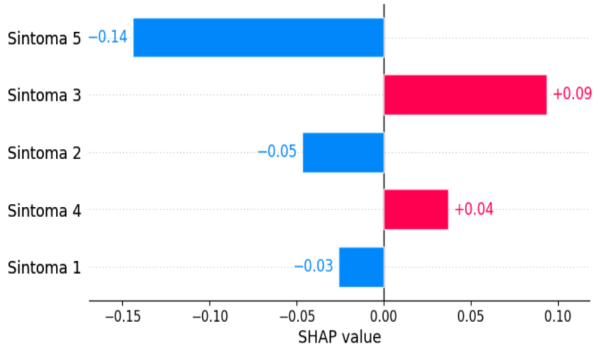
16

Abaixo estão algumas instâncias junto com a sua classificação real, a classificação dada pelo modelo e a visualização da explicação da predição gerada pelo método SHAP para serem observadas.

Instância 1:

Sintoma 1	0
Sintoma 2	0,48
Sintoma 3	0,45
Sintoma 4	0,47
Sintoma 5	0,04

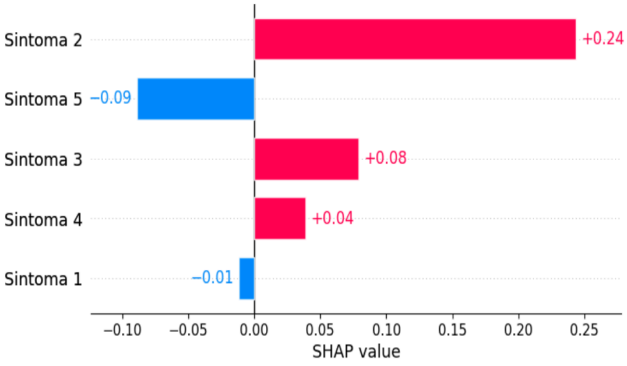
Classificação real	0
Classificação do modelo	0



Instância 2:

Sintoma 1	0.23
Sintoma 2	0.76
Sintoma 3	0.52
Sintoma 4	0.41
Sintoma 5	0.06

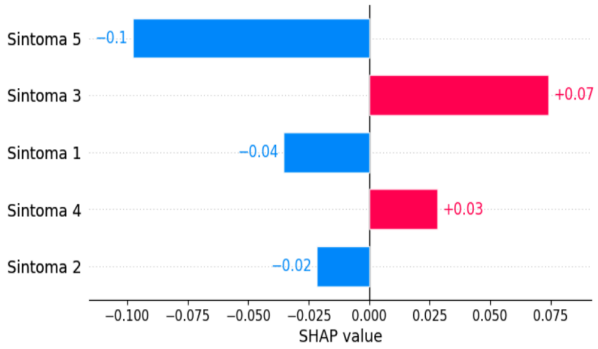
Classificação real	1
Classificação do modelo	1



Instância 3:

Sintoma 1	0,15
Sintoma 2	0,53
Sintoma 3	0,39
Sintoma 4	0,34
Sintoma 5	0,13

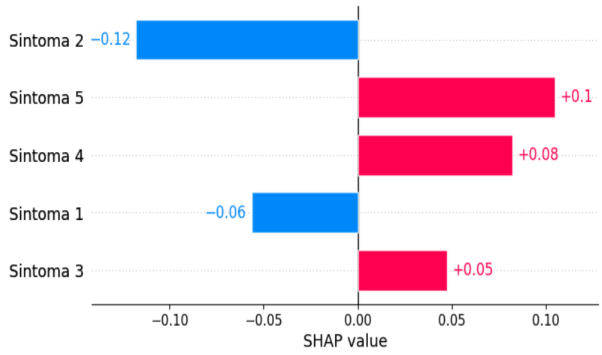
Classificação real	1
Classificação do modelo	0



Instância 4:

Sintoma 1	0,15
Sintoma 2	0,40
Sintoma 3	0,33
Sintoma 4	0,96
Sintoma 5	0,28

Classificação real	0
Classificação do modelo	1



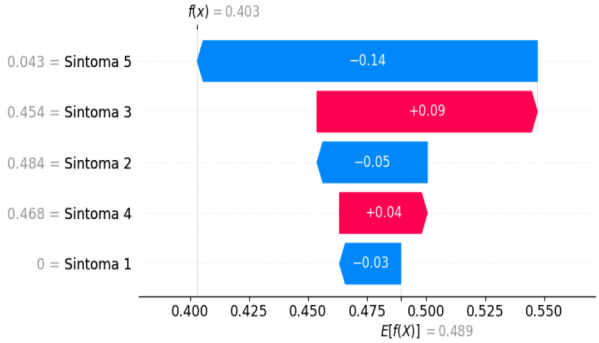
17

Abaixo estão algumas instâncias junto com a sua classificação real, a classificação dada pelo modelo e a visualização da explicação da predição gerada pelo método SHAP para serem observadas.

Instância 1:

Sintoma 1	0
Sintoma 2	0,48
Sintoma 3	0,45
Sintoma 4	0,47
Sintoma 5	0,04

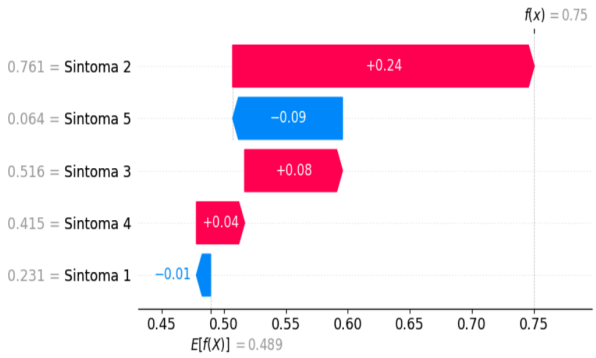
Classificação real	0
Classificação do modelo	0



Instância 2:

Sintoma 1	0.23
Sintoma 2	0.76
Sintoma 3	0.52
Sintoma 4	0.41
Sintoma 5	0.06

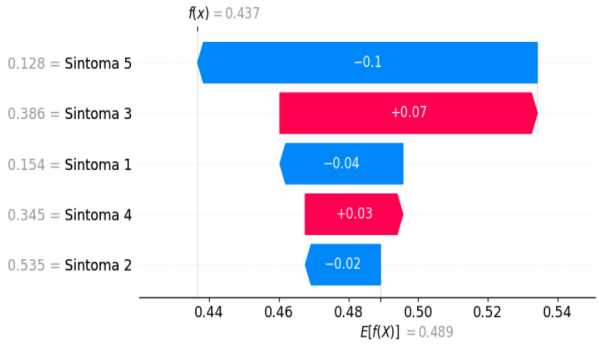
Classificação real	1
Classificação do modelo	1



Instância 3:

Sintoma 1	0,15
Sintoma 2	0,53
Sintoma 3	0,39
Sintoma 4	0,34
Sintoma 5	0,13

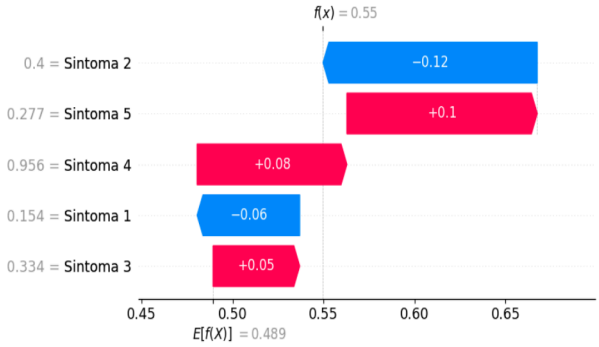
Classificação real	1
Classificação do modelo	0



Instância 4:

Sintoma 1	0,15
Sintoma 2	0,40
Sintoma 3	0,33
Sintoma 4	0,96
Sintoma 5	0,28

Classificação real	0
Classificação do modelo	1



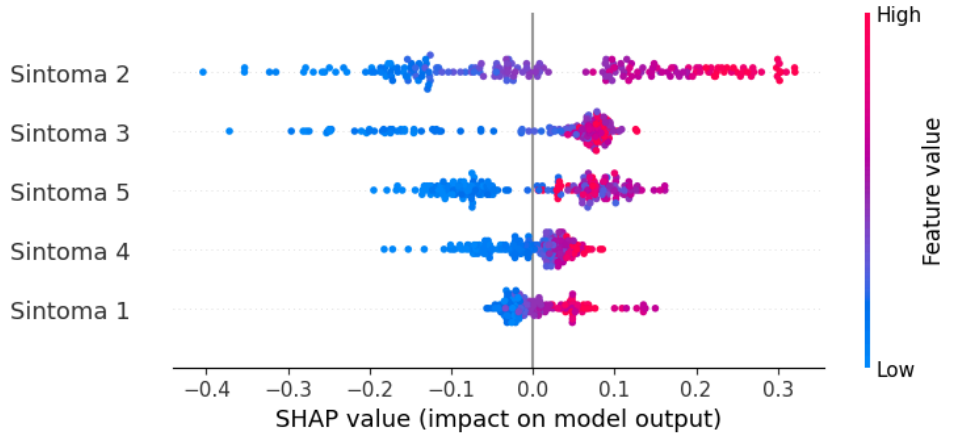
Perguntas predição com visualização da explicação 1

Grupo de perguntas feitas para o participante responder após observar registros com suas classes reais e classes preditas pelo modelo, fornecendo visualizações das explicações para ele.

18
{if(is_empty(randnumber1.NAOK),rand(1,3),randnumber1.NAOK)}

19

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:



Sintoma 1	0,08
Sintoma 2	0,64
Sintoma 3	0,32
Sintoma 4	0,59
Sintoma 5	0,04

*

Favor escolher apenas uma das opções a seguir:

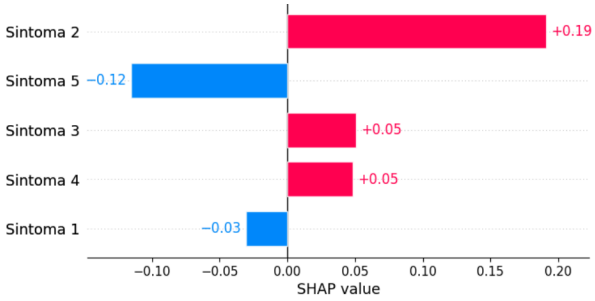
◯ Positivo

◯ Negativo

20

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:

Sintoma 1	0,08
Sintoma 2	0,64
Sintoma 3	0,32
Sintoma 4	0,59
Sintoma 5	0,04



*

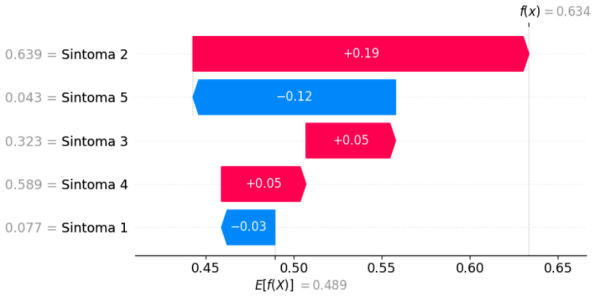
Favor escolher apenas uma das opções a seguir:

◯ Positivo

◯ Negativo

21
A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:

Sintoma 1	0,08
Sintoma 2	0,64
Sintoma 3	0,32
Sintoma 4	0,59
Sintoma 5	0,04



*

Favor escolher apenas uma das opções a seguir:

- ◯ Positivo
- ◯ Negativo

22 Quão confiante você está da sua classificação? *

Favor escolher apenas uma das opções a seguir:

- ◯ 1 - nada confiante
- ◯ 2 - não tão confiante
- ◯ 3 - relativamente confiante
- ◯ 4 - bem confiante
- ◯ 5 - extremamente confiante

23 O que te fez escolher essa classificação? (Explique que elementos dos dados ou da visualização te levaram a fazer essa escolha) *

Por favor, coloque sua resposta aqui:

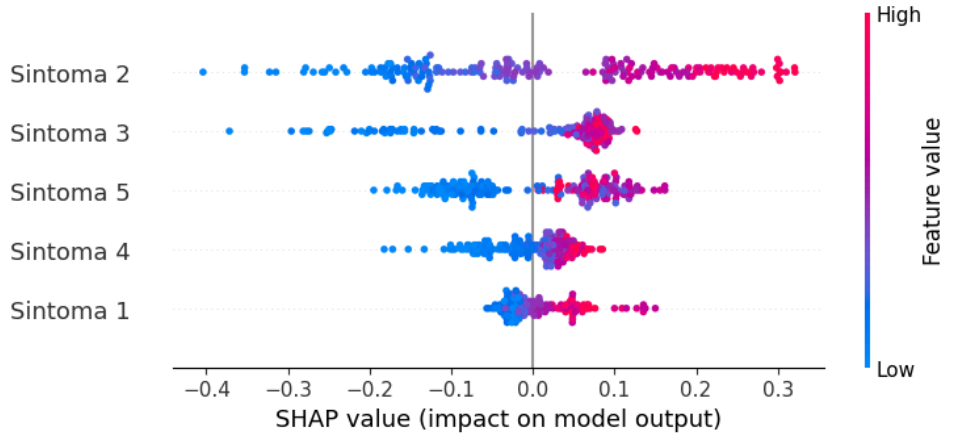
Perguntas predição com visualização da explicação 2

Grupo de perguntas feitas para o participante responder após observar registros com suas classes reais e classes preditas pelo modelo, fornecendo visualizações das explicações para ele.

24
{if(is_empty(randnumber1.NAOK),rand(1,3),randnumber1.NAOK)}

25

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:



Sintoma 1	0,08
Sintoma 2	0,41
Sintoma 3	0,47
Sintoma 4	0,24
Sintoma 5	0,06

*

Favor escolher apenas uma das opções a seguir:

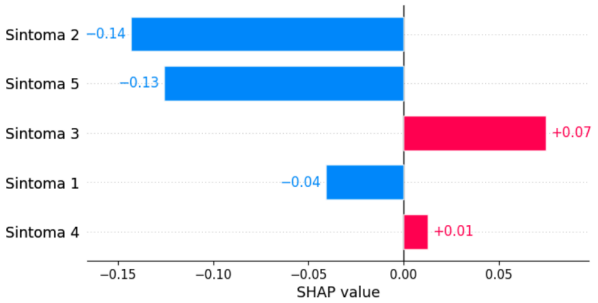
◯ Positivo

◯ Negativo

26

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:

Sintoma 1	0,08
Sintoma 2	0,41
Sintoma 3	0,47
Sintoma 4	0,24
Sintoma 5	0,06



*

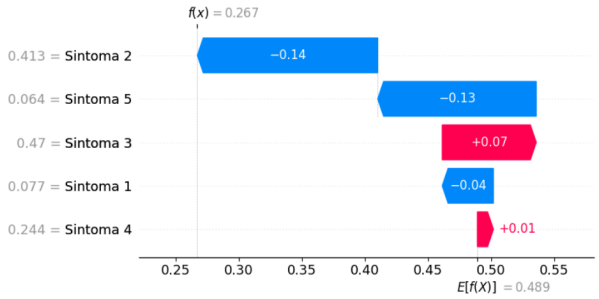
Favor escolher apenas uma das opções a seguir:

◯ Positivo

◯ Negativo

27
A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:

Sintoma 1	0,08
Sintoma 2	0,41
Sintoma 3	0,47
Sintoma 4	0,24
Sintoma 5	0,06



*

Favor escolher apenas uma das opções a seguir:

◯ Positivo
◯ Negativo

28 Quão confiante você está da sua classificação? *

Favor escolher apenas uma das opções a seguir:

◯ 1 - nada confiante
◯ 2 - não tão confiante
◯ 3 - relativamente confiante
◯ 4 - bem confiante
◯ 5 - extremamente confiante

29 O que te fez escolher essa classificação? (Explique que elementos dos dados ou da visualização te levaram a fazer essa escolha) *

Por favor, coloque sua resposta aqui:

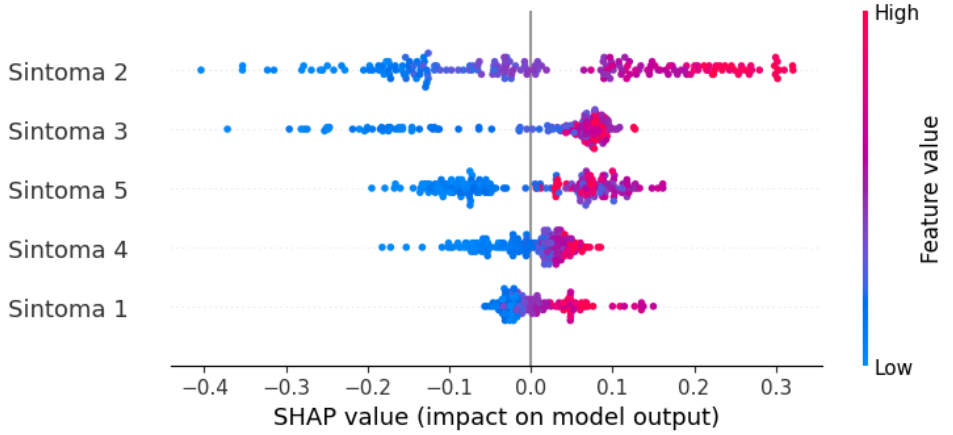
Perguntas predição com visualização da explicação 3

Grupo de perguntas feitas para o participante responder após observar registros com suas classes reais e classes preditas pelo modelo, fornecendo visualizações das explicações para ele.

30 {if(is_empty(randnumber1.NAOK),rand(1,3),randnumber1.NAOK)}

31

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:



Sintoma 1	0,54
Sintoma 2	0,91
Sintoma 3	0,45
Sintoma 4	0,24
Sintoma 5	0,34

*

Favor escolher apenas uma das opções a seguir:

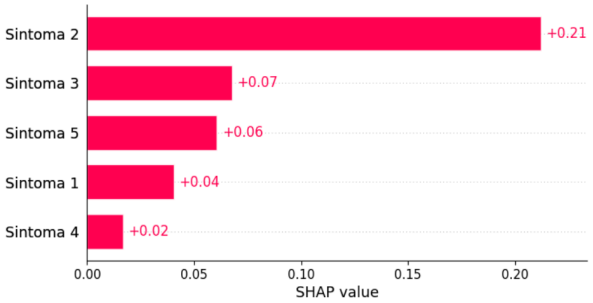
◯ Positivo

◯ Negativo

32

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:

Sintoma 1	0,54
Sintoma 2	0,91
Sintoma 3	0,45
Sintoma 4	0,24
Sintoma 5	0,34



*

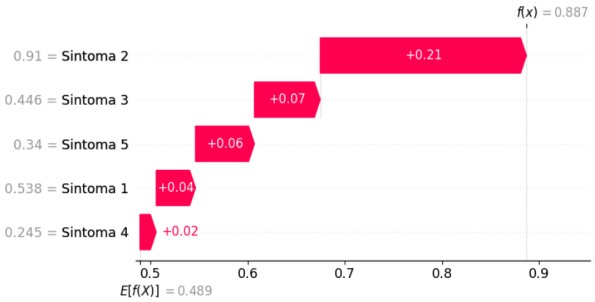
Favor escolher apenas uma das opções a seguir:

◯ Positivo

◯ Negativo

33
A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:

Sintoma 1	0,54
Sintoma 2	0,91
Sintoma 3	0,45
Sintoma 4	0,24
Sintoma 5	0,34



*

Favor escolher apenas uma das opções a seguir:

- ◯ Positivo
- ◯ Negativo

34 Quão confiante você está da sua classificação? *

Favor escolher apenas uma das opções a seguir:

- ◯ 1 - nada confiante
- ◯ 2 - não tão confiante
- ◯ 3 - relativamente confiante
- ◯ 4 - bem confiante
- ◯ 5 - extremamente confiante

35 O que te fez escolher essa classificação? (Explique que elementos dos dados ou da visualização te levaram a fazer essa escolha) *

Por favor, coloque sua resposta aqui:

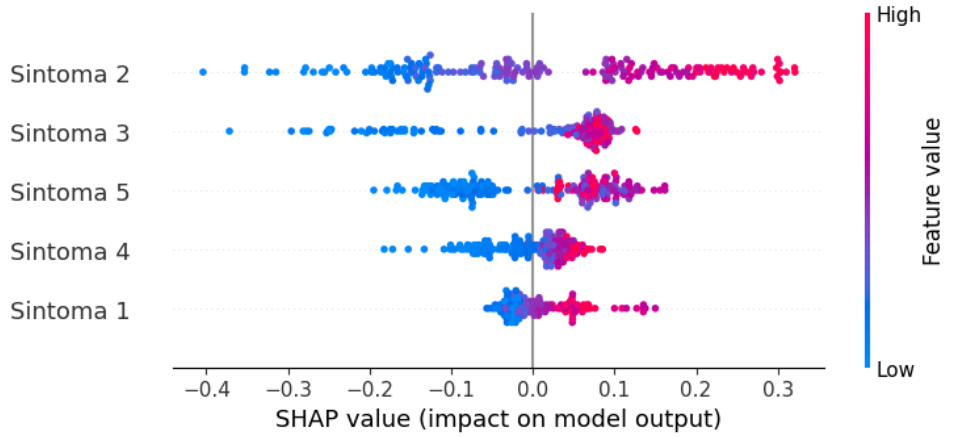
Perguntas predição com visualização da explicação 4

Grupo de perguntas feitas para o participante responder após observar registros com suas classes reais e classes preditas pelo modelo, fornecendo visualizações das explicações para ele.

