PUC RIO

Dissertação de Mestrado

Applications of machine learning methods in oil and gas for soft sensing, economic assessment, and multiphase flow simulation

Pedro Henrique Cardoso Paulo

Pontifícia Universidade Católica do Rio de Janeiro
Centro Técnico Científico
Departamento de Engenharia Mecânica

Rio de Janeiro, 13 de outubro de 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

Dissertação de Mestrado

Applications of machine learning methods in oil and gas for soft sensing, economic assessment, and multiphase flow simulation

Pedro Henrique Cardoso Paulo

Orientação: Professor Márcio da Silveira Carvalho
Coorientação: Professor Helon Vicente Hultmann Ayala

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Engenharia Mecânica pelo programa de Pós-Graduação em Engenharia Mecânica, no Departamento de Engenharia Mecânica

Rio de Janeiro, 13 de outubro de 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

Applications of machine learning methods in oil and gas for soft sensing, economic assessment, and multiphase flow simulation

Pedro Henrique Cardoso Paulo

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Engenharia Mecânica Aprovada pela Comissão examinadora abaixo:

Professor Márcio da Silveira Carvalho
Orientador
Departamento de Engenharia Mecânica – PUC-Rio

Professor Helon Vicente Hultmann Ayala
Co-Orientador
PUC-PR

Professor Florian Alain Yannick Pradelle
Departamento de Engenharia Mecânica – PUC-Rio

Dr. Geraldo Afonso Spinelli Martins Ribeiro
Petróleo Brasileiro S.A. – Petrobras

Rio de Janeiro, 13 de outubro de 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

Pedro Henrique Cardoso Paulo

Graduado em Engenharia Mecânica pela Universidade Federal do Rio de Janeiro, UFRJ (Rio de Janeiro – RJ, 2018), com pós-graduação em Engenharia Submarina pela Universidade Petrobras, UP (Rio de Janeiro – RJ, 2019). Atua profissionalmente como Engenheiro de Petróleo desde 2019, com foco em novos projetos, avaliação de ativos exploratórios, elevação artificial e garantia de escoamento, além de modelagem integrada de produção.

Ficha Catalográfica

Paulo, Pedro Henrique Cardoso

Applications of machine learning methods in oil and gas for soft sensing, economic assessment, and multiphase flow simulation / Pedro Henrique Cardoso Paulo ; orientação: Márcio da Silveira Carvalho ; coorientação: Helon Vicente Hultmann Ayala. – 2025.
115 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Mecânica, 2025.
Inclui bibliografia

1. Engenharia Mecânica - Teses. 2. Identificação de sistemas. 3. Medidores virtuais de vazão. 4. Avaliação econômica. 5. Escoamento multifásico. 6. Aprendizado de máquina. I. Carvalho, Márcio da Silveira. II. Ayala, Helon Vicente Hultmann. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Mecânica. IV. Título.

CDD: 621

To my wife, Isabela and my parents, Luiz and Fátima, for the constant
support during my whole life and career.

Acknowledgments

First, I thank both my advisor, Professor D.Sc. Márcio Carvalho and my co-advisor Professor D.Sc. Helon Ayala for the invaluable help in making possible the production of these results here presented. A special thanks for Professor Helon that, as my first official advisor in this work, guided most of the research and methods presented in this work, helping me at this start of my academic life with both teaching me how to research and constantly expanding my personal toolbox with methods and ideas to address many of the challenges I face as an engineer.

To my wife, Isabela, my gratitude and love for all the support and comprehension in this important but challenging cycle of my life.

I also thank my colleague in both Petrobras and PUC-Rio, and research partner Felipe da Costa Pereira, an outstanding engineer with whom I have the pleasure of sharing many publications. Let these works be only the firsts where we work together, both academically and professionally, to extract all the potential that data science has in petroleum engineering.

Many thanks for the colleagues in Petrobras that had significant impact in the works that I developed during the last couple of years and presented in this work. Special thanks for Gilberto Xavier, Saon Vieira and Anderson Faller, whose vast knowledge of data science and machine learning deep influenced the methods applied in this work and the analysis herein presented. I also thank my colleagues flow assurance and production engineers Rafael Fabricio Alves, Stella Cavalli and Sergio Paulo, provided me constant insights and suggestions that allowed to, in the midst of machine learning pipelines, always keep track of the physical behavior.

I also thank my colleagues Gustavo Gomes and Luiz Fofano that faced with me the challenges of starting this academic endeavor.

At last, I thank Petrobras for providing me with this opportunity for both personal and professional growth.

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

Abstract

Paulo, Pedro Henrique Cardoso; Carvalho, Márcio da Silveira (Advisor); Ayala, Helon Vicente Hultmann (Co-Advisor). **Applications of machine learning methods in oil and gas for soft sensing, economic assessment, and multiphase flow simulation**. Rio de Janeiro, 2025. 115p. Dissertação de Mestrado – Departamento de Engenharia Mecânica, Pontifícia Universidade Católica do Rio de Janeiro.

The use of machine learning has grown rapidly across various sectors—from healthcare and finance to energy—due to its ability to identify patterns and make predictions from complex datasets. Focusing on applications in the oil and gas industry, this work explores three main uses of machine learning: virtual flow metering, early-stage economic assessment of exploratory assets, and hybrid modeling for multiphase flow simulation. For virtual flow metering, different machine learning models were tested in combination with system identification techniques for flow rate prediction, along with strategies to enhance model performance. The use of more complex estimators yielded gains of up to 62% in R^2 compared to classical system identification approaches, while among the performance enhancement strategies, the inclusion of current-time data significantly improved results, maintaining the model's utility for forecasting and monitoring tasks with performance gains of around 15% in R^2. In economic assessment, an innovative approach using black-box classifiers trained on imbalanced datasets proved viable for fast asset evaluation, with investment data and oversampling strategies delivering relevant gains around 34% e 8%, respectively, albeit with trade-offs in recall and interpretability. In multiphase flow modeling, the proposed use of hybrid models—integrating commercial mechanistic models with data-driven estimators—consistently outperformed both physical and black-box models, achieving up to 71% error reduction in pressure gradient prediction. These results highlight the potential of machine learning to complement traditional engineering workflows, enhance decision-making, and address longstanding challenges in oil and gas operations.

Keywords

System Identification; Virtual Flow Meters; Economic Assessment; Multiphase Flow; Machine Learning.

Resumo

Paulo, Pedro Henrique Cardoso; Carvalho, Márcio da Silveira; Ayala, Helon Vicente Hultmann. **Aplicações de métodos de aprendizado de máquina na indústria de óleo e gás para sensores virtuais, avaliação econômica e simulação de escoamento multifásico**. Rio de Janeiro, 2025. 115p. Dissertação de Mestrado – Departamento de Engenharia Mecânica, Pontifícia Universidade Católica do Rio de Janeiro.

O uso de aprendizado de máquina tem crescido rapidamente em diversos setores — da saúde e finanças à energia — devido à sua capacidade de identificar padrões e realizar previsões a partir de conjuntos de dados complexos. Pensando em aplicações na indústria de óleo e gás, este trabalho explora três aplicações principais de aprendizado de máquina: sensoriamento virtual para medição de vazão, avaliação econômica preliminar de ativos exploratórios e modelagem híbrida para simulação de escoamento multifásico. Para o sensoriamento virtual, foram testados diferentes modelos de aprendizado de máquina combinados com a técnica de identificação de sistemas para a previsão de vazão, combinados com estratégias para melhoria de performance dos modelos. O uso de estimadores mais complexos trouxe ganhos de até 62% em R^2 comparando com abordagens clássicas de identificação de sistemas enquanto dentre as técnicas de melhoria de modelo a combinação com dados do tempo corrente melhorou significativamente os resultados obtidos, mantendo a utilidade do modelo para tarefas de previsão e monitoramento com ganhos da ordem de 15% em performance no R^2. Na avaliação econômica, uma abordagem inovadora de criação de classificadores caixa-preta treinados com conjuntos de dados desbalanceados provou-se viável para avaliação rápida de ativos, com dados de investimento e estratégias de oversampling gerando ganhos relevantes da ordem de 34% e 8%, respectivamente, embora com compromissos em termos de recall e interpretabilidade. Na modelagem de escoamento multifásico, a proposta de aplicação de modelos híbridos que integram modelos mecanicistas comerciais com estimadores baseados em dados superarou consistentemente tanto os modelos físicos quanto os modelos caixa-preta, obtendo até 71% de redução de erro na predição de gradiente de pressão. Esses resultados destacam o potencial do aprendizado de máquina para complementar fluxos de trabalho tradicionais de engenharia, aprimorar a tomada de decisão e enfrentar desafios históricos nas operações de óleo e gás.

Palavras-chave

Identificação de Sistemas; Medidores Virtuais de Vazão ; Avaliação Econômica; Escoamento Multifásico; Aprendizado de Máquina.

Table of contents

List of figures

List of tables

Data is the new oil

Clive R. Humby.

1 Introduction

Machine learning is a subset of artificial intelligence that has been growing in use especially for cases where pre-defining heuristic rules would lead to a proliferation of rules and exceptions (Bishop & Nasrabadi, 2006). This paradigm differs form the broader view of artificial intelligence as it allows for computer systems to learn from data without explicit programming (Mitchell & Mitchell, 1997).

Machine learning is being applied in different sectors, including different verticals of industry 5.0, such as health, manufacturing, logistics, finance and energy (Trivedi et al., 2024). AI applications have also became commonplace in the oil and gas industry, a significant sector still relevant for both global economy and energy supply (Inkpen & Moffett, 2011).

Some of the recent relevant applications of machine learning in the oil and gas industry focusing on reservoir engineering and geosciences include proxy modeling for reservoir simulation and production scenarios optimization (Chu et al., 2020; Bhattacharyya & Vyas, 2022; Koray et al., 2023), dimensionality reduction to improve history matching and reduce computational costs (Jo et al., 2022; Canchumuni et al., 2019), data generation using generative adversarial networks to simulate plausible geological scenarios and production forecasting under uncertainty (Zhong et al., 2021; Kang & Choe, 2020), and optimization of enhanced oil recovery (EOR) associated with CO_2 capture and storage (You et al., 2020; Khan et al., 2024). These applications demonstrate the versatility of machine learning in addressing complex subsurface challenges and enhancing decision-making across reservoir, production, and drilling operations. Applications of interest also include the problems of virtual from metering and soft sensors, early-stage exploration asset economic assessment and multiphase flow simulation and modeling.

1.1 Virtual Flow metering

Well monitoring and production rate estimations compose the reservoir management activity and have a major role in the oil and gas industry. In most offshore production facilities, the separation of the three phases (oil, gas, and water) of the wells is performed only for the total production of the unit. A second separation system is usually available to separate fluids from a single well during production tests to obtain an accurate measurement of its three-phase flow rates, but excluding production test periods, the production of each well is estimated as a fraction of the total production of the unit, making it uncertain. An alternative to well testing is the use of multiphase flow meters (Okotie et al., 2016), but obtaining precise data is challenging due to the complexity of multiphase flow.

Oil and gas production predictions typically use simulation models that rely on physical first principles. These models are created and updated by qualified engineers, using pressure and temperature data collected by sensors at key positions in the well, such as the bottom hole and wellhead. Adjusting the reservoir model to match the actual historical behavior is the process described by Alakeely & Horne (2022) as a history match. The dependency on well sensors for the history match is a challenge since there is always a risk of losing these sensors, and the data obtainde from said sensors is also subject to noise and bias errors. Camponogara et al. (2010) claimed that the downhole pressure is one of the most important variables to provide useful information for oil field management and recovery, but the PDG (permanent downhole gauge) sensor responsible for its measurement has a high failure rate. Freitas et al. (2021) also emphasized that replacing such equipment once it is damaged is not a common operation due to the high costs and operational risks. A typical sensor monitoring schema for a production well can be seen in Figure 1.1.

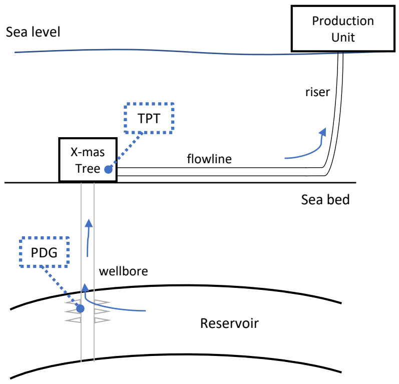


Figure 1.1: Typical offshore well and its pressure-temperature sensors: PDG (permanent downhole gauge) and TPT (temperature and pressure transducer).

Artificial intelligence techniques, such as machine learning, have been used in many studies to predict well rates and monitor reservoirs. Long-Short-Term Memory (LSTM) was widely studied for these applications due to its intrinsic time dependence formulation (Mercante & Netto, 2022; Song et al., 2022; Liu et al., 2022). Multi-Layer Perceptron (MLP) models were also evaluated as good rate predictors for several scenarios (Sabaa et al., 2023; Manami et al., 2023; Sandnes et al., 2021). Gradient Boosting algorithms such as XGBoost (Bikmukhametov & Jäschke, 2020; Sandnes et al., 2021), as well as random forest regressors (Song et al., 2022) were also considered for flow rate estimates.

Several studies have also implemented methods to enhance the quality of machine learning models. Various techniques are successful, including the utilization of autoencoders (Alakeely & Horne, 2021) as feature expansors of the input space. Mercante & Netto (2022) demonstrated that combining layers of various neural network architectures increased the accuracy of predictions. Liu et al. (2022) proposed the use of Echo State Networks combined with Genetic Algorithms to tune their parameters as short-term predictive models. Bikmukhametov & Jäschke (2020) proposed combining First-Principle Models, understood as models built based on physical principles and phisical modelling of the phenomenom, with a black-box approach to improve the quality of predictions. The sharing of data among wells, in a multi-task learning scheme, is proposed by Sandnes et al. (2021) as a way to improve model quality. The use of heterogeneous ANN ensemble models in which predictions are combined by using Simulated Annealing is reported by AL-Qutami et al. (2018).

1.2 Asset Appraisal in Oil and Gas

Modeling the economic behavior of systems and assets is essential for decision-making across various industries, as profitability is key to maintaining financial health and ensuring the continuous operation of companies. This is particularly relevant in economic sectors where outcomes depend on uncontrollable natural variables, introducing significant uncertainty — such as intermittent renewable energy sources, mining, and oil and gas.

In the oil and gas sector, the greatest uncertainty lies in exploratory assets, where limited information and inherently high uncertainty levels prevail, making it a high-risk venture (Suslick et al., 2009). Despite these challenges, accurately assessing the economic value of exploratory assets is crucial for companies in processes such as bidding for new areas, participating in farm-ins, and divesting from previously acquired regions. Improved evaluation of exploratory assets not only enhances portfolio management, ensuring a steady pipeline of promising new green fields to replenish oil reserves, but also helps companies avoid excessive spending on appraisal and data acquisition for areas that lack sufficient potential, given their scale and cost structure.

Conducting an economic evaluation of an asset requires calculating the net balance of investments and revenues over the asset's entire lifespan (Suslick et al., 2009). In the oil and gas sector, conducting this evaluation involves estimating the production profile, which influences revenue generation, while also assessing the capital requirements and operational costs for production facilities that should encompass expenses related to drilling, production pipelines, equipment, primary processing facilities, and hydrocarbon export systems. These tasks require the effort of multidisciplinary teams of reservoir, production, and facilities engineers to estimate the economic evaluation inputs for each scenario correctly. For exploratory assets, where the uncertainties are extremely relevant, this type of evaluation can require considerable resources and time, as the evaluation of different scenarios is needed to quantify and mitigate said uncertainties, and many different scenarios may require different production concepts. Even with solutions such as simplified metrics and methodologies, time and resource limitations often act as the deciding factors in defining the number of feasible scenarios in each evaluation (Suslick et al., 2009).

The importance of the economical factor in general industry may be seen in many recent works. For photovoltaic systems using batteries, Sandelic et al. (2022) built a model that predict the performance of a photovoltaic system taking into account environmental variables and residential payload and linked it to an economic model to prove the importance of correctly taking into account maintenance and replacement costs in quantifying the net present value (NPV) of an installation. In the context of energy storage, Domínguez-Jiménez et al. (2023) justified the value of a proposed electrical thermal storage for nordic communities based on quantifiable gains in cost to consumers and operators taking even into account the uncertainties of wind power generation. In applications making use of machine learning approaches, Jafary et al. (2024) proposed the application of artificial intelligence to perform land valuation tasks for tax and real estate applications, considering a range of possible features and obtaining promising results for XGBoost models. Soltani & Lee (2024) performed similar work focusing on the South Australian housing market, applying both tree-based ensembles and ANNs with relevant features from literature and applying Shapley analysis to understand feature importance. In finance, Kozina et al. (2023) presented a work of similar application focusing on default leasing contracts prediction and risk assessment, with significant improvements mostly in the recall.

In recent years, the oil and gas industry has made significant progress in applying machine learning and data-driven methods to assess and predict the behavior of its assets. Sircar et al. (2021) presented a general overview of applications of machine learning algorithms in the upstream sector with a focus on exploration, geology, drilling, and engineering, lacking applications on direct exploratory asset appraisal. A similar review was performed by Otchere et al. (2021), focusing on the most used algorithms for supervised learning in the oil and gas industry, like shallow neural networks and support vector machines (SVM), while classifying them for the area of application and variables used as inputs and outputs in past published works. Kuang et al. (2021) provided yet another application review, commenting on applications in well logging and seismic data. Kuang et al. (2021) also cited surface facilities engineering as an application, but focuses on digital twins and not on cost estimation as a possible usage for the technology.

Many majors in the oil and gas sector already make use of machine learning tools. A significant example is the workflow ALICE presented by Prochnow et al. (2022) and applied in Chevron, mostly in their tight rock reservoirs to predict hydrocarbon production, recovery, and resources more accurately than by using conventional proximity trends. Prochnow et al. (2022) claimed that the framework was being used to help with well landing zones determination and optimization, and exploration review assessment, while also presenting real data examples and applying Shapley interaction matrix analysis to infer the feature importance in the final model. Motivated by the area bidding process, Makhotin et al. (2022) proposed applying a data-driven approach with tree-based models to predict the oil recovery factor for water-flooded reservoirs as a faster and more accurate alternative to analogy and simulation techniques, but focusing solely on the recovery aspect of the appraisal.

1.3
Multiphase Flow Modeling

Steady-state multiphase flow simulation in pipelines is essential in the oil and gas industry to better determine and understand pressure-flow rate relations in production systems. By combining steady-state multiphase flow models with inflow performance relationships from reservoirs, it is possible to determine system production capacities, better design completion and surface equipment, and optimize production systems (Shippen & Bailey, 2012). Multiphase flow simulation in pipelines may also be used as first-principles models to estimate production flow rates in systems that lack multiphase flow metering, being relevant for production allocation and history matching processes (Pereira et al., 2025; Paulo et al., 2024).

Multiphase steady-state flow simulations in pipelines are typically performed with one-dimensional mass, momentum, and energy balances (Danielson et al., 2005; Bendlksen et al., 1991), either considering the mixed fluid or taking into account each phase individually. Due to the complexity of two-phase flow behavior, an empirical approach was first used to describe the phenomenon (Ansari et al., 1994). Some examples of said models include empirical models for narrow inclination ranges, as Hagedorn & Brown (1965), going to the more general empirical models as Beggs & Brill (1973), and drift-flux models as Bhagwat & Ghajar (2014). A more modern approach consists of applying mechanistic models, based on the physical nature of the flow and more applicable to a range of flow variables (Chaves et al., 2022). These models consist of flow pattern determination models, such as the Barnea model (Barnea, 1987) coupled with specific physics-based correlations to determine gas volume fractions and pressure drop according to the flow pattern (Ansari et al., 1994). Some examples of mechanistic models largely used in the oil and gas industry include the OLGA-S model (Bendlksen et al., 1991), the LEDA model (Danielson et al., 2005), and TUFFP model (Zhang et al., 2003), and the unified mechanistic model from Gomez et al. (2000). However, mechanistic models present inherent discontinuities due to numerical challenges or flow pattern transitions (Chaves et al., 2022), and while most mechanistic models were at first published in complete form, due to the complex nature of the implementation of said correlations and great reliance on proprietary test data, most of the industry standard models nowadays are made available as plug-ins (Shippen & Bailey, 2012), with many advancements in the results being kept as proprietary code. Thanks to that, recent literature on the subject tend to focus on developing new equations for estimating flow parameters (Alsarkhi et al., 2024), experimental investigations of specific flow regime and other parameters related (Hadzovic et al., 2025; Zhao et al., 2023) and model improvements and experiments for new challenges of the industry, such as carbon capture (Li et al., 2025) and dense gas flow (Diaz et al., 2024).

Even state-of-the-art models may present significant errors. Waltrich et al. (2019) performed an investigation of wellbore models to predict worst-case discharges and found errors of more than 100% for some parameters sets. Chaves et al. (2022) performed a screening of different correlations focusing on an experimental well dataset, getting errors close to 13% for the best performing model. In a general sense, errors around 10% are also to be expected when operating in regions where the models show good performance (Shippen & Bailey, 2012; Waltrich et al., 2019).

A way to better adjust models to empirical data is the application of machine learning methods. Many examples of applications of machine learning for multiphase flow parameter estimation may be found in the literature about multiphase flow metering, where Bahrami et al. (2024) provided an extensive literature review regarding the application of artificial neural networks in predicting a single flow parameter, like flow pattern, gas volume fraction, or fluid rate. In phase volume fraction predictions, Osman (2004) proposed a neural network to predict the liquid fraction obtaining mean average errors of 9.407%, but limited to horizontal pipes and neglecting fluid properties as inputs, while Malayeri et al. (2003) created a radial basis function (rbf) neural network model to predict gas volume fraction by using volume rate and density ratios and Weber number, obtaining a mean average error of 5.8% in the test dataset, but neglecting the impact of changes in fluid, diameter and roughness. In flow pattern predictions, AlSaif et al. (2022) performed a broad study, using over 8700 experimental samples for horizontal, vertical and inclined pipes to train a neural network, obtaining accuracies over 97%, while other works like Lin et al. (2020) and AL-Dogail et al. (2022) presented good results but were more limited in fluid variety and pipeline inclination respectively. Zhang et al. (2023) tried to improve flow pattern prediction by training a fully-connected neural network that, in intermediate stages, also predicts phase holdups, but cannot predict the pressure drop, a relevant parameter for multiphase flow simulations. Deep learning models such as transformers have already been tested for flow regime identification (Ruiz-Díaz et al., 2024), but still with low accuracy.

For pressure gradient predictions, most of the applications of machine learning come from literature motivated by pipeline dimensioning and production systems design (Chaari et al., 2020; Seong et al., 2020; Hafsa et al., 2024), as the prediction of pressure drop is essential for optimizing oil and gas field and designing optimal production systems (Alakoum & Ghorayeb, 2025). Most of the works on this front, however, tend to focus mostly on pressure gradient and pressure drop prediction, neglecting other parameters. A complete work on predicting pressure gradient, phase fraction, and flow regime was performed by Kanin et al. (2019) with the motivation to overcome limitations regarding closure relationships and limited ranges of applications for correlations. Results obtained outperformed classical and benchmark models and used as inputs classical adimensional numbers from literature.

While machine learning models have the advantage of low computational cost after training and being easily updated no new data (Bikmukhametov & Jäschke, 2020), the black box aspect of these models result in the lack of comprehensive physical interpretations and no guarantee of adequate prediction for points outside the training range, gaps that may be addressed with the adoption of hybrid models that incorporate physical behavior associated with black-box estimators (Ma et al., 2024). Bikmukhametov & Jäschke (2020) commented that the hybrid approach, combining first-principles models solely based on physcial modelling and understanding with data models, is a promising research frontier for multiphase flow metering, and it has been addressed by recent literature with different approaches like using neural networks to estimate choke discharge coefficients, to be used to estimate flow rates (Hotvedt et al., 2020), using data-driven models to update inflow reservoir data to be used in first-principles models (Vanvik et al., 2022) or using first-principles models to generate data for black-box model training and calibration (Ma et al., 2024). On the subject of mechanistic modeling of pressure drop, (Abdul-Majeed et al., 2022) proposed a ANN as a model for predicting phase fraction restricted to slug flow, outperforming all the existent classical models and generating better pressure drop results when coupled to mechanistic TUFFP model, which may be considered a black-box approach for phase fraction prediction associated with a hybrid approach to pressure drop calculation, but this approach demands access to the complete mechanistic model formulation in use, something that is less common in the current state-of-the-art commercial models (Shippen & Bailey, 2012). As hybrid models are still a recent research field, many new and possible hybrid model strategies, like the hybrid model strategy, where physics-based models are used to generate inputs for black-box models that either calculate the final output or the expected deviation between the physics and data models, have not yet been fully explored. The hybrid approach is of special interest in this work as it would allow the improvement of results regardless of access to the complete formulation of commercial models and it has been studied in many fields like plant transpiration estimation (Liu et al., 2023), evaporation in arid climates (Alsumaiei, 2025), vehicle dynamics modeling (Li et al., 2025), and closed-cycle drying moisture content estimation (Zhou et al., 2025), where it showed better results than basic black-box approaches.

1.4 Objectives

On the subject of virtual flow metering, the primary objective of this work is to conduct a case study on developing a black-box model that can predict well flow rates by utilizing available field measures and system identification techniques and investigate techniques to improve the results of said models. While this work primarily concentrates on liquid rate estimation, the techniques and framework presented in this study can be applied to any desired variable.

For the economic assessment of exploratory assets, this work proposes to create a predictive machine learning classifier capable of predicting the economic viability of a low maturity asset based on minimal information to speed up the process of appraisal for exploratory assets. This approach, according to the literature review presented, has not yet been studied in oil and gas assets, with most of the applications of machine learning on this field focusing on predicting production metrics, interpreting data acquired in the field and helping better plan the field development strategy over time, information that may help assess the economic viability of an asset but do not provide a direct assessment of the field. Better parallels for the proposed approach may be found in other segments, such as investments, renewable energies and real estate, but even in these fields, most works tend to focus on trying to predict the final asset value or model the final economic indicators through physical and economic models and cashflow , while this work focused on assessing its viability through classification, an approach that in this work was applied to oil and gas assets, but could easily be extended to other industrial segments.

Regarding multiphase flow modeling, the main objective of this work is to investigate hybrid model strategies that use commercial mechanistic models results as inputs for black box models that predict hybrid values out direct classification outputs for pressure gradient, gas fraction and flow regime prediction to fill the gap in literature, an advantageous strategy that do not demand access to specific parts of the physical model. Literature review shows that the application of machine learning methods to calculate multiphase flow parameters is a reality, but still allows for considerable improvements. Few works focus on more than one parameter estimation, and many studies are limited to the fluids considered and the inclination range. On the subject of hybrid models, the subject is new in the literature, and there are still gaps regarding the physics incorporation strategies and how to couple this approach with current commercial models.

1.5 Contributions

The contributions of this dissertation are detailed in chapters 3, 4, and 5. The specific contributions for each work are described below:

1. **System Identification Techniques for Soft Sensors and Multi-**

phase Flow Metering;

The main contribution of this work rests on the comparative evaluation of different strategies for improving the results of virtual flow meters by adding a system identification approach to the problem of rate estimation. This work discussed different possible input spaces and strategies to improve simple black-box system identification models such as time series decomposition and adding current-time data to the problem formulation to improve results.

The efforts made in this work contribute to the multiphase virtual flow metering field by showing that adding current-time data to a system identification model is overall the most effective strategy, allowing better cost-effective models and for high performance even in single-input cases, a strategy that reduces final model dependency from sensor data.

This work was presented at the 20th Symposium on System Identification of the International Federation of Automatic Control (IFAC), which took place in July 2024 at Northeastern University in Boston, MA.

- Reference: **PAULO, P. H. C.**; PEREIRA, F. D. C.; AYALA, H. V. H. **System Identification Techniques for Soft Sensors and Multiphase Flow Metering**. IFAC-PapersOnLine, v. 58, n. 15, p. 538-543, 2024.

2. **An Interpretable Data-Driven Framework for Economic Assessment of Oil and Gas Exploratory Assets**;

The main contribution of this work rests on showing that black-box machine learning methods are an alternative for early economic assessment of exploratory assets, an approach not yet tested for this kind of asset.

This work also contributes to the literature by assessing the issue of imbalanced datasets, something expected to be common when dealing to oil and gas assets, by applying oversampling techniques to the models and showing that, while oversampling techniques improve the overall model performance, they reduce the recall metric indicating a loss in the model's capability of predicting economic cases.

Regarding possible data available for the model creation, the work also contributes by showing that investments data were of great relevance for the final model, being more relevant than the oversampling techniques. Also, by applying explainability techniques to better understand the relevance of the features this work shows that the gains obtained with the inclusion of investments in the inputs may be related to assumed dependencies between operational and capital costs and production profile and reserves data, a behavior that is not desirable as it shows that the model is not capturing the expected effects of the investments and may be sensitive to overfitting.

