IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier

Anticipating Financial Risk: Machine Learning for Debt Management in Telecommunications

FILIPE O. F. ARSÉNIO[1, 2], ANTÓNIO RAIMUNDO[3, 4], JOÃO PEDRO C. B. B. PAVIA[3, 4]

[1]Department of Quantitative Methods for Management and Economics, ISCTE-University Institute of Lisbon, Avenida das Forças Armadas 1649-026 Lisboa, Portugal
[2]Department of Information Science and Technology, ISCTE-University Institute of Lisbon, Avenida das Forças Armadas 1649-026 Lisboa, Portugal
[3]Department of Digital Technologies, ISCTE-University Institute of Lisbon, Avenida Heliodoro Salgado nº 3, 2710-569 Sintra, Portugal
[4]Information Sciences, Technolgies and Architeture Research Centre (ISTAR), Avenida das Forças Armadas 1649-026 Lisboa, Portugal

Corresponding author: João Pedro C. B. B. Pavia (joao.pedro.pavia@iscte-iul.pt)

This article was (partially) supported ISTAR Projects UIDB/04466/2025 and UIDP/04466/2025

ABSTRACT The telecommunications industry is characterized by intense competition and rapid technological evolution, making financial stability a critical factor for sustained growth. This work focuses on leveraging machine learning techniques to analyze and predict customer payment behavior within a Portuguese telecommunications company, aiming to reduce financial losses associated with unpaid debts.

Using the CRISP-DM methodology, the project first develops supervised learning models to predict whether customers will remain good payers, based solely on internal data. Among the algorithms tested, Random Forest achieved the highest accuracy of 99%, enabling early identification of potential defaulters. Complementing this, unsupervised learning methods, specifically Principal Component Analysis for dimensionality reduction and K-Means clustering, uncover hidden behavioral segments within the customer base. The optimal clustering identified five distinct groups, some of which show near-homogeneous target values (close to 0 or 1), allowing for strong characterization of compliant and non-compliant profiles.

The findings demonstrate the effectiveness of combining supervised and unsupervised learning for risk analysis. Supervised models allow scenario testing by altering feature values to simulate changes in payment behavior. In unsupervised learning, analyzing ambiguous clusters through comparison with more definitive ones helps estimate likely client outcomes and supports proactive management.

Future work may explore focused clustering of non-compliant clients, alternative data preprocessing, and time series forecasting to further improve predictive accuracy and operational utility.

INDEX TERMS Supervised Learning, Unsupervised Learning, Financial data, Telecommunications, Risk Management

I. INTRODUCTION

The telecommunications industry plays a critical role in the global economy, serving as the backbone of connectivity and enabling numerous other industries to thrive. As stated by the authors of [1], "The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies." This extremely competitive industry is intensified by "rapid changes, market liberalization, technological advancements, and diverse attractive plans" [2], making financial stability crucial for sustaining growth and innovation. It is because of this competitiveness that the ability to analyze financial data effectively becomes a crucial factor for telecommunications companies to make informed decisions, optimize operations, and maintain competitiveness in order to be at the top of the industry [3].

Supervised learning, a subset of machine learning, has emerged as a valuable tool for analyzing financial data, enabling companies to make data-driven decisions [4]. These techniques have been widely applied in various domains, showcasing their versatility and effectiveness in addressing complex financial problems. Supervised learning provides a

structured framework to uncover patterns and relationships in financial data, allowing telecommunications companies to predict customer churn [5], optimize credit risk management [6], and enhance profitability forecasts [7], among others.

While supervised learning has been highly effective in financial data analysis, it is not without limitations. One of its primary challenges is the need for labeled data. This is where unsupervised learning comes into play, offering a complementary approach to uncover hidden structures in data without requiring explicit labels [8], and proving essential for analyzing complex datasets by enabling pattern discovery, dimensionality reduction, and clustering, with applications not only in telecommunications and finance, but also in medicine, education, biology, computer science, and social sciences [9]. Moreover, as highlighted by the authors of [8], such techniques are particularly valuable in domains like image recognition, speech processing, and anomaly detection, where extracting meaningful insights from unlabeled data is crucial.

In telecommunications, unsupervised learning is widely applied in churn prediction, as it helps companies identify patterns in customer behavior. By using dimensionality reduction and clustering techniques, businesses can group customers based on their usage patterns and service preferences, which in turn enables the development of more targeted retention strategies [10]. Typically, customer data encompasses various attributes, such as account duration, international usage, and customer service interactions, all of which contribute to identifying distinct behavioral segments [9]. By uncovering hidden patterns within these datasets, unsupervised learning allows companies to anticipate churn risk more effectively and, as noted by the authors of [9], "...improve their marketing strategies which indulge in profit maximization."

Given that financial data analysis is crucial for the sustainability and competitiveness of telecommunications companies [3], and that both supervised [5] and unsupervised learning techniques [8] provide valuable insights into data patterns, the combination of these elements becomes highly relevant.

The main objectives of this project are threefold:

- Develop a supervised learning model capable of predicting whether a customer is likely to remain a good payer or not.
- Reduce the dimensionality of the dataset and apply clustering techniques in order to uncover hidden patterns in the data using unsupervised learning methods.
- Explore how both supervised and unsupervised learning approaches can be used together in a complementary manner, which consists of predicting whether a customer is a good payer and then verifying which cluster they belong to, in order to assess whether they are grouped with typical good payers or bad payers. This allows for a dual confirmation of the prediction and opens the door to more informed actions. For instance, if a customer is predicted to be a good payer but falls within a cluster of bad payers, the company can take preventive measures such as closer monitoring. Conversely, if a customer is predicted to be a bad payer but belongs to a cluster of good payers, it may be worth offering incentives or support to help reverse the negative trend.

Regarding the innovation of this study, unlike the usual approaches that integrate unsupervised learning as a preprocessing step, this study introduces a complementary workflow where supervised models predict payment behavior first, and clustering is then applied to validate or challenge these predictions. This sequential design enables the identification of potential risk transitions (mainly from good payer to bad payer) and provides actionable insights for proactive debt management in the telecommunications sector.

The study was structured following the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, a structured and iterative framework widely recognized for its flexibility and applicability across different business domains [11], [12]. This metodology consists of six main phases (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment) and is represented as an iterative cycle, allowing continuous revisions between phases.

Finally, the structure of this article is organized as follows. Section II presents the state of the art on supervised learning, unsupervised learning, and hybrid approaches, with particular emphasis on financial data and applications within the telecommunications sector, highlighting how this study contributes to and extends the existing literature. Section III follows the CRISP-DM methodology by detailing the business understanding, data understanding, and data preparation phases, during which the dataset is examined and transformed to ensure suitability for modeling. Section IV covers the modeling and results, comprising two experiments: the first uses all available features to evaluate four supervised learning algorithms, followed by dimensionality reduction using PCA and t-SNE and clustering with four different algorithms; the second experiment employs a reduced set of features selected through feature importance metrics and KMO scores. Finally, Section V discusses the findings, reflects on their practical implications, and presents the main conclusions along with recommendations for future research.

II. LITERATURE REVIEW

The prediction of customer payment behavior and financial outcomes in the telecommunications sector has received considerable attention, driven by the increasing digitalization of financial services and the availability of high-volume operational data. The literature spans supervised, unsupervised and hybrid machine learning approaches, with applications in churn prediction, fraud detection, credit scoring and profitability forecasting.

A. SUPERVISED AND UNSUPERVISED LEARNING APPLIED TO FINANCIAL DATA

Supervised learning techniques are widely used in forecasting, risk assessment, and fraud detection, enabling institutions to extract actionable insights from historical datasets to predict future outcomes. Financial data are often high-dimensional, heterogeneous, and imbalanced, making the use of advanced supervised learning methods both challenging and valuable [13]–[15].

Credit scoring is a key area for supervised learning. While logistic regression remains popular for its interpretability, modern machine learning models such as Random Forest, Gradient Boosting Machine (GBM), and Support Vector Machine (SVM) offer higher accuracy by capturing non-linear relationships in the data [13], [14], [16].

Detecting fraudulent transactions is another critical application. Machine learning models, including Naïve Bayes, SVM, and deep learning architectures like Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN), are widely used for identifying anomalies in transaction patterns [15]. A notable example is the integration of Extreme Gradient Boosting (XGBoost) with Bidirectional Gated Recurrent Unit (BiGRU) and self-attention networks for internet loan fraud detection, which achieved exceptional accuracy and precision on benchmark datasets [17], [18]. These methods address challenges such as data imbalance by leveraging techniques like outlier scoring and feature engineering.

Supervised learning has also revolutionized financial market forecasting. Time-series models such as Support Vector Regression are used to predict stock prices, debt levels, and other financial indicators. In the Montenegrin electricity industry, this model outperformed traditional models by efficiently capturing non-linear dependencies in data, demonstrating its efficacy in debt risk prediction [13], [16].

Deep learning models have shown significant promise in financial applications. Architectures like LSTM and CNN are increasingly employed for processing sequential and high-dimensional data, as they are capable of capturing temporal dependencies and extracting hierarchical features, wich is particularly valuable in financial contexts where customer behavior evolves over time and data often contains diverse patterns. For instance, LSTM combined with attention mechanisms has proven effective in credit risk modeling by accurately identifying complex patterns in customer behavior [16], [17]. Similarly, hybrid approaches that combine feature selection with deep learning models have enhanced both interpretability and performance in financial forecasting [18].

The application of supervised learning extends to the insurance sector, particularly in predicting lapses in mortgage life insurance policies. Models such as Random Forest, XGBoost, and Artificial Neural Network (ANN) have been utilized to identify clients at higher risk of policy cancellation [19].