36 {if(is_empty(randnumber1.NAOK),rand(1,3),randnumber1.NAOK)}

37

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:



Sintoma 1	0
Sintoma 2	0,56
Sintoma 3	0,44
Sintoma 4	0,10
Sintoma 5	0

*

Favor escolher apenas uma das opções a seguir:

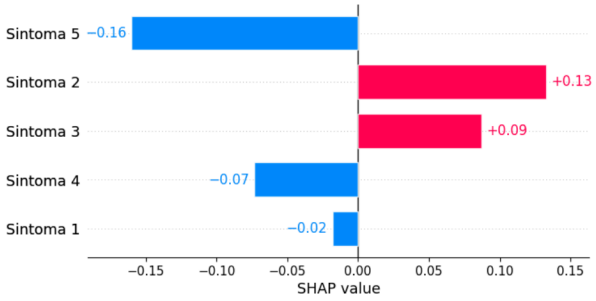
◯ Positivo

◯ Negativo

38

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:

Sintoma 1	0
Sintoma 2	0,56
Sintoma 3	0,44
Sintoma 4	0,10
Sintoma 5	0



*

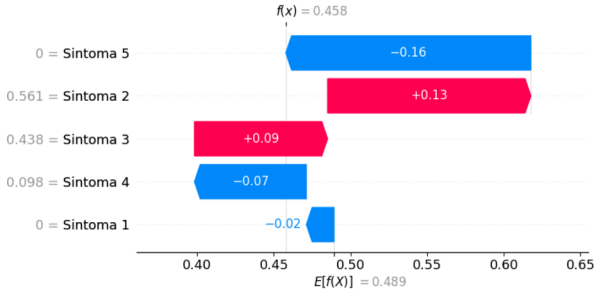
Favor escolher apenas uma das opções a seguir:

◯ Positivo

◯ Negativo

39
A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo e as visualizações das explicações geradas pelo método SHAP, classifique o registro a seguir:

Sintoma 1	0
Sintoma 2	0,56
Sintoma 3	0,44
Sintoma 4	0,10
Sintoma 5	0



*

Favor escolher apenas uma das opções a seguir:

- ◯ Positivo
- ◯ Negativo

40 Quão confiante você está da sua classificação? *

Favor escolher apenas uma das opções a seguir:

- ◯ 1 - nada confiante
- ◯ 2 - não tão confiante
- ◯ 3 - relativamente confiante
- ◯ 4 - bem confiante
- ◯ 5 - extremamente confiante

41 O que te fez escolher essa classificação? (Explique que elementos dos dados ou da visualização te levaram a fazer essa escolha) *

Por favor, coloque sua resposta aqui:

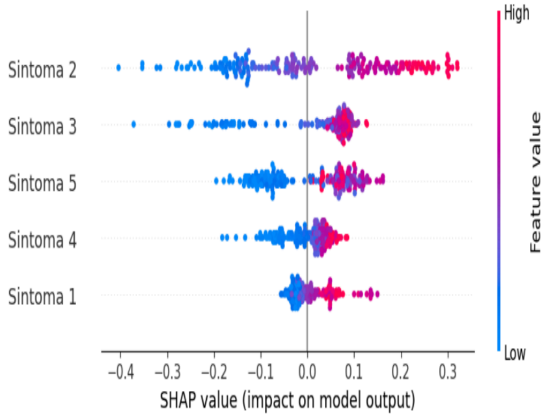
Observações com visualização e explicação textual

Observações com visualização e explicação textual

42
{if(is_empty(randnumber1.NAOK),rand(1,3),randnumber1.NAOK)}

43

Abaixo está a visualização da explicação do comportamento do modelo gerada pelo método SHAP, a sua descrição textual, que deve ser observada juntamente com algumas instâncias, suas classificações reais e as classificações dadas pelo modelo:



This plot shows the features sorted by magnitude of impact on the model in general, considering their absolute mean SHAP values. Each feature has a beeswarm and they are composed by dots that represent each observation and are distributed along the x-axis according to the observations' SHAP values. On the right side of the plot there is a vertical bar that gives the color shades of the dots that represent from high to low values of the features. Red dots represent high feature values, shades of purple represent medium values and blue dots represent low values. That means that along with the information of how much impact each feature has on the model output, it is possible to know what ranges of feature values have positive, negative or no impact. In this particular case, the 5 features that presented highest impact in the model were Sintoma 2, Sintoma 3, Sintoma 5, Sintoma 4 and Sintoma 1 respectively.

Instância 1:

Sintoma 1	0,54
Sintoma 2	0,80
Sintoma 3	0,54
Sintoma 4	0,51
Sintoma 5	0,40

Classificação real	1
Classificação do modelo	1

Instância 2:

Sintoma 1	0,46
Sintoma 2	0,45
Sintoma 3	0,26
Sintoma 4	0,12
Sintoma 5	0,96

Classificação real	0
Classificação do modelo	0

Instância 3:

Sintoma 1	0,38
Sintoma 2	0,73
Sintoma 3	0,31
Sintoma 4	0,09
Sintoma 5	0,89

Classificação real	0
Classificação do modelo	1

Instância 4:

Sintoma 1	0,54
Sintoma 2	0,40
Sintoma 3	0,22
Sintoma 4	0,16
Sintoma 5	0,17

Classificação real	1
Classificação do modelo	0

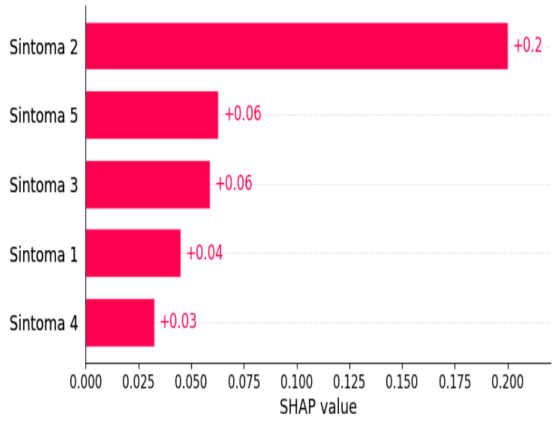
44

Abaixo estão algumas instâncias junto com a sua classificação real, a classificação dada pelo modelo, a visualização da explicação da predição gerada pelo método SHAP e a sua descrição textual para serem observadas.

Instância 1:

Sintoma 1	0,54
Sintoma 2	0,80
Sintoma 3	0,54
Sintoma 4	0,51
Sintoma 5	0,40

Classificação real	1
Classificação do modelo	1

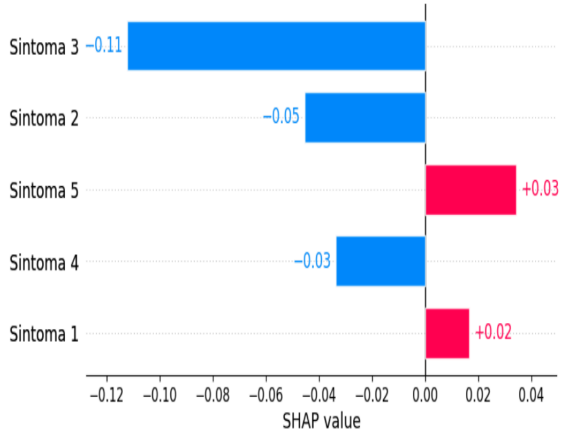


This plot shows the features sorted by magnitude of impact on the model considering their absolute SHAP values. It shows the magnitude of the impact as well as if the contribution was towards a class or another. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot and have their value displayed with a positive sign. The bars is blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot and have their value displayed with a negative sign. For the chosen observation, the features ['Sintoma 2' 'Sintoma 5' 'Sintoma 3' 'Sintoma 1' 'Sintoma 4'] had the highest contribution for the positive class respectively and the features [] had the highest contribution for the negative class respectively.

Instância 2:

Sintoma 1	0,46
Sintoma 2	0,45
Sintoma 3	0,26
Sintoma 4	0,12
Sintoma 5	0,96

Classificação real	0
Classificação do modelo	0



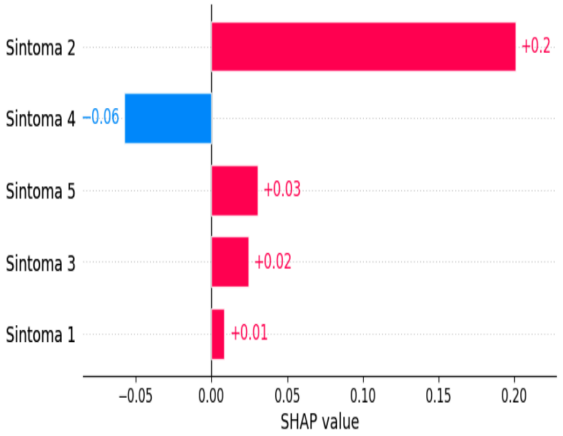
This plot shows the features sorted by magnitude of impact on the model considering their absolute SHAP values. It shows the magnitude of the impact as well as if the contribution was towards a class or another. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot and have their value displayed with a positive sign. The bars is blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot and have their value displayed with a negative sign. For the chosen observation, the features ['Sintoma 5' 'Sintoma 1'] had the highest contribution for the positive class respectively and the features ['Sintoma 3', 'Sintoma 2', 'Sintoma 4'] had the highest contribution for the negative class respectively.

Instância 3:

Sintoma 1	0,38
Sintoma 2	0,73
Sintoma 3	0,31
Sintoma 4	0,09
Sintoma 5	0,89

Classificação real	0
Classificação do modelo	1

\

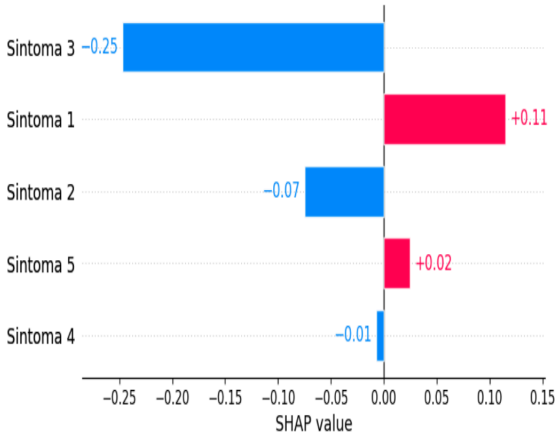


This plot shows the features sorted by magnitude of impact on the model considering their absolute SHAP values. It shows the magnitude of the impact as well as if the contribution was towards a class or another. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot and have their value displayed with a positive sign. The bars is blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot and have their value displayed with a negative sign. For the chosen observation, the features ['Sintoma 2' 'Sintoma 5' 'Sintoma 3' 'Sintoma 1'] had the highest contribution for the positive class respectively and the features ['Sintoma 4'] had the highest contribution for the negative class respectively.

Instância 4:

Sintoma 1	0,54
Sintoma 2	0,40
Sintoma 3	0,22
Sintoma 4	0,16
Sintoma 5	0,17

Classificação real	1
Classificação do modelo	0



This plot shows the features sorted by magnitude of impact on the model considering their absolute SHAP values. It shows the magnitude of the impact as well as if the contribution was towards a class or another. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot and have their value displayed with a positive sign. The bars is blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot and have their value displayed with a negative sign. For the chosen observation, the features ['Sintoma 1' 'Sintoma 5'] had the highest contribution for the positive class respectively and the features ['Sintoma 3', 'Sintoma 2', 'Sintoma 4'] had the highest contribution for the negative class respectively.

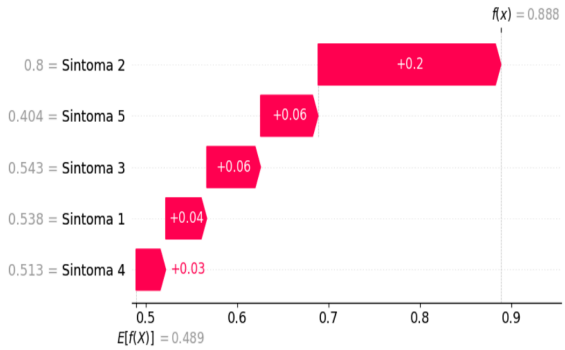
45

Abaixo estão algumas instâncias junto com a sua classificação real, a classificação dada pelo modelo, a visualização da explicação da predição gerada pelo método SHAP e a sua descrição textual para serem observadas.

Instância 1:

Sintoma 1	0,54
Sintoma 2	0,80
Sintoma 3	0,54
Sintoma 4	0,51
Sintoma 5	0,40

Classificação real	1
Classificação do modelo	1

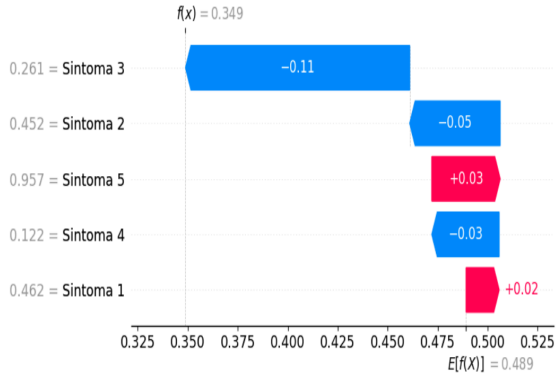


In this plot, the x-axis has the possible values for the target variable instead of the SHAP values, and the y-axis has the feature names. The features are sorted by magnitude of impact on the model considering their absolute SHAP values. In the x-axis there is also the representation of the target expected value E[f(X)], which is the mean target value of all the predictions. It shows how much each feature contributed for that prediction to have been higher or lower than the expected value. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot. The bars in blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot. For the chosen observation, the features ['Sintoma 2' 'Sintoma 5' 'Sintoma 3' 'Sintoma 1' 'Sintoma 4'], had the highest contribution for the positive class respectively and the features [] had the highest contribution for the negative class respectively.

Instância 2:

Sintoma 1	0,46
Sintoma 2	0,45
Sintoma 3	0,26
Sintoma 4	0,12
Sintoma 5	0,96

Classificação real	0
Classificação do modelo	0

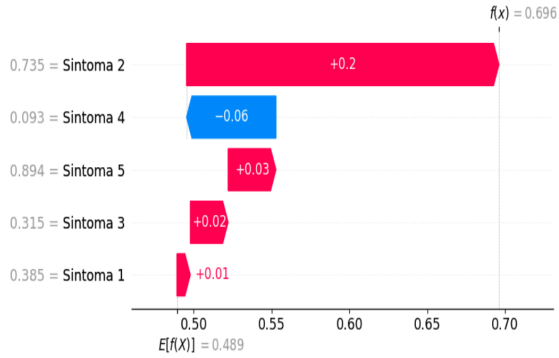


In this plot, the x-axis has the possible values for the target variable instead of the SHAP values, and the y-axis has the feature names. The features are sorted by magnitude of impact on the model considering their absolute SHAP values. In the x-axis there is also the representation of the target expected value E[f(X)], which is the mean target value of all the predictions. It shows how much each feature contributed for that prediction to have been higher or lower than the expected value. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot. The bars in blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot. For the chosen observation, the features ['Sintoma 5' 'Sintoma 1'], had the highest contribution for the positive class respectively and the features ['Sintoma 3', 'Sintoma 2', 'Sintoma 4'] had the highest contribution for the negative class respectively.

Instância 3:

Sintoma 1	0,38
Sintoma 2	0,73
Sintoma 3	0,31
Sintoma 4	0,09
Sintoma 5	0,89

Classificação real	0
Classificação do modelo	1

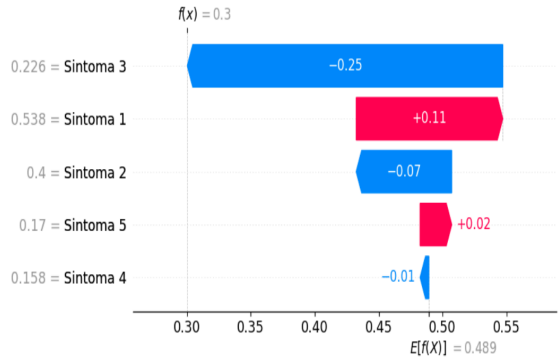


In this plot, the x-axis has the possible values for the target variable instead of the SHAP values, and the y-axis has the feature names. The features are sorted by magnitude of impact on the model considering their absolute SHAP values. In the x-axis there is also the representation of the target expected value E[f(X)], which is the mean target value of all the predictions. It shows how much each feature contributed for that prediction to have been higher or lower than the expected value. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot. The bars in blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot. For the chosen observation, the features ['Sintoma 2' 'Sintoma 5' 'Sintoma 3' 'Sintoma 1'], had the highest contribution for the positive class respectively and the features ['Sintoma 4'] had the highest contribution for the negative class respectively.

Instância 4:

Sintoma 1	0,54
Sintoma 2	0,40
Sintoma 3	0,22
Sintoma 4	0,16
Sintoma 5	0,17

Classificação real	1
Classificação do modelo	0



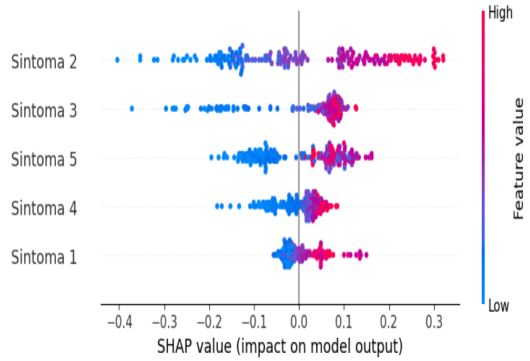
In this plot, the x-axis has the possible values for the target variable instead of the SHAP values, and the y-axis has the feature names. The features are sorted by magnitude of impact on the model considering their absolute SHAP values. In the x-axis there is also the representation of the target expected value E[f(X)], which is the mean target value of all the predictions. It shows how much each feature contributed for that prediction to have been higher or lower than the expected value. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot. The bars in blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot. For the chosen observation, the features ['Sintoma 1' 'Sintoma 5'], had the highest contribution for the positive class respectively and the features ['Sintoma 3', 'Sintoma 2', 'Sintoma 4'] had the highest contribution for the negative class respectively.

Perguntas predição com visualização e explicação textual 1

Grupo de perguntas feitas para o participante responder após observar registros com suas classes reais e classes preditas pelo modelo, fornecendo visualizações e explicações textuais para ele.

46
{if(is_empty(randnumber1.NAOK),rand(1,3),randnumber1.NAOK)}

47
A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo, as visualizações das explicações geradas pelo método SHAP e a explicação textual dessa visualização, classifique o registro a seguir:



This plot shows the features sorted by magnitude of impact on the model in general, considering their absolute mean SHAP values. Each feature has a beeswarm and they are composed by dots that represent each observation and are distributed along the x-axis according to the observations' SHAP values. On the right side of the plot there is a vertical bar that gives the color shades of the dots that represent from high to low values of the features. Red dots represent high feature values, shades of purple represent medium values and blue dots represent low values. That means that along with the information of how much impact each feature has on the model output, it is possible to know what ranges of feature values have positive, negative or no impact. In this particular case, the 5 features that presented highest impact in the model were Sintoma 2, Sintoma 3, Sintoma 5, Sintoma 4 and Sintoma 1 respectively.

Sintoma 1	0,15
Sintoma 2	0,66
Sintoma 3	0,27
Sintoma 4	0,19
Sintoma 5	0,17

*

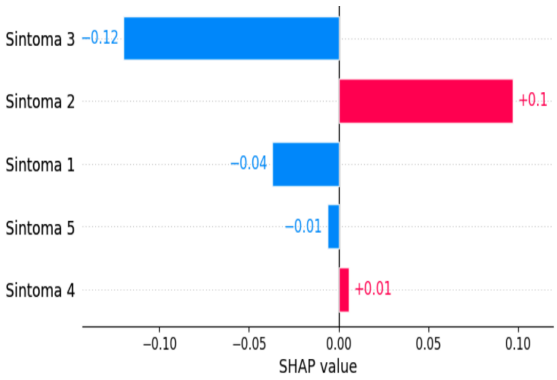
Favor escolher apenas uma das opções a seguir:

◯ Positivo
◯ Negativo

48

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo, as visualizações das explicações geradas pelo método SHAP e a explicação textual dessa visualização, classifique o registro a seguir:

Sintoma 1	0,15
Sintoma 2	0,66
Sintoma 3	0,27
Sintoma 4	0,19
Sintoma 5	0,17



This plot shows the features sorted by magnitude of impact on the model considering their absolute SHAP values. It shows the magnitude of the impact as well as if the contribution was towards a class or another. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot and have their value displayed with a positive sign. The bars is blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot and have their value displayed with a negative sign. For the chosen observation, the features ['Sintoma 2' 'Sintoma 4'] had the highest contribution for the positive class respectively and the features ['Sintoma 3', 'Sintoma 1', 'Sintoma 5'] had the highest contribution for the negative class respectively.

*

Favor escolher apenas uma das opções a seguir:

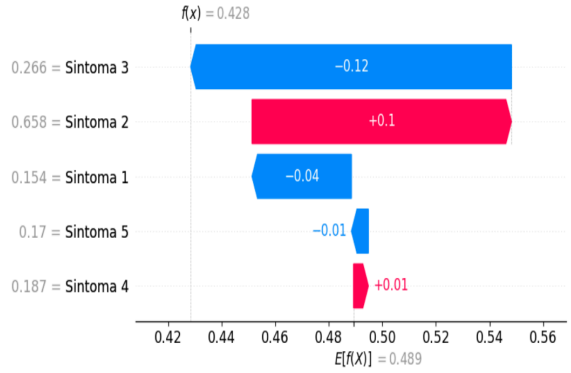
◯ Positivo

◯ Negativo

49

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo, as visualizações das explicações geradas pelo método SHAP e a explicação textual dessa visualização, classifique o registro a seguir:

Sintoma 1	0,15
Sintoma 2	0,66
Sintoma 3	0,27
Sintoma 4	0,19
Sintoma 5	0,17



In this plot, the x-axis has the possible values for the target variable instead of the SHAP values, and the y-axis has the feature names. The features are sorted by magnitude of impact on the model considering their absolute SHAP values. In the x-axis there is also the representation of the target expected value E[f(X)], which is the mean target value of all the predictions. It shows how much each feature contributed for that prediction to have been higher or lower than the expected value. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot. The bars in blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot. For the chosen observation, the features ['Sintoma 2' 'Sintoma 4'], had the highest contribution for the positive class respectively and the features ['Sintoma 3', 'Sintoma 1', 'Sintoma 5'] had the highest contribution for the negative class respectively.