3. **Hybrid machine learning models for improving state-of-the-art mechanistic flow models**;

   The main contributions of this work consist of proving the effectiveness of the strategy and its capability to provide improved results by comparing with analogous black-box models, with commercial correlations, and with the results obtained by Kanin et al. (2019) while considering the expected error estimated by cross-validation to ensure a valid comparison. The performance gains obtained ranged from 10% to over 100% depending on the tested model and output variable considered.

   Additional contributions of this work include extensive testing of possible base mechanistic models, data preprocessing and scaling strategies, and data-driven estimators, including classical machine learning models, neural networks, and the foundation model for tabular data TabPFN, proposed by Hollmann et al. (2025) and not yet tested for this task. The development and application of a metric to assess the individual contributions of physics-based and data-driven terms of the hybrid model and its application to better understand the impact of the base model and the data-driven model on the final response.

   Finally, as an incremental contribution, this work briefly presents a comparison of black-box models focusing on assessing the physical representativeness they provide to address the concerns regarding data-driven models not being able to correctly represent the physics and extrapolate the training data raised by other works (Ma et al., 2024; Bikmukhametov & Jäschke, 2020). This work shows that physical representativeness is achievable by applying explainability techniques, more specifically Shapley values, to the black-box models created.

2
Methods

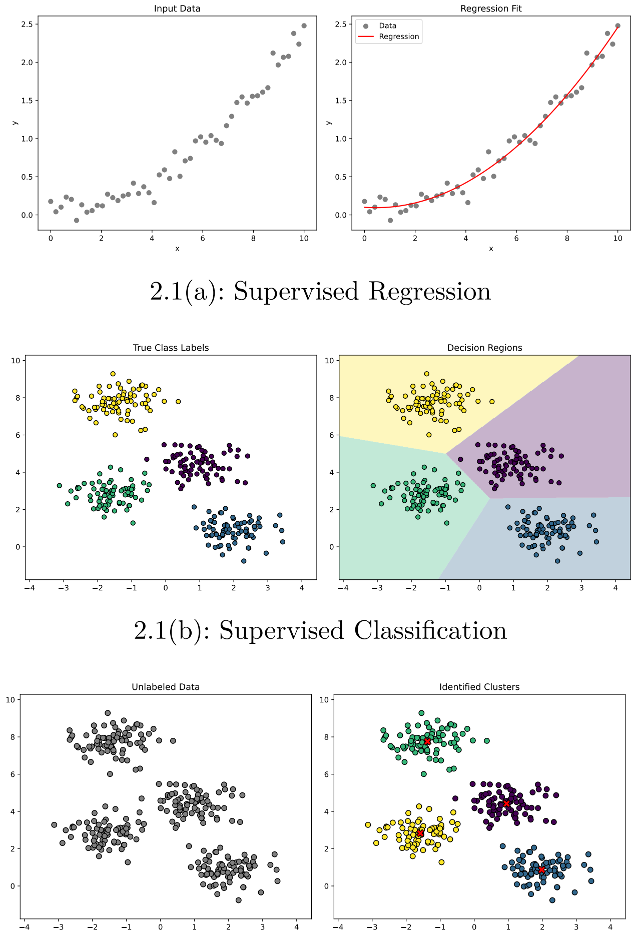
In this chapter, the relevant methods for the contributions elaborated is listed and developed.

2.1
Machine Learning

Artificial Intelligence (AI) is a broad term to classify machines that are able to mimic to some extent human intelligence and cognitive processes (IBM, 2023). As a broad term, this classification contains classical programming logic routines with basic, hard-coded decision-making logic, such as `if-then-else` structures. A subset of Artificial Intelligence is the Machine Learning (ML), that can be defined as the science of programming computers so they can learn from data or, in a broader sense, the field of study that gives the computers the ability to learn without being explicitly programmed (Géron, 2019). This differs from more general AI applications as machine learning enables writing programs that, through error minimization processes, determine relations in data by learning from data itself without the need of an explicit or "hard-coded" implementation of said relations.

Depending on the type of data available and how it is used, machine learning may be classified as supervised, unsupervised and reinforcement learning problems (IBM, 2023). Supervised learning problems are the ones where the algorithm training makes used of data where both the input variables and the desired output values are known. Most common applications of supervised learning are regression problems (illustrated in Figure 2.1(a)), where given the known inputs and desired continuous outputs, the main objective is to map a curve or hyperplane that best fits the known data, and classification problems (illustrated in Figure 2.1(b)), where for categorical outputs known it is of interest to determine the regions in the input space that should be attributed to each category.

Differently from supervised learning problems, both unsupervised learning and reinforcement learning problems do not demand prior knowledge of a paired input and output dataset (IBM, 2023). Unsupervised learning tries to find relations in datasets with no prior knowledge of outputs and is applicable



2.1(a): Supervised Regression

2.1(b): Supervised Classification

2.1(c): Unsupervised Learning

Figure 2.1: Simplified regression (a) and classification (b) supervised machine learning problems.

for clustering (illustrated in Figure 2.1(c)) and abnormality detection, while reinforcement learning tries to maximize the cumulative reward of the final response by a process of trial and error during predictions, being an interesting approach for robotics. As in this work, all datasets are comprised by inputs and output pairings, this section mainly focus in supervised learning algorithms.

2.1.1 Data-driven Model Pipeline

The first step in creating a data-driven model consists on loading the desired data that will be used to build the model. As the data loaded is not always ready to be used as input for the model training, this step is followed by data pre-processing, where different data sources may be merged, null values are treated, outliers are removed, categorical inputs are encoded, and procedures to remove the effects of dimensional variables such as standardization and normalization are performed (Géron, 2019).

Following the pre-processing step, an exploratory analysis and feature engineering are performed. In the exploratory analysis, the main objective is to better understand correlations in the data and identify tendencies in the dataset, normally making use of histograms, correlograms and crossplots to develop insights (Géron, 2019). The feature engineering step usually follows the exploratory analysis and the main goal usually is to reduce the number of input features by eliminating strongly correlated features identifies in the exploratory analysis or by applying some feature extraction techniques. It is also possible in the feature engineering step to increase the number of features either by decomposition strategies (usually for timeseries) or by applying some feature expansion techniques such as polynomial expansion or other mathematical operations. In cases where feature expansion is performed, it might be interesting to reassess the exploratory analysis in search for new insights and perform again the standardization or normalization in the new features.

After understanding the data, the process of model training and evaluation is performed. This is normally done by holding out a share of the data for testing and using the remaining data as training dataset. In this step, different models are tested and insights developed by analyzing the data may be incorporated to better select the adequate model family. It is also in this step that hyperparameters are selected through some search protocol, making use either of a validation dataset or cross-validation protocols (Géron, 2019). Once selected and trained, the model may be delpoyed into production to perform the designed task.

Regarding the dataset splitting into train and test, it is recommended to ensure that both datasets are statistically representative of the sample, something normally done by random sampling with stratification whenever different classes of samples may be determined. It is also important to avoid data leaks between the train and test dataset, including using only the train dataset to determine parameters for data standardization or normalization.

2.1.2 Supervised Machine Learning Algorithms

Creating a supervised learning model consists of, given an input set $\mathbf{X}$ and an output set $\mathbf{y}$, determining a function $f(\mathbf{X})$ that best approximates the expected outputs according to a pre-defined error metric, as illustrated in equation 2-1. This is done by some strategy of error minimization that may vary according to the base model assumed for the function f and the nature of the problem being modeled (classification or regression). No information besides the inputs and corresponding outputs is provided to the model, and all model "knowledge" is obtained from the data.

$$f(\mathbf{X}) \approx \mathbf{y} \tag{2-1}$$

Many different algorithms are available for the supervised learning task and as there is no optimal general-purpose approach to a supervised machine learning problem (Wolpert & Macready, 1997), it is necessary to evaluate multiple alternatives in order to determine the best fit for the problem. In this section, a brief overview of the machine learning algorithms applied in this work is presented. As most algorithms are applicable for both regression and classification tasks and it is of the scope of this work to address both problems, the adaptations and changes made in the algorithms to adapt for different types of problems are also commented.

2.1.2.1 Linear Models

Considering an input set $\mathbf{X}$ of matrix form, where each row represents a sample and each column represents a feature and a desired output column vector $\mathbf{y}$ that has the same number of rows than $\mathbf{X}$ as each row represents the desired output for the input set in the corresponding row of $\mathbf{X}$. For regression, the linear model consists on determining a vector of weights $\mathbf{w}$ that when multiplied by the input set $\mathbf{X}$ and added of a constant bias b generates the vector $\hat{\mathbf{y}}$ that best approximates the outputs vector $\hat{\mathbf{y}}$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} + b. \tag{2-2}$$

It is possible to prove that the above equation has an analytical solution by the least square method as a function of both $\mathbf{X}$, $\mathbf{y}$ and their sample-wise averages $\bar{\mathbf{X}}$ and $\bar{y}$

$$\mathbf{w} = \left[(\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{X} - \bar{\mathbf{X}})\right]^{-1} (\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{y} - \bar{y}), \tag{2-3}$$

$$b = \bar{y} - \bar{\mathbf{X}}\mathbf{w}. \tag{2-4}$$

It is important to bear in mind that, as $\bar{\mathbf{X}}$ and $\bar{y}$ are respectively a row vector and a scalar value, equation 2-3 has implicit broadcasting operations to ensure dimensional consistency.

In order to apply the linear regression concept to a classification problem, the concept of fitting the probability curve of a given instance belonging to a given class is adopted, with the addition of a non-linear function applied sample-wise in the output vector. For binary classification problems, the usual function is the logistic function

$$\sigma(z) = \frac{1}{1 - e^{-z}}. \tag{2-5}$$

As the function σ generates outputs in the interval $[0, 1]$, it is usually assumed that, for an output value $\sigma(z) > 0.5$, the output would be a positive for the classification problem, while $\sigma(z) < 0.5$ would be a negative. Combining the logistic function with the linear regression model, it is possible to estimate the probability column vector $\mathbf{p}$ by applying the function σ sample-wise

$$\mathbf{p} = \sigma(\mathbf{X}\mathbf{w} + b). \tag{2-6}$$

This formulation does not have an analytical solution, so the values of $\mathbf{w}$ and b are found through the minimization of an average loss function over the N samples

$$\min_{\mathbf{w},b} \frac{1}{N} \sum_{i=1}^{N} \left(-y_i \log(p_i) - (1 - y_i) \log(1 - p_i) \right). \tag{2-7}$$

It is important to notice that linear algorithms can be easily adapted for multi-output regression or multiclass optimization. This would only demand adding dimensions for both $\mathbf{w}$ and b and changing the activation function from the logistic to the softmax function, a generalization of the logistic function for more than one class.

2.1.2.2 Polynomial Models

Polynomial methods are usually achievable by applying linear methods to modified input matrices $\tilde{\mathbf{X}}$ containing n^{th} degree polynomial term combining the inputs. For example, considering the row $\mathbf{x}_i = \{x_{i1}, x_{i2}\}$ of the original input matrix $\mathbf{X}$ for a two-input variables problem, the same row in the modified input matrix considering a 2^{nd} degree polynomial with interaction terms would be $\tilde{\mathbf{x}}_i = \{x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2, x_{i1}x_{i2}\}$.

Once performed the feature expansion, the regression procedure is analogous to the one documented for linear models according to the type of problem and number of outputs.

2.1.2.3 Support Vector Machines

Support vector machines (SVM) are a group of supervised learning methods that perform classification and regression tasks by selecting appropriate hyperplanes in a n-dimensional space (Bishop & Nasrabadi, 2006). For classification tasks, the main goal is to select hyperplanes that best segregates different classes, thus maximizing the distance between correctly classified instances and the hyperplane while minimizing the distance between wrongly classified instances and the hyperplane that defines the correct classification region. For regression, the principles are similar, but the hyperplane itself is considered the representation of the desired response, so it searches for the hyperplane that minimizes the error for the predictions outside of a given margin (Smola & Schölkopf, 2004).

Focusing on classification, the base formulation of SVM selects the hyperplane that maximizes the separation from the nearest data points. This approach, known as hard margin, prioritizes the reduction of misclassifications in the train dataset, but may cause overfitting as the hyperplane gets distorted to better match the desired results. Alternatively, an approach one that minimizes this distance, referred to as soft margin classification, may be taken, allowing for some level of misclassifications in the search for a more general final model. The choice between soft and hard margin is determined by the regularization hyperparameter C (Smola & Schölkopf, 2004).

The objective function for the optimization process considering classification is given by

$$\min_{\mathbf{w},b,\boldsymbol{\zeta}}(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\zeta_i), \tag{2-8}$$

subject to

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, i = 1, \cdots, n,$$

where $y_i \in \{-1, 1\}$ is the output signal that indicates the desired output flag in a binary classification, $\mathbf{w}$ and b parametrize the hyperplane, $C > 0$ is the regularization hyperparameter, ζ_i is the margin distance, and ϕ is a kernel function that maps $\mathbf{x}_i$ into a non-linear space of higher dimension to better separate the data.

2.1.2.4
K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric supervised learning method based on the concept of proximity (Sklearn, 2023). It does not minimize any objective function and all new predictions are based solely on the existing training data provided to the model.

The base algorithm for the KNN method is relatively simple: given a distance metric previously defined, the k closest points to a new point where it is desired to infer the output. For classification problems, the class of the new point is the most frequent class among the k points selected, while for regression the output value is the average of said k points. It is also possible to attribute weights to the points in order to improve the results, normally giving more weights for the closest points from the new point (Sklearn, 2023). Any distance metric may be used to determine the k nearest points, with the most common being the Euclidean distance given by:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}, \tag{2-9}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the inputs for two different samples.

As the method does not assume any specific functional form or relationship between variables, it is suitable for complex, non-linear relationships (Taunk et al., 2019). On the other hand, the only way to improve the results obtained by the KNN method is the acquisition of more data to be provided as input and the tuning of the number of neighbors k to be used.

2.1.2.5
Multi-Layer Perceptron

The Multi-layer Perceptron (MLP) is a fully-connected neural network. As a neural network, it is composed of an input layer, an output layer and one or many hidden layers (Haykin, 1998), with all the neurons of a hidden layer receiving the results of all the neurons of a previous layer and being connected to all the neurons of the next layer.

Each layer i has its output $\mathbf{Z}^{[i]}$ calculated by applying a linear operation defined by the matrix $\mathbf{W}^{[i]}$ and the vector $\mathbf{b}^{[i]}$ to the previous layer $\mathbf{Z}^{[i-1]}$ followed by a element-wise non-linear activation function f_a, responsible for capturing the nonlinearities in the data, as described in equation 2-10. The number of output features in the matrix $\mathbf{Z}^{[i]}$ is determined by the number of neurons in the hidden layer and the term $\mathbf{b}^{[i]}$ is usually associated to an extra bias neuron.

$$\mathbf{Z}^{[i]} = f_a(\mathbf{Z}^{[i-1]}\mathbf{W}^{[i]} + \mathbf{b}^{[i]}) \tag{2-10}$$

The main difference between a MLP for classification and regression rests on the output layer: when used for regression purposes, the output layer consists of only the linear term exposed in equation 2-10, while for classification either a logistic function (binary classification) or a softmax function (multi-class outputs) is applied after the linear operation in the output layer. This is very similar to the treatment made in linear models, with the main difference that due to the nonlinear operations between hidden layers there is no analytical solution for both classification and regression problems. The training process consists of finding $\mathbf{W}^{[i]}$ and $\mathbf{b}^{[i]}$ for each layer that minimizes a loss function by using gradient-based optimization methods. This is only achievable algorithmically and efficiently thanks to the backpropagation method (Haykin, 1998) that applies chain rule derivatives to determine the gradient for each parameter in the network.

2.1.2.6 Decision Trees

Decision Trees are easy to interpret supervised methods that map the desired output values by applying simple decision rules according to understandable data characteristics (Kingsford & Salzberg, 2008). The basic approach of the algorithm consists on splitting the dataset according to the value of one input variable seeking to minimize a pre-defined error metric. Similar to the K-Nearest Neighbors process, for the created subsets the value attributed for the samples is the average value of all the samples contained in the subset for regression and the most frequent output class for classification. For classification problems, the Gini impurity is the error metric minimized in the splitting process while the mean squared error is the metric minimized in the regression problem.

The splitting process may be repeated in the subsets until a stopping criteria is reached, normally reaching a subset with a pre-defined minimum number of samples or a maximum number of subset levels (called tree depth) (Breiman et al., 2017). In the tree's hierarchical structure, each splitting decision is called a node and the final level where no more splitting is performed is called a leaf node. The inference of the value of a new sample is done by following the decision process mapped by the tree until a leaf node is found and the value of the output is mapped to be the most frequent output value of elements in the leaf (classification) or the average output value of said elements (regression).

The main advantages of the decision trees rest on the fact that they are easy to interpret, somehow emulating the decision-making process of human beings. Decision trees also need little to no data pre-processing, being able even to deal with categorical and null data as inputs. On the other hand, without pruning or limiting tree depth and leaf size decision trees may suffer from overfitting. Decision Trees also may present instabilities due to slight train data changes, that if in lower levels may change entirely the tree structures. These disadvantages, however, may also be mitigated by applying ensemble techniques.

2.1.2.7 Ensembles

Ensembles combine the predictions of several base estimators in order to improve generalization and robustness over a single estimator (Sklearn, 2023). The core concept is to improve results not by proposing stronger estimators, but to combine the predictions of several weak estimators ensuring that different estimators have better performance in different regions. This is ensured by applying adequate training algorithms.

Among the existing ensemble algorithms, the bagging and boosting algorithms are the most common. Bagging ensembles consist on training individual base estimators with a subset of the training dataset, randomly sampled with replacement (Breiman, 1996). The bagging approach counts on the randomness of training data for each estimator to improve results and allow, for example, for parallel model training. The boosting approach consists on training base estimators sequentially ensuring that the next trained estimator focus on improving the result of the previous one. Among the common boosting techniques, the Gradient Boosting focus on reducing a loss function with each new estimator added by means of a gradient descent algorithm (Friedman, 2001). Alternatively, the Adaptive Boosting (AdaBoost) technique reduces the error of the previous estimators by weighting incorrectly predicted samples in order to allow the next estimators to focus attention in improving these predictions (Schapire, 2003; Sklearn, 2023).

Decision trees are common base estimators used in ensemble techniques as ensembling helps mitigate most of the disadvantages of the estimator. They are usually applied in Gradient Boosting, AdaBoost and Bagging ensembles as base estimators, but also have specific ensemble techniques. An example of tree-based ensemble is the Random Forest technique, which consists in training a set of decision trees applying the bagging ensemble technique while also randomly omitting input parameters for some estimators, considering a random subset of features at each split in the decision tree (Ho, 1995). Another similar tree-based ensemble is the Extremely Randomized Trees (Extra Trees), which is similar to Random Forest but uses a modified Decision Tree as a base estimator that do not select the best split for each node but chooses the best split among randomly generated split candidates, adding more randomness for the model (Geurts et al., 2006)

In this work, as all the ensembles built used Decision Trees as base estimators, the term tree-based ensembles were used to generally described all the techniques listed in this section.

2.2
Physics-Data Hybrid Models

Different from a black-box model as presented in section 2.1.2, a hybrid model tries to add physical information to the black-box model in order to improve its performance by coupling it with a first-principles model. Liu et al. (2023) lists some strategies to achieve this purpose, starting with simply providing a physical model's outputs ($\hat{\mathbf{y}}$) as input features for a black-box model, as illustrated in equation 2-11. This strategy is a versatile one that can be easily applied to both classification and regression problems.

$$f(\mathbf{X}, \hat{\mathbf{y}}) \approx \mathbf{y} \tag{2-11}$$

Another strategy for hybrid model building consists of creating residual black-box models that are not trained to predict the output variable, but the expected deviation between the desired output and the physical model output, as illustrated in equation 2-12. This is also a powerful strategy, but of more limited application, as it cannot be easily adapted for classification problems, as a deviation between two output classes is not necessarily quantifiable to be used as an input for a black-box regression.

$$\hat{\mathbf{y}} + f(\mathbf{X}) \approx \mathbf{y} \tag{2-12}$$

It is also possible to combine the approaches from equations 2-11 and 2-12 to create residual models that use physical expected results as inputs, as illustrated in equation 2-13.

$$\hat{\mathbf{y}} + f(\mathbf{X}, \hat{\mathbf{y}}) \approx \mathbf{y} \tag{2-13}$$

2.3
System Identification

Dynamic systems are usually modeled through ordinary differential equations where time-derivatives of exogenous variables and the desired outputs are related. When this relationship is known, the dynamic response is usually obtained through transfer functions either on continuous or discrete time. In cases where this relationship is not known, system identification is a technique that may be applied to find a mathematical relationship between input and output using only available measurements, without prior knowledge regarding the structure, parameters, or physical principles of the model (Söderström & Stoica, 1989; Billings, 2013) by proposing a regression model that can estimate the value of the output variable y in the time instant k, as shown in equation 2-14

$$\hat{y}_k = f(\mathbf{X}_{k-1}) \,. \tag{2-14}$$

The input vector $\mathbf{X}$ is created by concatenating previous values of the input vector $\mathbf{u}$ and the output variable y considering a number n_u and n_y of previous values for $\mathbf{u}$ and y respectively, as illustrated in equation 2-15

$$\mathbf{X}_k = \begin{bmatrix} \mathbf{u}_k & \mathbf{u}_{k-1} & \dots & \mathbf{u}_{k-n_u} & -y_k & -y_{k-1} & \dots & -y_{k-n_y} \end{bmatrix} \,. \tag{2-15}$$

When the regression function f is linear, the method is called ARX **A**uto**R**egressive with E**X**ogenous variables). When the function is non-linear (usually a polynomial is used), the method is called NARX (**N**on-linear ARX). The classical non-linear approach for NARX cases is to apply polynomial expansions to the input vector $\mathbf{X}$, but due to the generic nature of the function f, any regressor can be used to create the model.

The main advantage of the system identification technique is the fact that it is able to adapt any available regressor to take into account time dependency. Also, the technique provides support for single input and single output (SISO), multiple input and single output (MISO) problems, single input and multiple output (SIMO), and multiple input and multiple output (MIMO) as long as the regressor chosen allows for multiple outputs. The main disadvantage of the method is that the base formulation counts on data sampled in regular time intervals, needing adaptations to use data with irregular time intervals.

2.4 Model Construction and Validation

2.4.1 Hyperparameter Tuning

In machine learning literature, a hyperparameter is a parameter of the learning algorithm (Géron, 2019). It is different form the model parameters as it is not estimated or affected by the training process, remaining constant during the model fit. Examples of hyperparameters include the maximum depth of Decision Trees, the C constant of SVM models and the k number of nearest neighbors to be considered in the KNN model.

As the results of all machine learning models may be deeply affected by the choice of hyperparameters, performing some sort of search to determine the best hyperparameters to use for the specific problem is advisable (Russell & Norvig, 2016). This can be done by either defining a hyperparameter grid of values to be searched, with each possible combination of pre-defined hyperparameter values being tested – a procedure called grid search – or to perform a random selection of hyperparameter values to test following pre-defined statistical distributions – a procedure called randomized search. Both hyperparameter sampling techniques are illustrated in Figure 2.2. For each hyperparameter combination one model is trained and the best model is selected according to a pre-defined error metric.

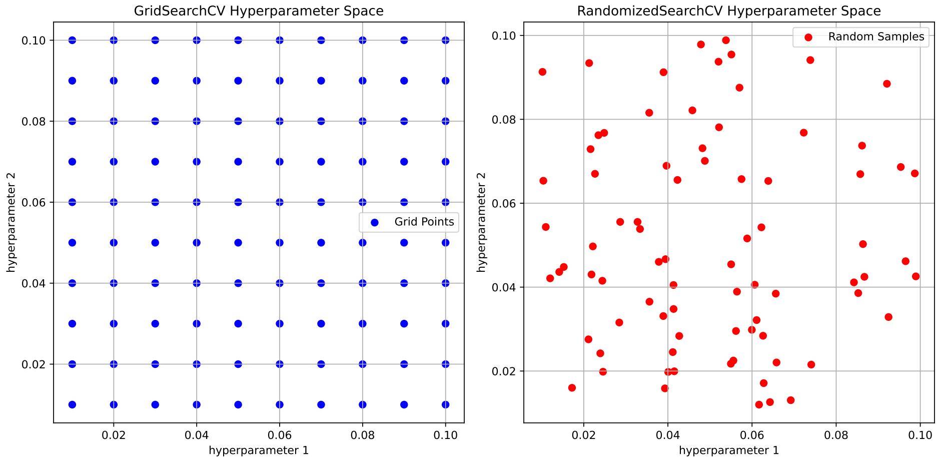


Figure 2.2: Examples of different hyperparameter sampling techniques for hyperparameter search: in blue, the grid approach where all hyperparameters to test have previously being selected and the possible test cases are the grid defined by them. In red, a randomized search where the values are selected at random from a uniform distribution.

Even though the grid search approach may be considered reliable as it fully explores the grid space proposed, it has the disadvantages of relatively elevated computational cost as all combinations of hyperparameters have to be tested, and the lack of power to explore hyperparameter values different form the ones used for generating the grid. Meanwhile, the randomized search approach has more exploratory potential and may provide more computational efficiency in better exploring the hyperparameter space with less samples due to its random nature.

2.4.2 Cross-validation

In order to select the best model in a hyperparameter search procedure, both performance metrics in the test dataset used and performance metrics in a holdout validation subset form the train data may be used. These approaches, however, may have the final results heavily influenced by the train dataset used, which is not desirable. In order to mitigate the dependency of the train dataset and ensure better model extrapolation, the cross-validation procedure is often a desirable approach for selecting between different competing models.

The cross-validation procedure consists of splitting training data into N folds at random and performing the train and evaluate procedure N times, holding one of the folds as a test dataset for evaluation purposes. The process of data splitting may be repeated M times, allowing for a final $N \times M$ metric score results to be obtained and compared, with the average of the obtained scores being considered for the model ranking procedure (Géron, 2019).

One additional advantage of the cross-validation technique is that it allows for an estimate of the confidence intervals ($\pm\delta$) for the final results by using the standard deviation (σ) of the scores obtained during cross-validation. This can be done by assuming that the metric result is a random variable of gaussian nature and applying the formula for the expected confidence interval of the mean error considering 95% confidence, presented in equation 2-16

$$\delta = \frac{1.96\sigma}{\sqrt{M \times N}}. \tag{2-16}$$

This approach was used in recent literature works (Kanin et al., 2019) to propose confidence intervals based on cross-validation.

2.4.3 Evaluation Metrics

Evaluation metrics are statistical quantities that try to quantify the model performance on a given dataset. There are specific metrics for regression and classification.

2.4.3.1 Evaluation metrics for regression

In this work the main error metrics used for comparing the performance of regression models were the coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE), calculated respectively by equations 2-17, 2-18 and 2-19

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{2-17}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{2-18}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{2-19}$$

In these equations, y_i represents the output value expected for the i^{th} sample on the dataset and $\hat{y}_i$ represents the value predicted by the proposed estimator.

2.4.3.2 Evaluation metrics for classification

For binary classification problems, where samples are only classified between positives and negatives, the evaluation metrics are calculated using as inputs the values obtained for the confusion matrix of a binary classifier, such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

The precision score is the ratio between all correctly classified positive samples and the total samples classified as positive. It is a metric adequate to estimate the confidence in the positive results obtained from the model. The formulation of the precision metric is shown in equation 2-20

$$\texttt{precision} = \frac{TP}{TP + FP}. \tag{2-20}$$

It is important to notice that the precision metric is only affected by the incorrect classification of negatives from the dataset as positives by the classifier, and thanks to that, it is unable to indicate if many positive samples are being classified as negatives by the model.

The recall score is the ratio between the correctly classified positive samples and all positive samples available in the input space. It measures how effective the classifier is in detecting the positives in the dataset as true positives, with higher values indicating that fewer positives will be labelled as negatives. The formulation for this metric is shown in equation 2-21

$$\texttt{recall} = \frac{TP}{TP + FN}. \tag{2-21}$$

The recall score is a complement to the precision score as the score decreases with the false negatives. However, higher recalls do not ensure that the classifier will be able to separate and correctly identify the negatives, since a classifier that predicts only positive as an output would, by definition, have a recall of 1 due to the lack of False Negatives (FN).