Unsupervised learning techniques are particularly useful in scenarios where labeled data is scarce or when identifying hidden structures can enhance decision-making. In churn prediction, these methods play a crucial role by revealing customer behavior patterns that are not explicitly labeled, allowing for more strategic interventions. A fundamental application of unsupervised learning in churn prediction is customer segmentation through clustering. By analyzing customer usage patterns and service preferences, clustering techniques such as K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and hierarchical clustering help classify users into distinct behavioral groups, enabling telecommunications companies to tailor retention strategies and improve customer satisfaction [9].

Beyond customer segmentation, anomaly detection is another critical application of unsupervised learning in churn prediction. By identifying irregular customer behaviors that may indicate an increased likelihood of churn, techniques like Rough Set Theory and density-based clustering help detect deviations in customer activity, providing early warnings for proactive intervention [20].

While supervised and unsupervised learning techniques each offer unique advantages in churn prediction, combining both approaches has proven to be highly effective in enhancing predictive accuracy and decision-making. Hybrid models leverage the strengths of supervised learning in classification and prediction, while incorporating unsupervised methods for feature engineering, data preprocessing and better interpretability. A hybrid approach that integrates supervised and unsupervised learning has shown great effectiveness in churn prediction, as shown by the authors of article [21], where clustering techniques were applied in order to segment customers based on their behavior, and then used supervised learning models, like SVM and Recursive PARTitioning, to predict churn within each segment, improving prediction accuracy.

Additionally, ensemble models that combine clustering and classification have been proposed to enhance churn prediction performance, as demonstrated by the authors of articles [22] and [23], where a combination of clustering and classification algorithms was used to improve the predictive power and robustness of churn prediction models in the telecom industry. These models first cluster customers based on shared behavioral traits and then use supervised classifiers to predict churn likelihood within each cluster, improving interpretability and precision.

By combining supervised and unsupervised approaches, telecom companies can build more accurate and adaptive churn prediction models, leading to better retention strategies and reduced customer attrition.

B. SUPERVISED AND UNSUPERVISED LEARNING TECHNIQUES APPLIED TO FINANCIAL DATA IN TELECOMMUNICATIONS

1) Supervised Learning

Regarding supervised learning techniques, three main research areas stand out: Fraud Detection [24], Profitability Prediction [7], and Churn Prediction [25].

Churn prediction refers to estimating the likelihood that a customer will leave a service provider. As stated by [25], "the churn customers are the customers, who decides to escape from the service provider and prepare to shift other competitors in the market." According to [26], effective churn modeling involves not only predicting whether a customer will churn but also uncovering the reasons behind it, though most research has focused on the first aspect. In the telecommunications sector, supervised learning methods are widely applied to churn prediction, with performance depending on the chosen algorithms and metrics. Decision tree-based models, such as Random Forest (RF) and GBM, are among the most used. RF frequently achieves metrics close to 99% in precision, recall, and F1-Score [5], [27], while GBM delivers more variable results, ranging from robust performance to modest outcomes such as an F1-Score of 60% and precision of 67% in [26], despite a good ROC score of 86%. Logistic regression provides moderate but competitive results, with an F1-Score of 81% reported in [2], though boosted variants may underperform slightly. Deep learning approaches, such as optimized convolutional neural networks, can achieve high precision (94.76% in [25]), yet other complex models like Fully Connected Layer Convolutional Neural Network - Long Short Term Memory have shown mixed performance, with a low F1-Score (50.93%) despite high precision (94.15%). XGBoost has also demonstrated strong results, with studies [1], [28] highlighting its effectiveness in handling large-scale tabular data. Finally, ensemble methods appear promising: in [29], an ensemble achieved the highest F1-Score (85%) among tested models, suggesting that combining algorithms may better capture underlying patterns.

Fraud, as defined by the authors of [24], is "...the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets." Detecting such activities is crucial due to their significant impact across sectors like online banking, e-commerce, and telecommunications, causing multi-billion dollar losses globally [30]. In telecom, fraud detection involves spotting anomalies in customer accounts, usage, or financial transactions. Machine learning techniques such as Random Forest, SVM, and XGBoost are commonly used to detect fraud in real-time by learning from historical data [24], [30], [31]. Advanced methods such as graph-based approaches, deep learning, and ensemble models have shown strong performance. The Bridge to Graph (BTG) method [31] achieved an Area Under Curve (AUC) of 92.45% and F1-Score of 87.01%, demonstrating the value of graph-based models in capturing relational data. Deep learning techniques, particularly Deep Convolutional Neural Network, reached 82% accuracy [30], while traditional methods like Random Forest and Gradient Boosting remained competitive. ANN also performed effectively, often outperforming classical algorithms like SVM and Decision Trees [24].

Profitability prediction forecasts a company's future financial performance using factors such as market conditions, customer behavior, and historical financial data [7], [32]. Supervised learning models, including SVM [7], Random Forest [32], Neural Networks [7], and k-Nearest Neighbors (k-NN) [7], have been applied to assess the factors that most influence financial outcomes. In particular, k-NN achieved a precision of 95.5%, recall of 94.7%, and F1-Score of 95.1%, while other models like Naïve Bayes, ANN, and SVM also showed good performance. Despite promising results, profitability prediction in telecommunications is relatively underexplored compared to churn or fraud detection. The limited number of studies constrains broader insights, highlighting the need for further research to evaluate the effectiveness of different machine learning techniques in this domain.

2) Unsupervised Learning

Unlike the articles focused on supervised learning techniques, the articles that focus on unsupervised learning techniques are only applied to churn prediction. A predominant approach is spectral clustering, as applied by the authors of articles [33], [34]. The authors of article [33] introduced an enhanced spectral clustering algorithm with S-Distance, achieving a Jaccard Index of 92.45% and an F-score of 87.01%, demonstrating that modified distance measures can improve clustering quality, while the authors of [34] proposed a scalable spectral clustering method using Nyström and Lanczos approximations, focusing on reducing computational complexity rather than predictive accuracy, highlighting a trade-off between efficiency and precision.

Another key trend is the integration of graph theory with clustering methods, particularly in social network analysis, according to the authors of article [35]. In this study, Ward's Minimum-Variance Clustering was combined with graph-based metrics, achieving a top decile lift of 2.5-10, depending on the model used. Their findings suggest that churn behavior is influenced by social connections, where influential customers trigger additional churn within their network.

Traditional clustering techniques, such as K-means and hierarchical clustering, have also been widely used, as the authors of article [10] compared K-means, K-medoids, FCM, GK, and Hierarchical Clustering, reporting an accuracy of 55.15% for FCM, while K-means obtained the highest Rand Index (50.50%) and Fowlkes Mallows Index (61.09%). These results emphasize the importance of selecting the appropriate clustering method based on the characteristics of the dataset.

Additionally, the authors of article [20] combined hierarchical clustering with rough set theory, stating that their approach proved effective in identifying customer behavior patterns. Similarly, the authors of article [9] applied K-means clustering for customer segmentation, confirming its effectiveness in detecting high-risk churn groups.

3) Supervised and unsupervised learning combined

The integration of supervised and unsupervised learning techniques in the telecommunications sector has been widely explored for customer churn prediction. A key observation from these studies is that hybrid approaches consistently

outperform individual techniques, demonstrating their effectiveness in enhancing predictive accuracy and model robustness. This improvement is evident across multiple studies where the authors compared the results obtained by using supervised and unsupervised learning in isolation versus in combination.

A distinct approach was identified by the authors of article [36], where unsupervised learning was applied after supervised learning. In this study, XGBoost was first employed to predict customer churn, achieving an accuracy of 85.1% and detecting approximately 11,000 churn customers per month. Subsequently, K-Means clustering was used to segment these customers, leading to a 0.93% reduction in churn by successfully retaining 9,000 customers. This highlights the utility of clustering as a post-processing tool to refine churn management strategies.

Moreover, the authors of study [37] have demonstrated that higher performance metrics might not always indicate methodological superiority but rather result from data preprocessing choices. For instance, in that study PCA was used to reduce dimensionality from 186 to 60 principal components before applying classification models. While this approach led to an extremely high reported accuracy (99.97%), the retention of a significant portion of the original data complexity suggests that the model still operated with substantial information, potentially inflating predictive performance. Additionally, the impact of class imbalance is often underexplored in these studies. Many real-world churn datasets exhibit highly imbalanced distributions, which can lead to misleadingly high accuracy scores if not properly addressed through resampling techniques or appropriate metric selection.

The role of clustering in these studies varies significantly. In article [38], the authors used K-Means as an exploratory tool for customer segmentation without direct integration into the classification models. In contrast, the authors of article [39] exemplify a more intertwined approach, where unsupervised learning is a critical preprocessing step that enhances the classification process. Their Multi-Objective Evolutionary Ensemble Clustering (MOEEC)-1 and MOEEC-2 models integrate clustering within ensemble learning frameworks, leading to increased classification accuracy and robustness.

Several trends emerge from these hybrid models. Regarding the choice of clustering methods, most studies employ centroid-based techniques such as K-Means, Adaptive K-Means, and K-Medoids, likely due to their simplicity and interpretability in customer segmentation. In terms of supervised learning models, the most frequently used classifiers include SVM, RF, Gradient Boosting Trees, and XGBoost. Ensemble techniques, such as AdaBoost and Stacking, also appear frequently, highlighting the advantages of leveraging multiple classifiers to enhance predictive accuracy. Finally, the effectiveness of hybrid models is evident across all studies, as the combination of clustering and classification techniques leads to improved churn prediction. For instance, the authors of [23] reported that RF outperformed SVM, achieving an accuracy of 95.44% and an F1-Score of 97% for non-churners and 84% for churners, emphasizing the advantage of hybrid approaches.

Additionally, some studies introduced novel methodologies beyond traditional clustering and classification. The authors of article [40] proposed a probabilistic possibilistic fuzzy clustering model combined with an artificial neural network, which allows for more flexible membership assignments compared to traditional K-Means clustering. This approach improves the model's ability to capture uncertainty in customer behavior. Meanwhile, the authors of article [41] employed a Bayesian hierarchical joint model with SVM, which incorporates both time-to-event and count data to refine churn prediction. These methodologies highlight the potential for hybrid models to go beyond standard techniques, integrating probabilistic and hierarchical modeling approaches for enhanced churn prediction.