*

Favor escolher apenas uma das opções a seguir:

◯ Aprovado

◯ Rejeitado

50 Quão confiante você está da sua classificação? *

Favor escolher apenas uma das opções a seguir:

- ◯ 1 - nada confiante
- ◯ 2 - não tão confiante
- ◯ 3 - relativamente confiante
- ◯ 4 - bem confiante
- ◯ 5 - extremamente confiante

51 O que te fez escolher essa classificação? (Explique que elementos dos dados, da visualização ou da explicação textual te levaram a fazer essa escolha) *

Por favor, coloque sua resposta aqui:

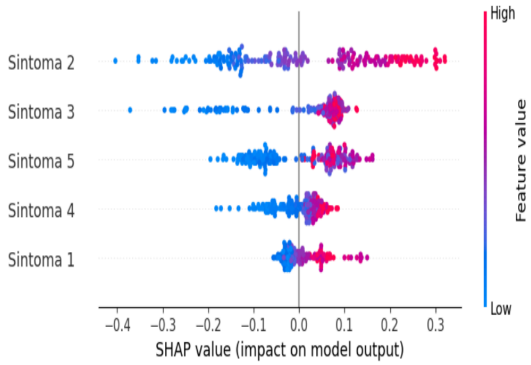
Perguntas predição com visualização e explicação textual 2

Grupo de perguntas feitas para o participante responder após observar registros com suas classes reais e classes preditas pelo modelo, fornecendo visualizações e explicações textuais para ele.

52 {if(is_empty(randnumber1.NAOK),rand(1,3),randnumber1.NAOK)}

53

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo, as visualizações das explicações geradas pelo método SHAP e a explicação textual dessa visualização, classifique o registro a seguir:



This plot shows the features sorted by magnitude of impact on the model in general, considering their absolute mean SHAP values. Each feature has a beeswarm and they are composed by dots that represent each observation and are distributed along the x-axis according to the observations' SHAP values. On the right side of the plot there is a vertical bar that gives the color shades of the dots that represent from high to low values of the features. Red dots represent high feature values, shades of purple represent medium values and blue dots represent low values. That means that along with the information of how much impact each feature has on the model output, it is possible to know what ranges of feature values have positive, negative or no impact. In this particular case, the 5 features that presented highest impact in the model were Sintoma 2, Sintoma 3, Sintoma 5, Sintoma 4 and Sintoma 1 respectively.

Sintoma 1	0,31
Sintoma 2	0,55
Sintoma 3	0,46
Sintoma 4	0,11
Sintoma 5	0,04

*

Favor escolher apenas uma das opções a seguir:

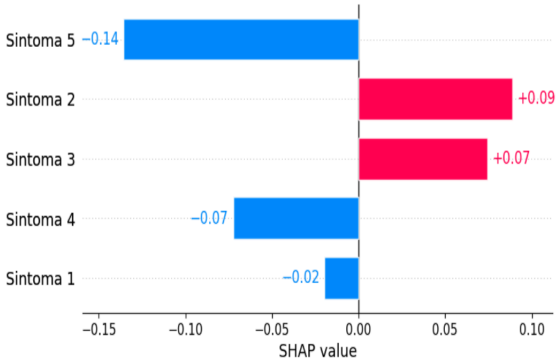
◯ Aprovado

◯ Rejeitado

54

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo, as visualizações das explicações geradas pelo método SHAP e a explicação textual dessa visualização, classifique o registro a seguir:

Sintoma 1	0,31
Sintoma 2	0,55
Sintoma 3	0,46
Sintoma 4	0,11
Sintoma 5	0,04



This plot shows the features sorted by magnitude of impact on the model considering their absolute SHAP values. It shows the magnitude of the impact as well as if the contribution was towards a class or another. The bars in red represent the SHAP values of features that contributed for the positive class, and therefore grow to the right side of the plot and have their value displayed with a positive sign. The bars is blue represent the SHAP values of features that contributed for the negative class, and therefore grow to the left side of the plot and have their value displayed with a negative sign. For the chosen observation, the features ['Sintoma 2' 'Sintoma 3'] had the highest contribution for the positive class respectively and the features ['Sintoma 5', 'Sintoma 4', 'Sintoma 1'] had the highest contribution for the negative class respectively.

*

Favor escolher apenas uma das opções a seguir:

◯ Aprovado

◯ Rejeitado

55

A partir da observação dos registros exibidos anteriormente, das suas classes reais, suas classes preditas pelo modelo, as visualizações das explicações geradas pelo método SHAP e a explicação textual dessa visualização, classifique o registro a seguir:

Sintoma 1	0,31
Sintoma 2	0,55
Sintoma 3	0,46
Sintoma 4	0,11
Sintoma 5	0,04

B
Study Session Material

The study session material consisted of a text explaining how the SHAP method works and a use case to illustrate its functioning, an example of instance of this use case along with the method's results, a list of all the information provided by the method and a list of questions to help in the construction of the visualizations. The questions are presented insection B.1).

B.1
Questions

1. How would you represent local explanations? (Local explanations give information about a prediction for a specific instance.)
2. How would you represent global explanations? (Global explanations clarify the behavior of the model as a whole, without focusing on specific instances.)
3. How would you represent the magnitude of impact of each feature?
4. How would you represent each feature's value?
5. How would you represent the expected value (base value)?
6. How would you represent the relationship between the shap values and the expected value to get to the prediction value?
7. How would you represent the relationship between the feature and its impact on the prediction?
8. How could you help the user understand which features are more important to the model and which are less important?
9. How could you help the user understand which features have a positive impact and which have a negative impact on the prediction?
10. If you have the training data, what would you like to see when the system makes a mistake? How would you like to see this information?

C
Study Material

The study material consisted of an informed consent form (Termo de Consentimento Livre e Esclarecido TCLE), a respondent profile questionnaire, and a simulation questionnaire, all presented insection C.1).

C.1
Study Form

Estudo sobre eficácia de representações visuais de explicações geradas pelo método SHAP de Explainable Artificial Intelligence

Olá!

Este questionário faz parte de um estudo acadêmico sobre a eficácia de representações visuais geradas pelo método **SHAP** para explicar os resultados de modelos de Machine Learning. O objetivo principal do campo de **Explainable Artificial Intelligence (XAI)** é criar métodos que tornem os modelos de Machine Learning complexos mais interpretáveis, transparentes e confiáveis.

O **SHAP (SHapley Additive exPlanations)** é um desses métodos, baseado em conceitos da teoria dos jogos, que atribui a cada atributo de um modelo um valor representando seu impacto em uma predição específica. Este estudo avalia a eficácia de representações visuais das explicações geradas pelo SHAP com base nos seguintes fatores:

- **Entendibilidade**: A explicação ajuda a compreender como o modelo toma decisões?
- **Utilidade**: A explicação é útil para tomar melhores decisões e ações?
- **Confiança**: A explicação aumenta a confiança no modelo?
- **Informatividade**: A explicação fornece informações suficientes para entender as decisões do modelo?
- **Satisfação**: A explicação atende às expectativas do usuário?

Contexto:

Imagine que você trabalha em um banco no setor de concessão de empréstimos. Cada solicitação é avaliada por um modelo de Machine Learning, que decide se o empréstimo deve ser aprovado ou negado. O modelo considera os seguintes atributos:

- Idade
- Renda anual
- Score de crédito
- Tempo (em anos) de experiência profissional
- Tempo (em anos) do histórico de crédito
- Indicador de inadimplência de empréstimos anteriores (binário: sim/não)
- Valor do empréstimo como percentual da renda anual
- Taxa de juros do empréstimo
- Valor total do empréstimo

Frequentemente, especialmente quando o empréstimo é negado, clientes solicitam explicações para entender os motivos por trás da decisão. Sem informações sobre como o modelo funciona, você não conseguiria fornecer respostas detalhadas. Além disso, o modelo foi treinado com dados históricos e, ocasionalmente, pode tomar decisões erradas ou até mesmo injustas.

Para melhorar sua compreensão das decisões do modelo e fornecer respostas mais precisas aos clientes, você começou a usar o método SHAP, que gera explicações sobre como o modelo chega às suas conclusões.

O que será solicitado no questionário:

Nas próximas etapas do questionário, você será guiado por diferentes cenários:

1. Inicialmente, algumas solicitações de empréstimo serão apresentadas sem explicação, junto com a decisão do modelo e a decisão correta (aprovado ou rejeitado). A decisão correta de uma solicitação é a decisão que um especialista que trabalha no setor de consessão de empréstimos tomaria caso estivesse analizando ela. Aqui, você tentará compreender as decisões sem o auxílio das explicações.
2. Em seguida, você verá novas solicitações e tentará prever qual seria a decisão do modelo para cada caso.
3. Depois, você será apresentado a solicitações acompanhadas de explicações geradas pelo método SHAP, junto com a decisão do modelo e a decisão correta. Essa etapa tem como objetivo ajudar você a entender melhor o funcionamento do modelo.
4. Em seguida, você tentará novamente prever as decisões do modelo em novos casos e avaliará as explicações fornecidas com base nos cinco fatores mencionados anteriormente: Entendibilidade, Utilidade, Confiança, Informatividade e Satisfação.
5. Por fim, iremos te pedir que compare dois conjuntos de visualizações com base nos mesmos fatores avaliados anteriormente.

Agradecemos imensamente sua participação neste estudo. Suas respostas são fundamentais para nossa pesquisa!

* Indicates required question

Termo de Consentimento Livre e Esclarecido

Apresentação e Termo de Consentimento

Você está sendo convidado(a) a participar de uma pesquisa que investiga a eficácia do método de explicação de modelos de aprendizado de máquina SHAP. A sua participação nesta pesquisa é totalmente voluntária.

Título da Pesquisa: Eficácia de representações visualis de explicações geradas pelo método SHAP em classificações.

Os pesquisadores responsáveis pelo estudo poderão fornecer qualquer esclarecimento sobre o mesmo, assim como tirar dúvidas.

Os dados para contato são os seguintes:

Pesquisador Responsável: Bianca Moreira Cunha

Endereço: bcunha@inf.puc-rio. br

1) Objetivo Geral: Os objetivos primários desta pesquisa são: (i) Avaliar a eficácia de visualizações geradas a partir das saídas de um método de explicação de modelos de machine learning

2) Objetivos Específicos: (a) Avaliar se as visualizações das saídas geradas por método de explicação de modelos de machine learning melhoram a interpretabilidade de modelos (b) Avaliar se visualizações geradas em conjunto com pesquisadores de visualização da informação são eficazes (c) Avaliar se visualizações que foram geradas em conjunto com especialistas de visualização são mais eficazes do que visualizações que não tiveram participação destes pesquisadores na sua elaboração.

3) Riscos: (a) Toda pesquisa realizada com seres humanos apresenta riscos. No entanto, os riscos apresentados nesta pesquisa são mínimos, pois as tecnologias utilizadas não alteram aspectos fisiológicos, psicológicos ou sociais dos participantes. Os questionários aplicados não tratam de aspectos relacionados à pessoa em si, mas sim às técnicas e processos utilizados. (b) Para evitar constrangimentos, sua identidade será mantida em sigilo. Além disso, você poderá ausentar-se do local do estudo a qualquer momento,

caso sinta-se desconfortável. (c) Todos os materiais necessários para realização da pesquisa serão fornecidos pelos próprios pesquisadores, de forma que não acarrete custos para os participantes. Além disso, está assegurado o direito a indenizações e qualquer tipo de assistência necessária para reparação a qualquer prejuízo causado pela pesquisa.

4) Benefícios: Através desta pesquisa espera-se identificar problemas e oportunidades de melhorias no uso de técnicas e processos de Interação Humano-Computador. Além disso, espera-se contribuir para a melhoria da aprendizagem destas técnicas por estudantes do ensino superior de computação, contribuindo desta forma para sua formação. Os benefícios gerados serão: (a) Materiais didáticos aprimorados e enriquecidos com base nas experiências práticas dos estudantes universitários em sala de aula, disponibilizados gratuitamente para todos os participantes do estudo e para estudantes de computação em geral; (b) Melhorias no uso das técnicas propostas, para que estas técnicas, posteriormente, possam ser mais bem adaptadas para serem utilizadas na indústria de software; (c) Melhor apoio ao desenvolvimento de softwares inovadores, com maior qualidade, mais fáceis de utilizar para os usuários e, em última análise, que melhorem a sociedade como um todo.

5) Procedimentos: Você responderá a um questionário com perguntas relacionadas à avaliação de método de explicação de modelos de machine learning. Todos os questionários respondidos serão descartados após a sua análise. O questionário será respondido de forma anônima, portanto seu nome não será utilizado em nenhum momento durante a análise ou apresentação dos resultados.

6) Tratamento de possíveis riscos e desconfortos: A sua participação consiste tão somente no preenchimento de um questionário. Ainda assim, serão tomadas todas as providências durante a coleta desses dados de forma a garantir a sua privacidade e seu anonimato. Os dados coletados durante o estudo destinam-se estritamente a atividades da pesquisa. Desta forma, não serão utilizados para qualquer forma de avaliação profissional ou pessoal.

7) Custos: Você não terá nenhum gasto ou ônus com a sua participação no estudo e, também, não receberá qualquer espécie de reembolso ou gratificação devido à participação nesta pesquisa.

8) Confidencialidade da Pesquisa: Toda informação coletada neste estudo é confidencial, e seu nome e o da organização não serão identificados de modo algum, a não ser em caso de autorização explícita para esse fim.

9) Participação: Sua participação neste estudo é muito importante e voluntária. Você tem o direito de não querer participar ou de sair deste estudo a qualquer momento, sem penalidades. Você também tem o direito de se recusar a responder a qualquer pergunta do questionário. Para participar deste estudo você deverá ser maior de idade (ter 18 anos ou mais).

10) Declaração de Consentimento: Li ou alguém leu para mim as informações contidas neste documento antes de assinar este termo de consentimento. Declaro que toda a linguagem técnica utilizada na descrição deste estudo de pesquisa foi explicada satisfatoriamente e que recebi respostas para todas as minhas dúvidas. Confirmo também que recebi uma cópia deste Termo de Consentimento Livre e Esclarecido. Compreendo que sou livre para me retirar do estudo em qualquer momento, sem qualquer penalidade. Declaro ter mais de 18 anos e dou meu consentimento de livre e espontânea vontade para participar deste estudo.

Toda a dúvida a respeito desta pesquisa, poderá ser perguntada diretamente para Bianca Cunha, através do e-mail: bcunha@inf.puc-rio.br . Enquanto, as objeções a respeito da conduta ética poderão ser questionadas ao Câmara de Ética em Pesquisa da PUC-Rio.

* Required

Por favor, concorde com o Termo de Consentimento Livre e Esclarecido (TCLE) abaixo para continuar. *

Se você discordar dos termos apresentados neste questionário sua participação será excluída deste estudo e seus dados não serão utilizados assim como não será necessário preencher as questões apresentadas neste questionário. A sua participação é de extrema importância para os autores desta pesquisa e todos os mesmos obedecem e seguem restritivamente os termos postados no TCLE devidamente assinado por você.O

1. Você concorda com o Termo de Consentimento Livre e Esclarecido? *

   Mark only one oval.

   - ◯ Concordo
   - ◯ Discordo

Perguntas de perfil

2. Qual é o seu grau de instrução?

   Mark only one oval.

   - ◯ Cursando graduação
   - ◯ Graduação completa
   - ◯ Cursando mestrado
   - ◯ Mestrado completo
   - ◯ Cursando doutorado
   - ◯ Doutorado completo

3. A sua área de estudo é STEM (Science, Technology, Engineering, Math)?

   Mark only one oval.

   - ◯ Sim
   - ◯ Não

4. Qual é o seu nível de conhecimento sobre Machine Learning/IA?

   Mark only one oval.

   - ◯ Nenhum
   - ◯ Básico
   - ◯ Intermediário
   - ◯ Avançado
   - ◯ Especialista

5. Qual é o seu nível de conhecimento sobre XAI (Explainable Artificial Intelligence)?

   Mark only one oval.

   - ◯ Nenhum
   - ◯ Básico
   - ◯ Intermediário
   - ◯ Avançado
   - ◯ Especialista

6. Qual é o seu conhecimento sobre Visualização da Informação?

   Mark only one oval.

   - ◯ Nenhum
   - ◯ Básico
   - ◯ Intermediário
   - ◯ Avançado
   - ◯ Especialista

7. Quais aspectos você considera mais importantes em uma representação visual de uma explicação de uma predição? (Selecione todos que se aplicam)

   Check all that apply.

   - ☐ Entendibilidade (A explicação ajuda a compreender como o modelo toma decisões.)
   - ☐ Utilidade (A explicação é útil para tomar melhores decisões e ações.)
   - ☐ Confiança (A explicação aumenta a confiança no modelo.)
   - ☐ Informatividade (A explicação fornece informações suficientes para entender as decisões do modelo.)
   - ☐ Satisfação (A explicação atende às expectativas do usuário.)

8. O que você espera de uma representação visual de uma explicação de uma predição gerada por um modelo de Machine Learning?

Observe as instâncias abaixo, suas classificações dadas pelo modelo e as classificações corretas.

Instância 1

Variável	Valor
Idade	24
Renda anual	36947,00
Tempo (anos) de experiência profissional	0
Valor do empréstimo	9500,00
Taxa de juros do empréstimo	13,92
Valor do empréstimo como percentual da renda anual	0,26
Tempo (anos) do histórico de crédito	2
Score de crédito	680
Indicador de inadimplência de empréstimos anteriores	0

Classificação do modelo
Aprovado

Classificação correta
Aprovado

Instância 2

Variável	Valor
Idade	29
Renda anual	129252,00
Tempo (anos) de experiência profissional	9
Valor do empréstimo	11760,00
Taxa de juros do empréstimo	11,18
Valor do empréstimo como percentual da renda anual	0,09
Tempo (anos) do histórico de crédito	10
Score de crédito	648
Indicador de inadimplência de empréstimos anteriores	0

Classificação do modelo
Rejeitado

Classificação correta
Rejeitado

Instância 3

Variável	Valor
Idade	30
Renda anual	73100,00
Tempo (anos) de experiência profissional	7
Valor do empréstimo	25257,00
Taxa de juros do empréstimo	13,47
Valor do empréstimo como percentual da renda anual	0,35
Tempo (anos) do histórico de crédito	6
Score de crédito	708
Indicador de inadimplência de empréstimos anteriores	0

Classificação do modelo
Aprovado

Classificação correta
Rejeitado

Instância 4

Variável	Valor
Idade	22
Renda anual	117469,00
Tempo (anos) de experiência profissional	0
Valor do empréstimo	14000,00
Taxa de juros do empréstimo	11,48
Valor do empréstimo como percentual da renda anual	0,12
Tempo (anos) do histórico de crédito	2
Score de crédito	686
Indicador de inadimplência de empréstimos anteriores	0

Classificação do modelo
Rejeitado

Classificação correta
Aprovado

A partir do que você observou nas instâncias acima, classifique as instâncias seguintes de acordo com o que você acredita que seria a classificação do MODELO.

9. Classifique a instância abaixo:

Variável	Valor
Idade	23
Renda anual	36812,00
Tempo (anos) de experiência profissional	3
Valor do empréstimo	7925,00
Taxa de juros do empréstimo	15,27
Valor do empréstimo como percentual da renda anual	0,22
Tempo (anos) do histórico de crédito	4
Score de crédito	660
Indicador de inadimplência de empréstimos anteriores	0

Mark only one oval.

- ◯ Aprovado
- ◯ Rejeitado

10. Quão confiante você está da sua classificação?

Mark only one oval.

	1	2	3	4	5	
Pou	◯	◯	◯	◯	◯	Muito confiante

11. Classifique a instância abaixo:

Variável	Valor
Idade	26
Renda anual	107633,00
Tempo (anos) de experiência profissional	1
Valor do empréstimo	8000,00
Taxa de juros do empréstimo	10,99
Valor do empréstimo como percentual da renda anual	0,07
Tempo (anos) do histórico de crédito	3
Score de crédito	542
Indicador de inadimplência de empréstimos anteriores	1

Mark only one oval.

- ◯ Aprovado
- ◯ Rejeitado

12. Quão confiante você está da sua classificação?

Mark only one oval.

	1	2	3	4	5	
Pou	◯	◯	◯	◯	◯	Muito confiante

13. Classifique a instância abaixo:

Variável	Valor
Idade	23
Renda anual	39925,00
Tempo (anos) de experiência profissional	1
Valor do empréstimo	12300,00
Taxa de juros do empréstimo	11,01
Valor do empréstimo como percentual da renda anual	0,31
Tempo (anos) do histórico de crédito	3
Score de crédito	537
Indicador de inadimplência de empréstimos anteriores	0

Mark only one oval.

- ◯ Aprovado
- ◯ Rejeitado

14. Quão confiante você está da sua classificação?

Mark only one oval.

	1	2	3	4	5	
Pou	◯	◯	◯	◯	◯	Muito confiante

15. Classifique a instância abaixo:

Variável	Valor
Idade	27
Renda anual	38607,00
Tempo (anos) de experiência profissional	6
Valor do empréstimo	6000,00
Taxa de juros do empréstimo	12,84
Valor do empréstimo como percentual da renda anual	0,16
Tempo (anos) do histórico de crédito	8
Score de crédito	704
Indicador de inadimplência de empréstimos anteriores	0

Mark only one oval.

- ◯ Aprovado
- ◯ Rejeitado

16. Quão confiante você está da sua classificação?

Mark only one oval.

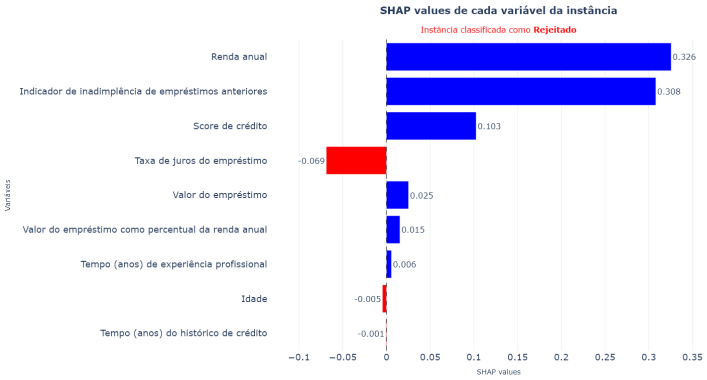
	1	2	3	4	5	
Pou	◯	◯	◯	◯	◯	Muito confiante

Nessa seção serão apresentadas visualizações de explicações de 4 NOVAS instâncias (4 visualizações para cada instância). Observe as representações e tente tirar um entendimento do comportamento do modelo através delas.