The F1 score is a harmonic mean of the precision and the recall score, which presents the main advantage of capturing both how effective the model is in classifying the positives and how trustworthy the positive results of the model are. The metric is calculated by the equation 2-22

$$F1 = \frac{2TP}{2TP + FP + FN}. \tag{2-22}$$

The accuracy is measured as the rate of correctly classified samples (both positives and negatives) and the total samples from the dataset. Similarly to the F1 score, it captures well the effectiveness in both classifying positive and negative cases, but in cases where the dataset is heavily imbalanced may be distorted in favor of classifiers that prioritize the most frequent class as an output. It is formulated by the equation 2-23

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}. \tag{2-23}$$

The balanced accuracy score is an alternative to the accuracy score that compensates for imbalanced datasets. It may be interpreted as the arithmetic mean of the recalls for positive and negative cases. The main advantage of this metric is that classifiers that only predict a single class as the output will have the same score as a random classifier, regardless of the frequency of said class in the original dataset. The formulation for the metric is shown in equation 2-24

$$BACC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right). \tag{2-24}$$

The metrics listed in this section can always be adapted for multi-class classification problems, but, in this work only the accuracy was used as it can be easily interpreted as the ratio between correctly predicted samples and the total dataset regardless of the number of different output classes. In cases where a metric that decreases with the performance is needed, the inaccuracy metric, defined by equation 2-25 was used

$$INACC = 1.0 - ACC. \tag{2-25}$$

2.5 Model Explainability

The search for model explainability started with the growing application of machine learning models in different business sectors and their increasing complexity that makes understanding the inner works of the decision-making process difficult to grasp, being an alternative to better understand various aspects of the model and addressing stakeholders' concerns about the drawbacks of models, and data-specific biases (Belle & Papantonis, 2021). Explainable AI has become especially relevant in the context of the industry 5.0, where the presence of AI in decision-making process makes imperative to these models to be interpretable in their results (Domínguez-Jiménez et al., 2023).

The search for explainable AI has stimulated the development of many industry standard tools to address the black-box model explainability issue, including the **SH**apley **A**dditive ex**P**lanation (SHAP) analysis (Lundberg & Lee, 2017), a unified approach to assess the input feature importance for a final prediction, providing explainability for complex black-box models in machine learning. The fundamentals fo SHAP are based on the concept of Shapley values based on game theory to determine the contributions for each player in an n-person game by attributing to each player a share of the final surplus (Shapley, 1951).

The typical SHAP output is, for each sample, a set of Shapley values attributed to each input in the sample that, when added with a fixed bias obtained from the dataset, provides the probability function predicted for the final model in that sample. This follows the additive feature attribution method, where the final model function $f(\mathbf{x})$ is approximated by a function $g(\mathbf{x}')$ of binary variables $\mathbf{x}'$ that can be mapped to $\mathbf{x}$ through the relation $\mathbf{x} = h_x(\mathbf{x}')$

$$f(\mathbf{x}) \approx g(\mathbf{x}') = \phi_0 + \sum_{i=1}^{N} \phi_i x_i' \tag{2-26}$$

Where $x_i' \in \{0, 1\}$ is a variable that can be interpreted as the feature presence, where 0 indicates that the feature is not present (or, at least, relevant according to the mapping function) and 1 that the feature is present in the sample. By ensuring the properties of local accuracy, missingness and consistency, it is possible to demonstrate that the values ϕ_i converge to the Shapley values (Shapley, 1951; Lundberg & Lee, 2017) in cooperative gaming theory, with the interpretation that ϕ_0 is the expected value of the model for cases where all simplified features are missing ($\mathbf{x}' = \mathbf{0}$) and each value ϕ_i represents the individual contribution of each input for the final model result.

The SHAP approach allows for both the identification of the most relevant features for each sample result and the identification if a given feature contributes positively or negatively for a given answer obtained, helping better interpret model trends for both classification and regression problems.

3
System Identification Techniques for Soft Sensors and Multiphase Flow Metering

Production rates estimation and forecasting is a complex problem in multiphase flow systems, such as oil wells. In the oil and gas industry, this task is normally performed by first-principles simulation models, which are costly to maintain and adjust. An alternative to this process could be black-box models based on historical data to perform this task.

This work proposes to perform an extensive evaluation of existing regressors to create a black-box system identification model to predict oil well liquid flow rates. Improvement strategies such as time series decomposition, and the use of current-time data are also evaluated. The final results demonstrated the importance of pressure and temperature data for the well flow rate determination and identified the use of current-time data as the best tested technique to improve the performance of the final model.

This work was presented at the 20th Symposium on System Identification of the International Federation of Automatic Control (IFAC), which took place in July 2024 at Northeastern University in Boston, MA. The reference to this publication is provided in section 1.5.

3.1
Case Study

The Volve field is an oilfield on the North Sea discovered in 1993 and operated by Equinor with production starting in 2008 and ending in 2016 (Sokkeldirektoratet, 2023). The Volve dataset (Trivedi, 2020), containing exploration and production data from this field, was disclosed by Equinor in 2018. The dataset includes reservoir and well production data, including pressure and temperature sensors, choke sizes, and daily averages of the production rates. Variables and definitions are listed in the table 3.1.

For this work, only data from well 15/9-F-1 C from the Volve field was used. This well was selected thanks to its dynamic behavior and multiple shut-ins during its life, making it an ideal case for system identification. Figure 3.1 shows the data for this well.

Table 3.1: Volve Dataset variables and definitions of interest for this case

Variable	Definition	Unit
BORE_LIQ_VOL	Average liquid flow rate on one day of production	m^3/d
AVG_DOWNHOLE _PRESSURE	Daily average of the downhole pressure measured by the permanent downhole gauge (PDG)	bara
AVG_DOWNHOLE _TEMPERATURE	Daily average of the downhole temperature measured by the permanent downhole gauge (PDG)	°C
AVG_WHP_P	Daily average of the wellhead pressure measured at the Christmas tree valve	bara
AVG_WHP_T	Daily average of the wellhead temperature measured at the Christmas tree valve	°C
AVG_CHOKE_SIZE_P	Daily average of the choke valve opening position (full opening percentage)	%

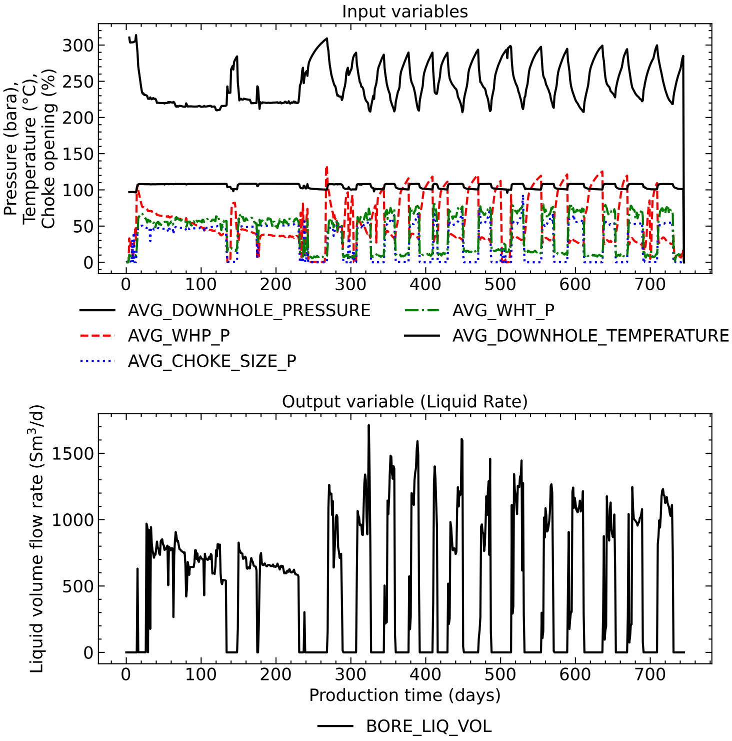


Figure 3.1: Example of well production data (pressure, temperatures, and flow rates) available on Volve dataset

3.2
Proposed method

The method proposed in this study consists of creating a system identification black-box model for simulating the dynamic behavior of a production well liquid flow rate. To do so, multiple estimators, lag orders, and different input spaces was tested and compared. The basic pipeline performed on this work is shown in Figure 3.2.

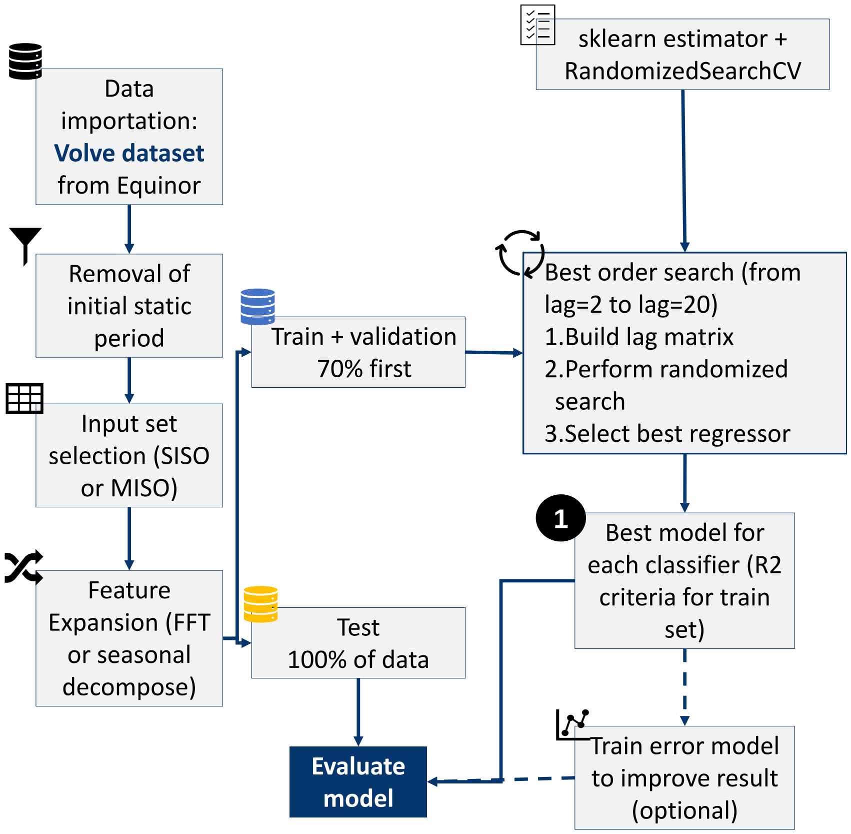

sklearn estimator + RandomizedSearchCV
Data importation: Volve dataset from Equinor
Removal of initial static period
Train + validation 70% first
Best order search (from lag=2 to lag=20)
1.Build lag matrix
2.Perform randomized search
3.Select best regressor
Input set selection (SISO or MISO)
1
Best model for each classifier (R2 criteria for train set)
Feature Expansion (FFT or seasonal decompose)
Test 100% of data
Train error model to improve result (optional)
Evaluate model

Figure 3.2: Schematic Representation of the optimal model selection procedure

3.2.1
Input features selection

To evaluate the importance of using multiple sensors to fit the model, the train data was fed to the pipeline on a single input single output (SISO) scheme and a multiple input single output (MISO) scheme, as summarized in Table 3.2.

For both cases, the average liquid volume flow rate (BORE_LIQ_VOL) was used as the output variable. It is important to note that the input space impacts the overall utility of the model. In a SISO case, only the choke opening was used by the model, a control variable that can be prescribed

Table 3.2: Evaluated input scenarios for the Volve Dataset model

Input space case	Inputs
SISO	AVG_CHOKE_SIZE_P
MISO	AVG_DOWNHOLE_PRESSURE AVG_WHP_P AVG_CHOKE_SIZE_P AVG_WHT_P AVG_DOWNHOLE_TEMPERATURE

through the prediction as a premise or as an output from another optimization process, making the final model useful both as a soft sensor and as a model for production forecasting. For the MISO case, variables that are also a consequence of the liquid flow rate, like pressures and temperatures, were used as inputs, turning the model dependent on real-time sensor data and unfit for forecasting purposes.

3.2.2 Data splitting and pre-processing

The production history from well 15/9-F-1 C was split into a train set composed of the first 70% of the production history and a test set composed of the remaining 30%. It was normalized by scaling all train data values to a [0, 1] interval.

Feature expansion based on time series decomposition was also tested in this work, with two strategies, applied to both input and output. The first one was decomposition using Short Time Fourier Transform (STFT), a variant of the classical FFT that takes into account the change in frequency distribution and phase in nonstationary signals. The second decomposition technique used was seasonal decomposition, which breaks time series into trend, seasonal, and residual components.

For the STFT, the implementation used in this work was the one presented in the `scipy` package from Virtanen et al. (2020). The STFT operation in the `scipy` package was used with default options. The series was split into two components using a threshold period of 25 days. For seasonal decomposition, the `scipy` package was used with a period of 30 days.

3.2.3 System identification approach

The base approach to allow considering time-dependency in the models evaluated was the application of system identification approach as described in 2.3. This was performed both with the original inputs and the expanded timeseries obtained through seasonal decomposition and STFT.

Normally, a system identification model always tries to predict the current time step value based on previous time step values to avoid violating causality relations, but it is possible to propose an alternative formulation, where current-time values of the input vector are considered in the model, as shown in equation 3-1

$$\hat{y}_k = f(\mathbf{X}_{k-1}, \mathbf{u}_k)\,. \tag{3-1}$$

Since many soft sensors are built only with current-time data as inputs, disregarding the time series for prediction purposes, this approach should not make the final model less useful for this objective. The same can be said for production forecasting purposes, bearing in mind the restrictions posed in section 3.2.1.

3.2.4 Regressor selection and hyperparameter tuning

For the regression task, this work made use of the `scikit-learn` (Pedregosa et al., 2011; Buitinck et al., 2013) package[1]. The sklearn regressors selected for this work are listed in the table 3.3. It is worth noting that the LinReg and PolyReg regressors correspond, respectively, to the ARX and NARX classical methods.

Table 3.3: `scikit-learn` regressors used for the system identification

Model name	sklearn regressor
LinReg (ARX)	LinearRegression
PolyReg (NARX)	PolynomialFeatures + LinearRegression
SVM	SVR
KNN	KNeighborsRegressor
DecisionTree	DecisionTreeRegressor
RandomForest	RandomForestRegressor
GradientBoosting	GradientBoostingRegressor
ExtraTrees	ExtraTreesRegressor
MLP	MLPRegressor

For cases with time series decomposition preprocessing, `MultiOutputRegressor` was used to allow multiple outputs with the normally single-output estimators used. A randomized search approach was used for hyperparameter tuning with a cross-validation of five folds and ten iterations. The randomized search was performed for 100 iterations.

[1] Source code for this work available in `https://github.com /prj-phcp/sklearn-sid/`

3.2.5 Error metrics

The main metric used to decide the best model in this work is the R^2 score. This error metric may be applied to two different types of simulations. The first type, called one-step ahead (OSA) is performed by using measured output data (y) from prior time steps to predict the next value of $\hat{y}$, as indicated by equations 2-14 and 2-15. The second approach is to perform a free-run simulation (FS), in which only initial values of y are used to start the prediction, and predicted values are incorporated into the lag vector defined in equation 2-15, and used to predict the next value of $\hat{y}$. Free-run simulations tend to deviate more from the original data since they are susceptible to error propagation but also are the most adequate type of simulation for long-term predictions and forecasting thanks to being less dependent on historical data.

In this work, these metrics were applied for the train data split or the total dataset (train and test), only for free-run simulations. The total dataset was chosen as the base for the main error metric because, even in the train dataset, free-run simulations are expected to introduce error thanks to error propagation along the predictions.

3.2.6 Lag order selection

The lag order was selected through a grid search. The best model was chosen for each lag order of 2 to 20 based on the highest R^2 from a free-run simulation on the train dataset.

3.2.7 Summary of improvement strategies

Table 3.4 presents a summary of terms used to identify the strategies applied in this work.

Table 3.4: Summary of applied strategies

Strategy	Definition
None	Simple system identification approach
u_k	System identification with current-time data
Seasonal	System identification with time series decomposed with seasonal decomposition
FFT	System identification with time series decomposed with STFT

3.3 Results and discussion

A summary of the best R^2 obtained for each input space and improvement strategy tested is presented in Figure 3.3. Figure 3.3 also presents the base estimator used to achieve the best result, and, for comparison purposes, the results obtained by applying classical literature models ARX and NARX. For some cases where time series decomposition was applied, it was not possible to train these simpler models due to numerical instabilities, hence no results are presented for the numerical decomposition cases. By analyzing Figure 3.3, it is possible to conclude that the application of more complex, nonlinear models was able to improve the final results when compared to the classical ARX and NARX models regardless of input space and strategy adopted. Results also show that, in cases where the same strategy was applied, the MISO input case had better performance than the SISO case. Even though this behavior is expected since the MISO case provides more physical information to the model, it is also worth mentioning that the MISO approach makes the model unfit for simulation purposes, limiting it to soft sensor applications thanks to the use of variables that are a consequence of the output, as described in Section 3.2.1.

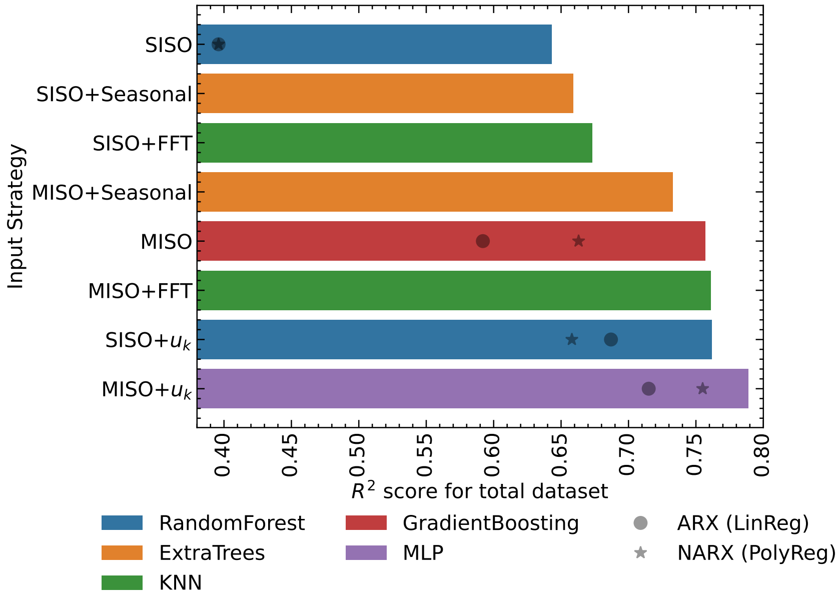


Figure 3.3: Comparison of R^2 metric on the total dataset for the best regressor for each input space and strategy

Still analyzing Figure 3.3, it can be concluded that the FFT expansion was the most successful feature expansion applied when associated with the KNN regressor in both MISO and SISO input spaces, indicating that the FFT components are adequate for measuring the distance between operational conditions in the input space. The seasonal decomposition expansion, however, was less effective, providing lower gains than the FFT in the SISO input space and even showing lower performance than not using feature expansion in the MISO space. This result indicates that little useful information can be extracted by analyzing the data seasonality for this problem.

Last, Figure 3.3 also shows that the best strategy evaluated in this work was adding current-time data to the input. This strategy was successful to the point that not only in the MISO input case it has the best performance evaluated, but in the SISO input case it outperformed every other MISO test performed, being the best model fit for production forecast purposes. It is also important to notice that, for production forecasting and soft sensor purposes, the approach of considering current-time data does not reduce the overall utility of the model, since in prediction forecasting the current-time SISO input is already prescribed and calculated to respect operational constraints and in soft sensors it is usual to use only current-time data in some models.

Figure 3.4 shows another evidence of the effectiveness of the current-time input strategy, showing that MISO with current-time inputs was the best strategy for every regressor tested except the KNN regressor, whose performance with the FFT feature expansion remained the best for the estimator.

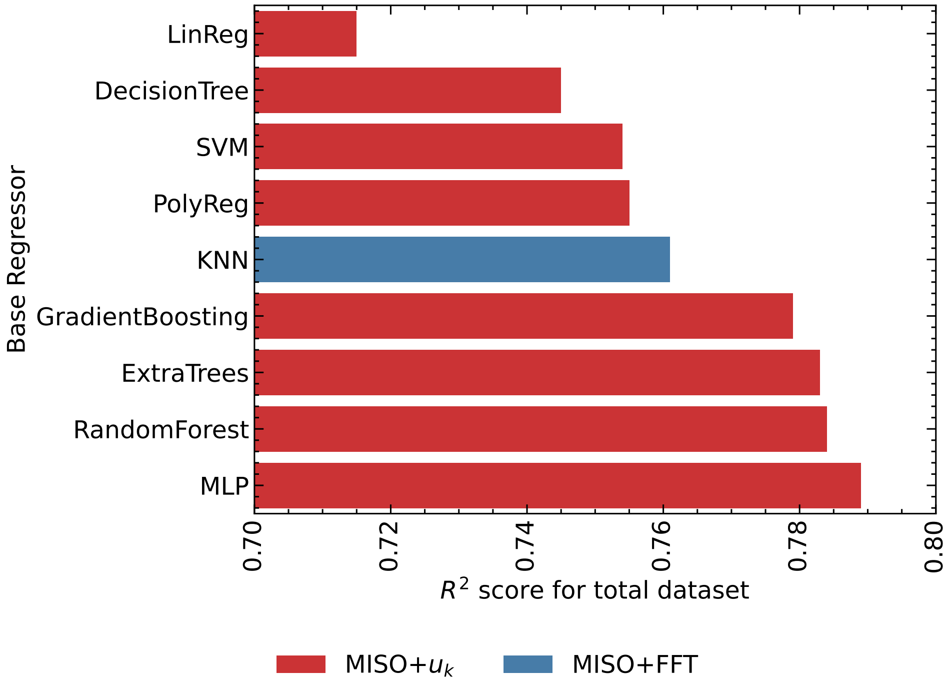


Figure 3.4: Comparison of R^2 metric on the total dataset for best strategy for each regressor tested

Figure 3.4 also shows that the current-time strategy allowed better results on some of the simpler models (as polynomial regression), which act as alternatives where computational performance is an issue. Figure 3.5. highlights this by showing that the current-time models with simpler regressors are the most cost-effective models considering the model size as an estimative of its complexity.

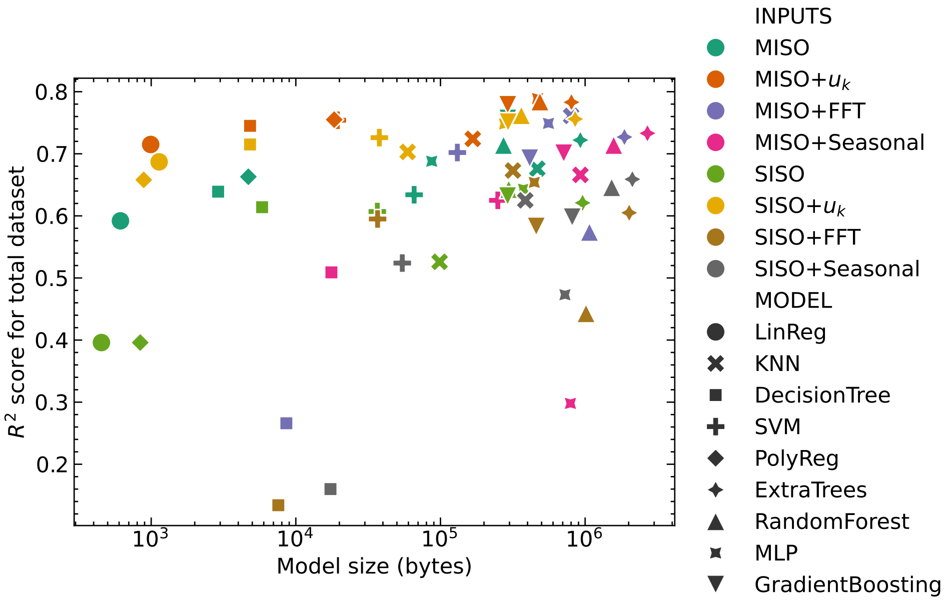


Figure 3.5: Model size compared with final R^2 score

Finally, the results obtained by the best-evaluated model are shown in Figure 3.6, where it is possible to see that the final model can represent adequately the tendencies of the original data, including shut-in periods.

3.4 Partial Conclusions

This work applied a system identification approach to creating a black box model of a production oil well, performing an extensive evaluation of regressors and improvement strategies. In search for the best model, this work built a Python class capable of using any sklearn-compatible regressor to perform the system identification and chose the best model by following a grid search approach for lag order selection and a randomized search approach for hyperparameter selection on the regressor. This approach was tested with both single input (SISO) and multiple input (MISO), with and without techniques to improve performance such as time series decomposition and usage of current time data to improve the predictions.

For the tests performed in this work, it was possible to conclude that the MISO input space and more complex models such as tree-based ensembles and neural networks presented the best results. Complex regressors allowed

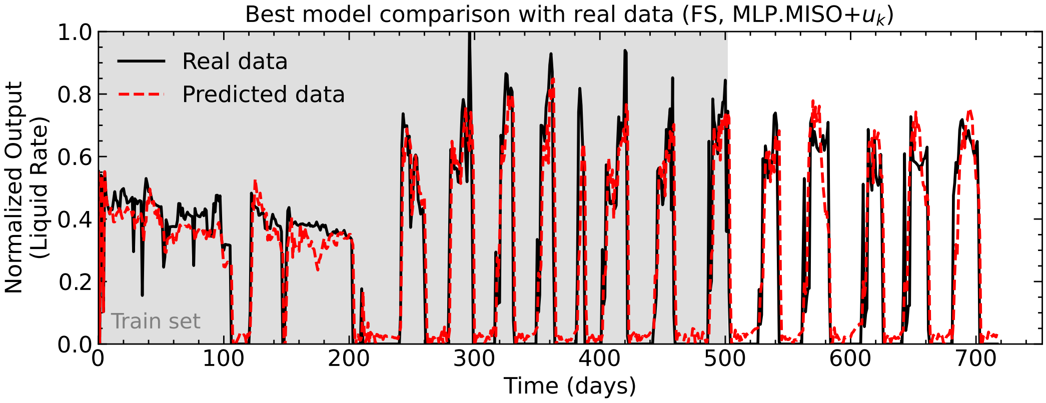


Figure 3.6: Comparison between best model tested and real data (free-run simulation)

for R^2 increases up to 62% when compared to classical system identification techniques while the MISO input space allowed for 15% R^2 increase when compared to SISO cases. Time series decomposition as a feature expansion strategy showed good results when pairing decomposition using STFT with the KNN regressor. The seasonal decomposition did not perform as well, indicating it may not be adequate for the problem.

The best strategy tested in this work was the addition of current time data to the system identification model, being this strategy able to significantly improve the results up to 15% of the R^2 and reduce the performance gap between the SISO and MISO models. This approach, however, should be used with care since it can reduce the overall utility of the model, even though it does not impede the use of the model for production forecasting and as a soft sensor, the main goals of this paper.

For future works, a better investigation of the effects of the current time data in the prediction and the possible improvements it can cause on the SISO models should be made, as well as testing other wells or similar datasets. Another possible work is the expansion of this study focusing on multi-output predictions, especially single input multiple output (SIMO) cases as the system identification approach still allows using the model outputs in previous timesteps as inputs for future predictions, allowing for models that only use control variables as inputs but still capitalize in the knowledge of sensor variables of interest.

4
An Interpretable Data-Driven Framework for Economic Assessment of Oil and Gas Exploratory Assets

Adequately evaluating industrial assets is of great strategic value for any business segment. This is also true for oil and gas exploratory assets, but their appraisal is not a trivial matter, especially due to the uncertainties involved and the number of scenarios to be evaluated for uncertainty mitigation. Although machine learning models are widely used in the oil and gas industry, and many business segments such as real estate have worked on proposing machine learning algorithms for asset appraisal, the application of said algorithms to perform economic assessment of exploratory assets in oil and gas is still an unexplored field.

In this chapter, it is proposed the creation of a black-box machine learning classifier capable of indicating the economic viability of an exploratory asset using a real project, heavily imbalanced dataset for model training and testing while evaluating estimators, possible inputs, and oversampling techniques to mitigate the dataset imbalance. The results showed that creating a model with said strategy is feasible, while the final results heavily dependent on the use or lack of oversamplers and whether investment data would be used as input for the model. The best results were obtained by using investments, followed by not using investments and applying oversamplers, but both strategies presented some trade-offs to be evaluated for each application.

4.1
Case study

To train the classification model, a database with over 4500 fields available for Petrobras was used. This database focuses on the economic assessment of projects and has both fields already in operation and possible prospects, assessed using metrics and preliminary cost estimation tools.

Table 4.1 lists the inputs available from the database used in this work. These inputs were selected for representing known factors that impact the overall cost and revenue of production development projects, such as contaminants, dominant hydrocarbon produced phase (oil or gas), area and volume of the reserves, and relevant challenges for the development of the field (being offshore, high reservoir depth, high water depth for the offshore case). Estimated investments data are also available on the dataset and might be used for the proposed studies, but as the main advantage of a pure machine learning approach is to avoid time-consuming studies such as estimating development costs during low-maturity phases such as the exploratory phase, using investments information that are highly uncertain for exploratory assets and demand some effort for its estimation is undesirable.

In this work, the database inputs described in Table 4.1 were divided into three major input sets to be used to infer some of the features' importance. The first one is the continuous inputs, which may be obtained from petrophysical, geological, or similar models developed for the sedimentary basin and analogous fields, like volumetric reserves, and fluid information. The second one consists of categorical inputs complementary to the continuous inputs, like identifiers of the most relevant hydrocarbon phase and of the presence of onshore and offshore development scopes. The third set consists of the estimated investments for the asset development, with information on capital, operational, and abandonment expenditures.