C. LITERATURE OVERVIEW AND PROPOSED APPROACH

The literature presents a diversified methodological landscape: supervised learning dominates predictive tasks (churn, fraud, profitability), unsupervised learning is central to segmentation and anomaly detection, and hybrid strategies frequently obtain the best empirical outcomes by combining both paradigms. Crucially, "every algorithm differs according to area of application and no algorithm is more powerful than the other in different scenarios" [42]. Therefore, methodological choices should be guided by dataset characteristics, evaluation protocols and operational constraints rather than by a priori preferences for specific algorithm families.

In studies using supervised learning techniques, it is common to propose the use of customer segmentation through machine learning and unsupervised learning techniques, as stated by the authors of studies [2], [28]. Furthermore, the authors of articles [26], [27] recommend using better balanced datasets to significantly improve results. In addition, the authors of article [26] suggest handling outliers carefully, as they can distort machine learning models, reducing accuracy and generalization.

A common limitation in articles is the small amount of available customer data, as noted by the authors of studies [27], [38], [41] , with more data allowing for better validation of machine learning methods. The authors of articles [7], [37], [39] recommend the use of more advanced techniques, such as deep learning, and the refinement of feature engineering, exploring additional techniques and features to enhance model performance. Additionally, the authors of [21] suggest determining the optimal number of clusters to achieve the best results.

Although the authors of [42] affirm that no algorithm is inherently more powerful than another, in different scenarios, it can be said, based on the analysis of studies that combine supervised and unsupervised learning, that this approach consistently outperforms the use of either technique individually.

Thus, the model proposed in this dissertation addresses all the limitations and recommendations identified in the analyzed studies. In addition to combining supervised and unsupervised learning techniques, this study benefits from a large dataset (millions of observations), employs advanced feature engineering techniques, such as balancing the dataset, and applies methods to determine the optimal number of clusters in unsupervised learning. By building upon these strengths, the proposed model is expected to achieve higher predictive accuracy and better customer segmentation compared to traditional approaches. Leveraging the advantages of both supervised and unsupervised learning, it not only improves classification performance but also uncovers hidden patterns within the data.

III. DATA ANALYSIS

Managing customer debt is a critical challenge for financial and telecommunications companies, as unpaid obligations directly affect cash flow and profitability. Traditional credit scoring approaches, which rely on historical payment behavior, are often ineffective for new customers or those lacking detailed credit histories. To address this limitation, this study develops a predictive framework that leverages supervised learning for debt repayment classification and unsupervised methods for dimensionality reduction and customer segmentation, enabling the discovery of behavioral patterns that enhance predictive accuracy.

The proposed system supports early detection of high-risk customers and automates risk assessment, allowing firms to implement preventive measures while reducing manual intervention. By doing so, it is expected to strengthen financial stability, improve debt recovery strategies, and lower operational, administrative, and legal costs associated with collection processes.

The dataset used in this study comes from a Portuguese telecommunications company and contains information on customer debt as of December 2023. It consists of 90 variables and 2,437,633 observations, with each observation representing a customer's contract account. The dataset provides a snapshot of the customer's financial situation at that time, covering different customer types (consumer, business, public sector, wholesale). It includes details on total debt, overdue and non-overdue amounts, debt segmented by different maturities, as well as indicators of indemnities, provisions, dissolution or insolvency status, installment payments for equipment purchases, payment methods (such as bank transfer), among others.

A significant portion of the variables in the dataset are purely informational (such as tax identification number, contract account number, company name, account type, etc.). These were excluded as they do not contribute to the models. This selection process was carried out with the assistance of experienced professionals who handle this data regularly, ensuring that no relevant information was discarded. After this filtering step, the dataset was reduced to 27 variables.

Prior to model development, extensive data preprocessing was conducted to ensure quality and consistency. Variables directly related to the construction of the target outcome were excluded to avoid data leakage, reducing the initial dataset to 25 predictors. Missing values were handled according to the nature of each variable: binary and event-based attributes (for example, insolvency, dissolution, credit sales) were encoded into categorical indicators, with null values representing non-occurrence, while numerical and temporal attributes were imputed using statistical measures such as the median or mode. New features were engineered, including customer account longevity and time since taxpayer identification creation, to capture behavioral and temporal dimensions relevant to credit risk. Redundant variables providing overlapping information were removed to reduce multicollinearity.

Outliers were addressed through winsorization [43], capping extreme values rather than excluding them, thereby preserving data representativeness while mitigating the influence of distortions. Min–Max normalization [44] was applied to numerical features to align scales across heterogeneous attributes, supporting model convergence and interpretability. Categorical variables were transformed using binary and one-hot encoding, ensuring compatibility with machine learning algorithms [45]. The target variable ("Debt Repayment") was defined as a binary indicator distinguishing between customers who honored or defaulted on their debt, based on overdue balances adjusted by short-term settlement windows.

Finally, as it is important to address class imbalance [46], the majority class ("Paid") was undersampled to match the minority class ("Not paid"), resulting in a balanced dataset. After preprocessing, the final dataset consisted of 1,087,666 observations and 61 features, encompassing financial, contractual, and behavioral information ready for predictive modeling.

IV. MODELING AND RESULTS

In this chapter, an initial experiment was conducted using all available features, where several models were compared across both supervised and unsupervised learning techniques. The best-performing models from each technique were then selected. Subsequently, a second experiment was carried out using the most relevant features identified from both approaches - through feature importance analysis in supervised learning and Kaiser-Meyer-Olkin (KMO) scores in unsupervised learning - with the aim of understanding whether the model could be reduced in dimensionality.

1) Risk of temporal leakage and validation strategies

The dataset used in this study represents a static snapshot of customers' financial status as of December 2023. This snapshot-based nature introduces a potential risk of temporal leakage, as some variables may implicitly encode information that is temporally downstream of the target outcome. To mitigate this risk, all features directly related to the target variable or derived from post-outcome information were excluded during preprocessing, ensuring that only contemporaneously available attributes were used for model training.

Regarding model validation, the data were split into training (80%) and testing (20%) sets using stratified sampling to preserve the original class distribution and enable a fair evaluation of model performance across classes. While a time-based validation scheme would be preferable to fully eliminate temporal leakage, such an approach was not feasible due to the lack of longitudinal historical data. Nonetheless, the adopted preprocessing and validation strategy reduces leakage risk and supports a reliable assessment of the proposed models.

A. SUPERVISED LEARNING MODELS

To predict customer debt repayment, four supervised classification algorithms were evaluated: Logistic Regression, Random Forest, k-Nearest Neighbors, and Neural Networks. Prior to modeling, a correlation analysis was performed to reduce multicollinearity in order to enhance model stability [47]. Highly correlated variable pairs (above 0.7 or below -0.7) were identified, and for each pair, the variable with the lower correlation to the target was removed. Correlations between remaining features and the target variable were also assessed, revealing that variables such as 'SEPA' (62.67%) and 'Serviço_Estado_Desativos' (-69.26%) provide strong predictive information.

The dataset was then split into training (80%) and testing (20%) sets using stratified sampling to maintain class distribution and ensure fair evaluation. Following these preprocessing steps, the dataset was ready for modeling.

1) Logistic Regression

A logistic regression model was trained to predict debt compliance and evaluated on both training and test sets. The model achieved 94% accuracy on the test set, with high F1-scores for both classes. Analysis of classification errors shows 9,536 false positives (customers predicted as payers who defaulted) and 3,148 false negatives (customers predicted as non-payers who actually paid). The recall for payers was 97% and precision was 92%, reflecting a good balance between correctly identifying payers and minimizing risky approvals. No significant overfitting was observed, as training and test metrics were similar, indicating that the model effectively generalizes to unseen data. Minimizing false positives is particularly important for reducing financial risk, while false negatives have a lower but still relevant impact on revenue and customer relations.

2) Random Forest

The Random Forest model was trained with 100 trees, maximum depth 20, and minimum samples per split/leaf set to control complexity. It achieved 99% accuracy on both training and test sets, with precision, recall, and F1-scores consistently high across both classes. Classification errors were low, with 2,165 false positives and 1,079 false negatives on the test set. Compared to logistic regression, Random Forest significantly reduces both types of errors, improving predictive reliability and minimizing financial risk. No signs of overfitting were observed.

3) k-Nearest Neighbors

A k-NN model with $k = 5$ was trained, achieving 97% accuracy on the training set and 96% on the test set. Precision, recall, and F1-scores were balanced for both classes. The test set included 3,880 false positives and 3,891 false negatives, slightly higher than Random Forest but still low relative to true positives and negatives. No overfitting was detected, and the model demonstrates reliable classification performance.

4) Neural Network

A neural network with an input layer of 64 neurons, one hidden layer of 32 neurons, 20% dropout, and a sigmoid output was trained using Adam optimizer and binary cross-entropy loss with early stopping. The model achieved 97% accuracy on the test set, with balanced precision, recall, and F1-scores. Test errors were 3,063 false positives and 2,766 false negatives, slightly improving over k-NN. Accuracy and loss curves indicate stable learning and no overfitting, confirming strong generalization to new data.

5) Best supervised learning model

In order to compare the performance of the four supervised learning models, Table 1 summarizes the key evaluation metrics, which are precision, recall, F1-score, and accuracy, on the test set for each model.

