ISSN 2007-9737

Optimizing Credit Risk Prediction in the Financial Sector Using Boosting Algorithms: A Comparative Study with Financial Datasets

Renzo Orlando Villanueva Mora[1,*], Edwin Jonathan Escobedo Cárdenas[1,2]

[1]Universidad de Lima, Carrera de Ingeniería de Sistemas,
Facultad de Ingeniería,
Peru

[2] Grupo de Investigación en Optimización Combinatoria e Inteligencia Artificial (OPTIMIA)
Peru

20193654@aloe.ulima.edu.pe, eescobed@ulima.edu.pe

Abstract. This paper presents a study of credit risk, which is a significant concern for financial institutions. Despite advances in predictive models, there is still room for improvement in accurately assessing credit risk. This study focuses on developing a methodological process to predict credit risk in the financial sector using algorithms based on boosting techniques, such as XGBoost, LightGBM and Boosted Random Forest. We found that datasets with good accessibility and an appropriate variable distribution are contained in the UCI Machine Learning Repository. These datasets have the potential to outperform results with different metrics, such as the F-Score and the Area Under the Curve. The datasets used include Statlog German Credit Data, Statlog Australian Credit Approval, Bank Marketing, Credit Approval, and South German Credit Data. The approach involves feature engineering, exploratory data analysis, and hyperparameter tuning. Furthermore, we propose a new strategy that involves adding a column based on an unsupervised algorithm such as K means. Our results indicate that XGBoost performs better than LightGBM and Boosted Random Forest in different scenarios. Finally, the performance of these boosting-based models is superior to that of Boosted Decision Trees and Factorization Machine models from previous studies. These findings are important for financial institutions seeking an effective methodology to improve the rate of credit risk prediction.

Keywords. XGBoost, lightGBM, boosted random forest, boosting algorithms, credit risk, credit score, financial sector.

1 Introduction

Credit risk is a sensitive topic for financial institutions for various reasons, primarily due to credit. [26] states that credit risk is subject to external scrutiny, as both central banks and auditors closely monitor the extent to which financial institutions comply with Basel guidelines and International Financial Reporting Standards. Furthermore, it is estimated that around 60% of people who have a bank account request approval for a financial loan [25]. The emerging growth of people who acquire financial loans is significant. [36] showed that the financial sector could increase its profits by 80% due to an accurate assessment of credit risk. Thus, banking competition can predict future behavior before a very high loan becomes evident.

Regarding the consequences of credit risk. [35] determined that the main one is the reduction in income for the bank and the increase in interest for the client, this percentage increase varies depending on the economic situation of both the country and the borrower. In addition, financial institutions can adjust their interest rates to compensate for credit risk, which could result in higher rates for borrowers with higher credit risk.

[9] asserts that the primary cause of credit risk is poor assessment of the borrower and

ISSN 2007-9737

inadequate loan monitoring. This increases the risk that borrowers may not meet their financial obligations. Furthermore, 15% of clients experience complications due to sudden increases in interest rates, which increase the financial burden on borrowers and, consequently, the credit risk.

In response to credit risk, several studies have emerged with the objective of implementing and testing machine learning algorithms to predict these events with the highest accuracy. Among them are those presented by [7, 31].

The gap to be addressed, according to the analysis carried out, lies in the exploration of algorithms based on boosting to develop predictive models aimed at improving the Area Under the Curve (AUC) and the F-Score in the evaluation of credit risk. The need for companies to have these predictive models arises because despite incorporating all credit variables, there is still room for improvement in the results. For example, data sets provided by the UCI Machine Learning Repository mentioned by [7] provide financial information for an entity German “Statlog German Credit Data” and an Australian financial institution “Statlog Australian Credit Approval ”, achieving better results of 81.08% and 94.03% AUC, respectively. Furthermore, the datasets mentioned by [31] from the UCI Machine Learning Repository include Portuguese credit marketing data “Bank Marketing” with the best result of 55.35%, and information from the credit card application process “Credit approval”with the best result of 96.47%. Furthermore, the data set from a German financial institution “South German Credit Data” achieved a better result of 96.47%. Score. Taking into account [29]'s suggestion to implement boosting-based algorithms due to its high usability in machine learning models to improve the accuracy of credit risk prediction and address the room for improvement in results. This research asks the following questions: What are the best momentum-driven machine learning models for predicting credit risk? And what variables are the most significant for predicting credit risk?

To address the research questions posed, the main objective is to design a methodological process to predict credit risk in the financial sector using boosting techniques such as LightGBM, XGBoost, and Boosted Random Forest, using customer credit information. This implies first identifying the most relevant variables that influence credit risk. Next, identify and select momentum algorithms for credit risk prediction. Finally, evaluate the performance of the selected boosting models using feature engineering, cross-validation, and hyperparameter tuning techniques.

The rest of this study is organized as follows. Section 2 provides a review of related work, Section 3 describes the data sets and the methodology applied, Section 4 presents the experimental results and Section 5 concludes.

2 Related Work

Financial loans have been increasing each day in the financial sector, so allowing for accurate credit risk prediction will improve the benefits by 80% [36]. Due to this, the collected articles were analyzed to achieve the research objective, which is to develop predictive models of credit risk in the financial sector using boosting algorithms based on client credit information.