Regarding the dataset, Figure 4.1 shows the correlation index magnitude matrix of the input variables, making evident that most of the variables considered are independent. The magnitude was chosen to focus on how strongly the variables are correlated, regardless of whether they are positively or negatively correlated. Strong correlations between variables are only found between the recoverable and in place volumes, normally expected to have similar tendencies as greater reserves in place tend to become greater recoverable volumes, the Onshore and Offshore variables, as most of the projects do not have a coexistence of both onshore and offshore scopes, and the Investments group variables. The relatively high correlation between investments is interesting to notice, as many simple techniques for estimating investments are limited to regressions for CAPEX, with all the other costs being defined as fractions of the Total CAPEX. It is also a correlation that naturally arises as all the costs tend to increase with more complex production systems. Complex production facilities tend to be more expensive, have greater operational costs, and have more costly and difficult abandonment procedures.

Table 4.1: List of input variables considered in the dataset, categorized into three groups consisting of reserves, fluid, and asset properties (Continuous inputs), classification variables describing the project characteristics (Categorical inputs), and Capital and Operational expenditures information (Investments)

Variable group	Variable name	Unit	Description
Continuous inputs	Water depth	m	Field water depth for offshore fields
	Reservoir depth	m	Reservoir depth below sea level
	First Production Year	-	Year of start of production (real or projected)
	API	°	Oil API degree
	CO2 Content	%	CO_2 content on reservoir fluid
	H2S Content	%	H_2S content on reservoir fluid
	GOR	scf/sbbl	Reservoir fluid GOR
	Field Total Area	km^2	Field total area
	Np_eq	boe	Recoverable hydrocarbons volume (estimated)
	Neq	boe	Hydrocarbons volume available in reservoir
	FR	%	Recovery factor (estimated)
Categorical inputs	Oil	-	Flag indicating if oil is the main product of the asset
	Onshore	-	Flag indicating if development plan has onshore scope
	Offshore	-	Flag indicating if development plan has offshore scope
Investments	Drilling Total CAPEX	MMUSD	Total drilling capital expenditure
	Facilities Total CAPEX	MMUSD	Total facilities capital expenditure
	Total ABEX	MMUSD	Total decommissioning and abandonment expenditure
	Total CAPEX	MMUSD	Total capital expenditure including exploratory phase
	Total OPEX	MMUSD	Total operational costs full life

Most companies perform asset appraisal by applying the net present value formulae described in equation 4-1, where a cash flow in a given period composed of costs (C) and revenues (R) in a given moment in time (t) is brought to present value by using a premised discount rate (r) adopted by the company. For oil and gas assets, the revenues come mostly from the predicted production of hydrocarbons, while the costs are comprised of capital and operational expenditures for the field development plan.

$$NPV = \sum_{t=0}^{N} \frac{R_t - C_t}{(1+r)^t} \tag{4-1}$$

The decision on whether an investment is economically viable is mostly dependent on whether or not the NPV of said investment is positive. Another metric that can be obtained from the NPV equation is the internal return rate (IRR), which is the discount rate that ensures NPV zero and, if superior to the premised discount rate for the NPV calculation, implies an NPV greater than zero.

As the dataset also presented information on NPV for each asset,

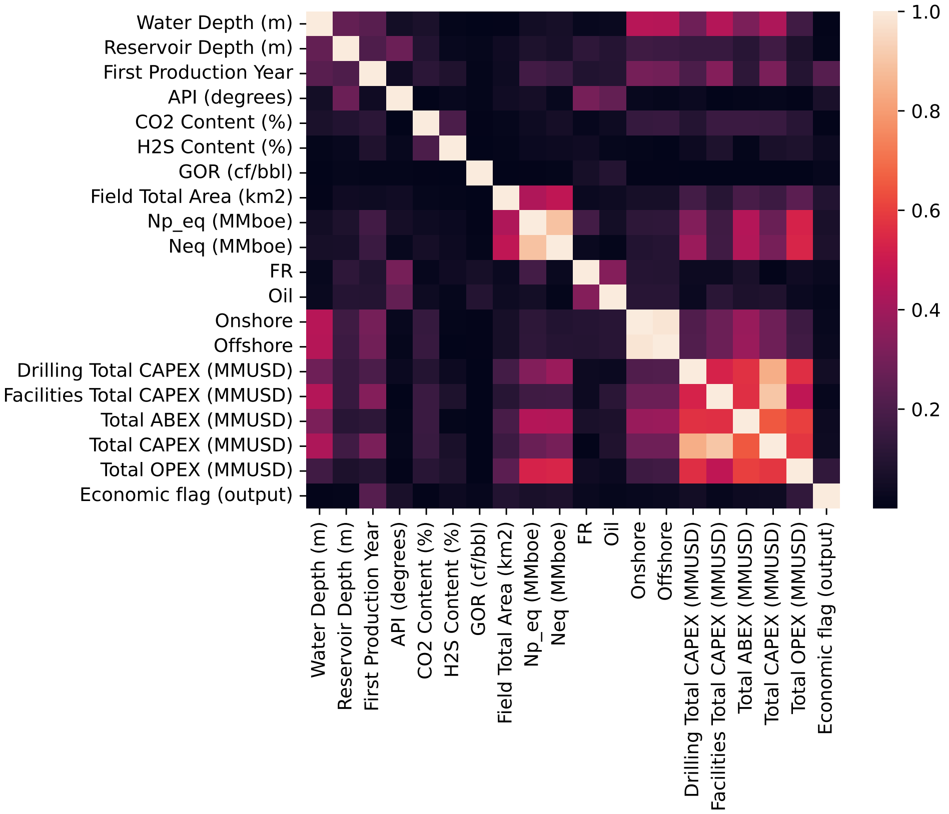


Figure 4.1: Correlation index magnitude matrix for all database input variables. It is possible to notice that most variables are independent, except the recoverable volume variables, the Offshore and Onshore variables, and the investments group variables.

considering a fixed return rate, for this work, it was assumed that every project with an NPV greater than zero was economical, and this rationale was used to create the output for the classification problem. It is important, however, to emphasize that, since the NPV used considered a default discount rate, the final classifier does not take it into account as an input, and the classifier created will always have the default rate as a premise. If it is desirable to have a classifier that can consider the discount rate as an input, either expanding the dataset to consider multiple discount rates or changing the classification strategy to a regression on the internal return rate instead of a classification would be possible strategies.

Another relevant premise that the dataset lacks is the oil reference price premise, normally considered through the price of the brent oil barrel. This premise is extremely relevant as it deeply influences the revenues that are obtained through the oil production and also holds some correlation with cost variables, but as the dataset does not allow for us to take into account the oil price as a possible input, it was not used in the creation of the model.

Figure 4.2 shows us the output classification distribution in the current dataset. It is important to notice that the dataset is imbalanced with of economic cases representing close to 70% of the database. Even though a predominance of economic cases is expected since they would be the focus of the business, this imbalance may distort the final classifier, causing it to prioritize economic cases as the output. To mitigate this risk, imbalance techniques, such as oversampling, were tested in this work.

4.2
Proposed Method

In this section, the methodology used for developing and evaluating the proposed classifiers is exposed. Figure 4.3 provides the graphical abstract of the proposed work pipeline adopted in this paper. For this work, Python 3.11.4 was used.

4.2.1
Data Loading and Pre-processing

The database was initially loaded in Python using the pandas package version 2.0.3. As the dataset consisted of multiple files, table merging was performed to prepare the dataset to be used in the classification problem.

After loading the dataset, data pre-processing steps were performed, with the removal of features from the database that are irrelevant to the problem. Lines containing NaN and outliers were also removed, reducing the complete

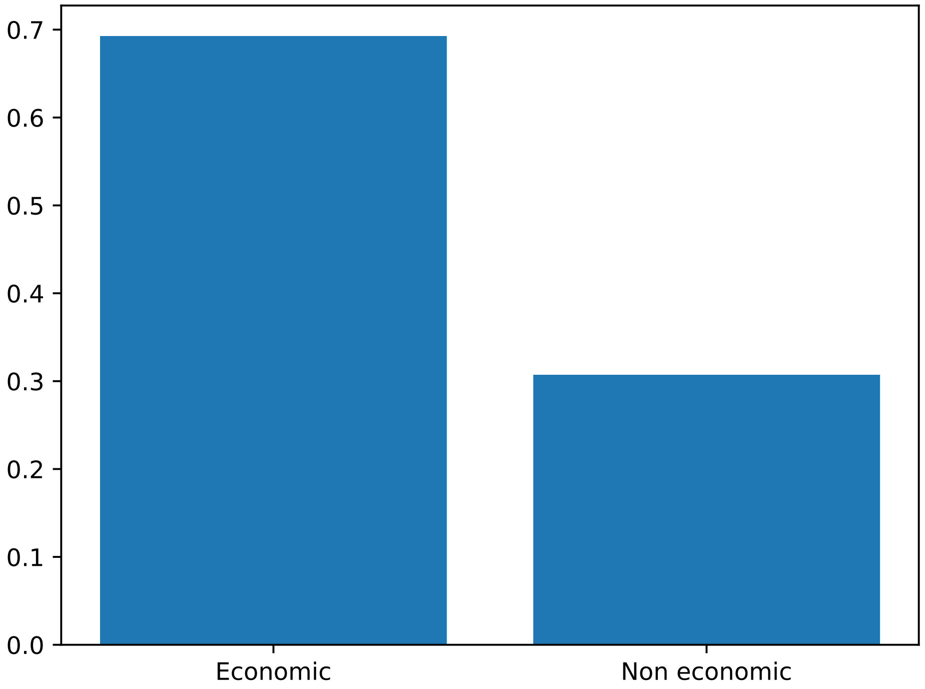


Figure 4.2: Database distribution according to the output case, showing that the dataset is heavily imbalanced in favor of economic cases (over 70% of the samples)

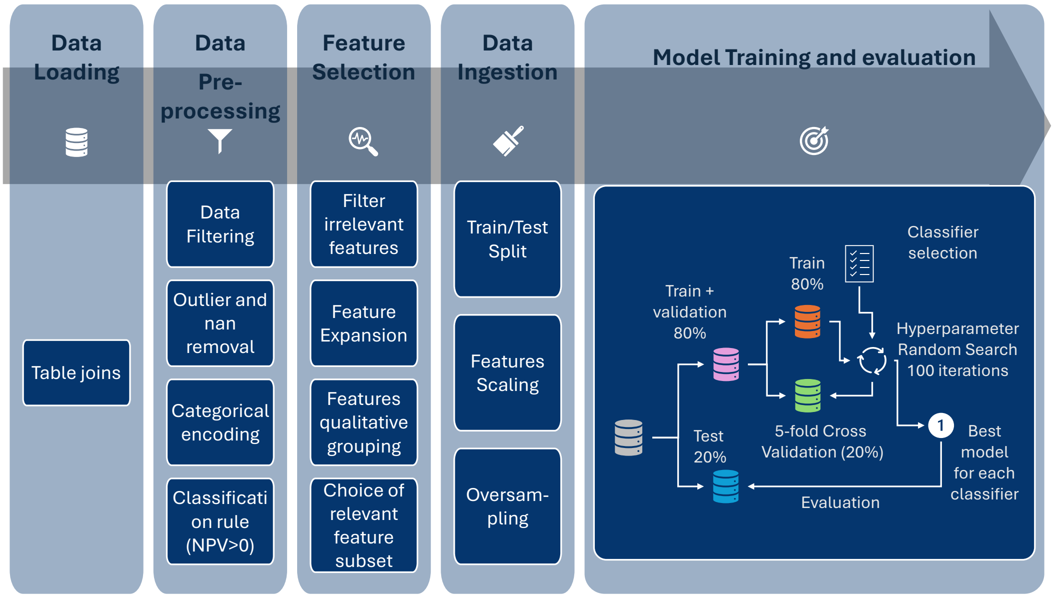


Figure 4.3: Data ingestion pipeline adopted for the economic classification problem proposed.

Table 4.2: Input spaces evaluated in feature selection and corresponding variable groups as declared in Table 4.2. These feature groups were used initially to perform a simplified feature selection process.

Input space	Variable groups used
X1	continuous inputs
X2	continuous inputs + categorical inputs
X3	continuous inputs + categorical inputs + investments

database to the 4681 samples used in this study. Finally, one-hot encoding was performed in the categorical inputs using the `OneHotEncoder` class from the `scikit-learn` Python package (Pedregosa et al., 2011). The output flag was also generated in this step using the rule of NPV lower than 0 as described in section 4.1.

4.2.2 Feature Selection

To assess the selected features' relevance, three subsets of the input space were created to be used as input space for a set of machine learning models without hyperparameter tuning. Table 4.2 relates the input spaces to be tested with the variable groups defined in Table 4.1.

The input spaces defined in Table 4.2 was used in the models with and without feature expansion and dimensionality reduction techniques to evaluate the effectiveness of these methods for the proposed case study. As feature expansion techniques, this work made use of the `PolynomialFeatures` expansor from the `scikit-learn` Python package and a set of custom functions normally used for simple regressions in engineering, such as inverses, logarithms, and the square root of the parameters provided as inputs. As a dimensionality reduction technique, the principal component analysis (PCA) implementation available in the `scikit-learn` package was used in this study.

4.2.3 Data Preparation

The data preparation consisted basically of splitting the train-test datasets and scaling the input features. The train dataset consisted of 80% of the total dataset, and the split was performed using the stratify option from the `train_test_split` function of the `scikit-learn` package, ensuring that both train and test datasets would have similar output class distributions. For scaling purposes, the `MinMaxScaler` was used to normalize the input space, ensuring that all the features in the train dataset would have values in the [0, 1] interval. The test dataset was normalized using the same scaler calibrated

Table 4.3: Main oversamplers considered in this work and respective reference. These oversamplers were all made available by the imblearn Python package.

Oversampler	Object in the API	Reference
Random	`RandomOverSampler`	Menardi & Torelli (2014)
SMOTE	`SMOTE`	Chawla et al. (2002)
ADASYN	`ADASYN`	He et al. (2008)
BlineSMOTE	`BorderlineSMOTE`	Han et al. (2005)
SVMSMOTE	`SVMSMOTE`	Nguyen et al. (2011)

for the train dataset, so the values in the test dataset may not be in the [0, 1] interval.

As exposed in section 4.1, the dataset considered is imbalanced, with around 70% of the samples being of the positive class. This work proposes evaluating, for some input spaces, oversampling techniques as a means to mitigate the effect of this imbalance on the final model response. For this task, the package `imbalance-learn` for Python (Lemaître et al., 2017) was used in this work. This package presents many advantages, such as compatibility with the `scikit-learn` API (Buitinck et al., 2013), pipeline objects enabling the addition of oversampling in the machine learning pipeline, and a set of ready-to-use oversamplers available in its implementation. Table 4.3 lists the oversamplers used in this work.

4.2.4 Model Training

The first step in training the machine learning models was to perform a hyperparameter search for each estimator tested. The hyperparameter search was performed by randomized search with cross-validation using the `RandomizedSearchCV` class from the `scikit-learn` package. For the randomized search, 100 different models with randomly generated hyperparameters were used against 50 randomly generated 5-fold splits of the training dataset. The best set of hyperparameters selected was the one with the best average results in the cross-validation step, considering the balanced accuracy score.

After selecting the best set of hyperparameters, the final model for each base estimator was then created by fitting the estimator using the selected hyperparameters and the total training dataset. For a list of the principal classifiers considered in this work, please refer to Figure 4.3.

4.2.5
Shapley Analysis

In this work, the model explainability was performed by applying the SHAP methodology proposed by Lundberg & Lee (2017) and discussed in chapter 2.5. For this task, the `shap` Python package developed by Lundberg & Lee (2017) was applied to determine the Shapley values for the best models of each final test performed to understand the final feature impact in the models.

This analysis is done on the probability output curve of the economic class in order to attribute to each output a continuous value of contribution to the final result. The SHAP Python package was used for this task, with the `PermutationExplainer` being selected for this study.

4.3
Results and Discussions

In this section, the initial feature engineering results are shown to explain the rationale for the input space selection for the hyperparameter search. The results for the best models after hyperparameter search are also shown, considering three principal case studies (no investments in the input, no investments in the input with oversampling, and investments in the input), and the final model ranking is presented. To better understand the feature importance in the final model and also the main strengths and weaknesses of the three case studies, the results of a SHAP analysis and the confusion matrices for the classifiers are presented.

Figure 4.4 shows a comparison of the results obtained during the feature engineering study performed for the dataset. By analyzing Figure 4.4, it is possible to conclude that the input space X1, containing only the continuous inputs listed in Table 4.2, showed very poor performance regardless of the classifier or the feature expansion strategy applied. Based on this result, for further tests, the input space X1 was not considered in this work.

Regarding the feature expansion techniques applied, consisting basically of polynomial expansion with simultaneous applying nonlinear functions to expand the input space, such as inverses, square roots and logarithms, Figure 4.4 makes evident that this strategy contributed very little for the improvement of the final results in the tested classifiers and dataset. The only considerable improvement obtained by using feature expansion techniques in this case was for the dataset X2, consisting of the continuous and categorical inputs from Table 4.1 for simpler estimators such as Logistic Regression, but even in this cases the final performance was worse than applying more complex classifiers to the unexpanded dataset. A similar conclusion can be obtained by analyzing

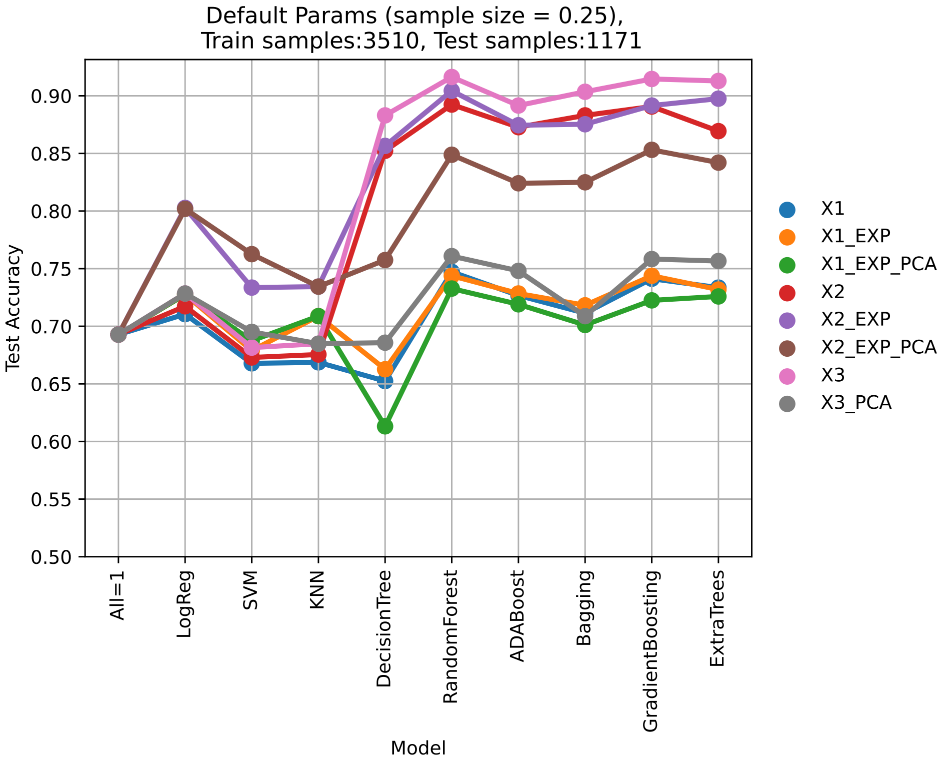


Figure 4.4: Feature engineering results obtained by fitting default estimators with different input spaces. Results show that the input space X1 is not adequate for the desired case study, while also making clear that feature expansion techniques and PCA application did not considerably improve the overall results.

the results where PCA was applied, associated with feature expansion, where it is possible to see that not only did PCA contribute very little to the final result, but for the dataset X2, it also reduced the performance for the more complex models. Based on these results, neither PCA nor feature expansion are further considered in this work.

Figure 4.5 shows the overall performance of the best tested classifiers obtained after the hyperparameter tuning step performed by using a randomized search. For this test, only the input spaces X2 (without investment data) and X3 (with investment data) were used, and their performance is compared in Figure 4.5. The impact of the oversampling techniques evaluated, only applied on the cases without investments as inputs, is also shown in Figure 4.5.

Figure 4.5 leads to the conclusion that neural networks and tree-based ensembles, especially GradientBoosting and ExtraTrees, were the best models for the proposed task, regardless of the input space and oversampling techniques. The only exceptions are the ADABoost and Bagging classifiers, which did not perform as well as other tree ensembles tested, even in cases where investments were used. It is also possible to notice by analyzing Figure 4.5 that, as a general rule, the use of investments in the input space tends to improve the results for the best classifiers tested, allowing balanced accuracies superior to 0.7 for the three best classifiers. For the less accurate classifiers, the use of investments led to results equivalent to or inferior to some oversampled cases, but never inferior to the tests performed without oversampling and without investments. This indicates that, for the studied application, investment data holds significant importance in determining if the prospect is economically viable or not. Since the main objective of this work is to propose a simple method for the appraisal of low maturity exploratory assets, the investment information is not always readily available or has significant uncertainties or errors associated with it, thus, it is desirable to focus on methods that are less dependent on investment data. On the other hand, as metrics and simpler methods for investments estimation are available in the oil and gas industry, such as the OAT (IPA, 2016) from Independent Project Analysis (IPA), investment-dependent classifiers are still viable solutions and should not be discarded, even if a critical analysis of the final solution is made necessary.

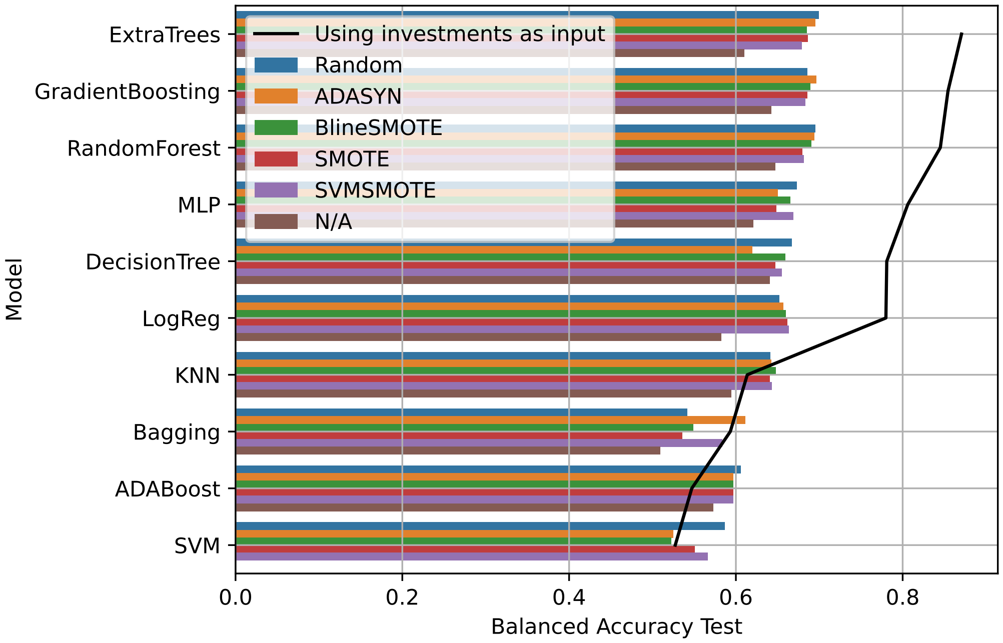


Figure 4.5: Performance comparison of classifiers obtained after hyperparameter search for the input space with Investments group variables and the input space without Investments group variables. For the no Investments case, the results with multiple oversampling techniques are also compared with the no oversampling case.

Table 4.4 shows more metric results only for the best models of each test (with investments, without investments and imbalanced, and without investments and not imbalanced). Results were ranked according to the best

Table 4.4: Results summary for the best classifiers obtained with and without the Investments group variables for the test dataset. In the cases without the investment variables, both the results without oversampling techniques and the results with the best-performing oversampling technique, considering the balanced accuracy score, are presented.

Model	Invest-ments	Over-sampler	*BACC*	*ACC*	Preci-sion	Recall	Rank
GradientBoosting	Yes	N/A	0.871	0.905	0.908	0.960	1
RandomForest	Yes	N/A	0.854	0.893	0.897	0.955	2
DecisionTree	Yes	N/A	0.845	0.875	0.899	0.923	3
ExtraTrees	Yes	N/A	0.806	0.867	0.861	0.963	4
MLP	Yes	N/A	0.781	0.825	0.858	0.895	5
ADABoost	Yes	N/A	0.780	0.862	0.838	0.994	6
ExtraTrees	No	Random	0.700	0.715	0.830	0.740	7
GradientBoosting	No	ADASYN	0.697	0.689	0.843	0.678	8
RandomForest	No	Random	0.695	0.709	0.829	0.730	9
MLP	No	Random	0.673	0.677	0.820	0.683	10
DecisionTree	No	Random	0.667	0.662	0.822	0.653	11
LogReg	No	SVMSMOTE	0.663	0.687	0.804	0.726	12
KNN	No	BlineSMOTE	0.648	0.623	0.820	0.584	13
RandomForest	No	N/A	0.647	0.748	0.769	0.909	14
GradientBoosting	No	N/A	0.643	0.742	0.767	0.900	15
DecisionTree	No	N/A	0.641	0.715	0.773	0.834	16
MLP	No	N/A	0.621	0.708	0.759	0.846	17
LogReg	Yes	N/A	0.614	0.736	0.749	0.932	18
Bagging	No	ADASYN	0.611	0.617	0.778	0.626	19
ExtraTrees	No	N/A	0.610	0.730	0.748	0.921	20
ADABoost	No	Random	0.606	0.653	0.760	0.729	21
KNN	No	N/A	0.595	0.682	0.745	0.821	22
KNN	Yes	N/A	0.593	0.673	0.746	0.801	23
SVM	No	Random	0.587	0.689	0.739	0.854	24
LogReg	No	N/A	0.583	0.721	0.732	0.943	25
ADABoost	No	N/A	0.573	0.701	0.729	0.906	26
Bagging	Yes	N/A	0.547	0.683	0.716	0.900	27
SVM	Yes	N/A	0.527	0.660	0.706	0.871	28
Bagging	No	N/A	0.509	0.695	0.697	0.991	29

balanced accuracy. Taking into account only the balanced accuracy metric in the test results, Table 4.4 made clear that the use of investments in the input space, with the best investment-based model having a balanced accuracy over 10% better than the best non-investment model with oversampling and over 34% better than the simple non-investment model, which only ranked 7th in the overall results. Also, considering only the imbalanced accuracy, it is possible to notice that the oversampling methods applied contributed significantly to improving the results obtained for non-investment classifiers, with the random oversampler being the overall most effective one.

Table 4.4 also indicates that, even though the oversampling techniques applied improved the balanced accuracy up to 8% and, generally, also improved the precision on the test dataset when compared to imbalanced non-investment cases, both accuracy and recall metrics were made worse, with the final recall being around 18% worse than the imbalanced case. That indicates that for oversampled cases, more economical prospects are being classified as non-economical, which is not desirable in a preliminary asset evaluation tool since it can lead to many missed opportunities. As a methodology intended for a quick screening of prospects that, in the future, would be detailed and analyzed with more data, it is preferable to have a greater number of false positives than a greater number of false negatives, reducing the loss of promising assets due to low recall while eliminating some non-economical assets before any further detailed work. Any non-economic assets that pass this first screening would have more studies conducted in future phases and eventually should also be eliminated. It is also important to emphasize that the classifiers built using investments in the inputs did not show this specific issue, having better performance in all the evaluated metrics, including the recall.