TABLE 1: Summary of Classification Metrics for Test Set across Models

Model	Class	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0	0.97	0.91	0.94	0.94
	1	0.92	0.97	0.94	
Random Forest	0	0.99	0.98	0.99	0.99
	1	0.98	0.99	0.99	
k-NN	0	0.96	0.96	0.96	0.96
	1	0.96	0.96	0.96	
Neural Network	0	0.97	0.97	0.97	0.97
	1	0.97	0.97	0.97	

As shown in Table 1, the random forest model clearly stands out as the best performer, achieving the highest scores in all evaluation metrics (99%). Although k-NN and the neural network showed similar results in performance metrics, random forest provides equally high performance with significantly faster training times when compared to k-NN. Additionally, the neural network, while performing well, it is not only slightly slower when it comes to training time, but also computationally expensive and can be challenging to optimize, especially in environments with limited resources. This makes Random Forest the most efficient choice in terms of cost-effectiveness, as it offers robust results without the high computational costs associated with neural networks.

As the random forest is the best performer, it is important to carefully analyse the performance metrics of this model, showed by the confusion matrix in Table 2.

TABLE 2: Confusion matrix for the Random Forest model

	Predicted 0	Predicted 1
True 0	106602	2165
True 1	1079	107688

The confusion matrix confirms the robustness of the model, with only 2,165 false positives (customers predicted as good payers who defaulted) and 1,079 false negatives (customers predicted as defaulters who actually paid) out of more than 215,000 predictions. These low error rates are particularly relevant for financial risk management, as false positives represent potential financial losses, while false negatives may lead to missed opportunities for customer retention. Additionally, the minority class ('Not Paid') achieved precision = 98%, recall = 99%, and F1-score = 99%, demonstrating that the model handles class imbalance effectively after applying undersampling. This is critical because misclassifying defaulters can have significant financial implications.

6) Feature importance analysis

To understand which variables most influence the predictions of the supervised learning models, the top 15 most important features of each model were analyzed. Feature importance provides insights into the key drivers of model performance and helps identify variables that consistently contribute to predicting debt compliance. By comparing feature importance across the four models studied, it was observed that several features appear consistently. Table 3 lists the variables that are among the top 15 in all four models, as well as those appearing in three out of the four models. These recurring features highlight the strongest predictors of debt repayment behavior, providing actionable insights for business decision-making.

TABLE 3: Variables that appear in 4 and 3 models

Variables that appear in 4 models	**Variables that appear in 3 models**
'Serviço_Estado_Desativos'	'Saldo Total'
'30 a 60'	'SAS - Ciclo_19.0'
'PPP'	'SAS - Ciclo_22.0'
'SEPA'	'VEP - Nº Prestações Venc'
'Longevidade'	'Data Criação NIF (Diferença Dias)'
'CAT_ CONTA_ CONTRATO_ MEO'	

B. UNUPERVISED LEARNING MODELS

Seeing that unsupervised models are based on pattern recognition without access to the target variable, before beginning the correlation analysis and modeling, the target variable was removed from the dataset to ensure that the patterns identified are derived solely from the features and not influenced by any prior knowledge of the outcome. Dimensionality reduction techniques, including PCA and t-SNE, were evaluated to determine the most suitable method. Subsequently, partitioning (k-Means), hierarchical (Ward and complete linkage), and density-based (DBSCAN) clustering approaches were applied and compared to identify the most effective method for grouping the data.

1) Dimensionality reduction

To assess whether PCA was suitable for the dataset, the Bartlett's test was applied to verify if the correlation matrix significantly differs from an identity matrix, confirming the suitability of the data for factor analysis when the result is significant ($p < 0.05$) [48]. The result showed a test statistic of 154,107,230.40 and a p-value close to zero, which led to the conclusion that the variables were sufficiently correlated to proceed with PCA. Furthermore, the KMO index was calculated. According to [49], a KMO score above 0.6 is recommended, but this dataset's score was only 0.1092, which is well below the threshold. This result indicated weak correlations within the dataset, suggesting that PCA might not be effective. Therefore, the individual KMO values for each variable were calculated, revealing that several variables had very low scores, further indicating little to no correlation with other variables.

Therefore, variables with individual KMO values below 0.5 were removed, resulting in a final set of 19 variables and an improved overall KMO of 0.7063. Bartlett's test subsequently confirmed that the correlation structure of the cleaned dataset was appropriate for PCA (p-value < 0.05).

The number of principal components was determined by analyzing eigenvalues, the scree plot, and cumulative explained variance.

Regarding eigenvalues, as the authors of [49] state, the rule of thumb consists of retaining components with eigenvalues greater than 1. However, none of the eigenvalues in this case exceed 1, indicating that none of the components explain a significant proportion of the variance in the dataset. Analyzing the scree plot is also important to determine the optimal number of Principal Component (PC), as stated by the authors of [49], and the scree plot shows an elbow at the fourth principal component, suggesting that three principal components should be chosen for further analysis. According to the authors of [49], the cumulative explained variance is important to analyze in order to understand how each principal component contributes to explaining the data variability, and in this case the first three components explained 66.97% of the total variance.

Therefore, the first three principal components were retained for further analysis, effectively capturing the majority of the data variability while reducing dimensionality. The three PC selected capture 39.17%, 16.80%, and 12.91% of the total variance, respectively. Each component was analyzed based on its loadings, which are presented in Table 4, in order to interpret its underlying meaning and assign an appropriate name.

TABLE 4: PC loadings

Variável	PC1	PC2	PC3
Cliente Estrangeiro	0.0004	-0.0001	-0.0014
Saldo Total	0.1887	0.4795	-0.2170
30 a 60 Dias Vencido	0.0106	-0.0287	-0.1259
Dissolução	0.0108	0.0020	0.0071
Insolvência	0.0287	0.0121	0.0106
Insolvência - PER	0.0007	0.0001	-0.0002
VEP	-0.0564	0.5453	-0.4853
VEP - Nº Prestações	-0.0142	0.2163	-0.2059
VEP - Nº Prestações Vencidas	0.0339	0.1303	-0.1176
PPP	-0.0070	-0.0046	-0.2007
PPP - Nº Prestações	-0.0059	0.0004	-0.1671
PPP - Valor do Plano	-0.0059	0.0013	-0.1692
Monitorização	-0.0018	-0.0180	-0.0233
Venda a Crédito	0.0000	-0.0011	-0.0003
SEPA (Pagamento Débito Direto)	-0.6282	0.5031	0.5543
Longevidade	0.2536	-0.0496	0.2216
Data Criação NIF (Diferença Dias)	-0.0427	-0.1154	0.0429
Estado do Serviço - Desativado	0.7056	0.3721	0.4337
Estado do Serviço - Interrupção	-0.0005	0.0140	-0.0389

So, based on these loadings, the following names were chosen to each PC:

- **PC1: Engagement vs. Churn Risk**
  PC1 was named 'Engagement vs. Churn Risk' due to its strong association with service status and payment behavior. The variable 'Serviço_Estado_Desativos' has the highest positive loading (0.71), indicating that customers with high PC1 scores are more likely to have had their service deactivated. Conversely, 'SEPA' has a strong negative loading (-0.63), suggesting that customers who actively use direct debit tend to have lower PC1 scores. Additionally, 'Longevidade' and 'Saldo Total' load positively, indicating that deactivated customers often have longer tenures and outstanding balances.
  This component distinguishes engaged, reliable customers (low PC1 scores - active service, SEPA users) from at-risk or churned customers (high PC1 scores - inactive, no SEPA usage). This interpretation aligns with operational expectations, where engaged customers maintain active services and use automated payment mechanisms, while deactivated customers may have accumulated debt over time.
- **PC2: Active Indebtedness**
  PC2 was named 'Active Indebtedness' as it is primarily influenced by variables that reflect current financial exposure. The variable 'VEP' shows a high positive loading (0.55), along with 'Saldo Total' (0.48) and 'SEPA' (0.50). These indicate that customers with high PC2 scores have considerable outstanding debt and are likely in the process of repayment, even if they use SEPA to manage those obligations. The loading for 'Serviço_Estado_Desativos' (0.37) suggests that some of these customers may already be deactivated, indicating unresolved debts.
  This component captures ongoing financial pressure: customers with high PC2 scores carry substantial debt and are often involved in payment arrangements. In contrast, customers with low PC2 scores show minimal or no active debt and are likely in a healthier financial position.
- **PC3: Payment Cycle Completion**
  PC3 was named 'Payment Cycle Completion' because it differentiates between customers who have likely concluded their financial obligations and those still engaged with repayment. 'SEPA' and 'Serviço_Estado_Desativos' both load positively (0.55 and 0.43 respectively), while 'VEP', 'Saldo Total', and all 'PPP' variables load negatively (for example, VEP = -0.49; PPP = -0.17 to -0.20).
  This indicates that high PC3 scores are typical of customers who used direct debit, cleared their debts, and subsequently deactivated their service, suggesting a completed financial cycle. On the other hand, low PC3 scores represent customers who are still repaying debts or have outstanding obligations, often without the structure of automated payments.

In addition to PCA, the t-SNE technique was evaluated for dimensionality reduction. Unlike PCA, t-SNE preserves local relationships without requiring eigenvalue analysis or a scree plot, automatically mapping the data to a lower-dimensional space based on point similarities [50].

In order to better analyse the formed clusters, t-SNE was applied to a small subset of 5,000 observations . While some agglomerations were detected, the number of distinct groups appeared high, and the boundaries between them were not clearly defined, providing limited insights for clustering. Therefore, the decision was made to proceed with the PCA technique, which facilitates the application of clustering algorithms.

2) Clustering the data

After deciding to use PCA to reduce data dimensionality to three principal components and naming them, clustering was performed to determine whether clients could be successfully grouped based on their characteristics. The performance of four different methods, namely K-Means, Ward's method, Complete Linkage, and DBSCAN, was analyzed to identify the most effective approach.

Firstly, the k-means algorithm was applied, and two methods were used, which were the elbow and silhouette methods, in order to determine the optimal number of clusters for K-Means. The elbow method identifies the point where adding more clusters has minimal impact [51]. Similarly, the silhouette method calculates silhouette scores for clusters within a

specific range, selecting the number with the highest score [51].