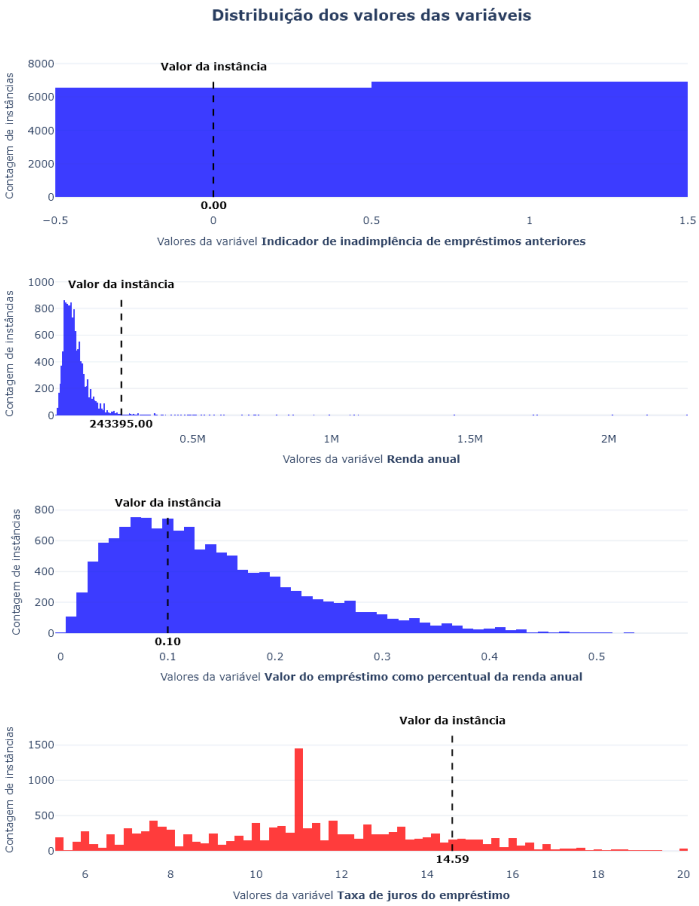
Instância 1
Classificação do modelo: REJEITADO

Classificação correta: APROVADO

Visualização A

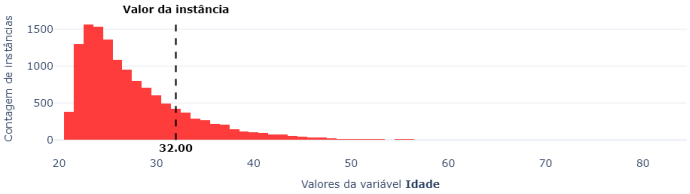


Visualização B


Distribuição dos valores das variáveis
Valor da instância
Contagem de instâncias
8000
6000
4000
2000
0
0.00
−0.5
0
0.5
1
1.5
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
1000
800
600
400
200
0
243395.00
0.5M
1M
1.5M
2M
Valores da variável Renda anual
Valor da instância
800
600
400
200
0
0.10
0
0.1
0.2
0.3
0.4
0.5
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
1500
1000
500
0
14.59
6
8
10
12
14
16
18
20
Valores da variável Taxa de juros do empréstimo


Valor da instância
800
600
400
200
0
Contagem de instâncias
694.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
1000
500
0
Contagem de instâncias
25000.00
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo

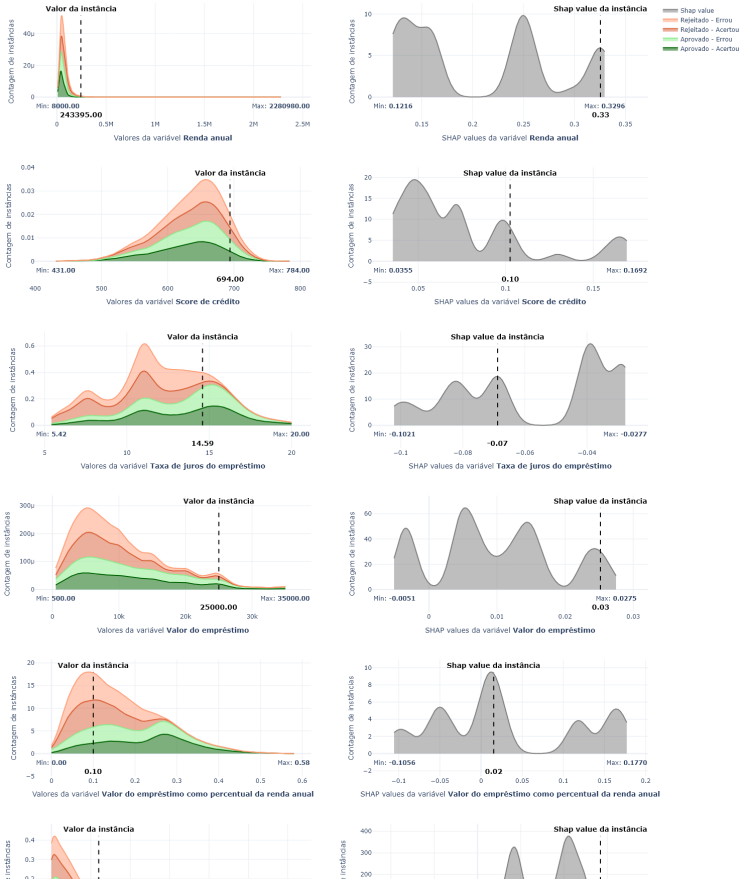

Valor da instância
1500
1000
500
0
Contagem de instâncias
32.00
20
30
40
50
60
70
80
Valores da variável Idade

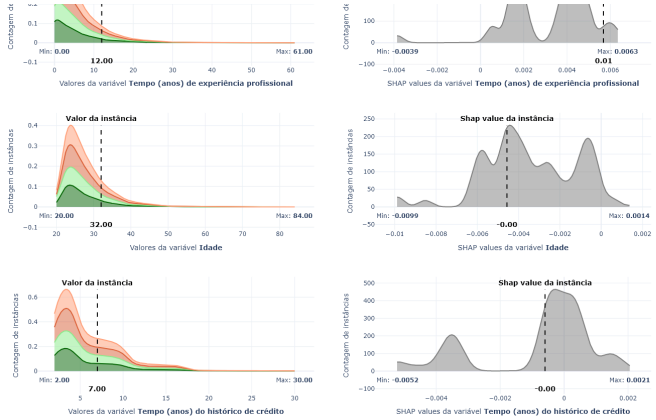

Valor da instância
3000
2000
1000
0
Contagem de instâncias
7.00
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito


Valor da instância
3000
ncias

Visualização C

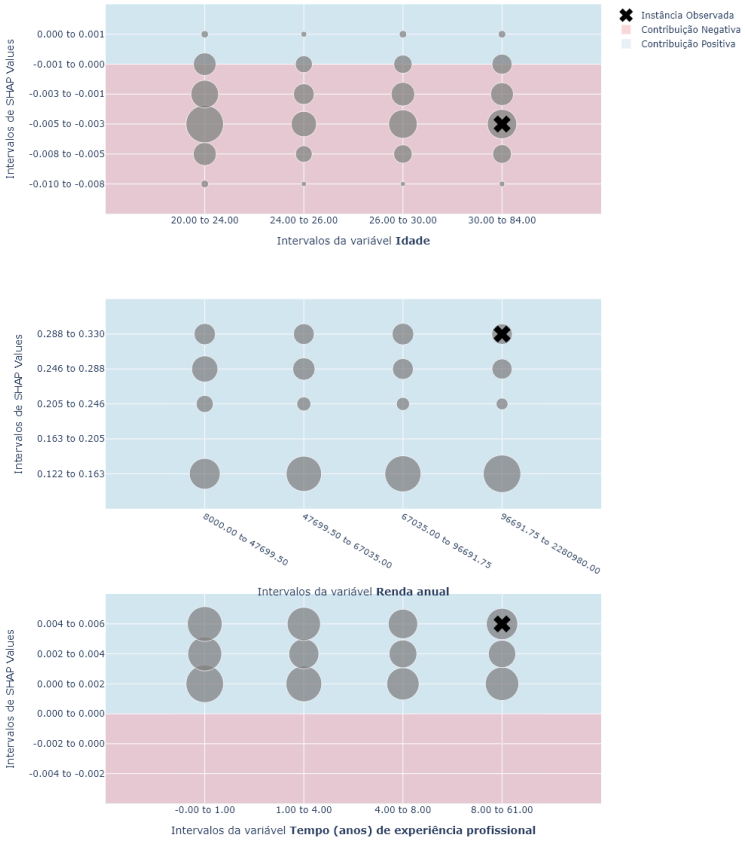
Distribuição de valores e de SHAP values de cada variável

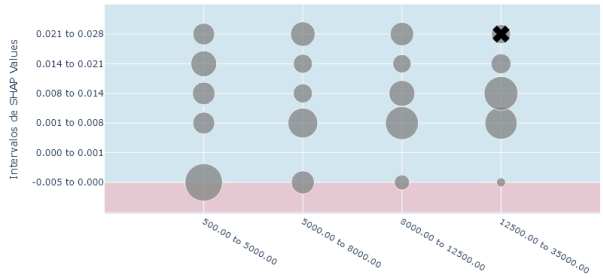


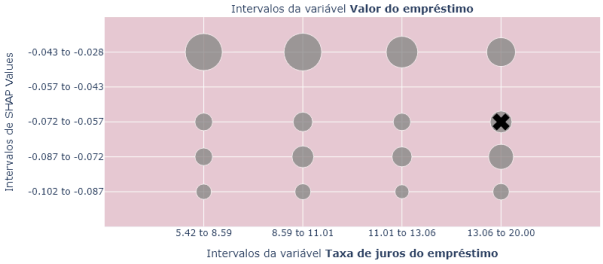


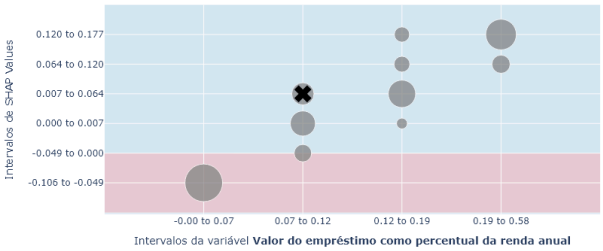
Visualização D

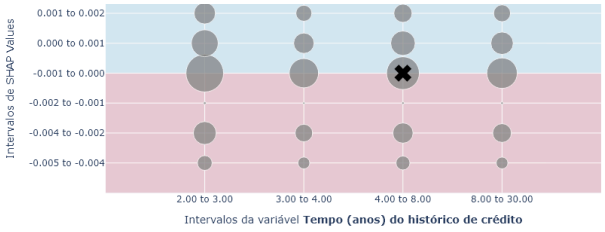
Concentração de instâncias por range de SHAP values e quartis das variáveis

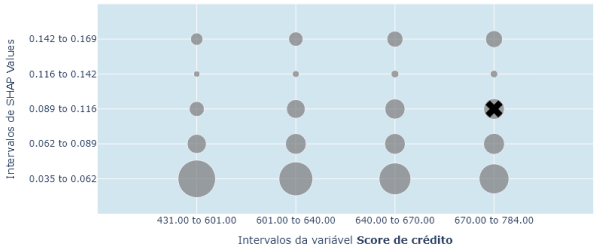



Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável **Valor do empréstimo**


Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável **Taxa de juros do empréstimo**


Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável **Valor do empréstimo como percentual da renda anual**



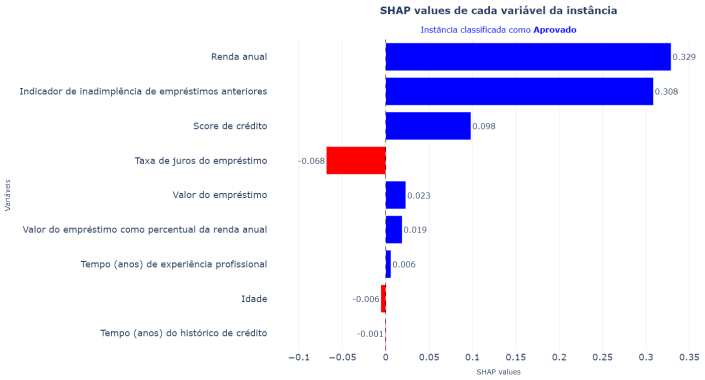


Nessa seção serão apresentadas visualizações de explicações de 4 NOVAS instâncias (4 visualizações para cada instância). Observe as representações e tente tirar um entendimento do comportamento do modelo através delas.

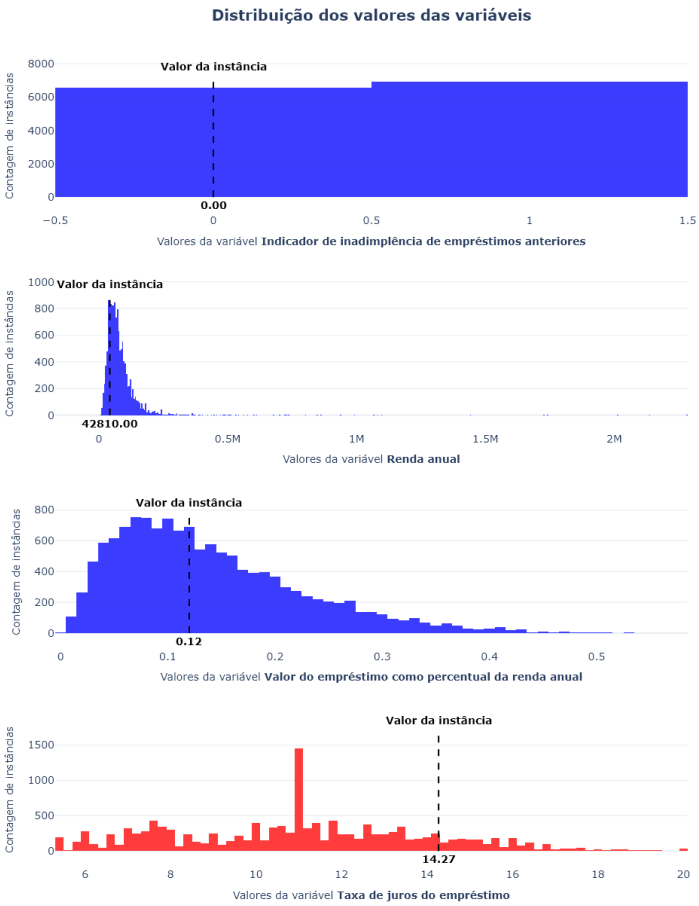
Instância 2
Classificação do modelo: APROVADO

Classificação correta: REJEITADO

Visualização A



Visualização B


Distribuição dos valores das variáveis
Valor da instância
0.00
Contagem de instâncias
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
42810.00
Contagem de instâncias
Valores da variável Renda anual
Valor da instância
0.12
Contagem de instâncias
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
14.27
Contagem de instâncias
Valores da variável Taxa de juros do empréstimo


Valor da instância
Contagem de instâncias
800
600
400
200
0
549.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
Contagem de instâncias
1000
500
0
5000.00
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo

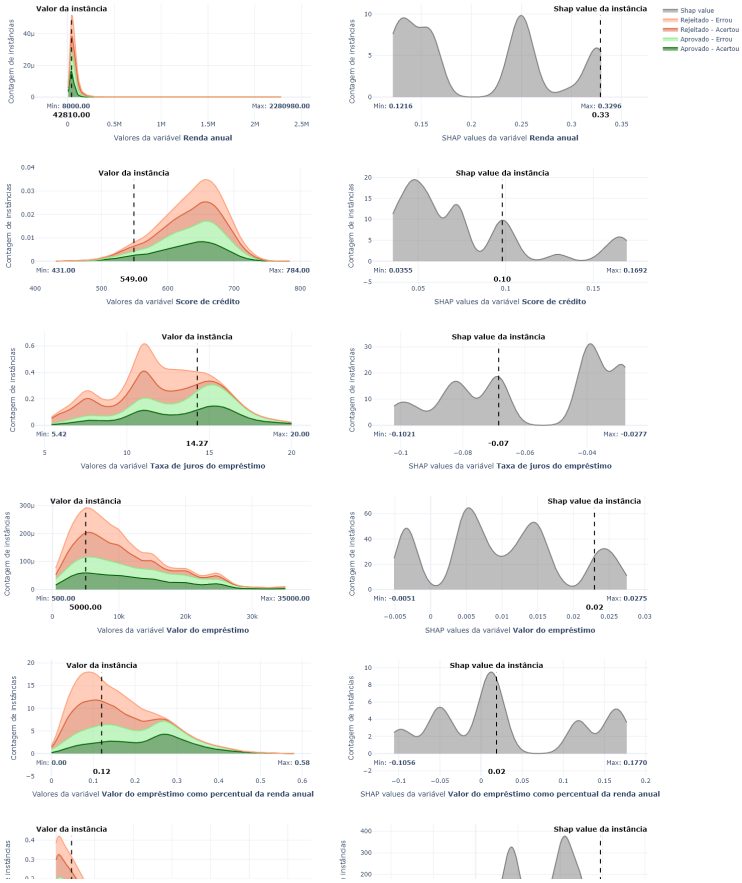

Valor da instância
Contagem de instâncias
1500
1000
500
0
28.00
20
30
40
50
60
70
80
Valores da variável Idade

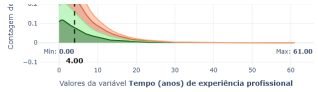

Valor da instância
Contagem de instâncias
3000
2000
1000
0
10.00
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito

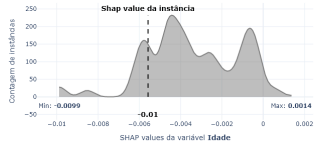

Valor da instância
3000

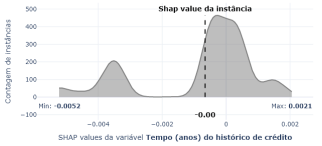
Visualização C

Distribuição de valores e de SHAP values de cada variável



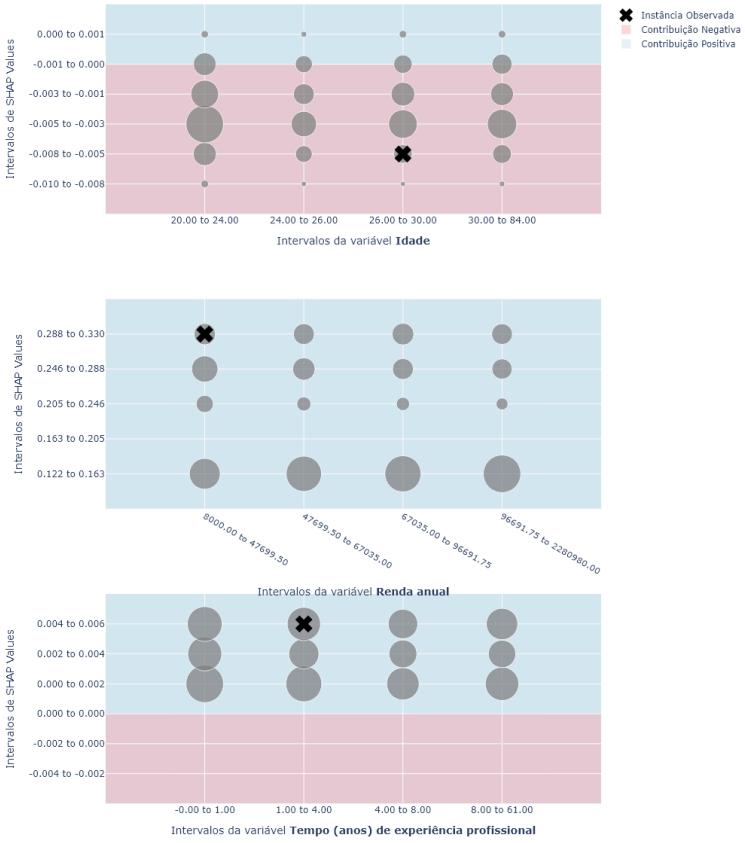


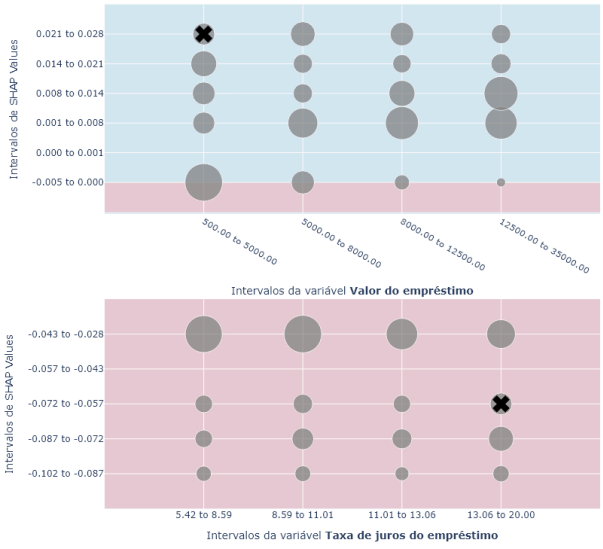


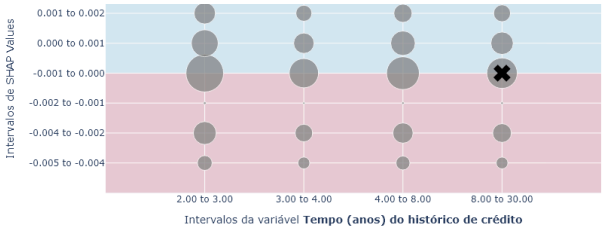


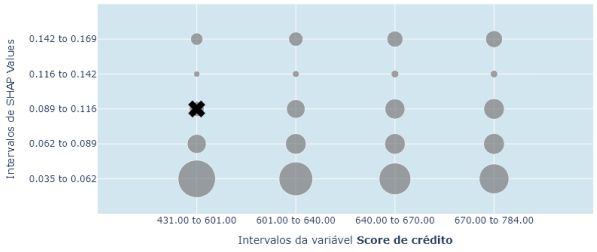
Visualização D

Concentração de instâncias por range de SHAP values e quartis das variáveis




Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável **Valor do empréstimo**
Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável **Taxa de juros do empréstimo**
Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável **Valor do empréstimo como percentual da renda anual**



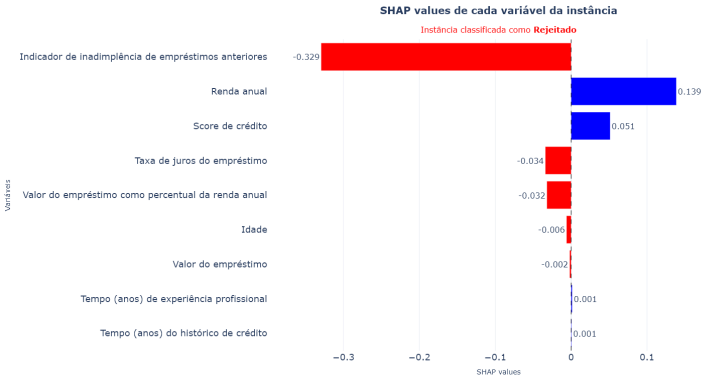


Nessa seção serão apresentadas visualizações de explicações de 4 NOVAS instâncias (4 visualizações para cada instância). Observe as representações e tente tirar um entendimento do comportamento do modelo através delas.