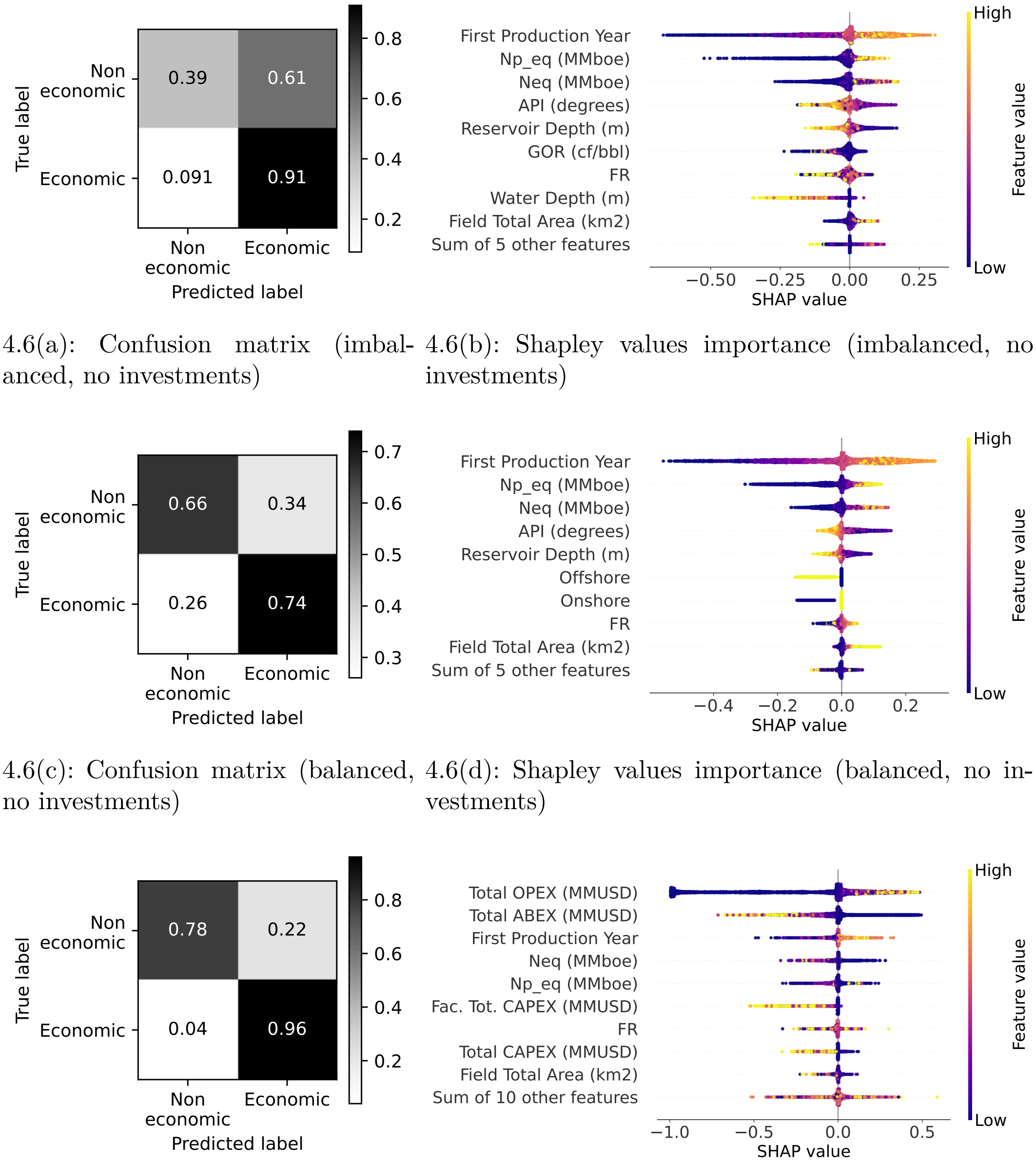
Figure 4.6 shows the confusion matrix and SHAP beeswarm plots for the best models considering the relevant case studies (without investments, without investments with oversampling, and with investments). The confusions matrices presented in Figure 4.6 once again made clear that the application of oversampling techniques on the non-investment case increased significantly the number of false negatives on the test set, indicating that to avoid missing opportunities due to false negatives, it may be preferable to use classifiers trained with the original imbalanced dataset. The confusion matrices also show that generally the estimators trained tend to perform better in classifying economic cases than non-economic ones, which is understandable when we consider that the base dataset has almost double the cases of economic assets when compared to non-economic ones. It is also relevant to notice that for the case with imbalanced inputs and no investments in the input (Figure 4.6(a)), the final result obtained for the non-economic cases is quite similar to a guess considering the original database distribution (or a dummy classifier).

Regarding the SHAP analysis performed, one may conclude by comparing the Figures 4.6(b) and 4.6(d) that little changed regarding the most relevant features for the final model when comparing the no investments cases with and without oversampling techniques to balance the input dataset. More than that, most of the tendencies displayed for the impact of the parameters on the final answer of the model being aligned to what would be expected on a complete asset evaluation, with high quality hydrocarbons (high API) and high recoverable volumes (Np and Np_eq) contributing to the economical aspect of the asset and characteristics that generally contribute to high investments demands, such as being offshore, deep reservoirs and higher water depths reduced the overall economical aspect of the asset. It is worth mentioning that for both cases as well the most relevant overall parameter was the first production year, which may be explained because assets that started producing in later dates should have more reserves still in place, allowing for better economic performance in the future.

Taking into account the results for the investments case, Figure 4.6(e) evidences once again that using investments is the best strategy to improve the accuracy of the final model, being able to improve both economic and non-economic cases detection. However, a more detailed analysis of Figure 4.6(f) shows that for the case where investments were used as inputs, they became the most relevant factors together with the First production year and the recoverable volume metrics. Even though this is expected, since a typical economical analysis considers the differences between revenue (controlled by the oil recoverable volumes) and the costs, some of the tendencies displayed in Figure 4.6(f) are against the expected tendencies, such as the improvement of the economical performance with the increasing of the total OPEX. As OPEX represents costs, it was expected that higher cost values would lead to worse economic results, indicating that this information is being used to infer other related asset properties known to be correlated to the OPEX, such as total production or facility size. Figure 4.1 indicates a relatively high correlation between OPEX and volume variables, such as Neq and Np_eq, which might help explain the unexpected tendency noted in the final OPEX contribution to the result. Once again, it is important to emphasize that high dependency of the final prediction on investments is also not a desirable behavior for exploratory assets assessment, since there is great uncertainty regarding these values, and they are not direct outputs from any geological or basin model.

4.4 Partial Conclusions

This work proposed to create a black box model capable of evaluating the economic viability of low maturity oil and gas assets, based on an imbalanced dataset of oil and gas projects and assets evaluated. This approach had not yet been tested, with the main contributions of this work being the application of the methodology to the proposed problem, the evaluation of the feasibility of the proposed model, a brief selection of possible input spaces for the problem, a brief study of feature expansion strategies for the dataset, the training of the model with randomized hyperparameter search for at least one dataset containing information about investments and one only counting on information regarding the geology and asset fluid, applying oversampling



4.6(a): Confusion matrix (imbalanced, no investments)

4.6(b): Shapley values importance (imbalanced, no investments)

4.6(c): Confusion matrix (balanced, no investments)

4.6(d): Shapley values importance (balanced, no investments)

4.6(e): Confusion matrix (imbalanced, investments)

4.6(f): Shapley values importance (imbalanced, investments)

Figure 4.6: Confusion matrices and SHAP values for the best models obtained for the three evaluated cases: no investments data and no oversampling techniques (4.6(a),4.6(b)), no investments data and oversampling techniques applied (4.6(c),4.6(d)), and investments data and no oversampling techniques (4.6(e),4.6(f)). The results are shown only for the best model of each case, considering the balanced accuracy score.

techniques to overcome imbalanced dataset limitations and applying SHAP analysis to evaluate the final best estimators and compare the final feature importance for the inputs.

For the initial feature selection and engineering studies proposed, tests showed that feature expansion techniques only improved results for very simple estimators, such as logistic regressions, while techniques such as PCA tended to reduce the final model performance. Regarding the possible input spaces previously selected, the most restricted one tended to perform poorly compared to the others, with the one containing investment information outperforming the others. The most restricted input was therefore eliminated from future tests, and feature expansion and PCA were not considered for the final model pipeline. After these tests, the classifiers were trained for the two best input spaces, applying oversampling techniques only for the input space without the investment variables, with tree-based ensembles and neural networks having the overall best performance in accordance with literature results for similar problems. Considering the imbalanced case with no investments as the base case, results showed that the use of investments as inputs produced a 34% balanced accuracy gain, outperforming every other case using input spaces without investments, while the application of oversampling techniques allowed 8% balanced accuracy improvements on the models that did not used investments as inputs, but closer inspection it was made clear that the recall metric was reduced up to 18% in oversampled cases, which may cause the loss of economically viable assets and indicate that some level of imbalance may be desirable on the dataset to prioritize economically viable assets.

Regarding the use of investments in the input, it improved both the balanced accuracy and the recall metrics of the final classifier, but closer inspection of the SHAP values for the classifiers indicated that some inputs had unexpected contribution tendencies with the final result, like cost variables impacting positively the final results for greater values. This is an indication of possible overfitting and that the model may be using cost variables to infer other field properties, which is undesirable. On that aspect, models without costs as input showed variations of the result more aligned with what is expected from the formulation of the economic problem.

For future works, developing models less dependent on imposed economic premises, such as regression models for internal return rate (IRR) instead of classifiers for the economic attractiveness, would allow the extension of this work to be more robust to changes in said premises. As the investment information improved the final results, both coupling the investments model with market or literature methodologies to estimate investments using more fundamental data, while also investigating the impact of this data in the final model, would be another possible line of work to develop after this paper. Documenting a full comparison of appraisal methods for a given asset, testing the methodology for different training datasets, or creating artificial datasets using already available metrics and the currently applied assessment techniques would also be possible future works. At last, due to the simplicity of the proposed methodology, applying it to other business segments may also be a possible future work.

5
Hybrid machine learning models for improving state-of-the-art mechanistic flow models

Steady-state multiphase flow simulation is key for production forecasting and allocation, but demands the estimation of relevant parameters, such as pressure gradient, gas fraction, and flow regime, to be executed. While literature exists on both physics-based modeling and black-box machine learning modeling to address this issue, hybrid models combining both approaches, with the potential to overcome most of each approach's limitations, are still scarce in the literature.

This chapter proposes a hybrid modeling approach that combines commercial mechanistic models with data-driven estimators, aiming to improve final prediction results for pressure gradient, gas fraction, and flow pattern. Performance improvements obtained with this strategy were 16-71% of the mechanistic RMSE for pressure gradient, 10-61% of the mechanistic RMSE for gas fraction, and 37-83% of the mechanistic accuracy for flow regime classification when compared to the benchmark mechanistic models. Compared to purely black- box models, the reduction was of 25-149% of the mechanistic RMSE, 9-85% of the mechanistic RMSE, and 3-11% of the mechanistic accuracy for the pressure gradient, gas fraction, and flow pattern prediction, respectively.

In this chapter, comparison with literature results was also performed, proving the effectiveness of the strategy, as a screening of different preprocessing strategies and estimators, including the TabPFN foundation model for tabular data proposed by (Hollmann et al., 2025). As the physical representativeness of black-box models for multiphase flow tasks is also a concern in the literature, a brief assessment was made using explainability techniques like Shapley values, demonstrating that physical representativeness is achievable but has to be verified during model selection.

5.1
Case Study

The case study consists of over 22,000 samples of experimental steady-state multiphase flow data. Figure 5.1 represents the typical layout of experi-

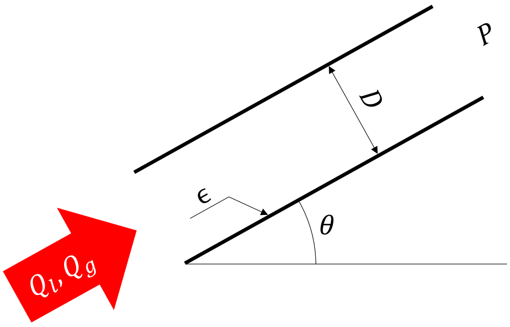


Figure 5.1: Schematic of controllable parameters during test procedures to determine multiphase flow behavior. For the variables' meaning, refer to Table 5.1

mental setups to raise this kind of data and the usual parameters that can be manipulated during tests. The dataset is proprietary and cannot be shared, but has been used in other similar works (Faller et al., 2023). The dataset has 14 features, where 11 can be considered inputs and 3 can be considered model outputs. The inputs consist basically of geometrical features, flow conditions, and fluid properties, while the outputs contain pressure gradient, gas fraction, and flow pattern information. Table 5.1 summarizes the parameters available in the dataset. It is important to comment that, while normally flow rates are the controllable parameters in experiments, it is usual in models and literature to work with superficial velocities, related to flow rates (Q) and diameters by equation 5-1.

$$j = \frac{4Q}{\pi D^2} \tag{5-1}$$

It is also important to notice that the dataset is heavily imbalanced, normally due to the natural laboratory and experimental setup limitations regarding space and resources. Figure 5.2 shows the distribution of geometrical features such as internal diameters and pipeline inclination in the dataset, evidencing that most of the experimental data obtained consists of diameters below 6 inches, which, considering most of Petrobras' scenarios, would be the lowest internal diameter considered for a multiphase production flowline. Inclination-wise, the majority of the dataset consists of either close to horizontal (near 0° inclination) or close to vertical ascending (90°) or descending (-90°), focusing on mapping the most common conditions in wells and horizontal pipelines.

Analyzing the fluid properties, Figure 5.3 displays that most of the data obtained is generated having either water or low-viscosity liquids, with many samples close to water density (close to 1000 kg/m^3). Regarding gas properties, air is the most commonly used gas with relatively low experimental pressures

Table 5.1: Variables available in the database. The inputs consist basically of fluid properties, geometrical features of the pipeline section, and its roughness. Meanwhile, the outputs consist of the pressure gradient, gas volume fraction, and a categorical variable representing the flow regime in the pipeline section, which has as possible values (1) stratified, (2) annular, (3) slug flow, and (4) disperse bubbles.

Variable group	Variable name	Symbol	Unit	Description
Inputs	DiaInt	D	m	Pipeline section internal diameter
	Angle	θ	°	Pipeline section angle with the horizontal axis
	Rough	ϵ	m	Pipeline section internal roughness
	Pres	P	Pa	Pipeline section internal pressure
	Vlnslp	j_1	m/s	Liquid superficial velocity in pipeline section
	Vgnslp	j_2	m/s	Gas superficial velocity in pipeline section
	RhoLiq	ρ_1	kg/m^3	Liquid density at pipeline section
	RhoGas	ρ_2	kg/m^3	Gas density at pipeline section
	FmuLiq	μ_1	Pa.s	Liquid dynamic viscosity at pipeline section
	FmuGas	μ_2	Pa.s	Liquid dynamic viscosity at pipeline section
	SigLiq	σ_{12}	N/m	Gas-liquid interfacial tension at pipeline section
Outputs	GRAD	dP/dL	Pa/m	Pressure gradient in the pipeline section
	ALPHA	α	-	Gas fraction in the pipeline section
	REGIME	ID	-	Flow pattern in the pipeline section

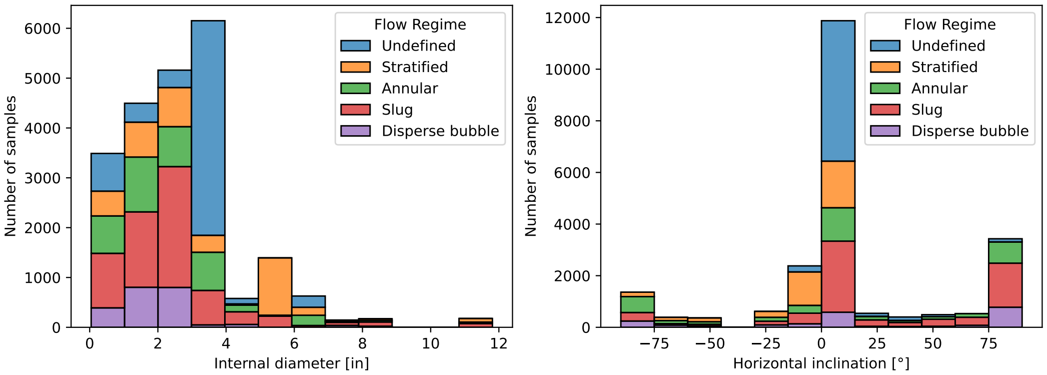


Figure 5.2: Geometrical features distribution in the dataset. It is possible to notice that most of the available samples consist of smaller internal diameters (<6”) and either vertical or horizontal flow.

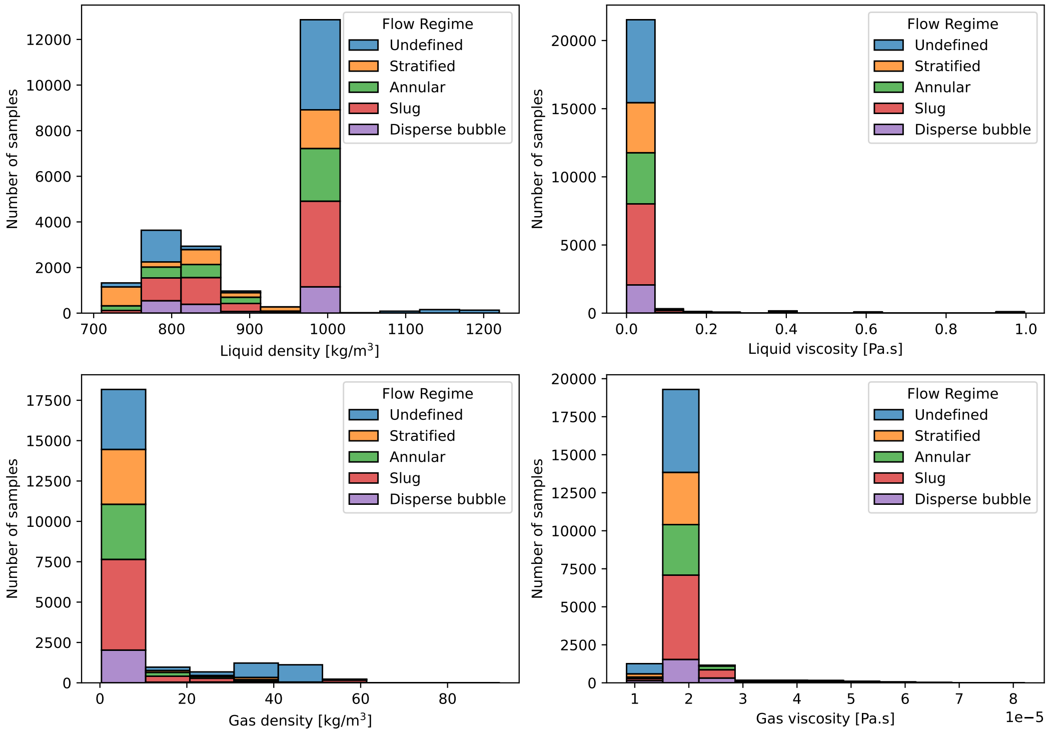


Figure 5.3: Fluid properties distribution in the database. This shows that most of the samples consist of experiments using water or light oils with low viscosity as the liquid phase and air or a similar gas at lower pressures as the gas phase.

close to 1 atm being the most common, creating a dataset with low gas viscosity variation and gas density very concentrated in low values (1000 times lower than the liquid).

Still discussing fluid properties, it is also important to remember that for gases at lower pressures, it is expected that some level of correlation exists between gas properties and the pressure, as dictated by the ideal gas law. This can be confirmed by the Figure 5.4, which represents the module of the correlation between variables and makes clear that fluid properties such as gas density and surface tension, and pressure are relatively correlated. As correlation between variables might pose a challenge for some machine learning algorithms, this aspect has to be taken into account.

Regarding the outputs normally obtained from multiphase flow correlations, it is important to bear in mind that these outputs comprise both continuous and categorical variables. While the Pressure gradient (GRAD) variable is continuous with no restriction to its value range and the void fraction (ALPHA) is continuous and restricted to a [0, 1] interval, the flow pattern is a categorical variable that represents the phase distribution inside the pipeline, that can be (1) stratified, (2) annular, (3) slug flow or (4) disperse bubbles and

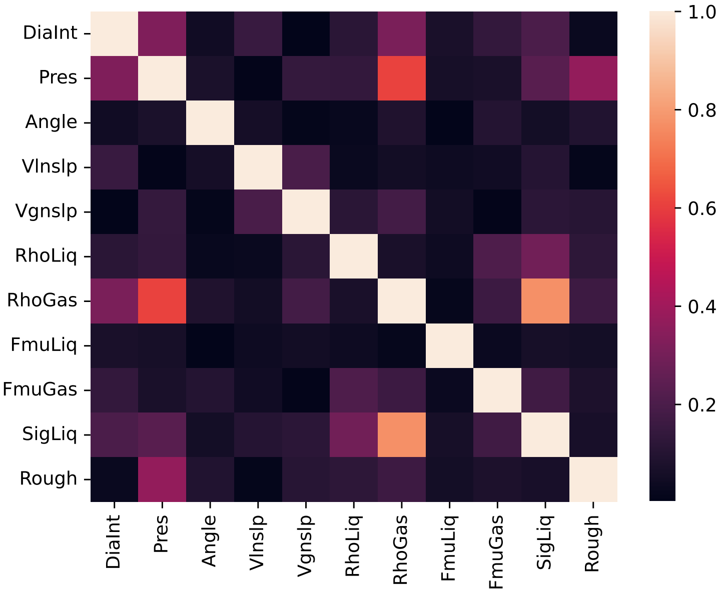
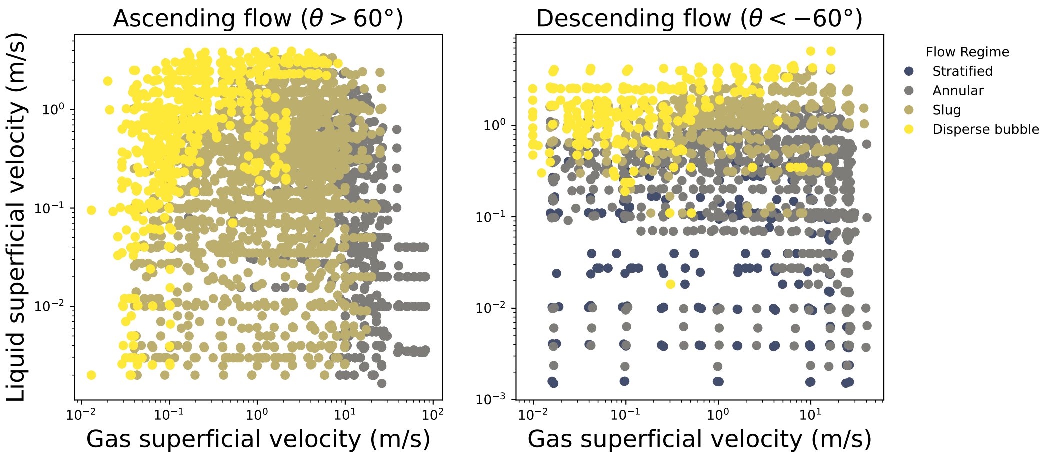


Figure 5.4: Correlation between all the inputs presented in the dataset. Relatively high correlations are found only between fluid properties related to the gas and the pressure variable

used by Kanin et al. (2019) in his black-box classification efforts. This implies that a multiphase flow correlation that is able to predict all the present outputs acts as both a regressor and a classifier at the same time, a task performed in mechanistic correlations by combining flow maps comprised of the stability criteria for each flow pattern with specialized pressure drop models for each flow regime.

Many works that focused on modeling flow pattern transition by using first-principles equations, like the works of Taitel & Dukler (1976) for horizontal pipes, the unified criteria presented by Barnea (1987), and the works of Gomez et al. (2000) for the unified model arrive in sets of equations that depend on geometrical parameters and fluid properties. For a fixed geometry and known fluid, a usual form of displaying the results is as phase maps, where the boundaries between regimes are displayed as a function of the logarithm of the superficial velocities. Figure 5.5 shows the dispersion of the flow pattern data for the dataset studied in this work for ascending and descending flow as a function of the log of said superficial velocities. As expected, it is possible to see that these properties allow for some level of separation between patterns, but as the dataset is not perfectly homogeneous in fluids and geometric characteristics, separating flow patterns becomes a non-trivial matter, demanding more information than just the relevant velocities. It is also interesting to notice that, especially for descending flow, there is a significant overlap of flow patterns, especially between the annular pattern and the slug and stratified patterns. This may prove to be a challenge for the classifiers tested to solve the problem.


Ascending flow ($\theta > 60°$)

Descending flow ($\theta < -60°$)

Figure 5.5: Flow regime dispersion of the dataset for ascending and descending flow. The log of the superficial velocities of the phases was selected according to the views presented in Barnea (1987). It is possible to see that the superficial velocities are not enough to fully separate the classes in the dataset used

At last, it is also important to comment that there are missing values for all the outputs in some samples, as not all the experiments were interested in determining or studying all the outputs of the database. For example, over 25% of the samples do not have flow regime information attached to them. Regarding flow regime, there is also an imbalance in the dataset, with more samples associated with slug flow (almost 30%) and fewer samples associated with disperse bubble flow (less than 10%). These behaviors should also be taken into account when treating the dataset and training the models.

5.2 Proposed Method

Figure 5.6 presents the graphical scheme of the proposed work. In this section, further details on the applied methodology is presented.

5.2.1 Database splitting and preprocessing

In order to evaluate the models to be tested, the database was split between a train dataset comprising 80% of the samples and a test dataset of the remaining 20% of the samples. As commented in section 5.1, the dataset presents many imbalances, especially in its inputs, but also in the classification output of the flow pattern. To ensure that the data splitting

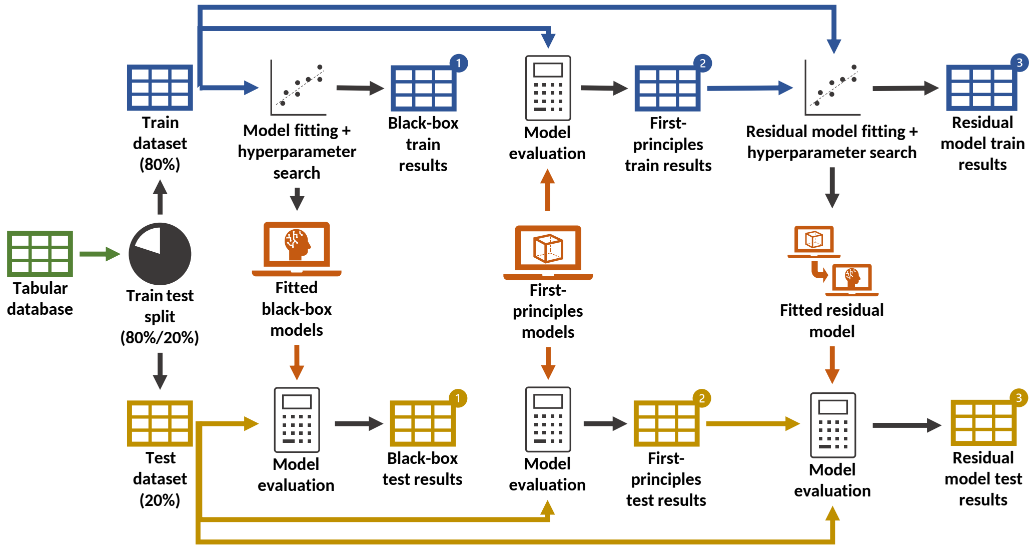


Figure 5.6: Schematic for the proposed study.

process would not affect the model training and evaluation process due to different flow regime distributions in the train and test datasets, the splitting was performed stratifying the flow patterns. As over 25% of the samples do not have flow pattern data, the no flow pattern data samples were also proportionally distributed between the train and test datasets.

It is possible to study strategies to improve the model's final results by feature engineering. Faller et al. (2023) obtained better results in black-box estimators by adding to the input space well-known adimensional quantities related to the multiphase flow phenomenon. In this work, however, no feature engineering was performed, with the ensembling with first-principles models being the main strategy to add physical knowledge to the model.

5.2.2 First-principles models evaluation

In this work, we define first-principles models as models created by using physical modeling of the multiphase flow phenomenon in pipelines. Even if these models may present closing relationships and parameters that might be fine-tuned depending on the experimental data available, this work assumes first-principle models as reference models that cannot be adjusted or changed, so no fitting process is performed in this step.

For this work, the first-principle models evaluation was made by using the commercial software PROSPER, a steady-state multiphase flow simulator developed by PE Limited. The software has an automation API that allows the manipulation of almost every input parameter, simulation running, and results reading. As this work aims to get only pressure gradient, gas volume fraction, and flow pattern, a simple 1-meter-long model, with a single pipe, was built to act as a virtual test bench for generating the results, and the Gradient calculation was performed in the software.

As commercial simulators usually take into account as inputs a fluid model, gas oil ratio of the fluid at standard conditions, a representative standard flow rate (usually liquid or gas), and water fraction in the liquid or water cut (WCT), some input manipulation was performed to reproduce the dataset results. First, a simple tabular fluid was defined, with constant fluid properties equal to the fluid properties present in the dataset for a large pressure range, including standard conditions. Water cut was assumed to be zero as the dataset is only 2-phase flow. As the fluid properties were defined as constant, the input GOR for the gradient calculation could be estimated using the superficial velocity definitions through equation 5-2.

$$GOR = \frac{j_2}{j_1} \tag{5-2}$$

For the flow rate to be imposed, this work considered the liquid rate to be the representative rate. The liquid rate input for gradient calculation was then calculated by using the liquid superficial velocity and pipeline diameter as inputs, according to equation 5-3.

$$Q_1 = j_1 \frac{\pi D^2}{4} \tag{5-3}$$

For this work, only the correlations Hydro2P, PetroleumExperts5, OLGAS2P, and OLGAS3PEXT were considered as first-principles models to be evaluated. These were selected for being the mechanistic models currently available in PROSPER for Petrobras. Other models available, such as OLGAS3P and Hydro3P, were discarded because they are only 3-phase versions of already contemplated models, and as the used dataset is only 2-phase, they should not present any differences from their 2-phase counterparts.