The Elbow Method indicated a noticeable inflection point around k = 4, while the Silhouette Method reached its peak at k = 6, although the difference between k = 5 and k = 6, and even k = 4, was minimal. Since no clear conclusion could be drawn, k = 3, 4, 5 and 6 were further evaluated using silhouette analysis. The results showed silhouette scores of 0.7088, 0.7632, 0.7938 and 0.8004, respectively. This indicates that 6 is the optimal number of clusters, although 5 is also a good choice because the silhouette values are very similar.

Secondly, hierarchical clustering was performed using the Ward linkage method. To determine the optimal number of clusters, a dendrogram was analyzed, which provides insight into how observations are grouped step by step, enabling the identification of natural divisions in the data. A significant jump in the distance between successive merges was observed just before forming five main clusters, indicating that cutting at this point would be a suitable choice. In addition, a silhouette analysis was performed to evaluate the optimal number of clusters, considering values of k ranging from 3 to 7. The average silhouette scores obtained for each k were as follows: 0.6657 for k=3, 0.7596 for k=4, 0.7925 for k=5, 0.7980 for k=6, and 0.7262 for k=7. The best score was obtained for k=6, however, as in the K-Means method, k = 5 produced a similar silhouette value.

Thirdly, hierarchical clustering was applied using the Complete Linkage method. In this approach, a clear separation of the data was observed at a dissimilarity level of approximately 1.5 to 2.0. Cutting the dendrogram at this height suggested the presence of three well-defined clusters, as further merges beyond this point involved a substantial increase in dissimilarity. Furthermore, a silhouette analysis was conducted for a range of cluster values (k = 3 to 7). The average silhouette scores obtained were 0.3618 for k=3, 0.5217 for k=4, 0.5210 for k=5, 0.4373 for k=6 and 0.4022 for k=7. The highest silhouette score was obtained for k=4, however, as in the K-Means and Ward method, k = 5 produced a similar silhouette value.

Lastly, regarding the DBSCAN algorithm, there are two essential parameters that must be manually defined: *eps* (the maximum distance between two points for them to be considered neighbors) and *min_samples* (the minimum number of points required to form a dense region). Due to the importance of these parameters to DBSCAN, a grid search technique — a method that systematically tests combinations of parameter values to find the best outcome — was employed to determine the optimal configuration, which returned the values of *eps* = 0.3 and *min_samples* = 11. With these parameters, the DBSCAN model produced 4 clusters, achieving a silhouette score of 0.6651.

In order to undestrand which method is best to cluster the data, table 5 summarizes the experiments conducted for each method with different numbers of clusters. For DBSCAN, there was no need for experiments, as the algorithm automatically selects the optimal number of clusters.

TABLE 5: Silhouette scores obtained by algorithm and number of clusters

Algorithm	**k = 3**	**k = 4**	**k = 5**	**k = 6**	**k = 7**
K-Means	0.7088	0.7632	0.7938	**0.8004**	–
Ward Linkage	0.6657	0.7596	0.7925	**0.7980**	0.7262
Complete Linkage	0.3618	**0.5217**	0.5210	0.4373	0.4022
DBSCAN	–	**0.6651**	–	–	–

Looking at the silhouette scores, it was observed that both K-Means and Ward Linkage performed best with 6 clusters, while Complete Linkage and DBSCAN, showed better performance with 4 clusters, although their silhouette scores remained lower compared to K-Means and Ward Linkage. Furthermore, although K-Means and Ward Linkage yielded the highest performance with 6 clusters, the silhouette values for 5 clusters were remarkably similar. Therefore, considering the trade-off between model complexity and minimal information loss, it seemed reasonable to opt for 5 clusters instead of 6, as this resulted in only a marginal decrease in the silhouette score. In terms of the best method, it was found that K-Means slightly outperformed Ward Linkage, making it the best choice for this dataset.

After choosing the best method and optimal number of clusters, the interpretation and naming of the clusters was made, both with the PC´s and the original variables.

Firstly, in order to better understand the characteristics of each cluster, the average values of the three PC´s and the debt fulfillment rate (proportion of good payers - the value 1 in the variable 'Cumprimento da dívida') were calculated for all observations grouped by cluster, which are presented in Table 6. The table also includes the percentage distribution of observations across clusters, providing insight into the relative size and prevalence of each group within the dataset.

TABLE 6: Average PC scores, debt fulfillment percentage, and observation distribution per cluster

Variable	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Engagement vs. Churn Risk	-0.0700	0.7393	-0.7248	-0.5025	0.7840
Active Indebtedness	-0.3998	0.1061	0.1136	0.7699	0.9957
Payment Cycle Completion	-0.2424	0.1998	0.3184	-0.5747	-0.5151
Cumprimento da dívida (%)	56.22%	0.22%	98.91%	75.51%	0.15%
Observation Percentage (%)	35.11%	30.33%	24.82%	7.46%	2.26%

- Cluster 0: Mixed profile (35.11% of observations)
  - Engagement vs. Churn Risk: -0.0700 (neutral)
  - Active Indebtedness: -0.3998 (low)
  - Payment Cycle Completion: -0.2424 (slightly incomplete)

  - Debt Fulfillment Rate: 56.22%

  This cluster represents a neutral or mixed group. Customers show low indebtedness but have not fully completed their payment cycles. With just over half paying on time, these profiles do not exhibit clear risk or reliability patterns.
- Cluster 1: High-risk churners (30.33% of observations)
  - Engagement vs. Churn Risk: 0.7393 (high churn risk)
  - Active Indebtedness: 0.1061 (moderate)
  - Payment Cycle Completion: 0.1998 (slightly complete)
  - Debt Fulfillment Rate: 0.22%

  These customers are likely deactivated and show poor payment behavior, despite not having the highest debt levels. Their combination of churn risk and almost nonexistent payment fulfillment makes them a priority for credit recovery or write-off.
- Cluster 2: Ideal customers (24.82% of observations)
  - Engagement vs. Churn Risk: -0.7248 (high engagement)
  - Active Indebtedness: 0.1136 (low–moderate)
  - Payment Cycle Completion: 0.3184 (mostly complete)
  - Debt Fulfillment Rate: 98.91%

  These are highly reliable and engaged customers, with regular SEPA usage and low-to-moderate indebtedness. They have largely completed their financial obligations and represent an ideal profile for retention and commercial campaigns.
- Cluster 3: Indebted but engaged (7.46% of observations)
  - Engagement vs. Churn Risk: -0.5025 (engaged)
  - Active Indebtedness: 0.7699 (high)
  - Payment Cycle Completion: -0.5747 (incomplete)
  - Debt Fulfillment Rate: 75.51%

  These customers are still involved and tend to pay, but carry significant debt and have not yet completed their repayment journey. With proper support, this group may migrate toward the ideal profile.
- Cluster 4: Disengaged and heavily indebted (2.26% of observations)
  - Engagement vs. Churn Risk: 0.7840 (high churn risk)
  - Active Indebtedness: 0.9957 (very high)
  - Payment Cycle Completion: -0.5151 (incomplete)
  - Debt Fulfillment Rate: 0.15%

  This segment includes the riskiest customers: likely deactivated, with high outstanding balances and poor payment performance. They require immediate attention for loss mitigation or collections.

Secondly, in order to complement the PCA-based clustering interpretation, the same clusters were analyzed using the original dataset with real, non-normalized values and without the transformation of categorical to numerical variables, which are presented in Table 7. This allowed for a more realistic and intuitive understanding of the characteristics of each group.

TABLE 7: Average original variable values per cluster (non-normalized)

Variable	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
'Saldo Total'	61.89	248.50	63.06	333.77	484.65
'30 a 60'	0.16	0.06	0.00	0.14	0.16
'SAS - Ciclo'	9.50	7.55	17.71	14.03	11.12
'VEP - Nº Prestações'	0.04	0.03	0.00	30.23	35.96
'VEP - Nº Prestações Venc'	0.00	0.00	0.00	1.11	13.20
'PPP - Nº Prestações'	0.38	0.76	0.02	0.76	0.01
'PPP - Valor do Plano'	7.90	0.02	0.36	16.24	0.31
'Longevidade'	345.15	801.63	77.89	37.64	993.93
'Data Criação NIF (Diferença Dias)'	4247.28	2881.60	3949.66	3044.93	2046.97
'Cumprimento da dívida' (%)	56.22%	0.22%	98.91%	75.51%	0.15%
'CAT_ CONTA_ CONTRATO'	PTC	MEO	PTC	PTC	PTC
'Sub-Segmento'	Consumo	Consumo	Consumo	Consumo	Consumo
'Estrangeiro'	N	N	N	N	N
'Dissolução'	N	N	N	N	N
'Insolvência'	N	N	N	N	N
'Insolvência - PER'	N	N	N	N	N
'VEP'	N	N	N	S	S
'PPP'	N	N	N	N	N
'Acompanhamento'	N	N	N	N	N
'Venda Créditos'	N	N	N	N	N
'SEPA'	N	N	S	S	N
'Serviço_ Estado'	Ativo	Desativos	Ativo	Ativo	Desativos

Cluster 0 *Previous interpretation:* Mixed group with moderate repayment rate (56.22%) and no clear pattern. *Real data insights:* Customers in this cluster have low average debt ('Saldo Total' $\approx$ 61.89€) and almost no active credit plans ('VEP' $\approx$ 0.04, 'PPP' $\approx$ 0.38). Their average tenure is 345 days, with relatively old NIFs (11.6 years). Categorical variables show they are typically 'PTC', non-'SEPA' users, non-'VEP' and their accounts are active. *Conclusion:* Consistent with the initial interpretation. A diverse and stable group with low debt and moderate repayment behavior, but no strong indicators of either risk or excellence.