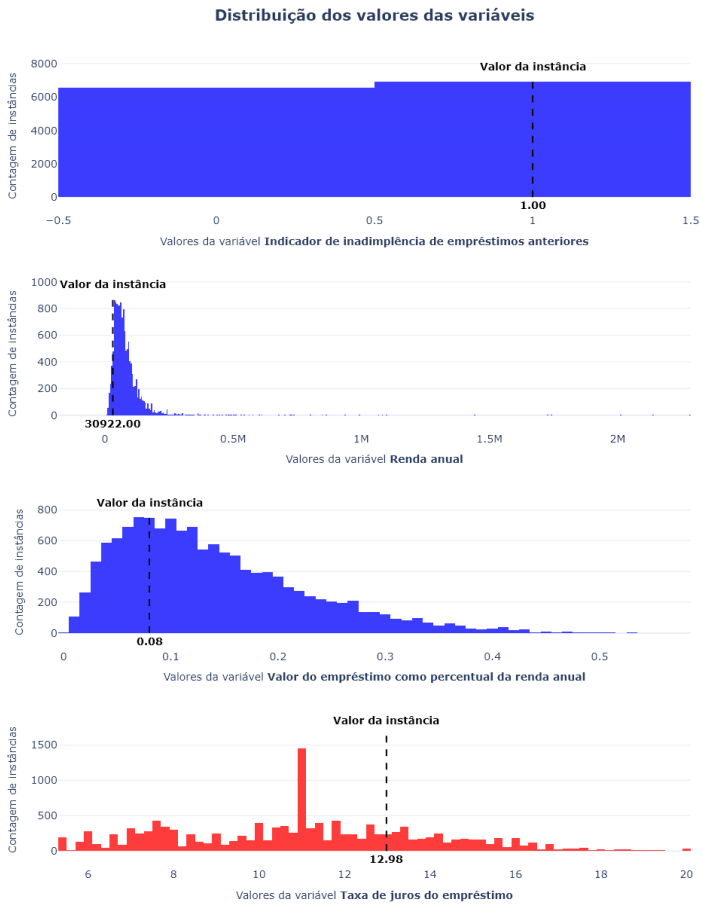
Instância 3
Classificação do modelo: REJEITADO

Classificação correta: REJEITADO

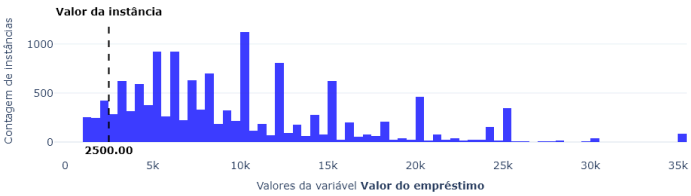
Visualização A



Visualização B


Distribuição dos valores das variáveis
Valor da instância
Contagem de instâncias
8000
6000
4000
2000
0
1.00
−0.5
0
0.5
1
1.5
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
1000
800
600
400
200
0
30922.00
0
0.5M
1M
1.5M
2M
Valores da variável Renda anual
Valor da instância
800
600
400
200
0
0.08
0
0.1
0.2
0.3
0.4
0.5
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
1500
1000
500
0
12.98
6
8
10
12
14
16
18
20
Valores da variável Taxa de juros do empréstimo


Valor da instância
800
600
400
200
0
Contagem de instâncias
690.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
1000
500
0
Contagem de instâncias
2500.00
0
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo

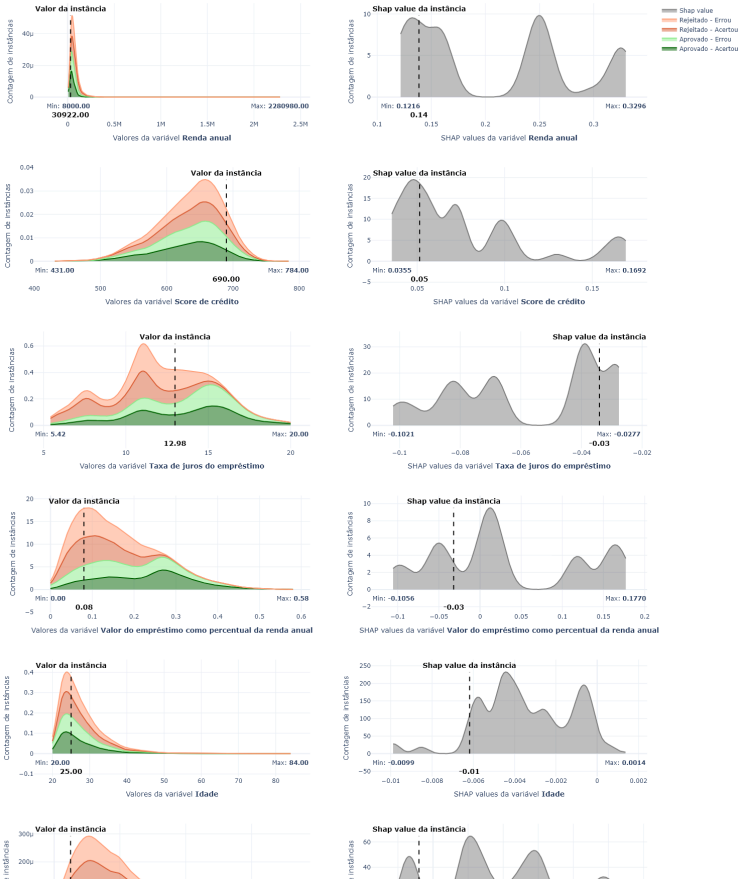

Valor da instância
1500
1000
500
0
Contagem de instâncias
25.00
20
30
40
50
60
70
80
Valores da variável Idade

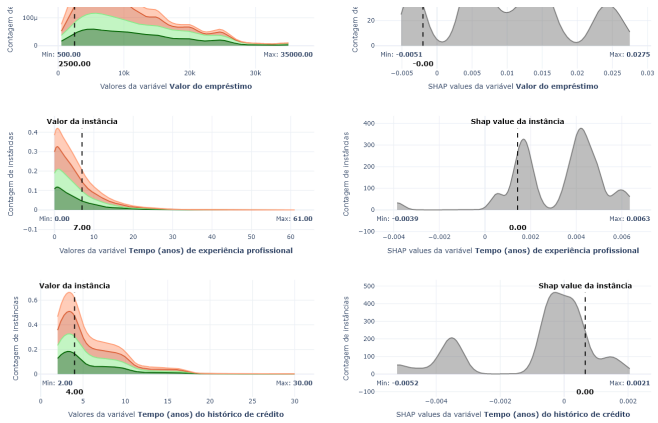

3000
Valor da instância
2000
1000
0
Contagem de instâncias
4.00
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito


Valor da instância
3000
ncias

Visualização C

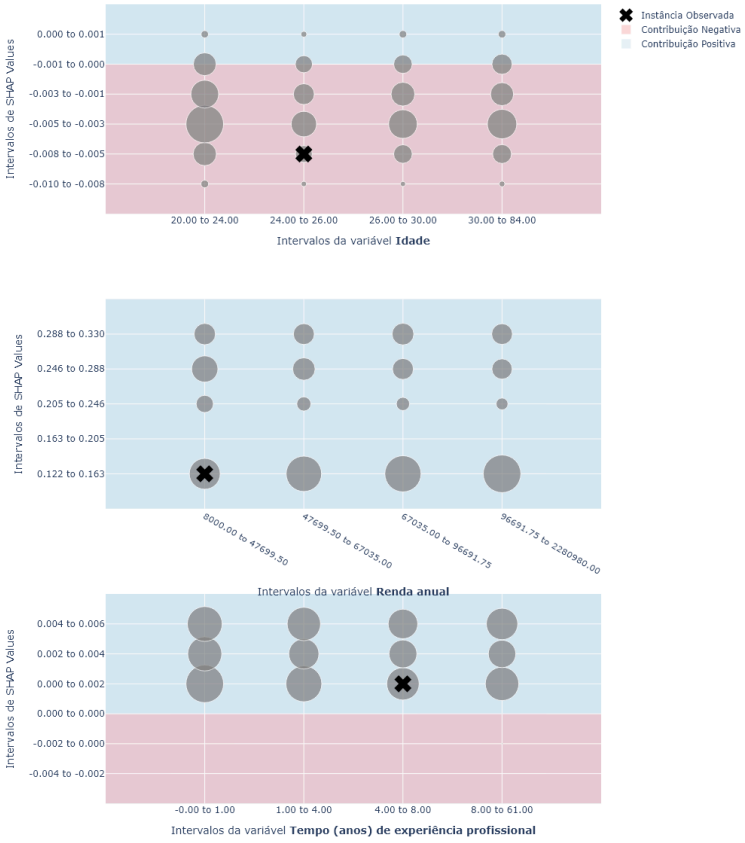
Distribuição de valores e de SHAP values de cada variável

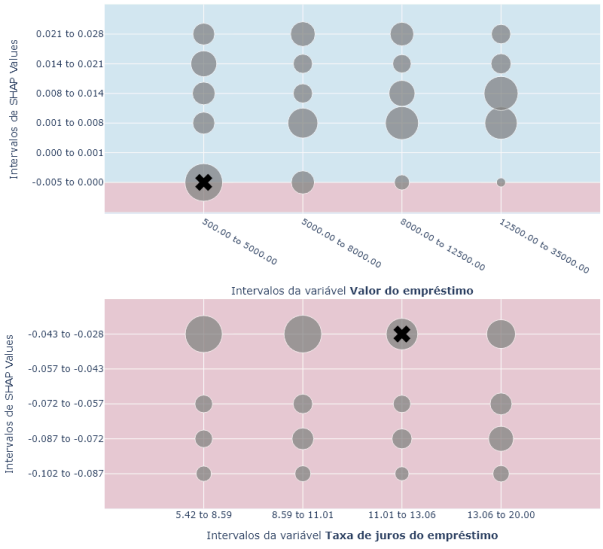


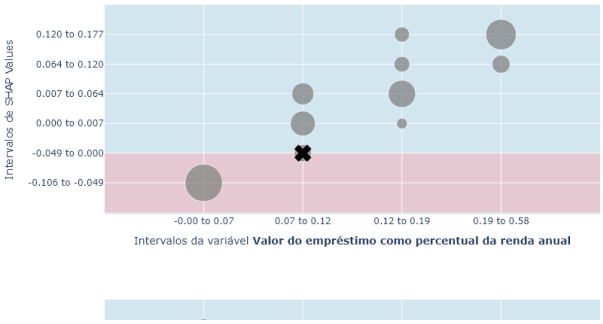


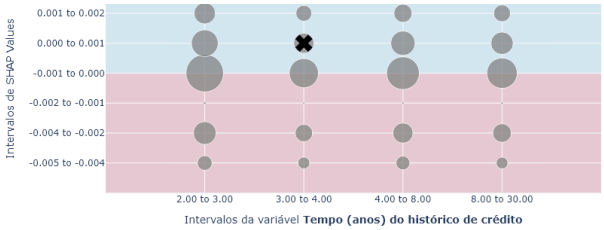
Visualização D

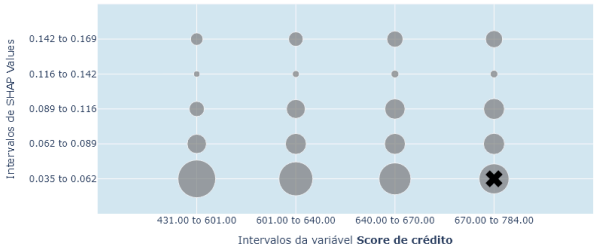
Concentração de instâncias por range de SHAP values e quartis das variáveis




Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável **Valor do empréstimo**
Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável **Taxa de juros do empréstimo**


Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável **Valor do empréstimo como percentual da renda anual**



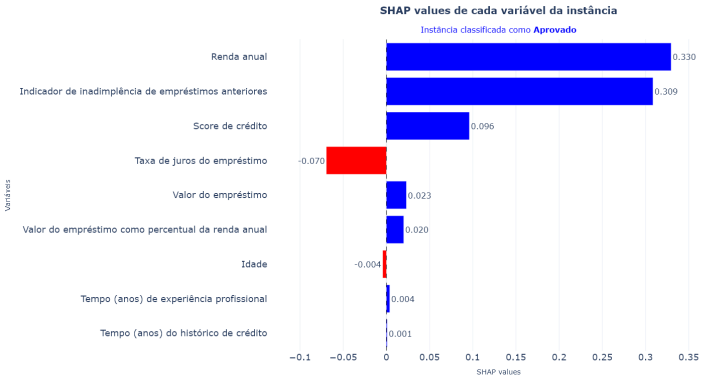


Nessa seção serão apresentadas visualizações de explicações de 4 NOVAS instâncias (4 visualizações para cada instância). Observe as representações e tente tirar um entendimento do comportamento do modelo através delas.

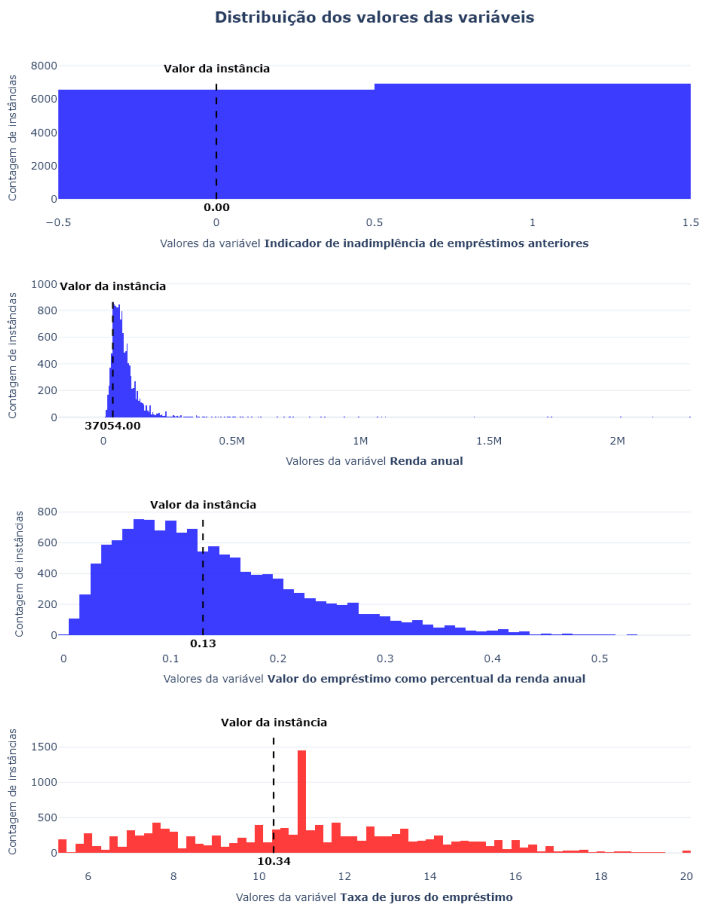
Instância 4
Classificação do modelo: APROVADO

Classificação correta: APROVADO

Visualização A

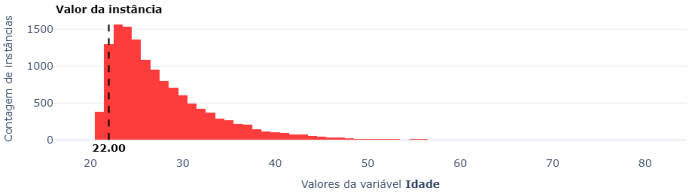


Visualização B

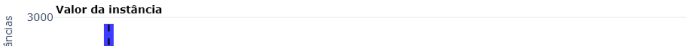

Distribuição dos valores das variáveis
Valor da instância
Contagem de instâncias
0.00
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
37054.00
Valores da variável Renda anual
Valor da instância
0.13
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
10.34
Valores da variável Taxa de juros do empréstimo

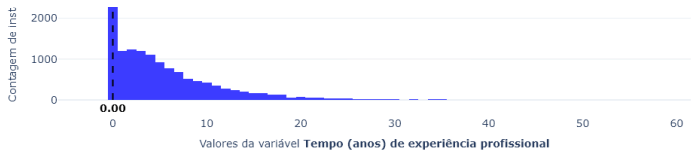

Valor da instância
800
600
400
200
0
Contagem de instâncias
562.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
1000
500
0
Contagem de instâncias
5000.00
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo


Valor da instância
1500
1000
500
0
Contagem de instâncias
22.00
20
30
40
50
60
70
80
Valores da variável Idade

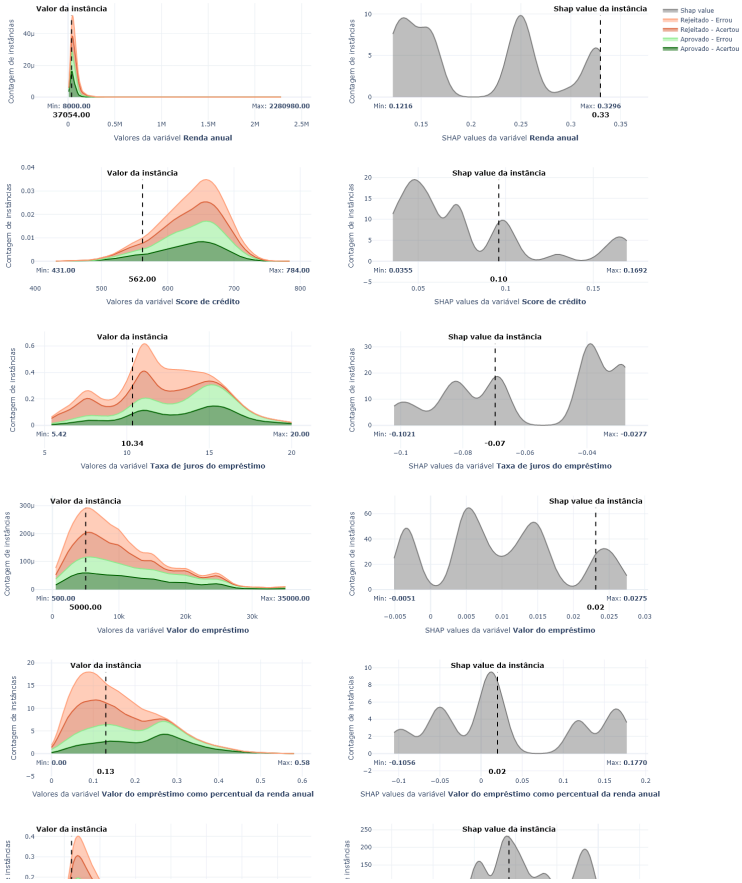

3000
Valor da instância
2000
1000
0
Contagem de instâncias
4.00
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito

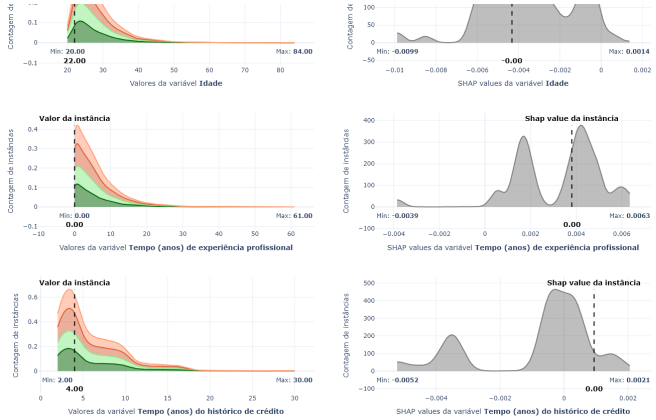

Valor da instância
3000
ncias



Visualização C

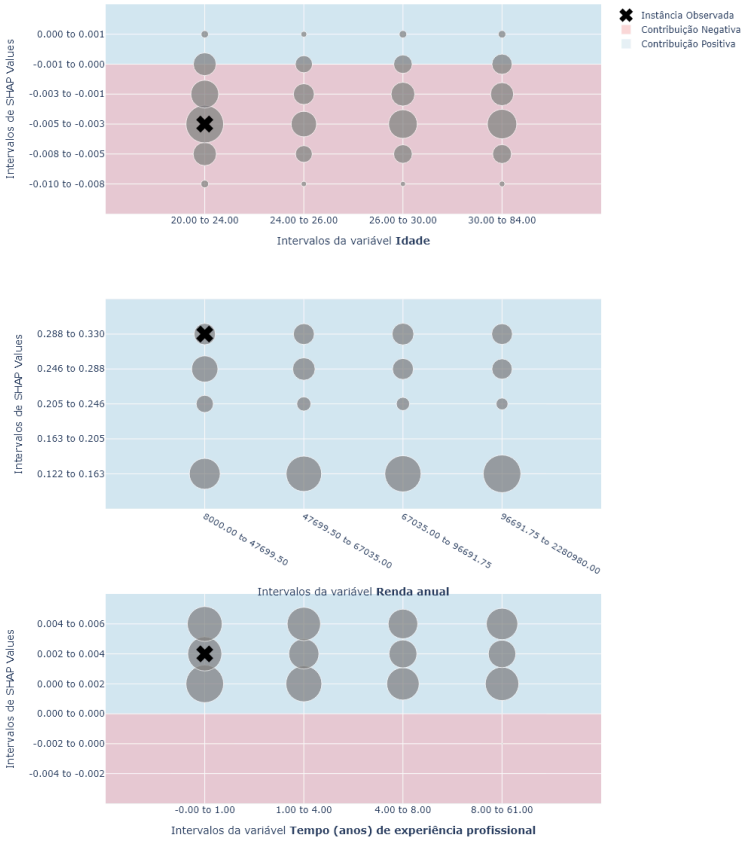
Distribuição de valores e de SHAP values de cada variável