5.2.3 Black-box model training

To serve as a comparison base to validate the gains of the hybrid model strategy, first, a pure black-box model approach was adopted using classical machine learning estimators available on the `scikit-learn` package (Pedregosa et al., 2011) following the data-driven pipeline presented in section 2.1. Table 5.2 lists the estimators considered for both the classification problem of the flow regime and the regression problems for pressure gradient and gas fraction and represents a great variety of classical machine learning algorithms, such as linear algorithms, support vector machines (SVM), neural networks,

Table 5.2: Estimators used for the black box regression, their adopted name, and the classes from the sklearn api (Buitinck et al. (2013)) for both the classification and regression problems. The TunedTabPFN demands the additional package tabpfn (Hollmann et al. (2025)) but it is also scikit-compatible

Estimator name	Classifier class	Regressor class
Linear	LogisticRegression	LinearRegression
Polynomial	PolynomialFeatures+ LogisticRegression	PolynomialFeatures+ LinearRegression
SVM	SVR	SVC
DecisionTree	DecisionTreeClassifier	DecisionTreeRegressor
RandomForest	RandomForestClassifier	RandomForestRegressor
GradientBoosting	GradientBoostingClassifier	GradientBoostingRegressor
ExtraTrees	ExtraTreesClassifier	ExtraTreesRegressor
MLP	MLPClassifier	MLPRegressor
TunedTabPFN	TabPFNRegressor	TabPFNClassifier

trees, and tree-based ensembles.

As the database is comprised of physical properties with different magnitudes, the input scale might exert some influence on the final model performance. Beyond the scale importance, it is also known that some inputs in the database should have some physical correlation between them, as it is expected from the pressure and gas properties as exposed by Figure 5.4. To evaluate the scale effects, all the models tested were trained with and without prior data scaling with either `StandardScaler`, `MinMaxScaler`, or `RobustScaler` available on the `scikit-learn` package. The models with no scaling and with each of the scalers tested were also trained with and without the application of principal component analysis (PCA) in the input dataset, to evaluate the relevance of correlation between inputs on the final model.

In addition to the `scikit-learn` estimators, a black-box model was also created by using the proposed TabPFN foundation model (Hollmann et al., 2025). This model is based on the attention mechanism largely used in transformers applied in large language models (LLM) and was created to predict tabular data up to 10000 samples for both classification and regression efficiently and with greater accuracy than classical models.

Hollmann et al. (2025) made its model available through the `tabpfn` Python package, which encapsulates the code in a sklearn-compatible interface, allowing for its used with `RandomizedSearchCV` object. In this work, both the `TabPFNRegressor` and `TabPFNClassifier` estimators made available in the tabpfn package were tested for the regression and classification tasks studied. Scalers and PCA were also evaluated with the TabPFN model. It is important to note that the dataset used in this work has more samples than the original limit proposed by (Hollmann et al., 2025), so the performance obtained may not be as good as the one obtained for other datasets in the original work.

5.2.4 Hybrid model training

The hybrid model training followed the approaches described in section 2.2. The approach from equation 2-13 was applied for regression problems as predicting pressure gradient and gas fraction, while the approach from equation 2-11 was applied for classification problems. Mechanistic models were used as physics-based models to generate the $\hat{\mathbf{y}}$ physical outputs from the first-principles models and all three possible outputs considered in this work (GRAD, ALPHA, REGIME) were provided as inputs for the hybrid model, but as flow pattern is a categorical variable the one-hot encoding technique was applied to it to avoid ordinal association between flow regimes.

In this work, all the possible combinations of base first-principle model and possible regressor from the `scikit-learn` API listed in Table 5.2 were tested alone and combined with different scalers and PCA.

5.2.5 Evaluation metrics

In this work, the main error metrics used for comparing the performance of regression models were the coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE), with R^2 being used for both hyperparameter search and for comparison with literature results.

Regarding the classification models, the main metric that was used to compare and rank models is the accuracy score (ACC). As an alternative to the accuracy score, the inaccuracy score was also used when comparison demands a metric that reduces with performance increase.

5.2.6 Randomized search and cross-validation

For the proposed study, 50 random models were generated for each estimator, with different hyperparameters each. For the cross-validation, the number of folds (N) adopted was 3, and the number of different random splits performed (M) was 5, generating a total of 15 model evaluations for each model tested. To allow comparison of the final results with the ones obtained by Kanin et al. (2019), the same confidence interval methodology considering 95% confidence and described in section 2.4.2 was used.

5.2.7 Metrics for quantifying physical and data contributions (PD-score)

Suppose that three different models have been trained to address a specific problem: a purely physical model, called model P, a data-driven model with no physical component, called model D, and a hybrid model that combines both data-driven and physical approaches, called model H. The three models are evaluated using an error metric that decreases with the improvement of the model's performance, such as mean square error or one minus the accuracy of the classification model, giving rise to the error values ϵ_P, ϵ_D, and ϵ_H for the models. It is possible to propose a gain from adding physical modeling to the problem by quantifying the difference between the error metric of the purely data-driven model and the error metric of the best performing model that incorporated physical information (either the purely physical or the hybrid model), as proposed in equation 5-4.

$$GP = \epsilon_D - \min(\epsilon_P, \epsilon_H) \tag{5-4}$$

Similarly, equation 5-5 exemplifies how the same methodology can be applied to calculate the data gain using the physical model and the best performing model that considered a data-driven approach. It is important to notice that both physical and data gains can be negative values. A negative physics gain indicates that the purely data-driven approach outperformed both the physical and hybrid models, while a negative data gain indicates the best-performing model was the purely physical one.

$$GD = \epsilon_P - \min(\epsilon_D, \epsilon_H) \tag{5-5}$$

As error metrics may be dimensional, it is desirable to convert the proposed gains to adimensional scores by normalizing them using a reference value. The easiest approach would be to consider the sum of both gains as a normalizing value, ensuring that way that the sum of both P and D scores would be 1. This approach, however, is not advantageous for the studied case since the gains proposed by equations 5-4 and 5-5 may be negative, leading to the possibility of small, negative, or zero denominators. To eliminate this issue, this work proposes using as normalizing value the sum of the maximum value between each gain and zero or, in another interpretation, the sum of the non-negative gains. This allows the definition of the physical score (P-score) as defined in equation 5-6, and the data score (D-score) may be defined similarly by equation 5-7.

$$P = \frac{GP}{\max(GP, 0) + \max(GD, 0)} \tag{5-6}$$

$$D = \frac{GD}{\max(GP, 0) + \max(GD, 0)} \tag{5-7}$$

This definition of both P-score and D-score, presented in equations 5-6 and 5-7 respectively, allows some interpretations. In cases where the hybrid model is the best performing model, both scores will be positive and higher the closer the respective non-hybrid model is to the final hybrid model result. In cases where one of the gains is negative, the model with a positive gain will have a score of 1, while the model with a negative gain will have a negative score that gets closer to the value of -1, the closer the hybrid module error metrics get to the negative score model metric. For simplicity, in this work, the set of both physical and data scores proposed are referred to as the PD-score.

5.2.8 Shapley analysis

In this work, the Shapley values was used to validate the physical tendencies of the black-box model generated to predict the continuous variables (pressure gradient and gas fraction). In order to do so, both models were evaluated in terms of pressure gradients, with the gas volume fraction being used to predict the hydrostatic (or gravitational) pressure gradient according to equation 5-8.

$$\left.\frac{dP}{dL}\right|_{hydrostatic} = -(\alpha\rho_g + (1-\alpha)\rho_l) g \sin\theta \tag{5-8}$$

In addition to the Shapley values for the total gradient and the hydrostatic gradient, having both models combined allows for an evaluation of the combined effects of friction and acceleration in the pressure gradient. As it is expected that the acceleration contribution to be small, both terms were considered as part of a general friction gradient, defined by equation 5-9.

$$\left.\frac{dP}{dL}\right|_{friction} = \frac{dP}{dL} - \left.\frac{dP}{dL}\right|_{hydrostatic} \tag{5-9}$$

The `shap` Python package developed by (Lundberg & Lee, 2017) was applied to determine the Shapley values for the best models of each final test performed.

5.3 Mechanistic correlations evaluation and base results values

In order to obtain reference values for the mechanistic commercial model's performance, the four correlations assumed for this work were tested against the test dataset, and the error metrics obtained were documented. These results are important as they serve as the basis of comparison for the proposed methodology in this work. Table 5.3 lists the best value obtained for each metric for the three output variables among the four correlations tested.

Table 5.3: Best error metric value obtained for each output variable among the mechanistic commercial models evaluated. It is interesting to notice that even the best available commercial models may present relatively high errors

Variable	Metric	Best value
GRAD	MAE	348.0
	R_2	0.915
	RMSE	860.0
ALPHA	MAE	0.048
	R_2	0.843
	RMSE	0.099
REGIME	Accuracy	0.598

It is worth noting that, even considering commercial models widely used in the industry, the obtained results may present significant errors when compared to actual experimental data. The mean average error for the pressure gradient would be equivalent to an error of 1 bar for each 300 meters of pipeline considered, and the mean average error for the gas volume fraction would indicate a possible error of 5% in gas inventory estimations in a flowline. For flow regime determination, the accuracy is lower than 60%. While physical models may present accuracies close to 80% in predicting flow patterns, as shown in Pereyra et al. (2012) as mechanistic models are discontinuous in some phase transitions, it is expected that some relaxation is performed on the flow regime to better predict the pressure gradient and gas volume fraction, variables of greater practical use for production prediction in oil and gas.

5.4 Results and Discussions

In this section, the results obtained for the selected models documented in Table 5.2 following the approach presented in section 5.2. All results and error metrics presented were calculated considering only the test dataset, and relative performance gains have been calculated always comparing with the original mechanistic model's performance presented in section 5.3. For each model result presented, the best preprocessing strategy tested, consisting of the best scaling strategy and possible application of PCA is considered, and only the results for the best model according to the randomized hyperparameter search are registered.

First, the gains of the hybrid models when compared to both the base mechanistic models and black-box models built in this work were presented, followed by the quantification of the physics and data contribution for each possible combination of black-box model and mechanistic model. The results obtained were then compared with the ones obtained by Kanin et al. (2019), considering the confidence interval estimated using the cross-validation technique. Both the black-box and mechanistic models were ranked according to the RMSE and accuracy metrics.

At last, a brief discussion regarding incremental contribution results was made, presenting conclusions regarding the best preprocessing strategy obtained and how physically representative black-box models can be, addressing a common literature concern (Ma et al., 2024; Bikmukhametov & Jäschke, 2020).

Table 5.4 presents the performance increase in the error metrics for the hybrid models when compared with both pure data-driven models and the commercial mechanistic correlations results, presenting the results as a percentage of the mechanistic correlation errors presented in Table 5.3, considering RMSE as the most representative metric for continuous variables and inaccuracy for categorical variables.

Table 5.4 shows that, for the three variables evaluated, in most cases the hybrid model outperformed both the mechanistic benchmarks and the black-box models created as references. This is true for almost all the models tested, but the TunedTabPFN model in both the gas volume fraction estimation and the flow regime prediction, as the best hybrid model for gas fraction did not outperform the best data-driven model, and the best hybrid model for the flow regime prediction was not able to outperform the best mechanistic model evaluated.

Table 5.4 also shows that, for the continuous variables, relatively simpler models such as the SVM and Linear model tend to present greater gains of the hybrid model when compared to the data model than when compared to the physics-driven model, which may indicate that the hybrid strategy provides the final model knowledge of the nonlinearities and complexity of the problem that otherwise the base regressor would not be able to infer form the available data. This trend changes when more complex models, such as neural networks, are used, as these models can capture complex relations in the data, and the overall gain from the hybrid model over the data model tends to be lower than the gain over the physics-driven model. Figure 5.7 corroborates this observation by providing a visualization of the PD-score metric proposed in section 5.2.7, where it is possible to see that more robust black-box models tend to increase the data (D) score, while reducing the physics (P) score proportionally.

As Figure 5.7 shows, the PD-score for all the tested combinations of black-box models and base mechanistic models tested, it is also possible to conclude that the main factor determining the final PD-score is the selected

Table 5.4: Error metric reduction for the hybrid model strategy when compared with both the physical model (mechanistic) and the data-driven model using the same estimator. he reduction is presented as a percentage of the mechanistic error metric. Error metrics considered were RMSE for continuous outputs and inaccuracy for categorical outputs

Variable	Model	Error reduction as a percentage of mechanistic model error	
		Reduction over physical model	Reduction over data model
GRAD (dP/dL)	TunedTabPFN	**71**%	25%
	MLP	69%	29%
	GradientBoosting	56%	35%
	ExtraTrees	61%	25%
	RandomForest	56%	27%
	DecisionTree	52%	39%
	SVM	49%	125%
	Polynomial	61%	34%
	Linear	16%	**149**%
ALPHA (α)	TunedTabPFN	61%	-1%
	MLP	**51**%	9%
	GradientBoosting	48%	27%
	ExtraTrees	51%	41%
	RandomForest	48%	30%
	DecisionTree	46%	47%
	SVM	28%	35%
	Polynomial	38%	23%
	Linear	10%	**85**%
REGIME (ID)	TunedTabPFN	-8%	10%
	MLP	**83**%	4%
	GradientBoosting	80%	4%
	ExtraTrees	64%	4%
	RandomForest	70%	3%
	DecisionTree	64%	5%
	SVM	82%	**11**%
	Polynomial	79%	8%
	Linear	37%	9%

black-box model, with the physics-driven model having a secondary effect. More than that, Figure 5.7 shows that overall, the gradient models are the ones that benefit the most from the physical models and are the ones more sensitive to this input, which is to be expected since most models aim to better represent this variable and are fine-tuned for it. In contrast, the flow regime is the variable that most consistently relies on the data models, having high D-scores regardless of the correlation and the black-box model, behavior that may be attributed to the relaxations proposed to ensure continuity of the mechanistic models (Gomez et al., 2000).

Figure 5.8 shows the comparison between the results of this work and the results from Kanin et al. (2019), considering R^2 as the metric for continuous variables and accuracy for the categorical ones. The reference work only tested for possible estimators models that can be considered equivalent to

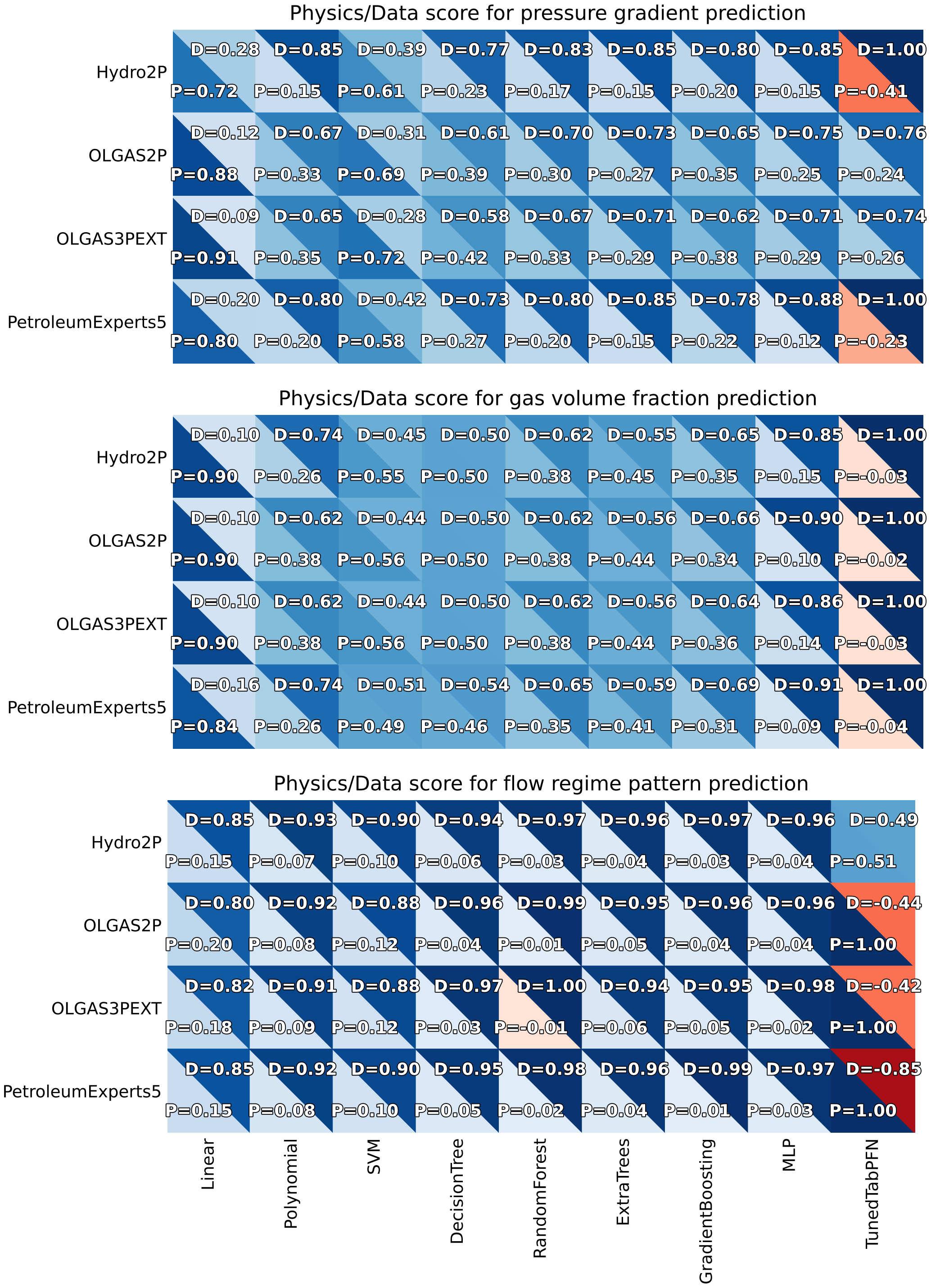


Figure 5.7: PD-score visualization for the three output variables evaluated. It is possible to see that, overall, for the best models high data contributions were observed. Physical contributions were relevant for the pressure gradient model, and were generally irrelevant for the flow regime pattern model

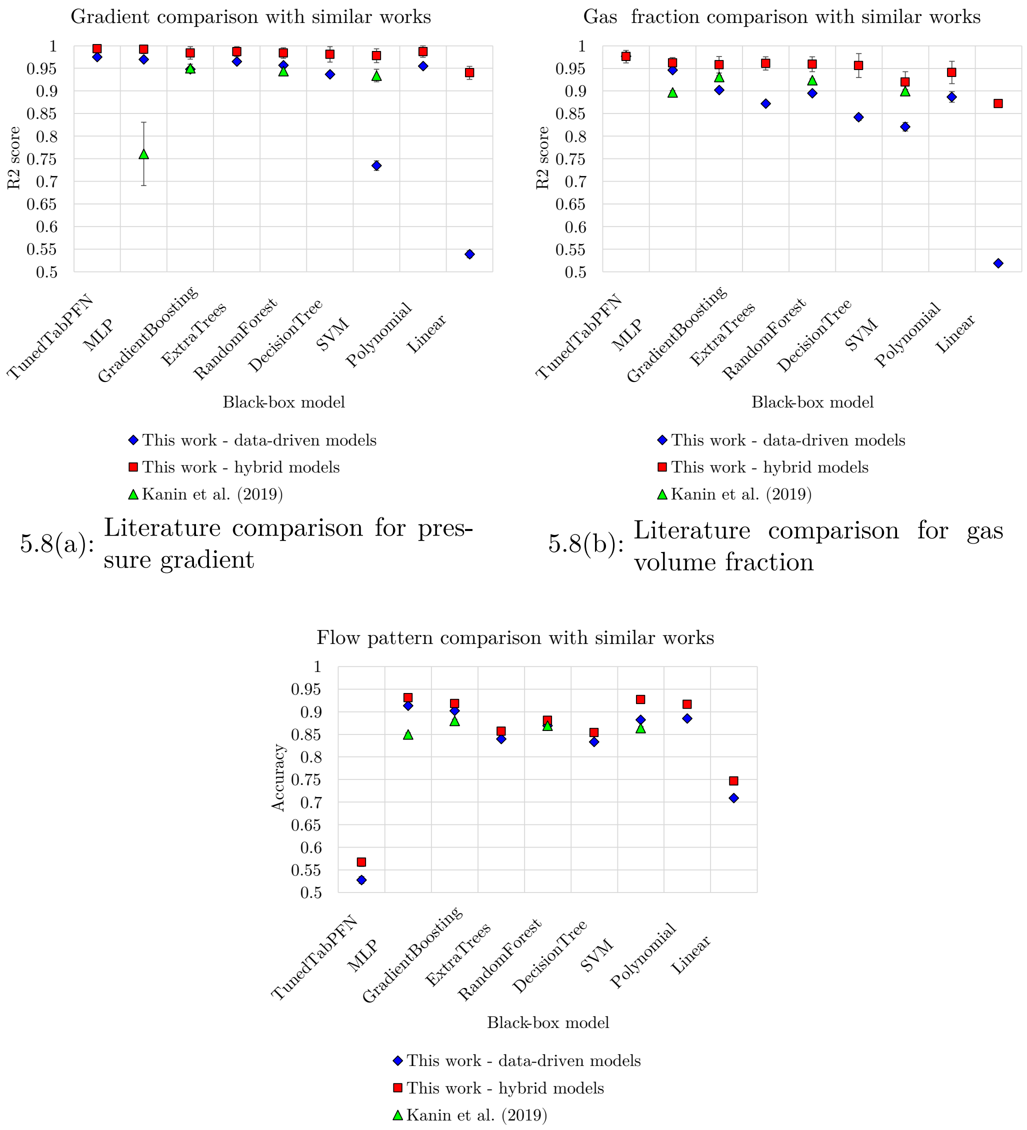
the MLP, GradientBoosting, RandomForest, and SVM models evaluated in this work. As both works had different input datasets and different feature engineering techniques, as Kanin et al. (2019) used adimensional numbers as inputs, the error bars are also presented being estimated by the same methodology presented in section 5.2.6, considering the standard deviation of the 3-fold, 5-iteration cross-validation performed during model training. While, for pure black-box models, there is no clear tendency of this work's models outperforming Kanin et al. (2019) results, with better results being obtained for specific models and output variables only, the hybrid models built in this work consistently outperformed both the data-driven models and the models presented by Kanin et al. (2019), proving the value of the proposed strategy for the multiphase flow area. By using the mean metric value, it is possible to see that the strategy improved literature results by 0.03-0.23 in R^2 for the pressure gradient, 0.02-0.06 in R^2 for the gas fraction, and 0.01-0.08 in accuracy for the flow regime.

It is also important to notice that this is true even considering that the error bars for the hybrid models were considerably larger than the error bars presented by Kanin et al. (2019), a phenomenon that is explained by the fact that the cross-validation performed in the reference work considered 100 repetitions, allowing for lower uncertainties for a similar standard deviation.

5.4.1 Analysis

The results already presented in this work show that the proposed hybrid strategy for hybrid models was able to outperform both simple black-box models and literature results. Considering the benchmark error for mechanistic models as a reference, hybrid models reduced the error when compared to purely physics-driven models by 16-71% for pressure gradient, 10-61% for gas fraction, and 37-83% for flow regime classification. Compared to purely black-box models, the reduction was of 25-149%, 9-85%, and 3-11% for the pressure gradient, gas fraction, and flow pattern prediction, respectively. This corroborates the claims of Bikmukhametov & Jäschke (2020) that adding physical information to machine learning models allows improvements in the final results when compared to pure data-driven models.

Only two models were exceptions to this trend, both using the TunedTabPFN regressor. The first one consists of the slight performance decrease in the hybrid model for gas volume fraction prediction, which is explained by the expected error bars presented in Figure 5.8. The low negative value for the P-score in Figure 5.7 also corroborates this argument as this



5.8(a): Literature comparison for pressure gradient

5.8(b): Literature comparison for gas volume fraction

5.8(c): Literature comparison for flow regime pattern

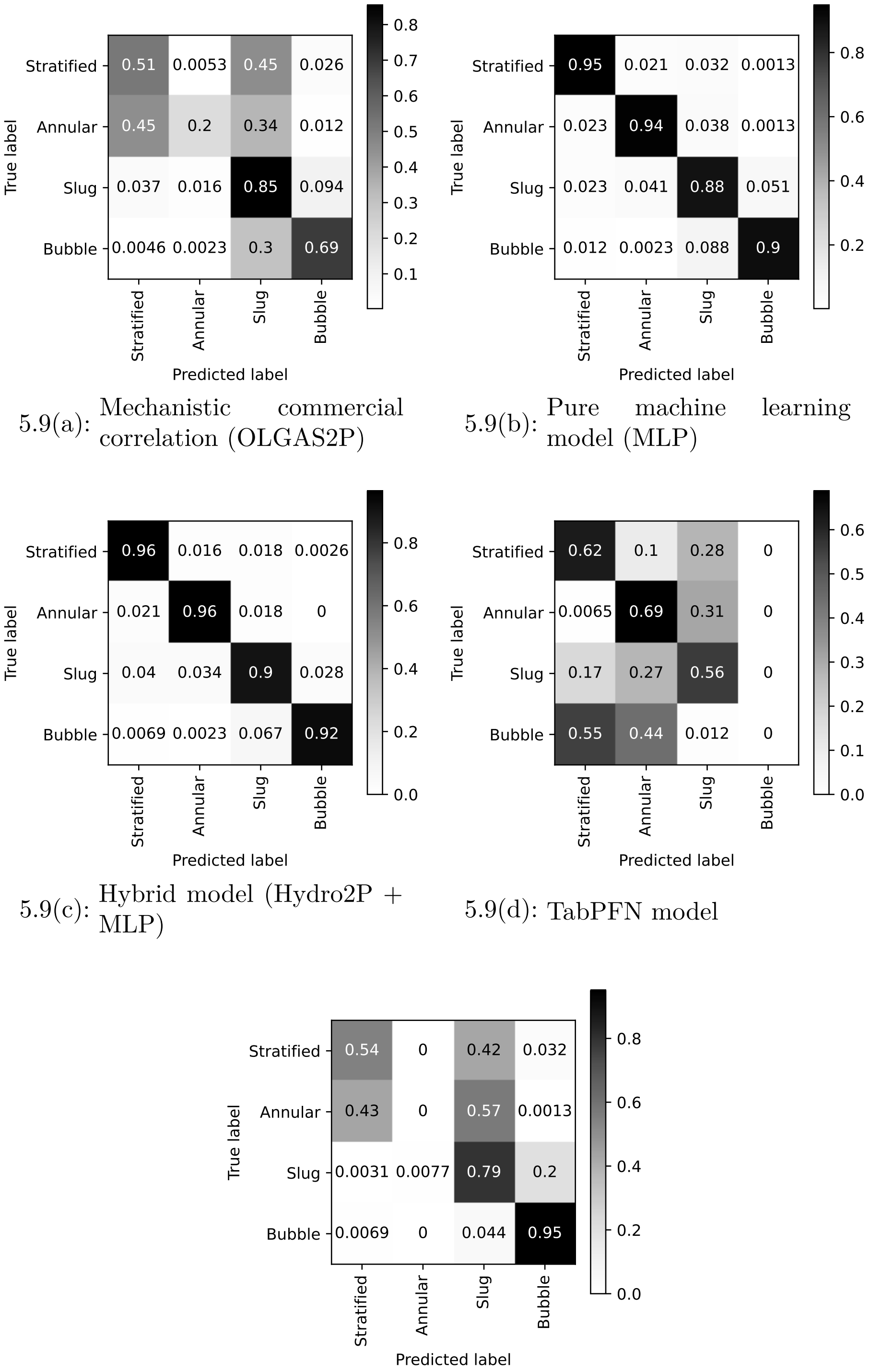
Figure 5.8: Comparison of the results obtained for the R^2 and accuracy metrics in this work with the ones presented by Kanin et al. (2019). Not all black-box models outperformed the reference results, but the hybrid models consistently showed better performance, even considering error bars obtained form the cross-validation performed in hyperparameter search

close-to-zero value may be interpreted as a model that relies mostly on the data part to perform the prediction.

The second model that did not follow the trend was the TunedTabPFN for the flow regime prediction, that was outperformed by the first-principles model. To better understand the reasons for this behavior, Figure 5.9 shows the confusion matrices for the best results for the best performing mechanistic, data-driven, and hybrid models, while also displaying the confusion matrices for the best TunedTabPFN results obtained as both black-box and hybrid model. It is possible to conclude by analyzing Figure 5.9 that, while both black-box and hybrid best models were able to represent the dataset with great accuracy for all flow patterns, the TunedTabPFN showed zero or near-zero accuracy for one flow pattern both in the black-box and hybrid cases. For the black-box case, the TunedTabPFN was unable to correctly identify the bubble pattern, which may be associated with the fact that this is the pattern with the lowest number of samples in the dataset, as commented in section 5.1. In hybrid models, the annular flow pattern was the one not correctly identified by the TunedTabPFN model, with this behavior being associated with the fact that the mechanistic correlations used as base models for the hybrid strategy show low accuracies for this specific flow pattern, as shown in Figure 5.9(a).