Cluster 1 *Previous interpretation:* High-Risk customers with extremely low repayment rate (0.22%). *Real data insights:* These clients have average debt ('Saldo Total' ≈ 248.50€), no active 'VEP' or 'PPP' plans, and high tenure (801 days), but relatively young NIFs. Most are 'MEO', 'SEPA'-compliant, non-'VEP' and their accounts are deactivated. *Conclusion:* Fully aligned with the previous interpretation. A highly disengaged and risky segment, despite average debt, these customers exhibit extremely low repayment rates, indicating poor payment behavior.

Cluster 2 *Previous interpretation:* Ideal customers (repayment rate 98.91%). *Real data insights:* These clients have low to moderate average debt ('Saldo Total' ≈ 63.06€), almost no active credit agreements ('VEP' ≈ 0.03, 'PPP' ≈ 0.00). Their average tenure is short (77 days), with relatively old NIFs (10.8 years). They are 'PTC', 'SEPA' users, non-'VEP' and accounts are active. *Conclusion:* Fully consistent with the previous analysis. A reliable and highly engaged group with excellent payment behavior, making them valuable for retention and commercial efforts.

Cluster 3 *Previous interpretation:* Indebted but engaged (repayment rate 75.51%). *Real data insights:* Clients in this cluster have significant average debt ('Saldo Total' ≈ 333.77€), several active installments ('VEP' ≈ 30.23, 'PPP' ≈ 0.76), moderate overdue installments (1.11) and very high plan value of 'PPP'. Their average tenure is short (38 days), with NIFs averaging 8.3 years. They are typically 'PTC', 'SEPA'-compliant, with 'VEP' and accounts active. *Conclusion:* Matches the initial interpretation. High-risk but potentially recoverable customers, indebted but still engaged and regularly paying. With adequate support, they may migrate towards the ideal profile.

Cluster 4 *Previous interpretation:* Heavily indebted and disengaged (repayment rate 0.15%). *Real data insights:* This group has the highest average debt ('Saldo Total' ≈ 484.65€), a large number of active and overdue installments ('VEP' ≈ 35.96, overdue ≈ 13.20), and almost no 'PPP' activity. Their tenure is long (994 days), with relatively young NIFs (5.6 years). They are mainly 'PTC', non-'SEPA', with 'VEP' and accounts deactivated. *Conclusion:* Fully confirms the initial interpretation. The most critical risk segment, requiring urgent attention for loss mitigation or collections.

The interpretation using real variables complements is consistent with the analysis conducted through PC´s, with both approaches identifying similar patterns within the clusters. For example, in Cluster 0, the PC´s interpretation pointed to a mixed and moderate profile, which is corroborated by the real variable analysis, showing customers with low debt and moderate payment behavior. In Cluster 2, the combination of a high debt fulfillment rate and low debt in the real variables reinforces the PC´s interpretation of an ideal group with excellent payment behavior.

Figure 1 shows the distribution of the clusters in space, based on the three principal components. This visualization

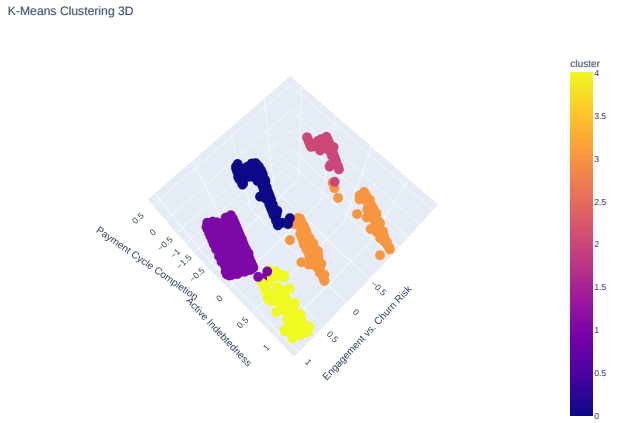


FIGURE 1: 3D visualization of K-Means clusters - top-down perspective

was generated using a sample of five thousand observations to improve clarity and interpretability. This 3D visualization of the clustering results reveals not only the general separation between groups, but also some potential substructures within the clusters themselves. For example, Cluster 3 appears to be spatially divided into two distinct subgroups, despite being labeled with the same cluster ID. This may indicate an internal heterogeneity within the cluster or a possible overlap with another cluster.

This visualization is particularly helpful for understanding the spatial distribution of the clusters. It highlights their relative positioning, reveals which ones are close to each other, and reinforces the observation of possible overlaps or internal divisions, such as the one noted in Cluster 3.

C. SECOND EXPERIENCE

While performing the first experience, it was noted that, in the analysis of feature importance, some variables consistently emerged as a Top15 feature throughout the supervised learning models, which are listed in table 3. Besides this, variables with a high KMO score in unsupervised learning can be interpreted as significant for unsupervised learning techniques. When comparing these variables across both supervised and unsupervised learning approaches, the following were found to be common to both: 'Longevidade', '30 a 60', 'Saldo Total', 'SEPA', 'PPP', 'VEP - Nº Prestações Venc', 'Data Criação NIF (Diferença Dias)', and 'Serviço_Estado_Desativos'. This is not a mere coincidence, rather their presence in both analyses suggests that these variables are not only highly informative in predicting debt payment (supervised learning) but also strongly correlated with other key features, effectively capturing underlying structures within the dataset (unsupervised learning).

Hence, a further analysis was conducted to evaluate the performance of all models using only these variables in order to understand if the dataset can be simplified.

Regarding supervised learning models, first, the correlations were analyzed. These remained consistent with the previous analysis, both in relation to the target variable and between the independent variables themselves. It is important to highlight that no pairwise correlations exceeded 0.7 or fell below -0.7, and therefore no variable was removed on the basis of multicollinearity. The analysis and comparison with the results obtained using the full set of features show that the overall performance loss across all models does not exceed 3% in the evaluation metrics. Considering this trade-off — a substantial reduction in dimensionality with only a slight decrease in performance — the use of only the common features proves to be a highly valuable and efficient approach.

Since the unsupervised analysis already involved dimensionality reduction through PCA, which reduced the feature space to three principal components, performing a second experiment using a pre-selected subset of features was deemed unnecessary, due to the fact that the results would be very similar. Therefore, further reducing the feature set prior to PCA would not provide additional insight or benefit to the clustering results.

D. COMPARISON WITH EXISTING LITERATURE

1) Supervised Learning

The results obtained from the supervised learning models were compared to previous studies related to churn detection, profitability prediction, and fraud detection in the telecommunications sector. Although these studies focus on slightly different problems than the one discussed in this work, they provide useful performance benchmarks. The best performing model in this study, Random Forest, achieved comparable performance to the authors of studies [27] and [5]. For the remaining studies, the predictive performance of the models in this research was generally superior.

2) Unsupervised Learning

Regarding the literature on unsupervised learning techniques, previous studies focused on using different clustering methods. A common pattern observed is the strong performance of K-Means clustering, which aligns with the findings of this research. Unlike this study, the reviewed works did not apply dimensionality reduction prior to clustering, so no direct comparison on that aspect is possible.

3) Supervised and Unsupervised Learning combined

Several studies in the literature have explored hybrid approaches that combine supervised and unsupervised learning methods. In most cases, the workflow involves applying dimensionality reduction techniques, such as PCA or t-SNE, to the dataset before applying supervised learning models. This reduces data complexity and improves model efficiency and generalization. Such integrated approaches have generally led to improved model robustness and accuracy.

In contrast, the present work applied supervised and unsupervised learning methods independently. Nevertheless, one experiment conducted in this work bears some resemblance to these hybrid methodologies. In a secondary analysis, dimensionality reduction was performed based on feature importance derived from supervised learning models and KMO analysis in the context of unsupervised learning. The performance of supervised models was then compared before and after the reduction, although this was not the primary focus of this work.

Although most studies reviewed in the literature apply unsupervised learning techniques prior to supervised learning, the approach taken by the authors of [36] is more similar to the one adopted in this work. In their work, XGBoost was first used to predict customer churn, followed by K-Means clustering to segment the predicted churners. This sequential approach enabled the company to target retention strategies more effectively, resulting in a 0.93% reduction in churn.

V. DISCUSSIONS AND CONCLUSIONS

A. OVERVIEW OF THE STUDY

This study aimed to analyze and model customer behavior regarding debt payment within a Portuguese telecommunications company. The process began with comprehensive data preprocessing, including handling missing values, outlier detection and treatment, feature engineering, normalization, categorical variable encoding, target variable creation, and dataset balancing.

Subsequently, various supervised learning algorithms were applied to predict the newly created target variable 'Cumprimento da divida'. The models tested were Logistic Regression, Random Forest, k-NN, and Neural Networks. Among these, the Random Forest model achieved the highest performance, with an accuracy of 99%, making it the most effective model for this task.

Following the supervised learning phase, unsupervised learning techniques were employed in two main steps. First, dimensionality reduction techniques were evaluated, with both PCA and t-SNE being tested, with PCA being selected as the preferred technique due to its superior interpretability and performance, and three principal components being retained and meaningfully named. Second, clustering techniques were applied to the PCA-transformed dataset, including K-Means, Ward's method, Complete Linkage, and DBSCAN. K-Means clustering outperformed the other methods, and the optimal number of clusters was determined to be five. These clusters were interpreted using both the principal components and the original dataset features to gain a deeper understanding of customer behavior.

B. DISCUSSION OF RESULTS

Although this study did not implement an integrated approach in the traditional sense, it demonstrated how supervised and unsupervised learning methods can be used in a complementary way. For example, after classifying clients as compliant or non-compliant using the supervised model, one can analyze their corresponding clusters to gain additional insight. If a compliant client falls into a cluster mostly com-

posed of non-compliant clients, this could signal a potential future risk.