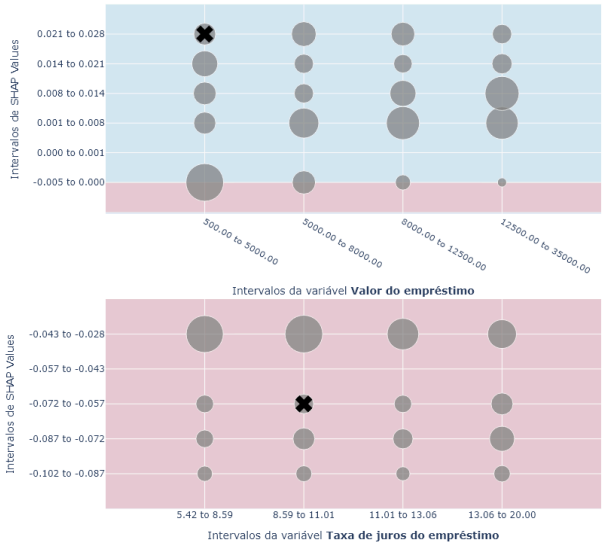


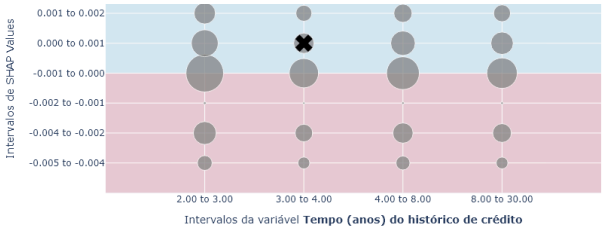


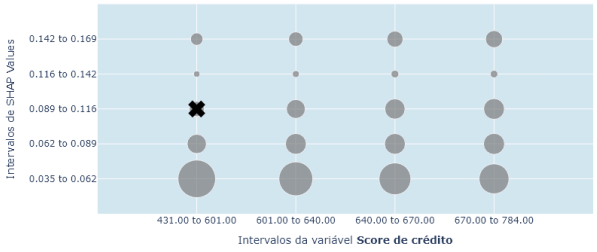
Visualização D

Concentração de instâncias por range de SHAP values e quartis das variáveis




Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável Valor do empréstimo
Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável Taxa de juros do empréstimo
Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável Valor do empréstimo como percentual da renda anual

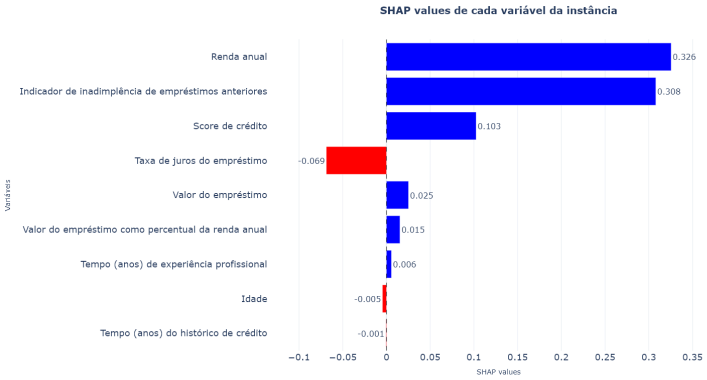




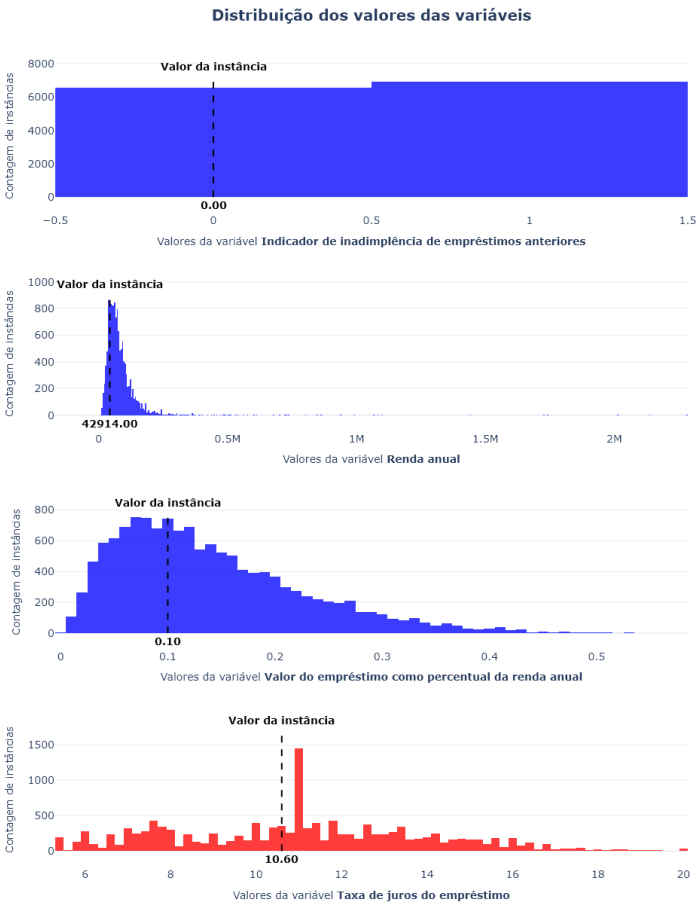
Nessa seção serão apresentadas visualizações de explicações de 4 NOVAS instâncias (4 visualizações para cada instância). Observe as representações e classifique cada uma delas como você acha que o MODELO as classificaria e depois responda a algumas perguntas sobre a sua percepção sobre elas.

Instância 1

Visualização A

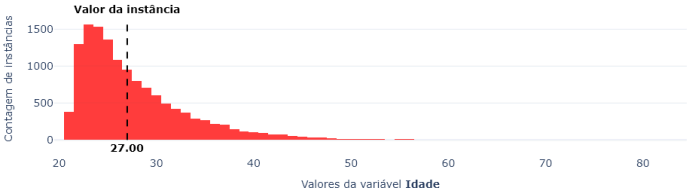


Visualização B


Distribuição dos valores das variáveis
Valor da instância
Contagem de instâncias
0.00
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
42914.00
Valores da variável Renda anual
Valor da instância
0.10
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
10.60
Valores da variável Taxa de juros do empréstimo


Valor da instância
Contagem de instâncias
800
600
400
200
0
591.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
Contagem de instâncias
1000
500
0
4197.00
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo

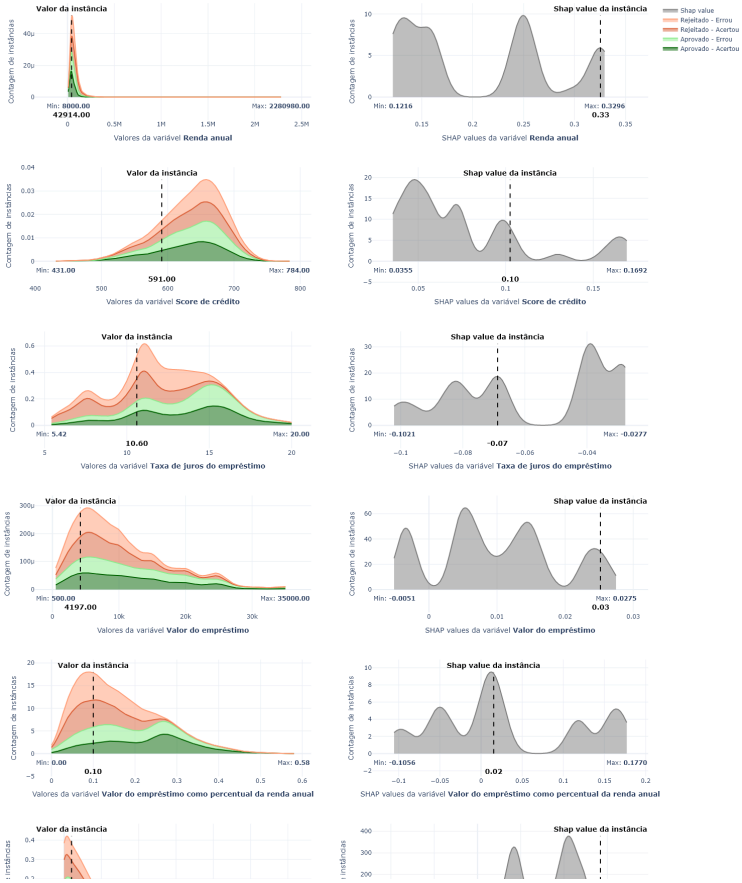

Valor da instância
Contagem de instâncias
1500
1000
500
0
27.00
20
30
40
50
60
70
80
Valores da variável Idade

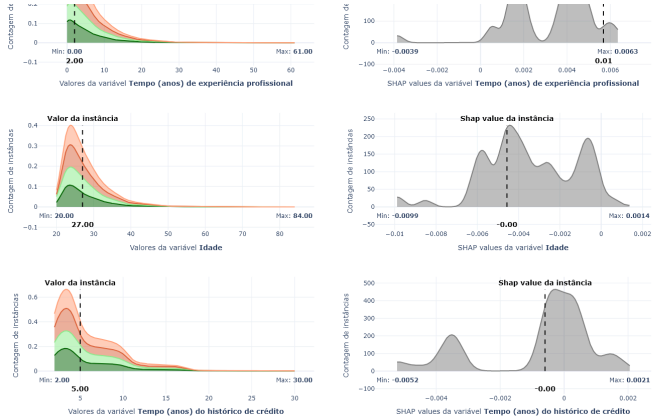

Valor da instância
Contagem de instâncias
3000
2000
1000
0
5.00
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito


Valor da instância
3000

Visualização C

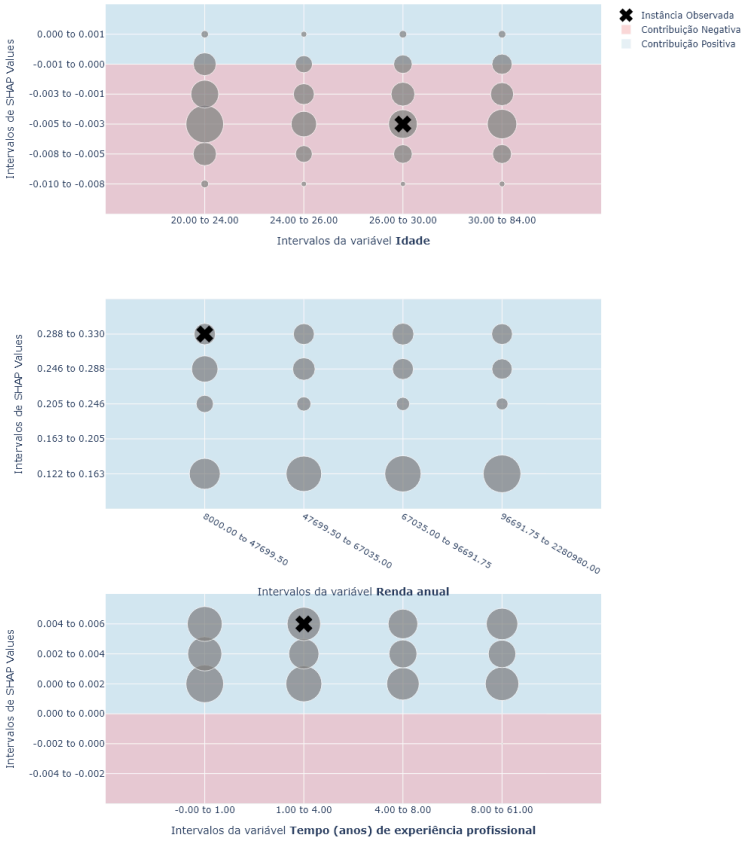
Distribuição de valores e de SHAP values de cada variável

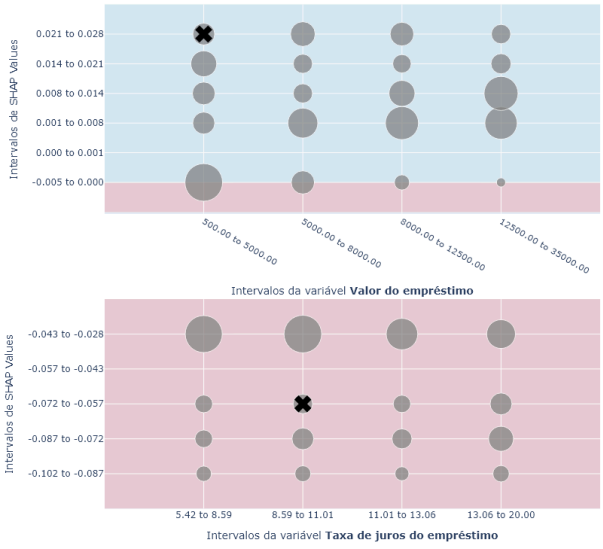


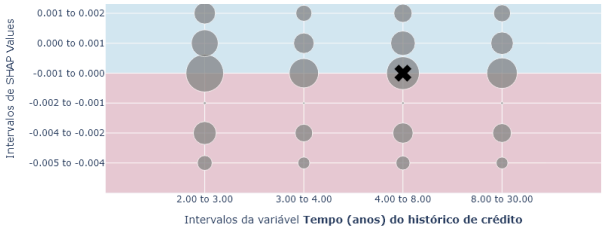


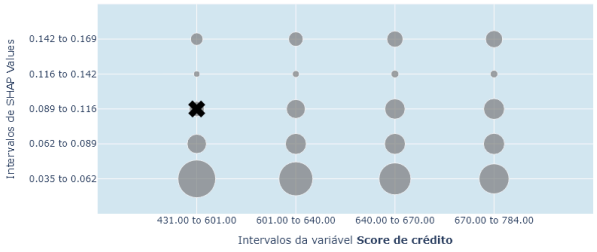
Visualização D

Concentração de instâncias por range de SHAP values e quartis das variáveis




Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável Valor do empréstimo
Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável Taxa de juros do empréstimo
Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável Valor do empréstimo como percentual da renda anual





17. Classifique a instância acima:

Mark only one oval.

- ◯ Aprovado
- ◯ Rejeitado

18. Quão confiante você está da sua classificação?

Mark only one oval.

	1	2	3	4	5	
Pou	◯	◯	◯	◯	◯	Muito confiante

19. A partir da representação visual da explicação, consigo entender como o modelo toma decisões.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

20. A explicação é útil para que eu tome melhores decisões ou ações.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

21. A explicação aumenta a minha confiança no modelo.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

22. A explicação fornece informações suficientes para explicar como o modelo toma decisões.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

23. Estou satisfeito/a com a explicação do modelo.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

24. Indique quanto a visualização A contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

25. Indique quanto a visualização B contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

26. Indique quanto a visualização C contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

27. Indique quanto a visualização D contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

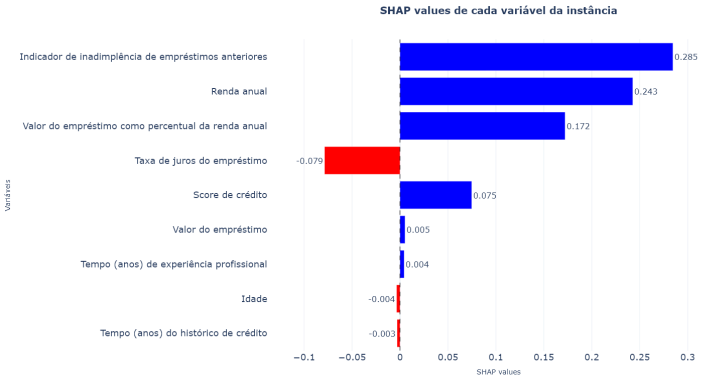
	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

28. Se possível, comente sobre a sua opinião sobre a entendibilidade, utilidade, confiança, informatividade e satisfação geradas por cada visualização ou combinação de visualizações.

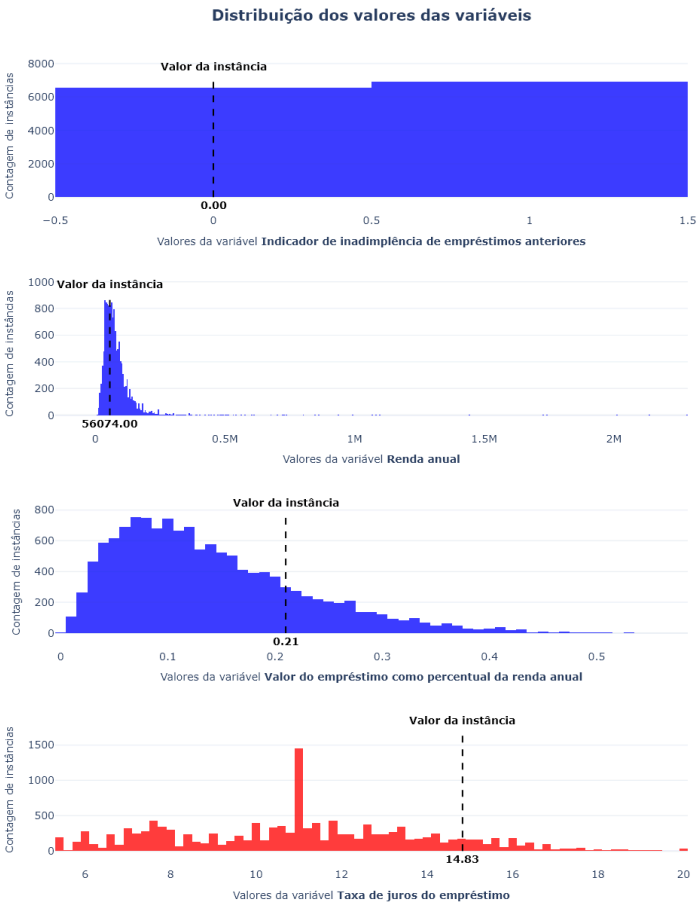
Nessa seção serão apresentadas visualizações de explicações de 4 NOVAS instâncias (4 visualizações para cada instância). Observe as representações e classifique cada uma delas como você acha que o MODELO as classificaria e depois responda a algumas perguntas sobre a sua percepção sobre elas.

Instância 2

Visualização A

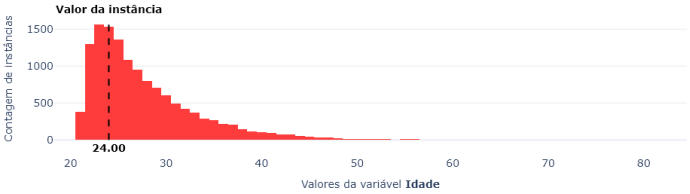


Visualização B


Distribuição dos valores das variáveis
Valor da instância
Contagem de instâncias
0.00
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
56074.00
Valores da variável Renda anual
Valor da instância
0.21
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
14.83
Valores da variável Taxa de juros do empréstimo

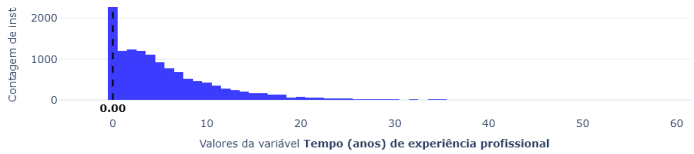

Valor da instância
Contagem de instâncias
800
600
400
200
0
688.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
Contagem de instâncias
1000
500
0
12000.00
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo


Valor da instância
Contagem de instâncias
1500
1000
500
0
24.00
20
30
40
50
60
70
80
Valores da variável Idade

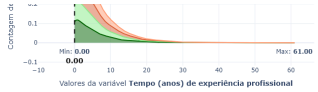

Valor da instância
Contagem de instâncias
3000
2000
1000
0
3.00
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito

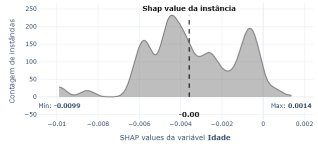

Valor da instância
3000

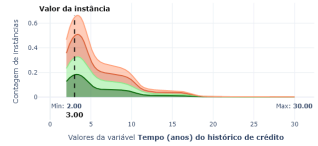


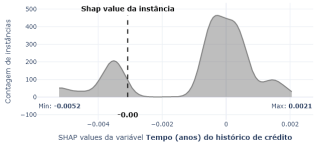
Visualização C

Distribuição de valores e de SHAP values de cada variável



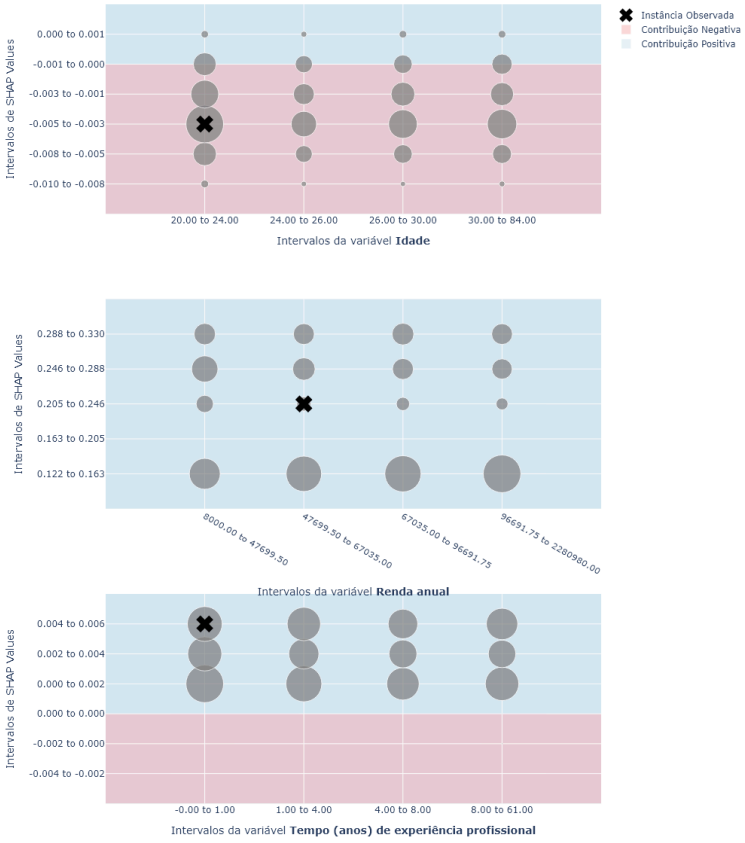


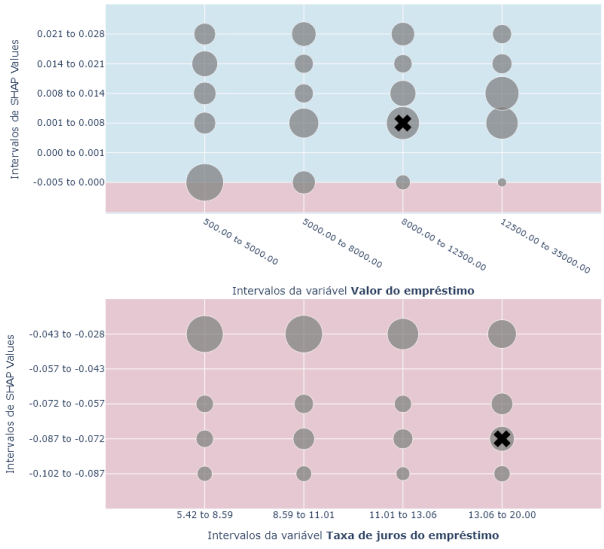


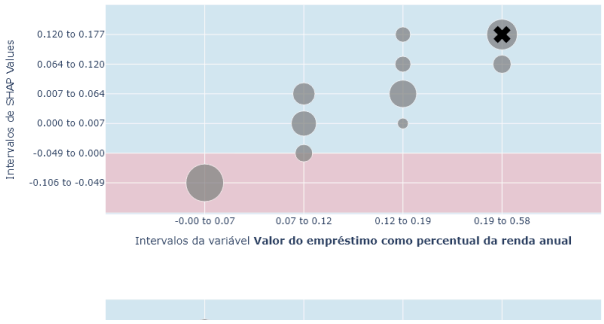


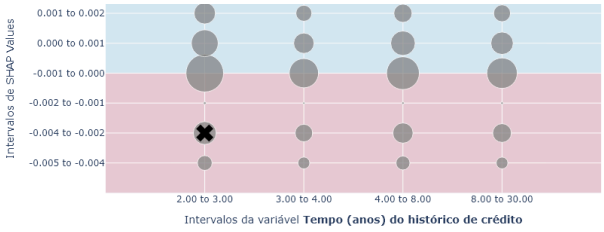
Visualização D

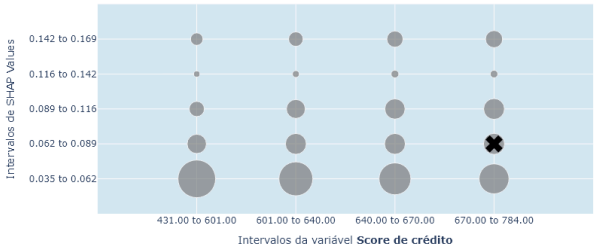
Concentração de instâncias por range de SHAP values e quartis das variáveis




Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável Valor do empréstimo
Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável Taxa de juros do empréstimo


Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável Valor do empréstimo como percentual da renda anual





29. Classifique a instância acima:

Mark only one oval.