Figure 5.9(a) also makes evident that the greatest inaccuracy in flow pattern predictions rests on the annular pattern detection. Even though Figure 5.9(a) shows the confusion matrix for only the best performing mechanistic model (OLGAS2P), the same behavior and error profile may be seen in the other tested models. Pereyra et al. (2012) in his work dedicated to evaluating the accuracy of first-principles flow pattern identification models also found a relatively low accuracy in annular pattern identification when compared to other flow patterns, but the purely physical models reached an accuracy around 70%, surpassing the commercial models' 20% accuracy obtained in this work. This may be attributed to the pressure gradient discontinuity present in mechanistic models due to the transition between annular and slug flow regimes, where there is also a tendency of near-zero thickness liquid films for slug flow near the annular transition, an issue that may cause numerical instabilities and is dealt through some relaxation criteria (Gomez et al., 2000). Commercial models also tend to use specific formulations to improve their results and ensure continuity of the pressure gradient, as adopting minimum slip velocity criteria for flow pattern determination (Bendlksen et al., 1991).

To complement Table 5.4, Tables 5.5, 5.6, and 5.7 present the ranking of both black-box and hybrid models for pressure gradient, gas fraction, and flow regime pattern, respectively. The ranking was performed considering the



5.9(a): Mechanistic commercial correlation (OLGAS2P)

5.9(b): Pure machine learning model (MLP)

5.9(c): Hybrid model (Hydro2P + MLP)

5.9(d): TabPFN model

5.9(e): TabPFN hybrid model (OLGAS2P correlation)

Figure 5.9: Confusion matrix for flow pattern comparison for best benchmark cases. One notices that the first-principles correlations present difficulties in discerning stratified from slug flow and identifying annular flow, while the TabPFN model was unable to identify the bubble pattern as a pure classifier and the annular pattern as an ensemble, with general accuracy below 60% in both cases

RMSE metric for regressors and the accuracy score for the classifier, and only the best results considering the tested preprocessing strategies and, for the hybrid models, the best base mechanistic correlation. Tables 5.5 and 5.6 show that, according to the ranking proposed, the black-box model used plays a key role in the final results both when applying hybrid strategies and building a pure black-box model. The rankings for both pressure gradient and gas fraction models follow a trend of deep learning and neural network models outperforming tree-based and polynomial models, and these models outperforming SVM and linear models.

Tables 5.5 and 5.6 make evident another benefit of the hybrid model strategy adopted in this work for low-performance black-box models: while in the black-box case some of the lowest performance models presented RMSE increases when compared to the mechanistic benchmark, all the hybrid models evaluated presented performance increase when comparing to the benchmark performance, including the Linear and SVM models that had the overall lowest regression performance. This is not necessarily true for the other metrics as, for the best model selection and ranking, only the RMSE was considered, so some of the lowest ranking models kept showing slightly lower performances than the benchmark, considering the MAE metric.

Tables 5.5 and 5.6 also show that the TunedTabPFN model for these cases was the one that showed the best performance, confirming the claims made in Hollmann et al. (2025) that the model can outperform classical machine learning algorithms for tabular data.

Table 5.5: Hybrid model ranking for pressure gradient, considering the RMSE as the ranking variable. This ranking considers both black-box and hybrid models and only the results obtained with the best preprocessing strategy adopted and the best base correlation for each estimator. In parentheses is presented the difference between the error metric and the best first-principles model result

Rank	Correlation	Model	Preprocessor	MAE	R^2	RMSE
1	N/A	TunedTabPFN	PCA, Standard	82.3 (-76%)	0.975 (+0.060)	465.2 (-46%)
2	N/A	MLP	PCA, Standard	190.8 (-45%)	0.970 (+0.055)	514.2 (-40%)
3	N/A	ExtraTrees	PCA, Standard	289.8 (-17%)	0.965 (+0.050)	549.8 (-36%)
4	N/A	RandomForest	PCA, Standard	314.3 (-10%)	0.957 (+0.042)	610.4 (-29%)
5	N/A	Polynomial	PCA, Standard	272.7 (-22%)	0.955 (+0.040)	628.4 (-27%)
6	N/A	GradientBoosting	PCA, Standard	325.9 (-6%)	0.948 (+0.033)	673.6 (-22%)
7	N/A	DecisionTree	PCA, Standard	310.3 (-11%)	0.937 (+0.022)	742.7 (-14%)
8	N/A	SVM	PCA, Standard	963.5 (+177%)	0.735 (-0.180)	1520.9 (+77%)
9	N/A	Linear	PCA, Standard	1103.2 (+217%)	0.539 (-0.376)	2006.1 (+133%)
1	OLGAS3PEXT	TunedTabPFN	None, Standard	55.6 (-84%)	0.993 (+0.078)	246.2 (-71%)
2	OLGAS3PEXT	MLP	PCA, Standard	101.1 (-71%)	0.992 (+0.077)	268.3 (-69%)
3	OLGAS2P	ExtraTrees	PCA, Standard	138.4 (-60%)	0.987 (+0.072)	337.6 (-61%)
4	OLGAS2P	Polynomial	PCA, Standard	169.8 (-51%)	0.987 (+0.072)	338.7 (-61%)
5	OLGAS3PEXT	RandomForest	PCA, Standard	170.4 (-51%)	0.984 (+0.069)	374.2 (-56%)
6	OLGAS3PEXT	GradientBoosting	PCA, Standard	172.6 (-50%)	0.984 (+0.069)	374.5 (-56%)
7	PetroleumExperts5	DecisionTree	PCA, Standard	168.5 (-52%)	0.981 (+0.066)	408.8 (-52%)
8	OLGAS2P	SVM	PCA, Standard	203.6 (-41%)	0.978 (+0.063)	442.7 (-49%)
9	OLGAS3PEXT	Linear	PCA, Standard	392.5 (+13%)	0.940 (+0.025)	726.5 (-16%)

Table 5.6: Model ranking for gas fraction prediction, considering the RMSE as the ranking variable. This ranking considers both black-box and hybrid models and only the results obtained with the best preprocessing strategy adopted and the best base correlation for each estimator. In parentheses is presented the difference between the error metric and the best first-principles model result

Rank	Correlation	Model	Preprocessor	MAE	R^2	RMSE
1	N/A	TunedTabPFN	PCA, Standard	0.014 (-0.034)	0.977 (+0.134)	0.038 (-0.061)
2	N/A	MLP	PCA, Standard	0.033 (-0.015)	0.946 (+0.103)	0.058 (-0.041)
3	N/A	GradientBoosting	PCA, Standard	0.040 (-0.008)	0.902 (+0.059)	0.078 (-0.021)
4	N/A	RandomForest	PCA, Standard	0.049 (+0.001)	0.895 (+0.052)	0.081 (-0.018)
5	N/A	Polynomial	PCA, Standard	0.044 (-0.004)	0.887 (+0.044)	0.084 (-0.015)
6	N/A	ExtraTrees	PCA, Standard	0.056 (+0.008)	0.872 (+0.029)	0.090 (-0.009)
7	N/A	DecisionTree	PCA, Standard	0.060 (+0.012)	0.842 (-0.001)	0.100 (+0.001)
8	N/A	SVM	PCA, Standard	0.084 (+0.036)	0.821 (-0.022)	0.106 (+0.007)
9	N/A	Linear	PCA, Standard	0.127 (+0.079)	0.519 (-0.324)	0.173 (+0.074)
1	OLGAS2P	TunedTabPFN	None, Standard	0.013 (-0.035)	0.976 (+0.133)	0.039 (-0.060)
2	Hydro2P	MLP	PCA, Standard	0.025 (-0.023)	0.962 (+0.119)	0.049 (-0.050)
3	Hydro2P	ExtraTrees	PCA, Standard	0.029 (-0.019)	0.961 (+0.118)	0.049 (-0.050)
4	Hydro2P	RandomForest	PCA, Standard	0.027 (-0.021)	0.959 (+0.116)	0.051 (-0.048)
5	OLGAS3PEXT	GradientBoosting	PCA, Standard	0.027 (-0.021)	0.958 (+0.115)	0.051 (-0.048)
6	OLGAS3PEXT	DecisionTree	PCA, Standard	0.027 (-0.021)	0.956 (+0.113)	0.053 (-0.046)
7	OLGAS3PEXT	Polynomial	PCA, Standard	0.034 (-0.014)	0.941 (+0.098)	0.061 (-0.038)
8	Hydro2P	SVM	PCA, Standard	0.054 (+0.006)	0.920 (+0.077)	0.071 (-0.028)
9	OLGAS2P	Linear	PCA, Standard	0.052 (+0.004)	0.872 (+0.029)	0.089 (-0.010)

For the classification task, it is possible to see in Table 5.7 that the black-box model selection is still a relevant factor for both the hybrid and data-driven model performance as also seen in Tables 5.5 and 5.6, but the model ranking presented changes as the SVM model showed, for the classification task, performance equivalent or superior to tree-based and polynomial models, and the TunedTabPFN that outperformed all the other models for the regression tasks was the model with the lowest performance both as black-box and hybrid model, in line with the fact that it was usually able to only detect three of the four regimes considered in this work, as exposed in Figure 5.9. This indicates that for future improvements of the TunedTabPFN model, the problem of flow pattern classification should be treated as a problem of interest. It is also important to note that a possible reason for this low performance may rest on the fact that the TabPFN model was only tested in datasets with 10000 samples or fewer, with the currently applied dataset having more samples to be considered.

5.4.2 Incremental contributions

In this section, additional topics related to the preprocessing strategies studied and physical representativeness of black-box models for extrapolation purposes are presented.

Regarding the preprocessing strategies adopted, Tables 5.5, 5.6, and 5.7 make evident that the application of StandardScaler combined with PCA was the most successful strategy adopted, being the one selected for the best performing model in all the estimators evaluated. The only exception to this behavior is the TunedTabPFN models in hybrid models that had only the StandardScaler as preprocessing strategy, which is explained by the fact that these models did not consider PCA as an option during training as hardware limitations did not allow for it in the training phase, but having StandardScaler for these models as the most still evidences that this method of data scaling is the most effective one for the dataset studied.

Figure 5.10 provides insights into why the StandardScaler showed better performance than other scalers tested in this work by comparing the explained variance distribution across the principal components of a PCA analysis for the considered dataset, comparing the effects of the scaling strategy in said distribution. It is possible to conclude in Figure 5.10 that the StandardScaler is the one that ensures a more homogeneous distribution of explained variance in all the principal components. The second best scaling strategy in that aspect is the MinMaxScaler as it also ensures explained variance distribution

Table 5.7: Model ranking for flow pattern prediction, ranked by accuracy. This ranking considers both black-box and hybrid models and only the results obtained with the best preprocessing strategy adopted and the best base correlation for each estimator. In parentheses is presented the difference between the error metric and the best first-principles model result

Rank	Correlation	Model	Preprocessor	Accuracy
1	N/A	MLP	PCA, Standard	0.913 (+0.315)
2	N/A	GradientBoosting	PCA, Standard	0.902 (+0.304)
3	N/A	Polynomial	PCA, Standard	0.885 (+0.287)
4	N/A	SVM	PCA, Standard	0.882 (+0.284)
5	N/A	RandomForest	PCA, Standard	0.869 (+0.271)
6	N/A	ExtraTrees	PCA, Standard	0.840 (+0.242)
7	N/A	DecisionTree	PCA, Standard	0.833 (+0.235)
8	N/A	Linear	PCA, Standard	0.709 (+0.111)
9	N/A	TunedTabPFN	PCA, Standard	0.528 (-0.070)
1	Hydro2P	MLP	PCA, Standard	0.931 (+0.333)
2	OLGAS2P	SVM	PCA, Standard	0.927 (+0.329)
3	OLGAS3PEXT	GradientBoosting	PCA, Standard	0.918 (+0.320)
4	OLGAS3PEXT	Polynomial	PCA, Standard	0.916 (+0.318)
5	Hydro2P	RandomForest	PCA, Standard	0.881 (+0.283)
6	OLGAS3PEXT	ExtraTrees	PCA, Standard	0.857 (+0.259)
7	Hydro2P	DecisionTree	PCA, Standard	0.854 (+0.256)
8	Hydro2P	Linear	PCA, Standard	0.747 (+0.149)
9	OLGAS2P	TunedTabPFN	None, Standard	0.567 (-0.031)

across all the features, but still concentrates most of the explained variation in the first half of the principal components. Other strategies, such as the RobustScaler and no scaling of the data, concentrate all the explained variance in a single feature, indicating poor capabilities of capturing data importance and variation.

Last, to address the concerns of many authors regarding how physically representative and adequate for data extrapolation data-driven models are (Ma et al., 2024; Bikmukhametov & Jäschke, 2020), this work proposes, as an additional contribution, an analysis of two of the black-box models created to determine how physically representative they are. For this task, the MLP and Polynomial models were chosen as examples of the black-box models presented in this work.

The first representative comparison consists of a visual comparison of flow regime maps generated by both models, presented in Figure 5.11 for vertical ascending, horizontal, and vertical descending flow. Properties besides superficial velocities and inclination angle were considered as the average value from the dataset. It is possible to see in Figure 5.11 that the MLP model was able to generate physically consistent regime maps when compared to unified mechanistic models such as the Barnea model (Barnea, 1987). Evidences of this consist of the overall regime boundaries detected and in the fact that a

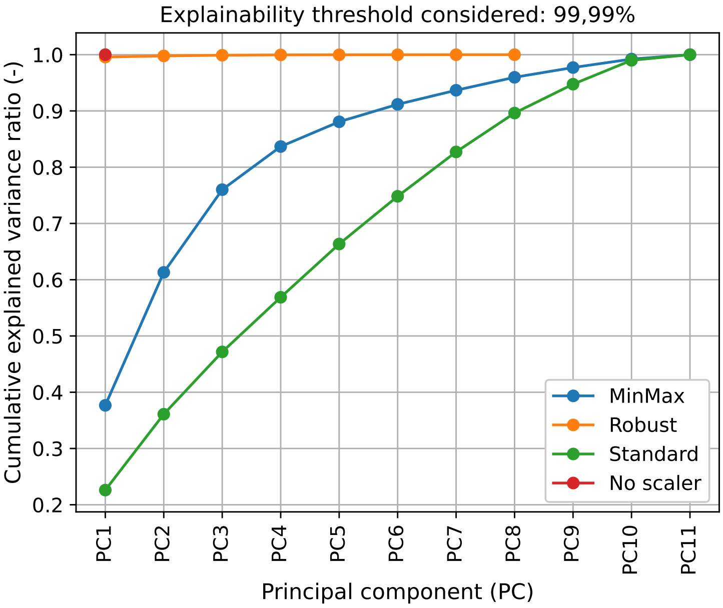


Figure 5.10: Cumulative explained variance as a function of principal components for different scaling methods in the dataset. One may notice that the StandardScaler better allows a more even total variance distribution between principal components, while other scalers concentrate most of the variance on the first ones

stratified flow pattern was not detected in vertical flow, which is expected. On the other hand, the results obtained for the Polynomial model do not seem to bear any physical resemblance with the ones from the Barnea model (Barnea, 1987), indicating that while it showed significant good performance, it may not be an adequate model for extrapolations.

Another interesting aspect to comment on in Figure 5.11 is about the main differences between the MLP flow regime map and the equivalent ones presented by Barnea (1987). First, for ascending vertical flow, in the high superficial velocities area, there is a considerably irregular boundary between annular and slug flow. While this might be understood as an unphysical behavior, the Barnea model (Barnea, 1987) attributes this region to a flow regime not considered in this work (churn) that may be understood as a chaotic transition between annular and slug flows that appear when gas velocity makes the slugs unstable. As this flow pattern was not considered, the observed behavior may simply be interpreted as the classifier trying to fit a complex transition zone to the best of the knowledge provided by the data. Another significant difference is in the descending vertical flow for the low superficial velocities. The Barnea model (Barnea, 1987) predicts that this region should be of annular flow, while the MLP model predicted a dispersed bubble regime. This difference may be associated mostly to the lack of data in this region

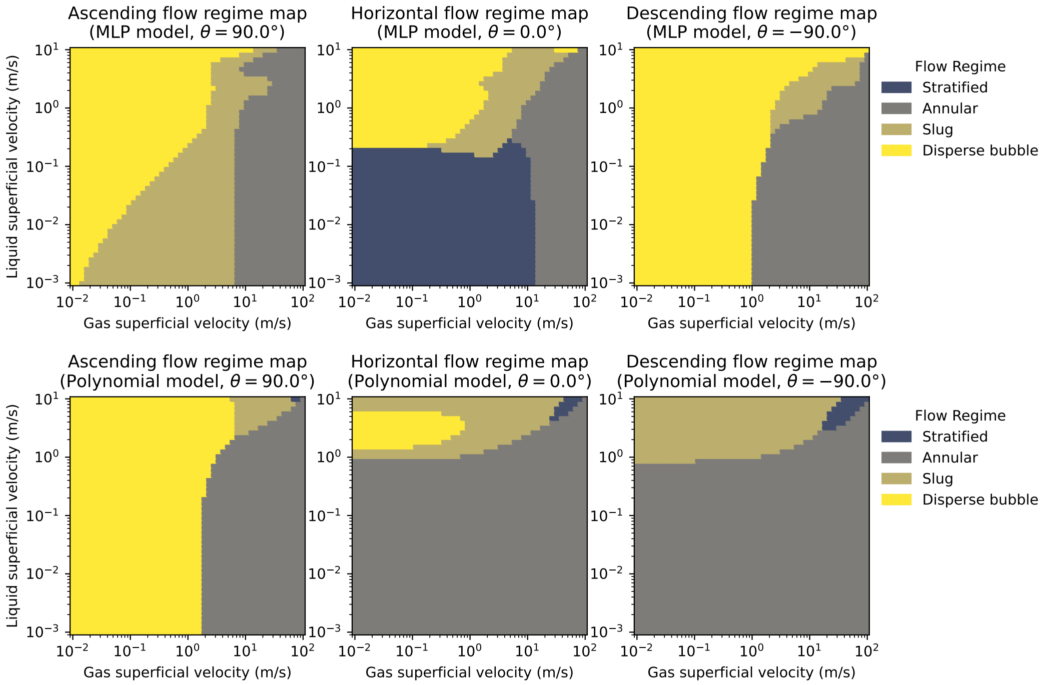


Figure 5.11: Flow regime map plots generated by a set of black-box models evaluated in this study. It is interesting to notice that some models were able to capture physical behaviors described by first-principles models, such as the general model proposed by Barnea (Barnea, 1987)

as presented in Figure 5.5 in section 5.1, which would make difficult for the classifier to correctly learn the flow regime in this region, and providing more experimental data to cover this gap would probably help in reducing model differences, especially since other literature works show that downhill flow overall tend to present greater errors in black-box models when compared to horizontal and uphill flows (Kanin et al., 2019).

To evaluate the physical representativeness of the continuous variables (pressure gradient and gas fraction), the Shapley analysis technique described in section 5.2.8 was applied for the total, hydrostatic and frictional gradient, calculated using the models trained and equations 5-8 and 5-9. Both MLP and Polynomial models were submitted to this analysis, and 5.12 shows the obtained results for the Shapley values for gas superficial velocity in ascending flow, with the value associated with each coordinate being the average Shapley value obtained for all the observations in the dataset contained in that region. The vertical ascending flow was selected as it is expected to show opposite trends in hydrostatic and frictional gradients, with the frictional gradient reducing for greater gas superficial velocities, indicating greater pressure drops, and the hydrostatic gradient increasing for greater gas velocities due to the expected increase in gas fraction. Figure 5.12 shows that the MLP model

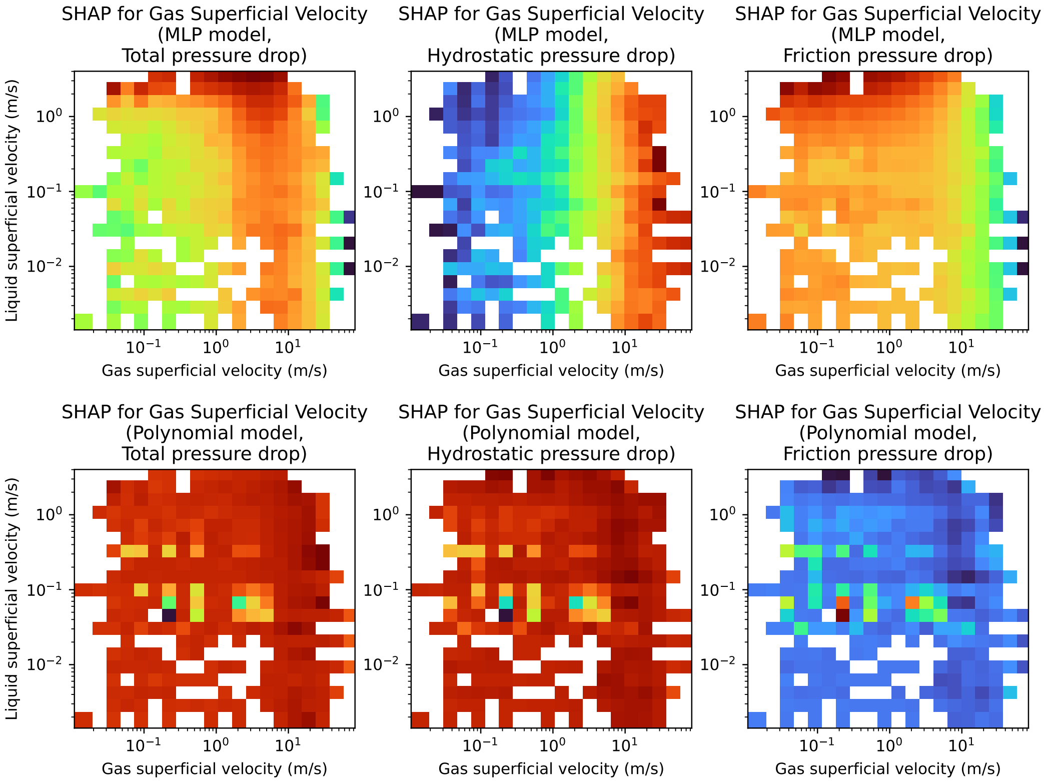


Figure 5.12: Shapley analysis for ascending flow in the total dataset for the gas superficial velocities for the prediction of total pressure gradient, hydrostatic gradient and friction gradient. The color red indicates more positive Shapley values, which, for pressure gradient prediction, indicates a pressure increase

displays that the MLP model has the expected physical tendency, while also presenting a maximum Shapley value region in the total gradient that represents the point where the hydrostatic effects are replaced by the friction effects as the most relevant ones. Meanwhile, the Polynomial model does not capture this physical trend, once again indicating less physical behavior and extrapolation capabilities.

Figures 5.11 and 5.12 address the literature concerns regarding if black-box models can capture the physical behavior in the data (Ma et al., 2024; Bikmukhametov & Jäschke, 2020) by comparing the MLP and Polynomial models' responses with what is expected from the multiphase flow phenomenon. Both models obtained good results according to the error metrics evaluated. The overall result indicates that it is possible to obtain purely data-driven models that correctly capture the physical trends, but at the same time validates the concern by showing that while the MLP was able to correctly capture expected trends, the Polynomial model showed less physical responses regardless of the overall metric results obtained.

5.5 Partial Conclusions

Reliable estimation of multiphase flow parameters is of crucial importance in the oil and gas industry, both in simulation and virtual flow metering tasks. Advancements both in physical models and data-driven models have been made in the last years, but hybrid models to predict multiphase flow behavior are still a new research field that presents many gaps to be explored. This work tries to address some of these gaps by proposing a hybrid model approach obtained combining a black-box estimator with a commercial mechanistic model to improve final model's performance and compare these approach with both mechanistic correlation benchmarks and purely data-driven approaches for the three most relevant variables generated as outputs form mechanistic models: pressure gradient, gas fraction and flow pattern. For this task, an experimental dataset with over 22,000 samples was used.

Results in this work showed that the hybrid model strategy proposed outperformed both the data-driven and the pure mechanistic approaches in all cases tested. Comparison with literature works presented by Kanin et al. (2019) also proved the effectiveness of the strategy, even considering error bars estimated by means of cross-validation. The application of the proposed PD-score metric to quantify the contributions of the physical and data-driven terms of the final model also proved to be useful in demonstrating that not only the gains observed were true mostly regardless of the mechanistic model used as a base model but confirming that commercial mechanistic models, normally optimized for pressure gradient prediction, allowed for greater physical contributions in the pressure gradient models than for the other variables.

Another important result obtained was showing that the final model ranking was mostly associated with the selection of an adequate black-box model for both the data-driven and hybrid cases, with the hybrid strategy having little impact on the ordering of the estimators. This work also showed that neural networks were the models that presented the best results overall, with the TabPFN model presenting the best results for the regression variables, but failing as an adequate estimator for flow pattern classification by not being able to correctly identify all the flow patterns considered in this work.

Regarding the preprocessing strategies tested in this work, applying the StandardScaler was the most successful strategy as it allowed for a more even distribution of the explained variance among the variables. Finally, one last contribution of this work was to address concerns regarding the physical representativeness of black-box models. This was made by both inspecting the final model results and applying explainability techniques like Shapley values to determine feature impact in the final response. Results showed that while black-box models may be able to physically represent the multiphase flow phenomenon, this representativeness is a concern that should be taken into account while selecting the final model, as models with good error metric results may fail in extrapolations.

For future works, there is considerable margin for improving the results obtained with TabPFN, especially in flow pattern prediction. Many of the difficulties found in this work were also associated with imbalances in the flow pattern samples or the lack of data for some specific conditions, so trying oversampling techniques to overcome imbalance issues and develop more experimental data to fill the gaps, especially in descending flow and annular flow pattern, is also a possible future work. The application of feature engineering techniques in the inputs, especially by using literature adimensional numbers and the application of more complex neural network architectures, capable of dealing with multiple outputs simultaneously, is also a possible future work.

6 Conclusion

In the presented work three, case studies of interest of the oil and gas industry were addressed with the application of classical machine learning models, proving that these methods are still applicable and fit for a multitude of engineering and businesses issues. These methods enabled addressing problems not yet treated, like the direct early assessment of exploratory assets using a direct economic classification approach and even allowed for satisfactory time series regressions and extrapolations when paired with the system identification approach. These models also enabled the creation of simple hybrid models, capable of taking physics into account to improve state-of-the -art engineering models while taking into account physical trends.

Regarding the tools already available, as this work used mostly of open-source packages for the data-driven tasks, it proves the already existing tools enable tackling many different engineering tasks in a very straightforward fashion. The existing standard package for classical machine learning `scikit-learn` allows for an easy inclusion of data-driven regressors and classifiers in engineering workflows while also having many different functionalities for model selection and evaluation. Beyond the model, packages for dealing with imbalanced data, something usual in engineering tasks, and packages for inferring the feature contribution for the final model response, of extreme relevance for ensuring physical adherence of the model's answers, are also available and can be used for engineers that need to create data-driven models.

Finally, considering the methods evaluated, it is possible to conclude that neural networks and tree-based ensembles are usually the most versatile and powerful classical machine learning methods regardless of the problem studied, as these methods claimed great performances in all the case studies performed. This work, however, also made clear that the best performance should not only be judged by error metrics, as exemplified by both the economic classification with investments and the SHAP analysis performed on the pressure gradients for the multiphase flow models.

Specifically, we can draw the following conclusions regarding each of the contributions:

1. **System Identification Techniques for Soft Sensors and Multi-**

phase Flow Metering

- Models using multiple input (MISO) configurations achieved performances up to 15% in R^2 better than SISO cases.
- Complex regressors such as tree-based ensembles and neural networks outperformed classical ARX and NARX up to 62% inR^2 for the pure SISO case.
- Time series decomposition using Short-Time Fourier Transform (STFT) improved results when combined with the KNN regressor, while seasonal decomposition was ineffective for this problem.
- Incorporating current time data into the system identification model significantly improved prediction accuracy up to 15% in R^2 and narrowed the performance gap between SISO and MISO models.
- Although beneficial for accuracy, the use of current time data may reduce the general utility of the model and should be applied cautiously.

2. **An Interpretable Data-Driven Framework for Economic Assessment of Oil and Gas Exploratory Assets**

- Feature expansion techniques improved performance only for simple models like logistic regression, while PCA generally reduced model performance.
- Input spaces containing investment information significantly outperformed more restricted input configurations.
- Tree-based ensemble models and neural networks achieved the best overall performance, consistent with literature findings for similar tasks.
- Oversampling techniques improved balanced accuracy for models without investment inputs up to 8%, but led to a 18% recall reduction, potentially causing economically viable assets to be misclassified.
- Some level of dataset imbalance may be beneficial to prioritize economically viable assets in classification tasks.
- Including investment variables improved both balanced accuracy up to 34% and recall up to 5%, but SHAP analysis revealed unexpected positive contributions from cost variables, suggesting possible overfitting.
- Models excluding cost variables showed more interpretable and economically consistent behavior, aligning better with the problem's formulation.