This dual-layer analysis, prediction followed by segmentation, offers an enriched perspective, similar to that employed by the authors of article [36]. Furthermore, the use of PCA allowed the identification of underlying patterns and dimensionality reduction, providing interpretable axes for visual cluster exploration and reducing computational complexity, as shown in Figure 1.

From a business perspective, this approach enables the company to test hypothetical scenarios by altering specific client features to assess their potential impact on payment behavior. This can be particularly useful for high-risk clients with substantial debt, allowing the company to make informed decisions, such as whether to continue or terminate the contract, or to implement targeted interventions.

For instance, imagine a company receives a business client in March who has debt but is a good payer. Hypothetical scenarios could be tested in order to predict if the client will continue to be a good payer. For example, if the client does not have a '30 a 60' value in March but has one in April, can the company expect the client to remain a good payer? Or, if the client used to pay via 'SEPA' in March but switches to a different payment method in April, will they continue to be a good payer? To test these scenarios, one would simply alter the value of the '30 a 60' variable or the 'SEPA' payment method and observe the supervised model's prediction, keeping other variables constant. As this is a more manual analysis, it would be most useful for clients with the highest debt, as they are causing the most significant financial loss to the company. If such analyses reveal that a good payer might turn into a bad payer if their 'Saldo Total' increases, the company could take preventive actions, such as terminating the contract or encouraging the client to avoid accumulating a higher balance.

Unsupervised analysis also offers value in identifying clients within ambiguous clusters such as clusters 0 and 3, where compliance behavior is less conclusive. By comparing these clients' characteristics with those from more definitive clusters, one can estimate whether they are likely to remain compliant or not. For example, if a client in cluster 0 is a good payer and has characteristics similar to those in cluster 2 (good payers), it is likely that the client will continue to be a good payer. Another example is a client in cluster 0 who is a non-payer but has characteristics similar to clients in clusters 1 or 4 (non-payers), indicating a higher likelihood of transitioning from a non-payer to a good payer.

C. CONCLUSION

In conclusion, this study successfully demonstrated how both supervised and unsupervised learning techniques can be used to understand and predict client payment behavior in the telecommunications sector. The findings are not only consistent with existing literature but also offer practical applications that can support more informed decision-making of the telecom company. Although initially applied independently, the subsequent joint analysis of both learning paradigms proved to be complementary, providing a robust analytical framework for future business applications and research developments.

D. RECOMMENDATIONS FOR FUTURE WORK

Several directions for future research can be proposed based on the findings of this study:

- **Focused Clustering on Non-Compliant Clients:** Conduct clustering exclusively on non-compliant clients to identify subgroups with distinct behavioral profiles, which could enhance the segmentation strategy and risk assessment.
- **Alternative Preprocessing Strategies:** Experiment with different preprocessing techniques (for example, normalization methods, outlier handling) to improve the performance of clustering algorithms, especially with the goal of increasing the silhouette score.
- **Application of Time Series Analysis:** Utilize time series forecasting models to predict the target variable based on historical debt behavior. This could offer an automatic and dynamic alternative to the static approach used in this study.
- **Simulation of Feature Changes:** Expand on the idea of testing client profiles by simulating changes in individual features to evaluate their impact on predicted compliance, supporting proactive risk management strategies.

REFERENCES

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," J Big Data, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0191-6.

[2] H. Jain, A. Khunteta, and S. Srivastava, "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost," in Procedia Computer Science, Elsevier B.V., 2020, pp. 101–112, doi: 10.1016/j.procs.2020.03.187.

[3] J. Gupta, "Data Analytics in Telecommunications," 2024. [Online]. Available: https://www.quantexa.com/resources/data-analytics-in-telecommunications/. [Accessed: Dec. 23, 2024].

[4] N. Nazareth and Y. V. Ramana Reddy, "Financial applications of machine learning: A literature review," Elsevier Ltd., Jun. 2023, doi: 10.1016/j.eswa.2023.119640.

[5] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," Results in Control and Optimization, vol. 14, Mar. 2024, doi: 10.1016/j.rico.2023.100342.

[6] C. Wang, C. Rao, F. Hu, X. Xiao, and M. Goh, "Risk assessment of customer churn in telco using FCLCNN-LSTM model," Expert Syst Appl, vol. 248, Aug. 2024, doi: 10.1016/j.eswa.2024.123352.

[7] F. Naz and F. Popowich, "Mining Retail Telecommunication Data to Predict Profitability," 2019, Unpublished.

[8] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An Unsupervised Machine Learning Algorithms: Comprehensive Review," International Journal of Computing and Digital Systems, vol. 13, no. 1, pp. 911–921, 2023, doi: 10.12785/ijcds/130172. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.085161879817&doi=10.12785%2fijcds%2f130172&partnerID=40&md5=4c8f17e978de881b785605ff53d8e635.

[9] V. Renuka Devi, G. Bharathi, and G. V. S. N. R. V. Prasad, "Prediction of customer churn in telecom sector using clustering technique," Int. J. Eng. Adv. Technol., vol. 8, no. 6 Special Issue 2, pp. 826–832, 2019, doi: 10.35940/ijeat.F1207.0886S219. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85073727612&

doi=10.35940%2fijeat.F1207.0886S219&partnerID=40&md5=ec4e8302e0601fb9f7bb76f0534ce10b.

[10] H. Wibowo and K. P. Sinaga, "Telecommunication Analytics Based on Customer Segmentation Using Unsupervised Algorithms," in 3rd Int. Conf. on Cybernetics and Intelligent Systems, ICORIS 2021, 2021, doi: 10.1109/ICORIS52787.2021.9649598. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124365145&doi=10.1109%2fICORIS52787.2021.9649598&partnerID=40&md5=e0f24f3e74cd869fc3ac525ac1b38c90.

[11] Data Science Central, "CRISP-DM: A Standard Methodology to Ensure a Good Outcome," 2020. [Online]. Available: https://www.datasciencecentral.com/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome/. [Accessed: Mar. 3, 2025].

[12] Data Science Process Alliance, "CRISP-DM: The Standard Data Mining Process," 2021. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/. [Accessed: Mar. 3, 2025].

[13] M. Đukanović, L. Kašćelan, S. Vuković, I. Martinović, and M. Ćalasan, "A machine learning approach for time series forecasting with application to debt risk of the Montenegrin electricity industry," Energy Reports, vol. 9, pp. 362–369, Sep. 2023, doi: 10.1016/j.egyr.2023.05.240.

[14] M. Abdoli, M. Akbari, and J. Shahrabi, "Bagging Supervised Autoencoder Classifier for Credit Scoring," 2021. [Online]. Available: http://arxiv.org/abs/2108.07800. [Accessed: Aug. 2021].

[15] S. Tyagi, "Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions," 2022. [Online]. Available: https://doi.org/ Accessed: Apr. 1, 2025.

[16] G. Li, H. D. Ma, R. Y. Liu, M. Di Shen, and K. X. Zhang, "A two-stage hybrid default discriminant model based on deep forest," Entropy, vol. 23, no. 5, 2021, doi: 10.3390/e23050582.

[17] V. L. N. Gorle and S. Panigrahi, "A semi-supervised Anti-Fraud model based on integrated XGBoost and BiGRU with self-attention network: an application to internet loan fraud detection," Multimed Tools Appl, vol. 83, no. 19, pp. 56939–56964, 2024, doi: 10.1007/s11042-023-17681-z.

[18] R. Najem, M. Fakhouri Amr, A. Bahnasse, and M. Talea, "Advancements in Artificial Intelligence and Machine Learning for Stock Market Prediction: A Comprehensive Analysis of Techniques and Case Studies," Procedia Computer Science, Elsevier B.V., pp. 198–204, 2024, doi: 10.1016/j.procs.2023.12.193.

[19] C. Manteigas and N. António, "Understanding and predicting lapses in mortgage life insurance using a machine learning approach," Expert Systems with Applications, vol. 255, 2024, doi: 10.1016/j.eswa.2024.124753.

[20] M. A. Khan, M. A. I. Khan, M. Aref, and S. F. Khan, "Cluster & rough set theory based approach to find the reason for customer churn," Int. J. Appl. Bus. Econ. Res., vol. 14, no. 1, pp. 439–455, 2016. [Online]. Available:https://www.scopus.com/inward/record.uri?eid=2-s2.0-84971264795&partnerID=40&md5=e11059c47a9363faa40261de98666166.

[21] M. Gök, T. Özyer, and J. Jida, "A case study for the churn prediction in Turksat internet service subscription," in Proc. IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining (ASONAM), 2015, pp. 1220–1224, doi: 10.1145/2808797.2808821.

[22] S. F. Bilal, A. A. Almazroi, S. Bashir, F. H. Khan, and A. A. Almazroi, "An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry," PeerJ Computer Science, vol. 8, 2022, doi: 10.7717/PEERJ-CS.854.

[23] M. Bagri, J. K. Singh, M. K. Abhilash, R. Sunitha, and S. Kumar, "Churn Analysis in Telecommunication Industry," in 2018 Int. Conf. on Automation and Computational Engineering (ICACE 2018), pp. 126–132, 2018, doi: 10.1109/ICACE.2018.8686852. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064883375&doi=10.1109%2fICACE.2018.8686852&partnerID=40&md5=a934853ddac1e65ab5b84850d19b0297.

[24] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," Journal of Network and Computer Applications, vol. 68, pp. 90–113, Jun. 2016, doi: 10.1016/j.jnca.2016.04.007.

[25] B. Garimella, G. V. S. N. R. V. Prasad, and M. H. M. K. Prasad, "Churn prediction using optimized deep learning classifier on huge telecom data," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 3, pp. 2007–2028, Mar. 2023, doi: 10.1007/s12652-021-03413-4.

[26] S. S. Poudel, S. Pokharel, and M. Timilsina, "Explaining customer churn prediction in telecom industry using tabular machine learning models," Machine Learning with Applications, vol. 17, p. 100567, Sep. 2024, doi: 10.1016/j.mlwa.2024.100567.