- ◯ Aprovado
- ◯ Rejeitado

30. Quão confiante você está da sua classificação?

Mark only one oval.

	1	2	3	4	5	
Pou	◯	◯	◯	◯	◯	Muito confiante

31. A partir da representação visual da explicação, consigo entender como o modelo toma decisões.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

32. A explicação é útil para que eu tome melhores decisões ou ações.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

33. A explicação aumenta a minha confiança no modelo.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

34. A explicação fornece informações suficientes para explicar como o modelo toma decisões.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

35. Estou satisfeito/a com a explicação do modelo.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

36. Indique quanto a visualização A contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

37. Indique quanto a visualização B contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

38. Indique quanto a visualização C contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

39. Indique quanto a visualização D contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

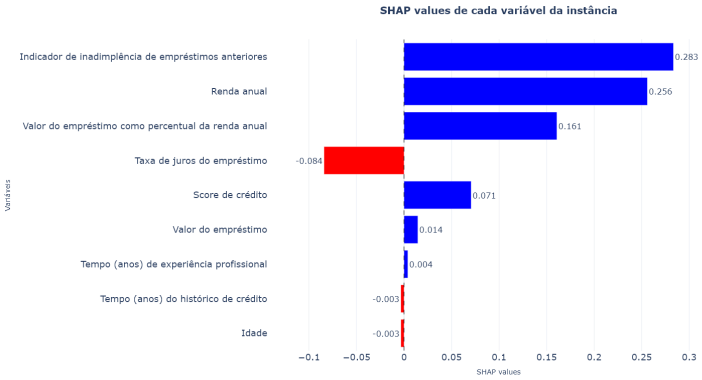
	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

40. Se possível, comente sobre a sua opinião sobre a entendibilidade, utilidade, confiança, informatividade e satisfação geradas por cada visualização ou combinação de visualizações.

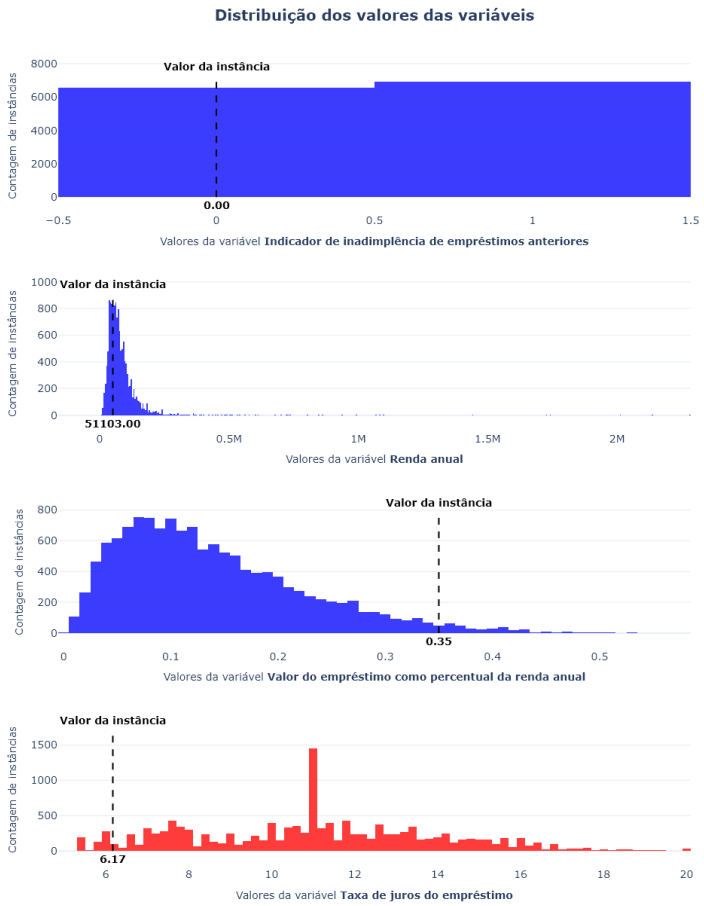
Nessa seção serão apresentadas visualizações de explicações de 4 NOVAS instâncias (4 visualizações para cada instância). Observe as representações e classifique cada uma delas como você acha que o MODELO as classificaria e depois responda a algumas perguntas sobre a sua percepção sobre elas.

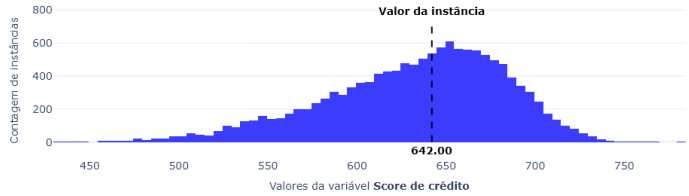
Instância 3

Visualização A

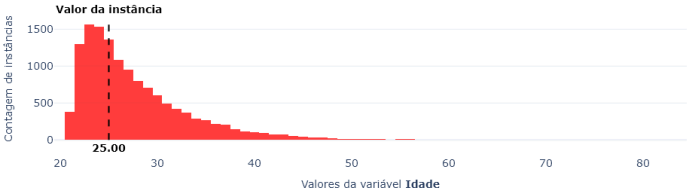


Visualização B


Distribuição dos valores das variáveis
Valor da instância
Contagem de instâncias
0.00
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
51103.00
Valores da variável Renda anual
Valor da instância
0.35
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
6.17
Valores da variável Taxa de juros do empréstimo


Valor da instância
Contagem de instâncias
800
600
400
200
0
642.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
Contagem de instâncias
1000
500
0
18000.00
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo

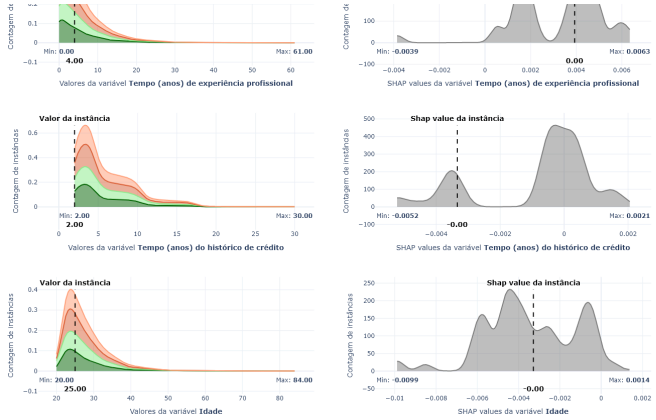

Valor da instância
Contagem de instâncias
1500
1000
500
0
25.00
20
30
40
50
60
70
80
Valores da variável Idade


Valor da instância
Contagem de instâncias
3000
2000
1000
0
2.00
0
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito


Valor da instância
3000

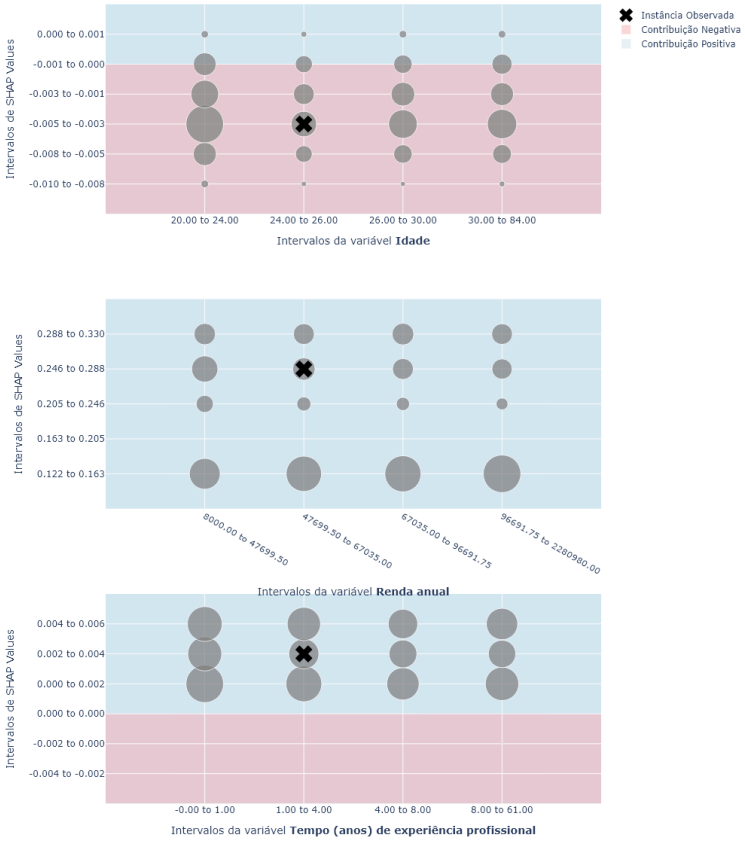
Visualização C

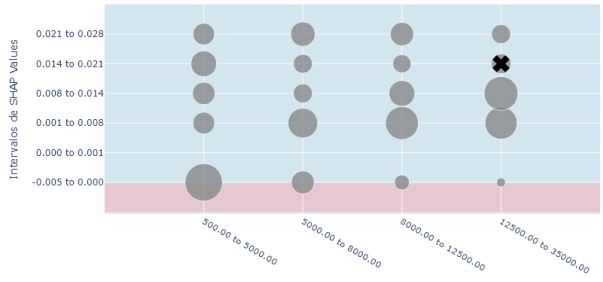
Distribuição de valores e de SHAP values de cada variável

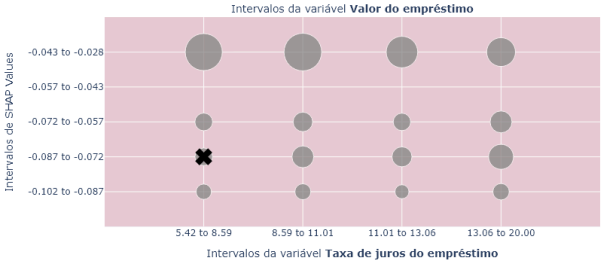


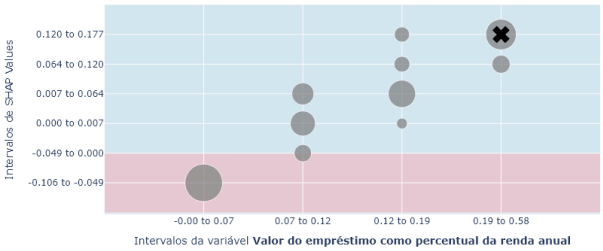
Visualização D

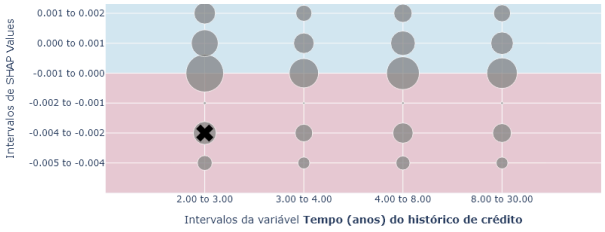
Concentração de instâncias por range de SHAP values e quartis das variáveis

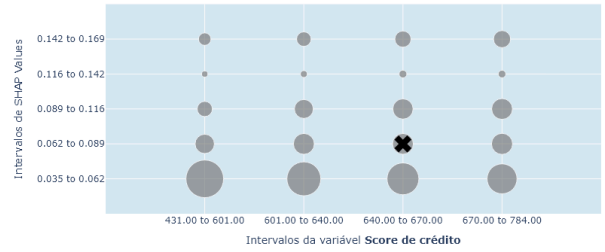



Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável **Valor do empréstimo**


Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável **Taxa de juros do empréstimo**


Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável **Valor do empréstimo como percentual da renda anual**





41. Classifique a instância acima:

Mark only one oval.

- ◯ Aprovado
- ◯ Rejeitado

42. Quão confiante você está da sua classificação?

Mark only one oval.

	1	2	3	4	5	
Pou	◯	◯	◯	◯	◯	Muito confiante

43. A partir da representação visual da explicação, consigo entender como o modelo toma decisões.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

44. A explicação é útil para que eu tome melhores decisões ou ações.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

45. A explicação aumenta a minha confiança no modelo.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

46. A explicação fornece informações suficientes para explicar como o modelo toma decisões.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

47. Estou satisfeito/a com a explicação do modelo.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

48. Indique quanto a visualização A contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

49. Indique quanto a visualização B contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

50. Indique quanto a visualização C contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

51. Indique quanto a visualização D contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

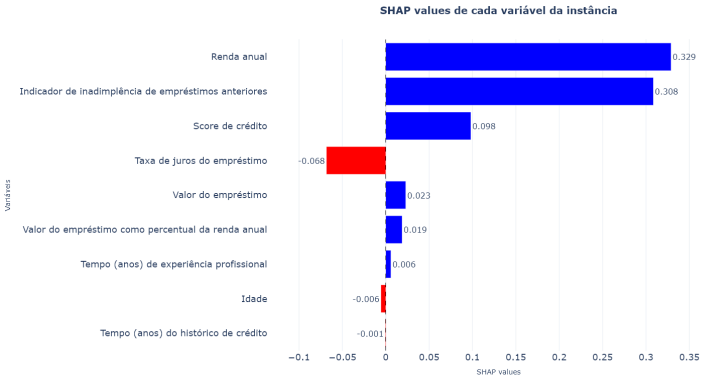
	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

52. Se possível, comente sobre a sua opinião sobre a entendibilidade, utilidade, confiança, informatividade e satisfação geradas por cada visualização ou combinação de visualizações.

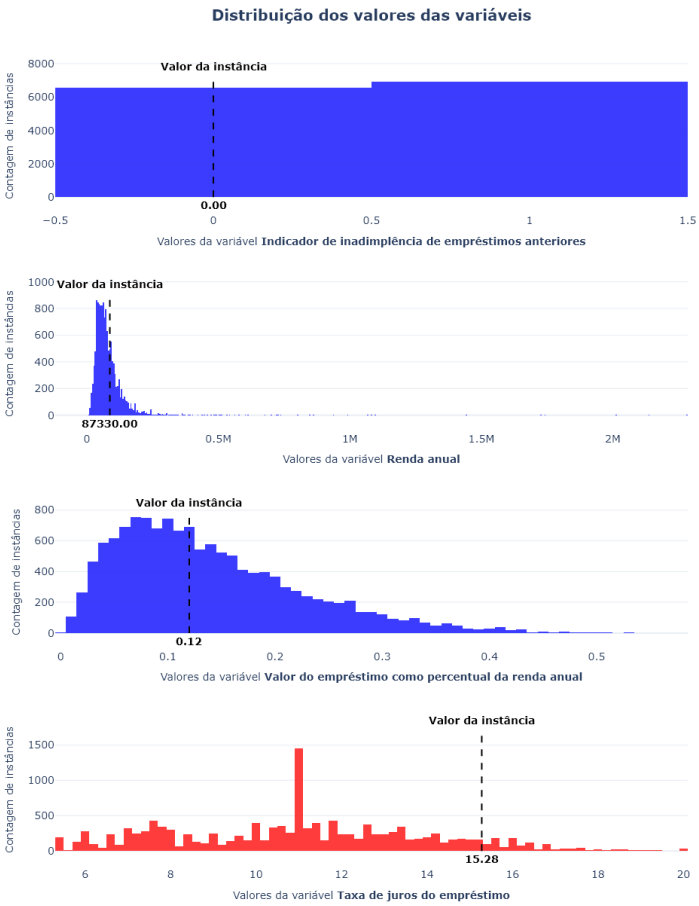
Nessa seção serão apresentadas visualizações de explicações de 4 NOVAS instâncias (4 visualizações para cada instância). Observe as representações e classifique cada uma delas como você acha que o MODELO as classificaria e depois responda a algumas perguntas sobre a sua percepção sobre elas.

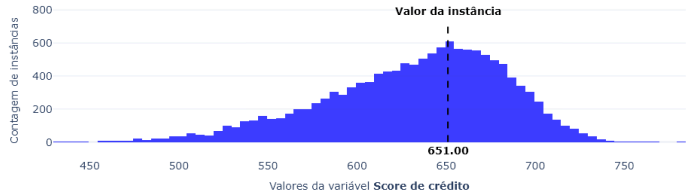
Instância 4

Visualização A



Visualização B


Distribuição dos valores das variáveis
Valor da instância
0.00
Contagem de instâncias
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
87330.00
Contagem de instâncias
Valores da variável Renda anual
Valor da instância
0.12
Contagem de instâncias
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
15.28
Contagem de instâncias
Valores da variável Taxa de juros do empréstimo


Valor da instância
800
600
400
200
0
Contagem de instâncias
651.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
1000
500
0
Contagem de instâncias
10500.00
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo

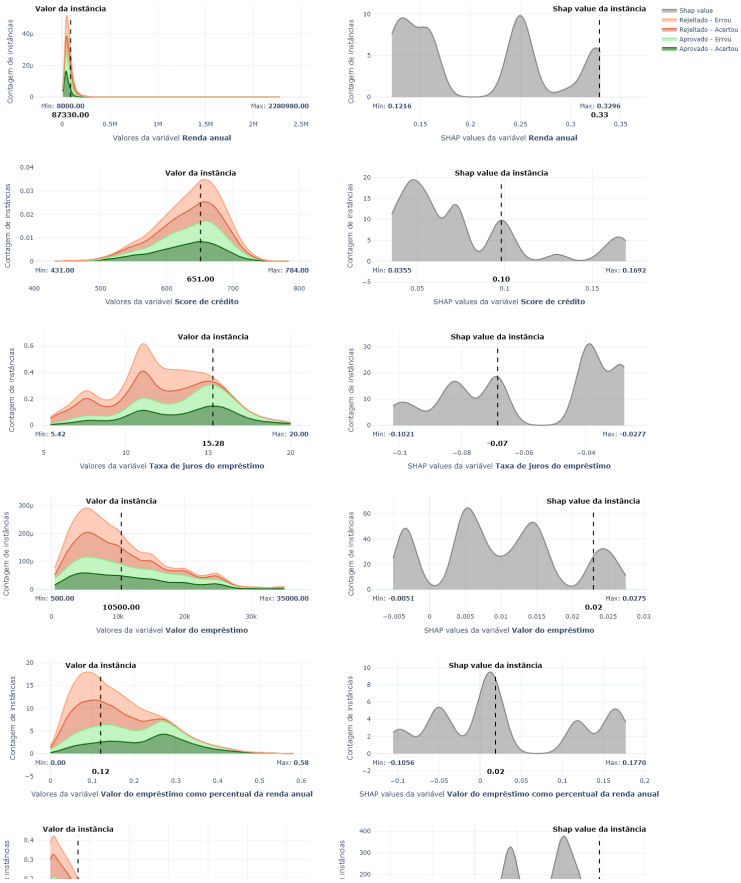

Valor da instância
1500
1000
500
0
Contagem de instâncias
26.00
20
30
40
50
60
70
80
Valores da variável Idade

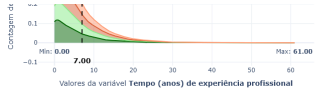

3000
Valor da instância
2000
1000
0
Contagem de instâncias
2.00
0
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito

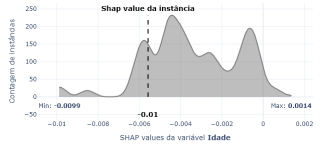

3000
Valor da instância
ncias

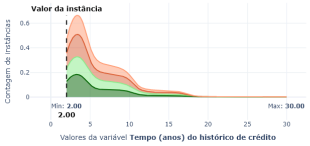
Visualização C

Distribuição de valores e de SHAP values de cada variável



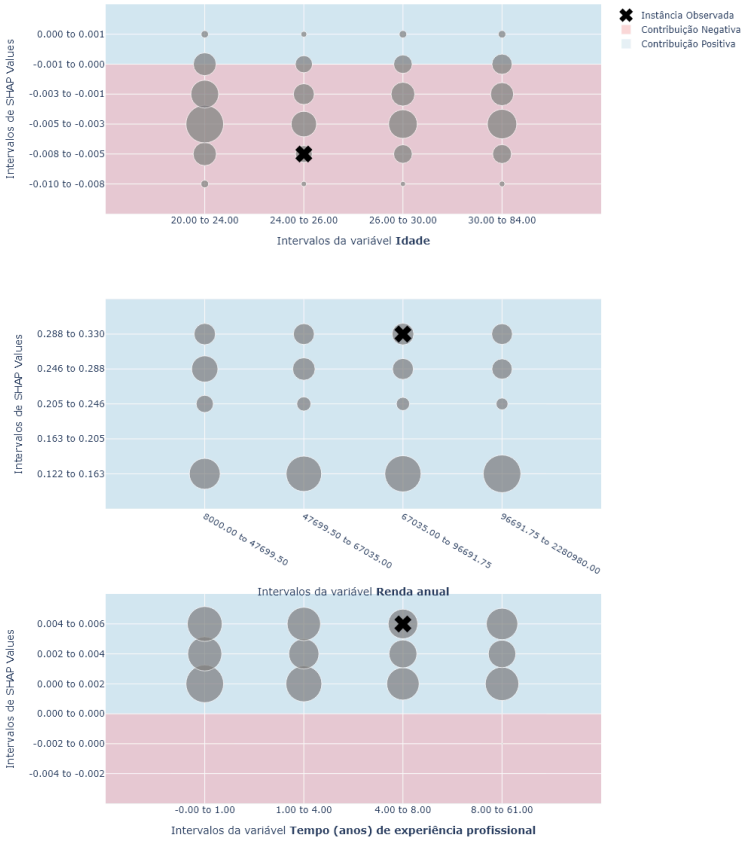


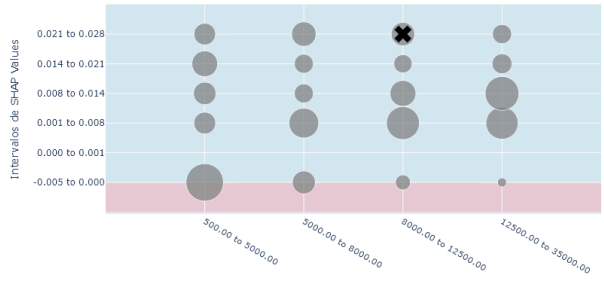


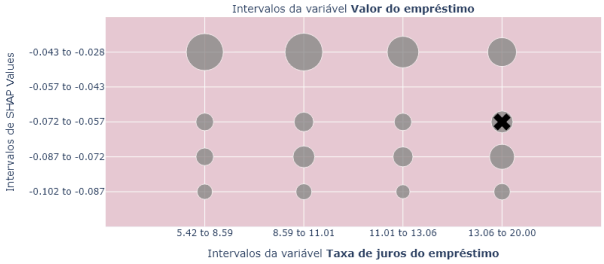


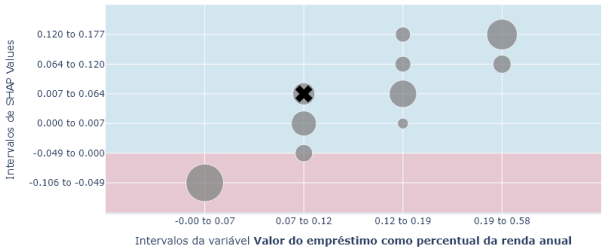
Visualização D

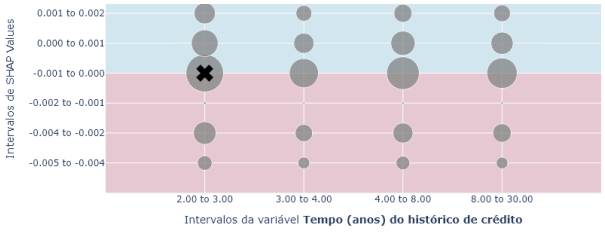
Concentração de instâncias por range de SHAP values e quartis das variáveis

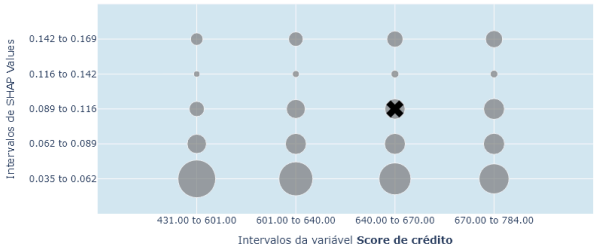



Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável **Valor do empréstimo**


Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável **Taxa de juros do empréstimo**


Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável **Valor do empréstimo como percentual da renda anual**





53. Classifique a instância acima:

Mark only one oval.

- ◯ Aprovado
- ◯ Rejeitado

54. Quão confiante você está da sua classificação?

Mark only one oval.

	1	2	3	4	5	
Pou	◯	◯	◯	◯	◯	Muito confiante

55. A partir da representação visual da explicação, consigo entender como o modelo toma decisões.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

56. A explicação é útil para que eu tome melhores decisões ou tome medidas.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

57. A explicação aumenta a minha confiança no modelo.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

58. A explicação fornece informações suficientes para explicar como o modelo toma decisões.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

59. Estou satisfeito/a com a explicação do modelo.

Mark only one oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo

60. Indique quanto a visualização A contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

61. Indique quanto a visualização B contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

62. Indique quanto a visualização C contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

63. Indique quanto a visualização D contribuiu para as suas respostas às perguntas anteriores:

Mark only one oval.

	1	2	3	4	5	
Não	◯	◯	◯	◯	◯	Contribuiu totalmente

64. Se possível, comente sobre a sua opinião sobre a entendibilidade, utilidade, confiança, informatividade e satisfação geradas por cada visualização ou combinação de visualizações.

65. As suas expectativas em relação às explicações foram atendidas? Se não, de que você sentiu falta?

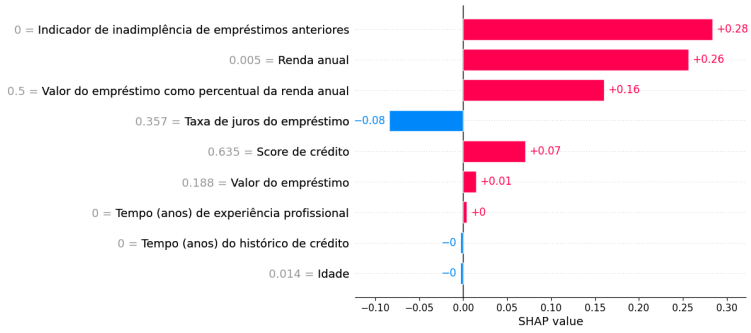
Observe os dois conjuntos de visualizações abaixo. Ambos representam a explicação da mesma instância.

Classificação do modelo: APROVADO

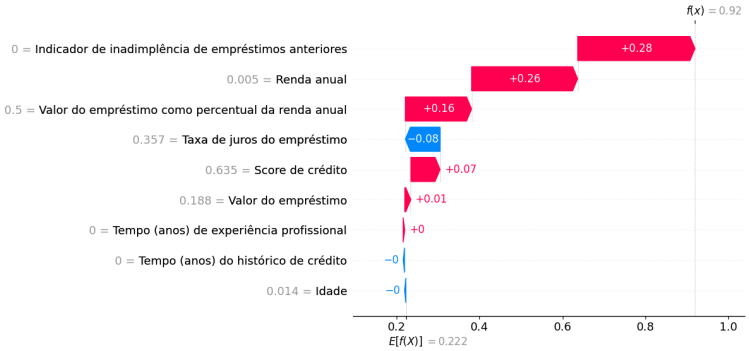
Classificação correta: APROVADO

Conjunto de visualizações 1

Visualização A

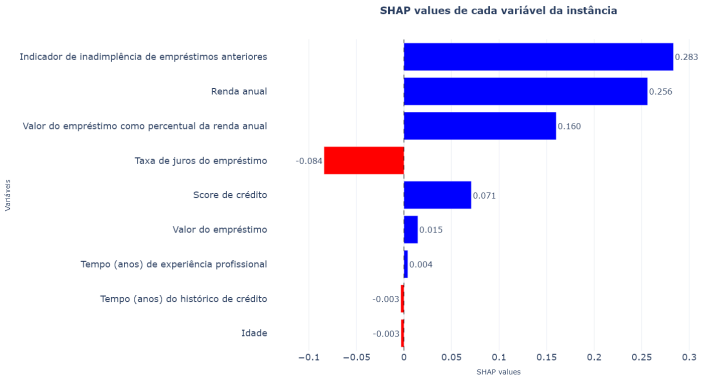


Visualização B

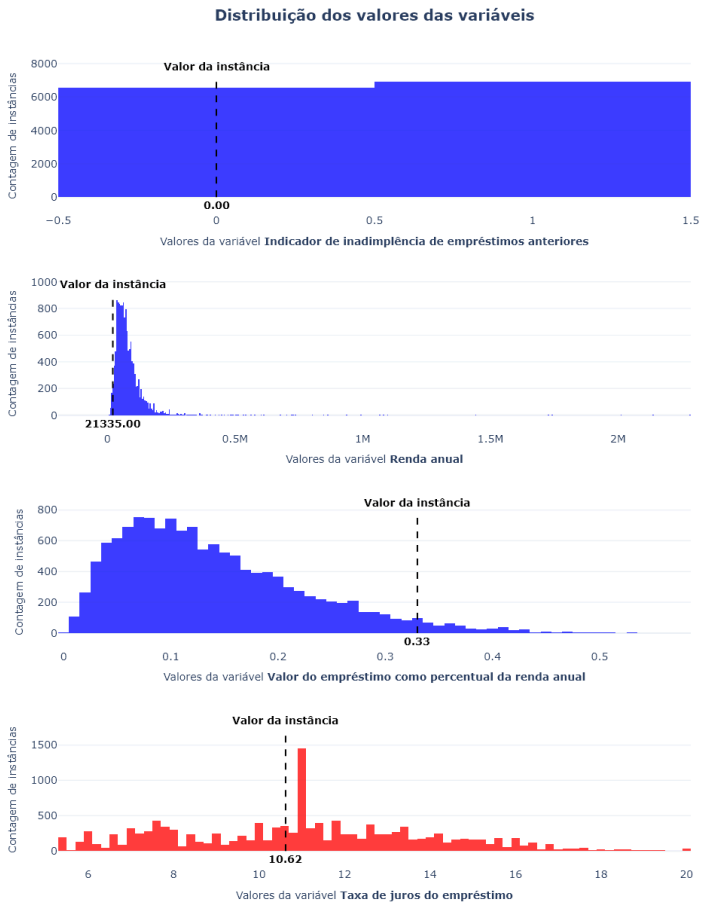


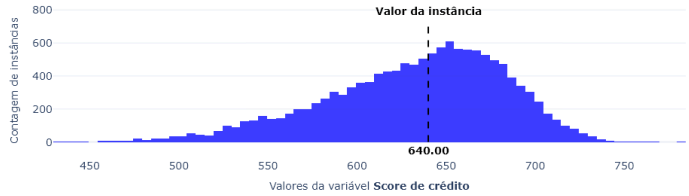
Conjunto de visualizações 2

Visualização A



Visualização B


Distribuição dos valores das variáveis
Valor da instância
Contagem de instâncias
0.00
Valores da variável Indicador de inadimplência de empréstimos anteriores
Valor da instância
21335.00
Valores da variável Renda anual
Valor da instância
0.33
Valores da variável Valor do empréstimo como percentual da renda anual
Valor da instância
10.62
Valores da variável Taxa de juros do empréstimo

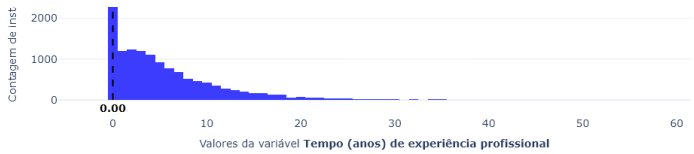

Valor da instância
Contagem de instâncias
800
600
400
200
0
640.00
450
500
550
600
650
700
750
Valores da variável Score de crédito


Valor da instância
Contagem de instâncias
1000
500
0
7000.00
5k
10k
15k
20k
25k
30k
35k
Valores da variável Valor do empréstimo


Valor da instância
Contagem de instâncias
1500
1000
500
0
21.00
20
30
40
50
60
70
80
Valores da variável Idade

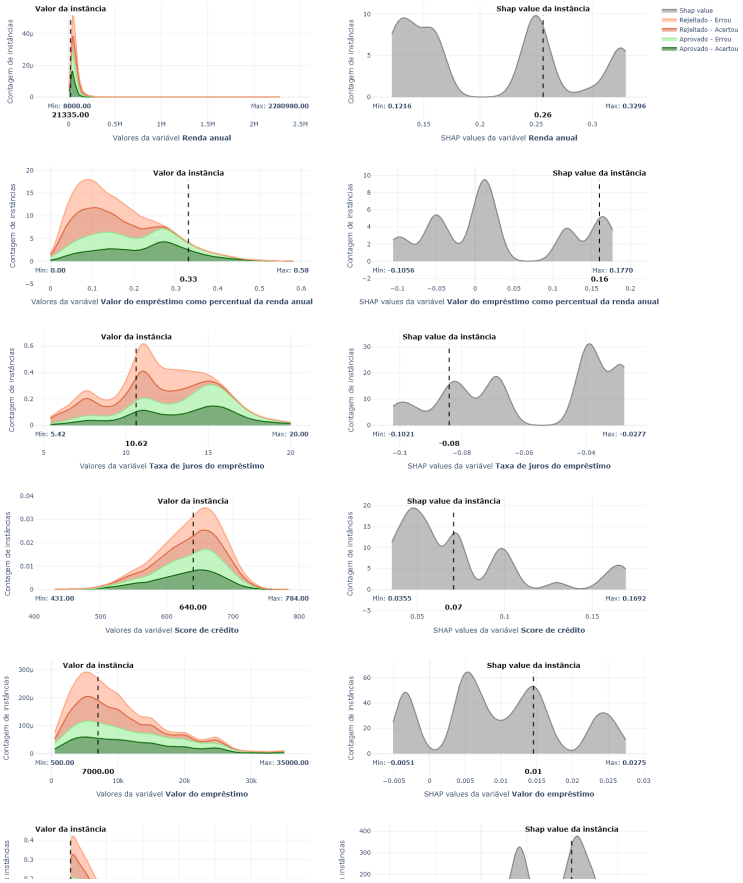

Valor da instância
Contagem de instâncias
3000
2000
1000
0
2.00
0
5
10
15
20
25
30
Valores da variável Tempo (anos) do histórico de crédito

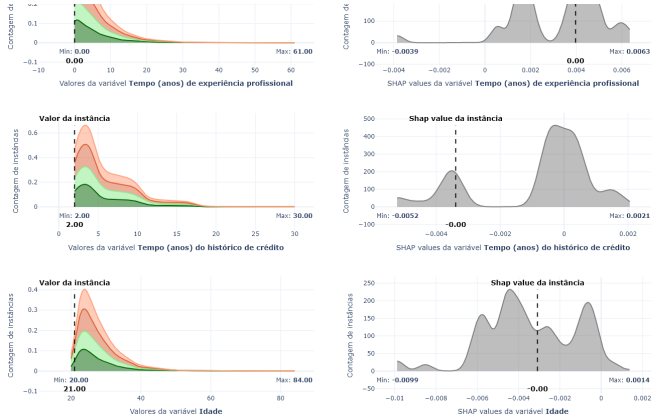

Valor da instância
3000



Visualização C

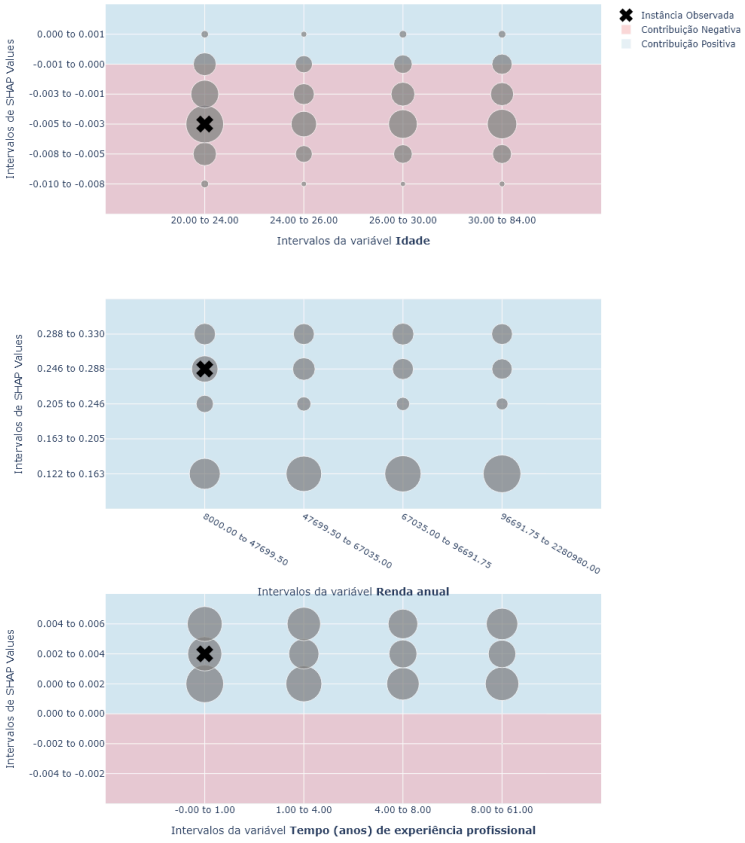
Distribuição de valores e de SHAP values de cada variável

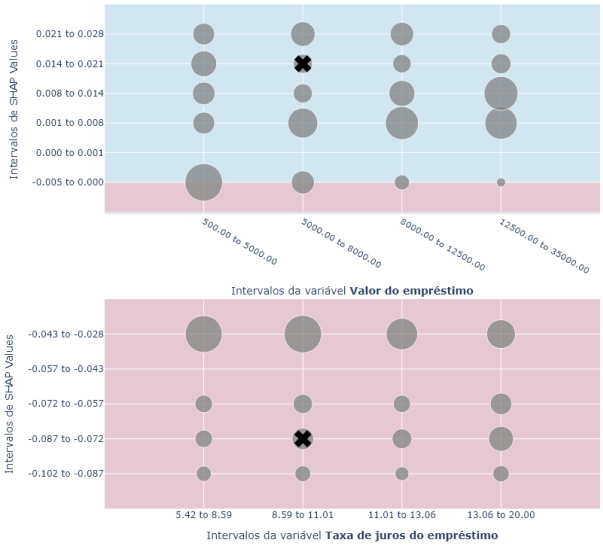


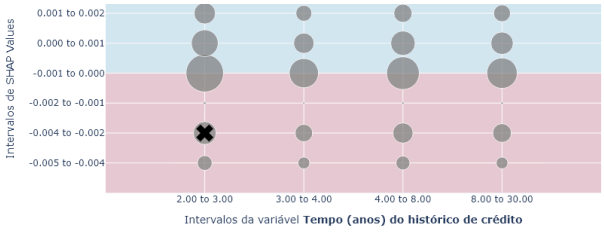


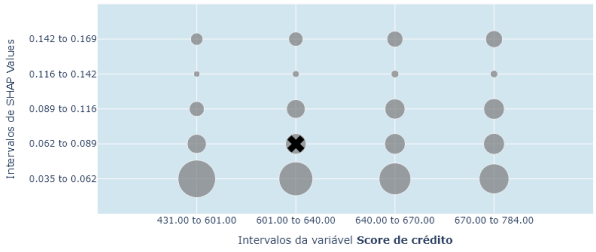
Visualização D

Concentração de instâncias por range de SHAP values e quartis das variáveis




Intervalos de SHAP Values
0.021 to 0.028
0.014 to 0.021
0.008 to 0.014
0.001 to 0.008
0.000 to 0.001
-0.005 to 0.000
500.00 to 5000.00
5000.00 to 8000.00
8000.00 to 12500.00
12500.00 to 35000.00
Intervalos da variável Valor do empréstimo
Intervalos de SHAP Values
-0.043 to -0.028
-0.057 to -0.043
-0.072 to -0.057
-0.087 to -0.072
-0.102 to -0.087
5.42 to 8.59
8.59 to 11.01
11.01 to 13.06
13.06 to 20.00
Intervalos da variável Taxa de juros do empréstimo
Intervalos de SHAP Values
0.120 to 0.177
0.064 to 0.120
0.007 to 0.064
0.000 to 0.007
-0.049 to 0.000
-0.106 to -0.049
-0.00 to 0.07
0.07 to 0.12
0.12 to 0.19
0.19 to 0.58
Intervalos da variável Valor do empréstimo como percentual da renda anual





66. Qual dos dois conjuntos de visualizações acima você acha que faz você entender melhor as decisões do modelo?

Mark only one oval.

- ◯ Conjunto 1
- ◯ Conjunto 2
- ◯ Ambos
- ◯ Nenhum

67. Qual dos dois conjuntos de visualizações acima você acha que é mais útil para que você tome melhores decisões ou medidas?

Mark only one oval.

- Conjunto 1
- Conjunto 2
- Ambas
- Nenhum

68. Qual dos dois conjuntos de visualizações acima você acha que aumenta a sua minha confiança no modelo?

Mark only one oval.

- Conjunto 1
- Conjunto 2
- Ambos
- Nenhum

69. Qual dos dois conjuntos de visualizações acima você acha que fornece informações suficientes para explicar como o modelo toma decisões?

Mark only one oval.

- ◯ Conjunto 1
- ◯ Conjunto 2
- ◯ Ambos
- ◯ Nenhum

70. Qual dos dois conjuntos de visualizações acima te deixa mais satisfeito/a?

Mark only one oval.

- ◯ Conjunto 1
- ◯ Conjunto 2
- ◯ Ambos
- ◯ Nenhum

71. Se possível, comente sobre a sua opinião sobre a entendibilidade, utilidade, confiança, informatividade e satisfação geradas por cada conjunto de visualizações.

_____

_____

_____

_____

_____

Obrigada pela sua participação!