3. **Hybrid machine learning models for improving state-of-the-art mechanistic flow models**

   - The herein proposed hybrid model outperformed both purely data-driven and mechanistic approaches across all tested variables: pressure gradient, gas fraction, and flow pattern.
   - Error reductions with the hybrid strategy ranged upt to 149% of the mechanistic correlation error when compared to data-driven models and 83% when compared to first-principles models.
   - Comparison with literature benchmarks confirmed the effectiveness of the hybrid strategy, even under cross-validation error analysis.
   - The PD-score metric successfully quantified the contributions of physical and data-driven components, showing that mechanistic models contributed more significantly to pressure gradient predictions.
   - Model performance ranking was primarily influenced by the choice of black-box estimator, with the hybrid strategy having limited impact on the ordering of models.
   - Neural networks achieved the best overall results, with the TabPFN model excelling in regression tasks but underperforming in flow pattern classification.
   - Among preprocessing techniques, StandardScaler yielded the most balanced distribution of explained variance across variables.
   - Concerns regarding the physical representativeness of black-box models were addressed using explainability tools like Shapley values, highlighting that good error metrics do not guarantee reliable extrapolation.

For future works, thanks to the continuous advancements made in the machine learning area, especially with currently available deep learning tools, there is significant space to further develop the presented applications through more problem-specific models, making use of more complex deep learning architectures that may provide better support for different physical problems, such as time series regressions.

Both the economic assessment and the multiphase flow modeling case studies made clear that incorporating engineering studies outputs-like investments or first-principles models outputs-as model inputs allows for performance increases. This opens a path for future works focusing on incorporating metrics and first-principle models in the formulation of more robust data-driven models, like physics informed neural networks (PINN) or more complex hybrid approaches.

Finally, as machine learning models always need reliable experimental data, it is possible still to contribute to literature by performing experimental studies to address dataset gaps. An example seen in this work is the lack of data for multiphase descending flow for low superficial velocities. More diverse data is also mandatory for creating robust foundation models as the TabPFN studied in this work, allowing for more general models with extrapolation powers that surpass what can be obtained with a single dataset.

For the contributions presented in this work, possible specific future works may be summarized as it follows:

1. **System Identification Techniques for Soft Sensors and Multiphase Flow Metering**

   – Deeply investigate the impact of current-time data on prediction accuracy, particularly for SISO models.
   – Test the proposed modeling approach on additional wells or similar datasets to assess generalizability.
   – Include multi-output prediction scenarios, with emphasis on Single Input Multiple Output (SIMO) configurations.

2. **An Interpretable Data-Driven Framework for Economic Assessment of Oil and Gas Exploratory Assets**

   – Develop models less dependent on predefined economic premises by creating regression models for internal return rate (IRR) instead of economic attractiveness classifiers.
   – Integrate investment estimation models with market-based or literature-derived methodologies to generate investment data from more fundamental sources.
   – Test the proposed methodology across varied training datasets to assess robustness and generalizability.
   – Create artificial datasets using existing metrics and current assessment techniques to expand the scope of model validation.
   – Apply the proposed methodology to other business segments beyond oil and gas, leveraging its simplicity and adaptability.

3. **Hybrid machine learning models for improving state-of-the-art mechanistic flow models**

   – Improve the performance of TabPFN models, particularly in flow pattern prediction tasks.

- Apply oversampling techniques to address class imbalance issues in flow pattern datasets.
- Develop additional experimental data to fill gaps in underrepresented conditions, especially for descending flow and annular flow patterns.
- Explore feature engineering strategies, including the use of dimensionless numbers from the literature to enrich model inputs.
- Investigate the use of more advanced neural network architectures capable of handling multi-output prediction tasks.

References

AL-DOGAIL, A.; GAJBHIYE, R.; ALNAJIM, A. ; ALNASER, M. **Dimensionless artificial intelligence-based model for multiphase flow pattern recognition in horizontal pipe**. SPE Production & Operations, 37:244–262, 5 2022.

AL-QUTAMI, T. A.; IBRAHIM, R.; ISMAIL, I. ; ISHAK, M. A. **Virtual multiphase flow metering using diverse neural network ensemble and adaptive simulated annealing**. Expert Systems with Applications, 93:72–85, 2018.

ABDUL-MAJEED, G. H.; KADHIM, F.; ALMAHDAWI, F. H.; AL-DUNAINAWI, Y.; ARABI, A. ; AL-AZZAWI, W. K. **Application of artificial neural network to predict slug liquid holdup**. International Journal of Multiphase Flow, 150:104004, 5 2022.

ALSAIF, A.; AL-SARKHI, A.; ISMAILA, K. ; ABDULKADIR, M. **Road map to develop an artificial neural network to predict two-phase flow regime in inclined pipes**. Journal of Petroleum Science and Engineering, 217, 10 2022.

ALAKEELY, A.; HORNE, R. **Simulating oil and water production in reservoirs with generative deep learning**. SPE Reservoir Evaluation & Engineering, 25(04):751–773, 2022.

ALAKEELY, A. A.; HORNE, R. N. **Application of deep learning methods to estimate multiphase flow rate in producing wells using surface measurements**. Journal of Petroleum Science and Engineering, 205:108936, 10 2021.

ALAKOUM, D.; GHORAYEB, K. **Machine learning-based multiphase pipeline network modeling**. Engineering Applications of Artificial Intelligence, 158:111517, 10 2025.

ALSARKHI, A.; SARICA, C. ; PEREYRA, E. **Novel correlations for the liquid holdup in a gas-liquid slug flow**. Geoenergy Science and Engineering, 237:212825, 6 2024.

ALSUMAIEI, A. A. **Hybrid residual modeling of pan evaporation in hyper-arid climates: Benchmarking interpretable neural architectures against physical drivers**. Journal of Hydrology: Regional Studies, 60:102572, 8 2025.

ANSARI, A. M.; SYLVESTER, N. D.; SARICA, C.; SHOHAM, O. ; BRILL, J. P. **A comprehensive mechanistic model for upward two-phase flow in wellbores**. SPE Production & Facilities, 9(02):143–151, 1994.

BAHRAMI, S.; ALAMDARI, S.; FARAJMASHAEI, M.; BEHBAHANI, M.; JAMSHIDI, S. ; BAHRAMI, B. **Application of artificial neural network to multiphase flow metering: A review**. Flow Measurement and Instrumentation, 97:102601, 7 2024.

BARNEA, D. **A unified model for predicting flow-pattern transitions for the whole range of pipe inclinations**. International Journal of Multiphase Flow, 13:1–12, 1 1987.

BEGGS, D.; BRILL, J. **A study of two-phase flow in inclined pipes**. Journal of Petroleum Technology, 25:607–617, 5 1973.

BELLE, V.; PAPANTONIS, I. **Principles and practice of explainable machine learning**. Frontiers in Big Data, 4, 7 2021.

BENDLKSEN, K. H.; MALNES, D.; MOE, R. ; NULAND, S. **The dynamic two-fluid model olga: Theory and application**. SPE Production Engineering, 6:171–180, 5 1991.

BHAGWAT, S. M.; GHAJAR, A. J. **A flow pattern independent drift flux model based void fraction correlation for a wide range of gas–liquid two phase flow**. International Journal of Multiphase Flow, 59:186–205, 2 2014.

BHATTACHARYYA, S.; VYAS, A. **Application of machine learning in predicting oil rate decline for bakken shale oil wells**. Scientific Reports, 12:20401, 2022.

BIKMUKHAMETOV, T.; JÄSCHKE, J. **Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models**. Computers & Chemical Engineering, 138:106834, 2020.

BIKMUKHAMETOV, T.; JÄSCHKE, J. **First principles and machine learning virtual flow metering: A literature review**. Journal of Petroleum Science and Engineering, 184:106487, 2020.

BILLINGS, S. A. **Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains**. Chichester, West Sussex, United Kingdom: John Wiley & Sons, 2013.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. v. 4 New York, United States: Springer, 2006.

BREIMAN, L. **Bagging predictors**. Machine learning, 24:123–140, 1996.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. A. ; STONE, C. J. **Classification and regression trees**. New York, United States: Routledge, 2017.

BUITINCK, L.; LOUPPE, G.; BLONDEL, M.; PEDREGOSA, F.; MUELLER, A.; GRISEL, O.; NICULAE, V.; PRETTENHOFER, P.; GRAMFORT, A.; GROBLER, J.; LAYTON, R.; VANDERPLAS, J.; JOLY, A.; HOLT, B. ; VAROQUAUX, G. **API design for machine learning software: experiences from the scikit-learn project**. In: ECML PKDD WORKSHOP: LANGUAGES FOR DATA MINING AND MACHINE LEARNING,, 2013, p. 108–122.

CAMPONOGARA, E.; PLUCENIO, A.; TEIXEIRA, A. F. ; CAMPOS, S. R. **An automation system for gas-lifted oil wells: Model identification, control, and optimization**. Journal of petroleum science and engineering, 70(3-4):157–167, 2010.

CANCHUMUNI, S.; EMERICK, A. ; PACHECO, M. **Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother**. Computers & Geosciences, 128:87–102, 2019.

CHAARI, M.; HMIDA, J. B.; SEIBI, A. C. ; FEKIH, A. **An integrated genetic-algorithm/artificial-neural-network approach for steady-state modeling of two-phase pressure drop in pipes**. SPE Production & Operations, 35:628–640, 8 2020.

CHAVES, G. S.; KARAMI, H.; FILHO, V. J. M. F. ; VIEIRA, B. F. **A comparative study on the performance of multiphase flow models against offshore field production data**. Journal of Petroleum Science and Engineering, 208:109762, 1 2022.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O. ; KEGELMEYER, W. P. **Smote: Synthetic minority over-sampling technique**. Journal of Artificial Intelligence Research, 16:321–357, 6 2002.

CHU, M.; MIN, B.; KWON, S.; PARK, G.; KIM, S. ; HUY, N. **Determination of an infill well placement using a data-driven multi-modal convolutional neural network**. Journal of Petroleum Science and Engineering, 195:106805, 2020.

DANIELSON, T. J.; BANSAL, K. M.; HANSEN, R. ; LEPORCHER, E. **Leda: The next multiphase flow performance simulator**. In: INTERNATIONAL CONFERENCE ON MULTIPHASE PRODUCTION TECHNOLOGY, Barcelona, Spain, 2005.

DIAZ, C. M. R.; POSTAL, A. T. ; RODRIGUEZ, O. M. H. **Experimental insights into dense-gas/liquid two-phase flow in horizontal and inclined pipes**. In: SPE BRAZIL FLOW ASSURANCE TECHNOLOGY CONGRESS, Rio de Janeiro, Brazil, November 2024.

DOMÍNGUEZ-JIMÉNEZ, J.; HENAO, N.; AGBOSSOU, K.; PARRADO, A.; CAMPILLO, J. ; NAGARSHETH, S. H. **A stochastic approach to integrating electrical thermal storage in distributed demand response for nordic communities with wind power generation**. IEEE Open Journal of Industry Applications, 4:121–138, 2023.

FALLER, A. C.; PAULO, P. H. C.; VIEIRA, S. C.; FABRO, A. T. ; DE CASTRO, M. S. **A machine learning approach on two-phase flow characterization and calculation based on a large experimental dataset**. In: THE 7TH MULTIPHASE FLOW JOURNEY ANNALS, p. 35–36, Rio de Janeiro, Brazil, 2023.

FREITAS, L.; BARBOSA, B. H. ; AGUIRRE, L. A. **Including steady-state information in nonlinear models: An application to the development of soft-sensors**. Engineering Applications of Artificial Intelligence, 102:104253, 6 2021.

FRIEDMAN, J. H. **Greedy function approximation: A gradient boosting machine**. The Annals of Statistics, 29(5):1189–1232, 2001.

GEURTS, P.; ERNST, D. ; WEHENKEL, L. **Extremely randomized trees**. Machine learning, 63:3–42, 2006.

GOMEZ, L. E.; SHOHAM, O.; SCHMIDT, Z.; CHOKSHI, R. N. ; NORTHUG, T. **Unified mechanistic model for steady-state two-phase flow: Horizontal to vertical upward flow**. SPE Journal, 5:339–350, 9 2000.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2. ed. Sebastopol, CA, United States: O'Reilly Media, 2019.

HADZOVIC, B. B.; ÆSØY, E.; LEINAN, P. R. ; DAWSON, J. R. **Characterising slug flow in a horizontal pipe using bubble image velocimetry**. International Journal of Multiphase Flow, 188:105196, 7 2025.

HAFSA, N.; RUSHD, S.; ALZOUBI, H. ; AL-FAIAD, M. **Accurate prediction of pressure losses using machine learning for the pipeline transportation of emulsions**. Heliyon, 10:e23591, 1 2024.

HAGEDORN, A. R.; BROWN, K. E. **Experimental study of pressure gradients occurring during continuous two-phase flow in small-diameter vertical conduits**. Journal of Petroleum Technology, 17:475–484, 4 1965.

HAN, H.; WANG, W.-Y. ; MAO, B.-H. **Borderline-smote: A new over-sampling method in imbalanced data sets learning**. In: Huang, D.-S.; Zhang, X.-P. ; Huang, G.-B., editors, ADVANCES IN INTELLIGENT COMPUTING, p. 878–887, Berlin, Heidelberg, 2005 Springer Berlin Heidelberg.

HAYKIN, S. **Neural networks: a comprehensive foundation**. USA: Prentice Hall PTR, 1998.

HE, H.; BAI, Y.; GARCIA, E. A. ; LI, S. **Adasyn: Adaptive synthetic sampling approach for imbalanced learning**. In: 2008 IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE), p. 1322–1328, Hong Kong, 6 2008 IEEE.

HO, T. K. **Random decision forests**. In: PROCEEDINGS OF 3RD INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, v. 1, p. 278–282 vol.1, Montreal, QC, Canada, 1995.

HOLLMANN, N.; MÜLLER, S.; PURUCKER, L.; KRISHNAKUMAR, A.; KÖRFER, M.; HOO, S. B.; SCHIRRMEISTER, R. T. ; HUTTER, F. **Accurate predictions on small data with a tabular foundation model**. Nature, 637:319–326, 1 2025.

HOTVEDT, M.; GRIMSTAD, B. ; IMSLAND, L. **Developing a hybrid data-driven, mechanistic virtual flow meter - a case study**. IFAC-PapersOnLine, 53:11692–11697, 2020.

IBM **Ai vs. machine learning vs. deep learning vs. neural networks**., July 2023 URL: https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks, Accessed: 2025-08-17.

IPA **Improve E&P Decision-Making With the Opportunity Assessment Toolkit (OAT)**., 2016 URL: https://www.ipaglobal.com/news/article/improve-ep-decision-making-with-the-opportunity-assessment-toolkit-oat/, Accessed: 2024-11-20.

INKPEN, A. C.; MOFFETT, M. H. **The global oil & gas industry: management, strategy & finance**. Tulsa, OK, United States: PennWell Corp., 2011.

JAFARY, P.; SHOJAEI, D.; RAJABIFARD, A. ; NGO, T. **Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors**. Cities, 151:105115, 2024.

JO, S.; AHN, S.; PARK, C. ; KIM, J. **Generative geomodeling based on flow responses in latent space**. Journal of Petroleum Science and Engineering, 211:110177, 2022.

KANG, B.; CHOE, J. **Uncertainty quantification of channel reservoirs assisted by cluster analysis and deep convolutional generative adversarial networks**. Journal of Petroleum Science and Engineering, 187:106742, 2020.

KANIN, E.; OSIPTSOV, A.; VAINSHTEIN, A. ; BURNAEV, E. **A predictive model for steady-state multiphase pipe flow: Machine learning on lab data**. Journal of Petroleum Science and Engineering, 180:727–746, 9 2019.

KHAN, W. A.; RUI, Z.; HU, T.; LIU, Y.; ZHANG, F. ; ZHAO, Y. **Application of machine learning and optimization of oil recovery and co2 sequestration in the tight oil reservoir**. SPE Journal, 29(06):2772–2792, 06 2024.

KINGSFORD, C.; SALZBERG, S. L. **What are decision trees?** Nature biotechnology, 26(9):1011–1013, 2008.

KORAY, A.; BUI, D.; AMPOMAH, W.; KUBI, E. ; KLUMPENHOWER, J. **Application of machine learning optimization workflow to improve oil recovery**. In: SPE OKLAHOMA CITY OIL AND GAS SYMPOSIUM, Oklahoma City, United States, 2023.

KOZINA, A.; ŁUKASZ KUŹMIŃSKI; NADOLNY, M.; MIAŁKOWSKA, K.; TUTAK, P.; JANUS, J.; PŁOTNICKI, F.; WALASZCZYK, E.; ROT, A.; DZIEMBEK, D. ; KRÓL, R. **The default of leasing contracts prediction using machine learning**. Procedia Computer Science, 225:424–433, 2023.

KUANG, L.; LIU, H.; REN, Y.; LUO, K.; SHI, M.; SU, J. ; LI, X. **Application and development trend of artificial intelligence in petroleum exploration and development**. Petroleum Exploration and Development, 48:1–14, 2 2021.

LEMAÎTRE, G.; NOGUEIRA, F. ; ARIDAS, C. K. **Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning**. Journal of Machine Learning Research, 18(17):1–5, 2017.

LI, C.; LIAO, Y.; JIA, W.; YANG, F.; HE, J. ; WANG, X. **A two-fluid model incorporating droplet entrainment to predict pressure drop and holdup for co2 two-phase flow in horizontal and vertical pipe**. Journal of Pipeline Science and Engineering, p. 100304, 6 2025.

LI, Q.; ZHANG, B.; HE, H.; WANG, Y.; HE, D. ; MO, S. **A hybrid physics-data driven approach for vehicle dynamics state estimation**. Mechanical Systems and Signal Processing, 225:112249, 2 2025.

LIN, Z.; LIU, X.; LAO, L. ; LIU, H. **Prediction of two-phase flow patterns in upward inclined pipes via deep learning**. Energy, 210:118541, 11 2020.

LIU, B.; HE, X. ; LIU, N. **Hybrid residual deep learning models with physical knowledge for improving plant transpiration estimation**. Computers and Electronics in Agriculture, 212:108135, 9 2023.

LIU, Y.; SHAN, L.; YU, D.; ZENG, L. ; YANG, M. **An echo state network with attention mechanism for production prediction in reservoirs**. Journal of Petroleum Science and Engineering, 209:109920, 2 2022.

LUNDBERG, S. M.; LEE, S.-I. **A unified approach to interpreting model predictions** In: Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S. ; Garnett, R., editors, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30, p. 4765–4774 Curran Associates, Inc., 2017.

MA, H.; HAN, G.; ZHU, Z.; WANG, B.; XIANG, X. ; LIANG, X. **Hybrid virtual flow metering on arbitrary well patterns for transient multiphase prediction driven by mechanistic and data model**. Geoenergy Science and Engineering, 243:213335, 12 2024.

MAKHOTIN, I.; ORLOV, D.; KOROTEEV, D.; BURNAEV, E.; KARAPETYAN, A. ; ANTONENKO, D. **Machine learning for recovery factor estimation of an oil reservoir: A tool for derisking at a hydrocarbon asset evaluation**. Petroleum, 8:278–290, 6 2022.

MALAYERI, M.; MULLERSTEINHAGEN, H. ; SMITH, J. **Neural network analysis of void fraction in air/water two-phase flows at elevated temperatures**. Chemical Engineering and Processing, 42:587–597, 8 2003.

MANAMI, M.; SEDDIGHI, S. ; ÖRLÜ, R. **Deep learning models for improved accuracy of a multiphase flowmeter**. Measurement, 206:112254, 2023.

MENARDI, G.; TORELLI, N. **Training and assessing classification rules with imbalanced data**. Data Mining and Knowledge Discovery, 28:92–122, 1 2014.

MERCANTE, R.; NETTO, T. A. **Virtual flow predictor using deep neural networks**. Journal of Petroleum Science and Engineering, 213:110338, 2022.

MITCHELL, T. M.; MITCHELL, T. M. **Machine learning**. Número 9 em McGraw-Hill International Editions New York, United States: McGraw-hill, 1997.

NGUYEN, H. M.; COOPER, E. W. ; KAMEI, K. **Borderline over-sampling for imbalanced data classification**. International Journal of Knowledge Engineering and Soft Data Paradigms, 3:4, 2011.

OKOTIE, V. U.; HOLLAENDER, F.; YOUNES, M. K.; ZIDAN, M. F.; UIJTTENHOUT, M. G. ; AL-KAABI, F. O. **Multiphase flowmeter performance: A critical piece of an offshore well management toolkit**. Abu Dhabi International Petroleum Exhibition & Conference, November 2016.

OSMAN, E.-S. A. **Artificial neural network models for identifying flow regimes and predicting liquid holdup in horizontal multiphase flow**. SPE Production & Facilities, 19:33–40, 2 2004.

OTCHERE, D. A.; GANAT, T. O. A.; GHOLAMI, R. ; RIDHA, S. **Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ann and svm models**. Journal of Petroleum Science and Engineering, 200:108182, 5 2021.

PAULO, P. H.; PEREIRA, F. C. ; AYALA, H. V. **System identification techniques for soft sensors and multiphase flow metering**. IFAC-PapersOnLine, 58:538–543, 2024.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M. ; DUCHESNAY, E. **Scikit-learn: Machine learning in Python**. Journal of Machine Learning Research, 12:2825–2830, 2011.

PEREIRA, F. D. C.; PAULO, P. H. C.; CARVALHO, M. D. S. ; AYALA, H. V. H. **Virtual flow metering using stacking ensemble techniques and nonlinear system identification**. Fuel, 394:134893, 8 2025.

PEREYRA, E.; TORRES, C.; MOHAN, R.; GOMEZ, L.; KOUBA, G. ; SHOHAM, O. **A methodology and database to quantify the confidence level of**

methods for gas–liquid two-phase flow pattern prediction. Chemical Engineering Research and Design, 90:507–513, 4 2012.

PROCHNOW, S. J.; RATERMAN, N. S.; SWENBERG, M.; REDDY, L.; SMITH, I.; ROMANYUK, M. ; FERNANDEZ, T. **A subsurface machine learning approach at hydrocarbon production recovery & resource estimates for unconventional reservoir systems: Making subsurface predictions from multimensional data analysis**. Journal of Petroleum Science and Engineering, 215:110598, 8 2022.

RUIZ-DÍAZ, C. M.; PERILLA-PLATA, E. E. ; GONZÁLEZ-ESTRADA, O. A. **Two-phase flow pattern identification in vertical pipes using transformer neural networks**. Inventions, 9:15, 1 2024.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**.: pearson, 2016.

SABAA, A.; ABU EL ELA, M.; EL-BANBI, A. H. ; SAYYOUH, M. H. M. **Artificial Neural Network Model to Predict Production Rate of Electrical Submersible Pump Wells**. SPE Production & Operations, 38(01):63–72, 02 2023.

SANDELIC, M.; SANGWONGWANICH, A. ; BLAABJERG, F. **Impact of power converters and battery lifetime on economic profitability of residential photovoltaic systems**. IEEE Open Journal of Industry Applications, 3:224–236, 2022.

SANDNES, A. T.; GRIMSTAD, B. ; KOLBJØRNSEN, O. **Multi-task learning for virtual flow metering**. Knowledge-Based Systems, 232:107458, 2021.

SCHAPIRE, R. E. **The Boosting Approach to Machine Learning: An Overview**, p. 149–171 Springer, New York, NY, United States, 2003.

SEONG, Y.; PARK, C.; CHOI, J. ; JANG, I. **Surrogate model with a deep neural network to evaluate gas–liquid flow in a horizontal pipe**. Energies, 13:968, 2 2020.

SHAPLEY, L. S. **Notes on the n-person game — ii: The value of an n-person game** Research Memorandum RM-670, RAND Corporation, Santa Monica, CA, United States, 1951 Originally published in 1951; PDF version created April 24, 2008.

SHIPPEN, M.; BAILEY, W. J. **Steady-state multiphase flow—past, present, and future, with a perspective on flow assurance**. Energy & Fuels, 26:4145–4157, 7 2012.

SIRCAR, A.; YADAV, K.; RAYAVARAPU, K.; BIST, N. ; OZA, H. **Application of machine learning and artificial intelligence in oil and gas industry**. Petroleum Research, 6:379–391, 12 2021.

SKLEARN **Api reference - scikit-learn 1.2.0 documentation**., 2023 URL: https://scikit-learn.org/stable/modules/classes.html, Accessed: 2023-01-11.

SMOLA, A. J.; SCHÖLKOPF, B. **A tutorial on support vector regression**. Statistics and computing, 14:199–222, 2004.

SOKKELDIREKTORATET **Field: Volve**., 2023 URL: https://www.norskpetroleum.no/en/facts/field/volve/, Accessed: 2024-01-11.

SOLTANI, A.; LEE, C. L. **The non-linear dynamics of south australian regional housing markets: A machine learning approach**. Applied Geography, 166:103248, 5 2024.

SONG, S.; WU, M.; QI, J.; WU, H.; KANG, Q.; SHI, B.; SHEN, S.; LI, Q.; YAO, H.; CHEN, H. ; GONG, J. **An intelligent data-driven model for virtual flow meters in oil and gas development**. Chemical Engineering Research and Design, 186:398–406, 2022.

SUSLICK, S. B.; SCHIOZER, D. ; RODRIGUEZ, M. R. **Uncertainty and risk analysis in petroleum exploration and production**. Terrae, 6(1):30–14, 2009.

SÖDERSTRÖM, T.; STOICA, P. **System Identification**. Prentice-Hall international series in systems and control engineering Uppsala, Sweden: Prentice Hall, 1989.

TAITEL, Y.; DUKLER, A. E. **A model for predicting flow regime transitions in horizontal and near horizontal gas-liquid flow**. AIChE Journal, 22:47–55, 1 1976.

TAUNK, K.; DE, S.; VERMA, S. ; SWETAPADMA, A. **A brief review of nearest neighbor algorithm for learning and classification**. In: 2019 INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND CONTROL SYSTEMS (ICCS), p. 1255–1260, Madurai, India, 2019.

TRIVEDI, A. **What information is available in the volve dataset?**, 2020 URL: https://discovervolve.com/2020/04/02/___how_to_access_volve/, Accessed: 2023-01-11.

TRIVEDI, C.; BHATTACHARYA, P.; PRASAD, V. K.; PATEL, V.; SINGH, A.; TANWAR, S.; SHARMA, R.; ALUVALA, S.; PAU, G. ; SHARMA, G. **Explainable ai for industry 5.0: Vision, architecture, and potential directions**. IEEE Open Journal of Industry Applications, 5:177–208, 2024.

VANVIK, T.; HENRIKSSON, J.; YANG, Z. ; WEISZ, G. **Virtual flow metering for continuous real-time production monitoring of unconventional wells**. In: PROCEEDINGS OF THE 10TH UNCONVENTIONAL RESOURCES TECHNOLOGY CONFERENCE, Houston, TX, United States, 2022 American Association of Petroleum Geologists.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; VAN DER WALT, S. J.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; POLAT, İ.; FENG, Y.; MOORE, E. W.; VANDERPLAS, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; VAN MULBREGT, P. ; SCIPY 1.0 CONTRIBUTORS **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python**. Nature Methods, 17:261–272, 2020.

WALTRICH, P. J.; CAPOVILLA, M. S.; LEE, W.; DE SOUSA, P. C.; ZULQARNAIN, M.; HUGHES, R.; TYAGI, M.; WILLIAMS, W.; KAM, S. ; ARCHER, A. **Experimental evaluation of wellbore flow models applied to worst-case-discharge calculations for oil wells**. SPE Drilling & Completion, 34(03):315–333, 2019.

WOLPERT, D.; MACREADY, W. **No free lunch theorems for optimization**. IEEE Transactions on Evolutionary Computation, 1(1):67–82, 1997.

YOU, J.; AMPOMAH, W. ; SUN, Q. **Development and application of a machine learning based multi-objective optimization workflow for co2-eor projects**. Fuel, 264:116758, 2020.

ZHANG, X.; HOU, L.; ZHU, Z.; LIU, J.; SUN, X. ; HU, Z. **Flow pattern identification of gas-liquid two-phase flow based on integrating mechanism analysis and data mining**. Geoenergy Science and Engineering, 228:212013, 9 2023.

ZHANG, H.-Q.; WANG, Q.; SARICA, C. ; BRILL, J. P. **A unified mechanistic model for slug liquid holdup and transition between slug and dispersed bubble flows**. International Journal of Multiphase Flow, 29:97–107, 1 2003.

ZHAO, X.; XU, Q.; WU, Q.; CHANG, Y.; CAO, Y.; ZOU, S. ; GUO, L. **Effect of high pressure on severe slugging and multiphase flow pattern transition in a long pipeline-riser system**. Experimental Thermal and Fluid Science, 148:110976, 10 2023.

ZHONG, Z.; SUN, A.; REN, B. ; WANG, Y. **A deep-learning-based approach for reservoir production forecast under uncertainty**. SPE Journal, 26(3):1314–1340, 2021.

ZHOU, M.; WANG, R.; CHENG, R.; SUN, Q.; YU, Q.; XIA, L. ; SUN, X. **A physics-constrained hybrid residual neural network for the prediction of moisture content in a closed-cycle drying system**. The Canadian Journal of Chemical Engineering, 103:2204–2217, 5 2025.