[27] J. Kristian Vieri, T. Ai Munandar, D. Budi Srisulistiowati, D. Handayani, A. No, and T. Sri Lestari, "Comparative Study of Classification Algorithms for Customer Decisions on Telecommunication Products Using Supervised Learning," 2023.

[28] S. M. Shrestha and A. Shakya, "A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal," in Procedia Computer Science, Elsevier B.V., 2022, pp. 652–661, doi: 10.1016/j.procs.2022.12.067.

[29] L. Dheekollu et al., "Modeling approaches for Silent Attrition prediction in Payment networks," in Proc. 20th IEEE Int. Conf. on Machine Learning and Applications (ICMLA 2021), Institute of Electrical and Electronics Engineers Inc., 2021, pp. 409–414, doi: 10.1109/ICMLA52953.2021.00070.

[30] A. Chouiekh and E. H. I. El Haj, "ConvNets for fraud detection analysis," in Procedia Computer Science, Elsevier B.V., 2018, pp. 133–138, doi: 10.1016/j.procs.2018.01.107.

[31] X. Hu, H. Chen, S. Liu, H. Jiang, G. Chu, and R. Li, "BTG: A Bridge to Graph machine learning in telecommunications fraud detection," Future Generation Computer Systems, vol. 137, pp. 274–287, Dec. 2022, doi: 10.1016/j.future.2022.07.020.

[32] A. Gaikwad, T. Ghodke, A. Jadhav, M. Pande, and S. Mirchandani, "Profit Prediction for Businesses using Machine Learning Algorithms," in Proc. 8th Int. Conf. on Communication and Electronics Systems (ICCES 2023), Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1039–1044, doi: 10.1109/ICCES57224.2023.10192797.

[33] K. K. Sharma, A. Seal, E. Herrera-Viedma, and O. Krejcar, "An enhanced spectral clustering algorithm with s-distance," Symmetry, vol. 13, no. 4, 2021, doi: 10.3390/sym13040596. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104180766&doi=10.3390%2fsym13040596&partnerID=40&md5=ab8493bac9ae480e0b7d38afbceca2a5.

[34] K. I. Shahid and S. Chaudhury, "Scalable clustering and applications," in ACM Int. Conf. Proceeding Series, 2016, doi: 10.1145/3009977.3010073. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85014856715&doi=10.1145%2f3009977.3010073&partnerID=40&md5=aa9e6d31608b37f8462d4f119593fc5d.

[35] S. M. Kostić, M. I. Simić, and M. V. Kostić, "Social network analysis and churn prediction in telecommunications using graph theory," Entropy, vol. 22, no. 7, 2020, doi: 10.3390/e22070753. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088565953&doi=10.3390%2fe22070753&partnerID=40&md5=18119c498a2915d2a410af4b1c19d062.

[36] Y. Gu, A. R. Malicdem, J. S. Dela Cruz, and T. D. Palaoag, "Using big data analysis to retain customers for telecom industry," in ACM Int. Conf. Proceeding Series, pp. 38–43, 2019, doi: 10.1145/3330482.3330510. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071128980&doi=10.1145%2f3330482.3330510&partnerID=40&md5=fb5b1d286ab7611b24283314e56f1449.

[37] H. Nutalapati, A. Hayat, R. Zheng, C. H. Li, N. Prakoso, and N. M. Tiglao, "Machine Learning-Based Predictive Analytics for Customer Churn in the Telecom Industry," in 2024 Int. Symp. on Networks, Computers and Communications (ISNCC 2024), 2024, doi: 10.1109/ISNCC62547.2024.10758995. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85212228512&doi=10.1109%2fISNCC62547.2024.10758995&partnerID=40&md5=c6d2b84bf471c9d1024c26a30e3be74c.

[38] S. Preetha and R. Rayapeddi, "Predicting customer churn in the telecom industry using data analytics," in Proc. 2nd Int. Conf. on Green Computing and Internet of Things (ICGCIoT 2018), pp. 38–43, 2018, doi: 10.1109/ICGCIoT.2018.8753096. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069518962&doi=10.1109%2fICGCIoT.2018.8753096&partnerID=40&md5=a97635ff406fb7bc2f5c7f98bdb192fc.

[39] K. F. Googerdchi, S. Asadi, and S. M. Jafari, "Customer churn modeling in telecommunication using a novel multiobjective evolutionary clustering-based ensemble learning," PLoS ONE, vol. 19, no. 6, Jun. 2024, doi: 10.1371/journal.pone.0303881. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195438770&doi=10.1371%2fjournal.pone.0303881&partnerID=40&md5=95a03e3becbfe9d75e28b60edd5840c3.

[40] E. Sivasankar and J. Vijaya, "Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network," Neural Comput. Appl., vol. 31, no. 11, pp. 7181–7200, 2019, doi: 10.1007/s00521-018-3548-4.

[Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047660042&doi=10.1007%2fs00521-018-3548-4&partnerID=40&md5=d4769472159ee4956b1cdfe6e26b5bad.

[41] R. A. Moral, Z. Chen, S. Zhang, S. McClean, G. R. Palma, B. Allan, and I. Kegel, "Profiling Television Watching Behavior Using Bayesian Hierarchical Joint Models for Time-to-Event and Count Data," IEEE Access, vol. 10, pp. 113018–113027, 2022, doi: 10.1109/ACCESS.2022.3215682. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140712174&doi=10.1109%2fACCESS.2022.3215682&partnerID=40&md5=6e4d5547e48c8f237cb24b3fe5561b2a.

[42] H. K. Gianey and R. Choudhary, "Comprehensive Review On Supervised Machine Learning Algorithms," in Proc. 2017 Int. Conf. on Machine Learning and Data Science (MLDS 2017), IEEE, 2017, pp. 38–43, doi: 10.1109/MLDS.2017.11.

[43] DataCamp, "Winsorized Mean: What You Need to Know to Handle Outliers," 2025. [Online]. Available: https://www.datacamp.com/tutorial/winsorized-mean. Accessed: Mar. 7, 2025.

[44] DataCamp, "Normalization in Machine Learning: A Complete Guide," 2023. [Online]. Available: https://www.datacamp.com/tutorial/normalization-in-machine-learning. Accessed: Mar. 7, 2025.

[45] M. Rafie, "Techniques for Converting Categorical Data into Numerical Data," 2023. [Online]. Available: https://medium.com/@rafiemon71/techniques-for-converting-categorical-data-into-numerical-data-f1c9d0a3863f. Accessed: Mar. 10, 2025.

[46] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," in 2020 11th Int. Conf. on Information and Communication Systems (ICICS), pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[47] Investopedia, "Multicollinearity," 2023. [Online]. Available: https://www.investopedia.com/terms/m/multicollinearity.asp. Accessed: Mar. 12, 2025.

[48] H. Arsham and M. Lovric, "Bartlett's Test," Int. Encycl. of Statistical Science, vol. 2, pp. 20–23, Mar. 2011, doi: 10.1007/978-3-642-04898-2_132.

[49] A. Shkeer and Z. Awang, "Exploring the Items for Measuring the Marketing Information System Construct: An Exploratory Factor Analysis," Int. Rev. of Management and Marketing, vol. 9, pp. 87–97, Oct. 2019, doi: 10.32479/irmm.8622.

[50] H. Liu, J. Yang, M. Ye, S. C. James, Z. Tang, J. Dong, and T. Xing, "Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data," J. Hydrol., vol. 597, p. 126146, 2021, doi: 10.1016/j.jhydrol.2021.126146.

[51] Ankita, "K-Mean: Getting the Optimal Number of Clusters," Analytics Vidhya, May 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/. Accessed: Apr. 1, 2025.

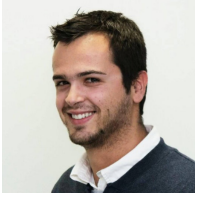
FILIPE O. F. ARSÉNIO received the B.Sc. degree in Economics from the Lisbon School of Economics and Management (ISEG) in 2023, and is currently pursuing the M.Sc. degree in Business Analytics at ISCTE – University Institute of Lisbon. His dissertation focuses on machine learning applications for financial risk prediction in the telecommunications sector.

He is currently working as a Financial Data Scientist and Support Analyst at MEO, a major Portuguese telecommunications company, where he applies data-driven methods to financial process optimization and risk analysis. His research interests include machine learning, financial analytics, risk modelling, telecommunications data, and applied data science.

JOÃO PEDRO C. B. B. PAVIA received his PhD in Information Sciences and Technologies in 2022 and he is currently an Assistant Professor in the Digital Technologies Department of the School of Applied Digital Technologies at ISCTE - University Institute of Lisbon, where he coordinates the Bachelor Degree in Digital Technologies and Information Security.

He is a member of the IEEE and is also a member of COST - European Cooperation in Science and Technology, where he collaborates with professionals from various countries to develop solutions focused on the field of intelligent radio communications for seamless inclusive interactions. In addition, another of his collaborations is related to the development of physical layer security solutions for reliable and resilient 6G systems.

He has been involved as a local organizer and reviewer at several conferences and symposia. He is also a reviewer for several journals in the field of information sciences and technologies. His research interests include topics related to cybersecurity, wireless communication networks, artificial intelligence, and big data.



ANTÓNIO RAIMUNDO , who holds a PhD in Information Science and Technology from Iscte - University Institute of Lisbon, Portugal, is an Assistant Professor at the School of Applied Digital Technologies at Iscte Sintra, where he also coordinates the Bachelor's degree in Digital Technologies and Artificial Intelligence.

Currently, he is an Integrated Researcher at ISTAR-IUL, focusing his research on areas such as Machine Learning / Deep Learning, Large Language Models, Computer Vision, and Data Science. With several years of teaching experience, he has had the opportunity to be a trainer in short courses, postgraduate programs, and a speaker at various business events.