2.1 Datasets related to predict credit risk

As per the reviewed articles, different datasets with relevant variables for predicting credit risk using machine learning models were identified. A freely accessible website called the UCI Machine Learning Repository was identified, which contains a large number of datasets related to credit risk prediction, with more than six datasets available. This repository was referenced by [7] in the dataset of a German financial entity “Statlog German Credit Data” and an Australian financial entity “Statlog Australian Credit Approval”. Additionally, [31] references the website in the dataset of a German financial entity “South German Credit Data”, a marketing campaign of a Portuguese financial entity “Bank Marketing”, and credit card application information “Credit Approval”. These datasets, mentioned by [7, 31], are accessible and have a suitable distribution of variables.

ISSN 2007-9737

Regarding size and relevance, the Lending Club dataset, cited by [3, 5, 10, 23] provided by a financial services company is the largest with around 2 million records. Despite its breadth, the dataset contains outliers and missing values, which require greater computational cost and proper data treatment.

On the other hand, [18] offers a set of Tunisian financial data designed to project the future turnover of its affiliated companies. This dataset is intended to facilitate decision-making regarding loan approval and includes 10 features. [8] also identified and mentioned a dataset of small financial services companies in Italy called AIDA (Analisi Informatizzata Delle Aziende Italiane). This dataset contains information, records and financial ratios of all Italian companies that need to present their accounts. The total number of observations is approximately one million.

In another study, [27] employs a financial dataset from Taiwan containing 30,000 instances, of which 6636 are debtor cases, and includes variables such as education level and credit limit, among others. Finally, the American company Orange has a dataset with 3333 records, recognized by [2, 20]. Although the latter has an adequate number of records, this may not be sufficient due to its accessibility.

2.2 Metrics and algorithms used to evaluate a predictive model using machine learning

Researchers from the collected studies used various classification algorithms, such as the Multilayer Perceptron and XGBoost, to obtain accurate predictions about credit risk. They employed multiple metrics to evaluate the effectiveness of these algorithms. Moreover, researchers widely use the precision metric to evaluate the efficiency of their predictive models. Precision refers to the proportion of correctly identified positives out of all positive identifications [15]. It is important to note that precision can be applied to all classification algorithms to evaluate their performance.

In the studies presented by [17, 22, 37] the results indicate that the models achieve an accuracy of more than 83%. When reviewing the algorithms, it is highlighted that tree-based models, such as Random Forest, show high levels of accuracy. However, it is worth highlighting the implementation of boosting techniques such as Gradient Boosting, which [37] uses to achieve an accuracy of 99%.

[22] complement their precision analysis with other metrics like recall and F-score, offering a more comprehensive understanding of their models' performance. In their study, the XGBoost-KNearestNeighbor algorithm achieves an F-Score of 98.7%, indicating a high harmonic mean between precision and recall. On the other hand, the study presented by [12] reports a very low area under the curve. Specifically, the Logistic Regression model and the KNN model achieved areas under the curve of 50% and 54%, respectively. This indicates that the models have a deficiency in distinguishing between positive and negative classes.

In another of the analyzed studies, [6] implement various metrics such as the coefficient of determination (R^2), mean squared error (MSE), and mean absolute error (MAE) to evaluate the performance of the Multilayer Perceptron, XGBoost, and TabNet models. These metrics were used to measure the ability to fit the data and to quantify the accuracy of the predictions made by the models. On the other hand, in the study by [11], the F1-Score, cross-entropy and akaike information criterion are observed for measuring the Random Forest, which obtains an F1-Score of 91.90%. Subsequently, in the study presented by [21], metrics like Accuracy are used for Logistic Regression, Decision Trees, and Random Forest models. The best accuracy, which was 81.12%, was obtained in the regression model.

Finally, in the study presented by [31], various metrics like accuracy, matthews correlation coefficient, precision, recall, F-Score, True-positive Rate, True-negative Rate, False-negative Rate, False-positive Rate, area under the curve, and G-mean are observed to evaluate the proposed Factorization Machine model, obtaining a best F-Score of 96.47% and a lowest F-Score of 55.35%.

2.3 Preprocessing techniques and methodologies used in the financial sector for credit risk

Due to the impurity of different datasets, there is a need to perform cleaning. This involves applying different data preprocessing techniques in order to use these sets appropriately. [39] show that data preprocessing is critical to improving data quality, which in turn can lead to more accurate and efficient models.

[24] uses a combined methodology, the classifier assembly method and Bootstrap Aggregating (bagging). First, multiple bootstrap samples are generated from the dataset to train various classifiers such as Multilayer Perceptron, SVM, Decision Tress, and KNN. The final prediction is then obtained through a majority vote. Otherwise, to prepare the data it uses value normalization.

Alternatively, the most used technique is feature selection, which was used by [7, 14, 16, 22, 23]. Taking into account the relevance of selected features is crucial for model performance. Based on this, the methodology implemented by [22] is highlighted as consisting of two main stages: First, XGBoost is used for feature selection to retain only the most important features. Second, a multilayer perceptron (MLP) with ReLU activation functions is employed for credit risk classification, ensuring efficient convergence. The accuracy of the model is validated with a test suite. Furthermore, the performance of various classifiers, including KNN, NB, DT, RF, and SVM, is compared to evaluate their effectiveness in predicting credit risk.

The last technique used is remove repeated data, which was used by [2, 4, 14, 20]. Highlighting the methodology implemented by [4] which consists of preprocessing the dataset with label coding and handling of missing values, and the application of one-hot coding for categorical variables. Several classification models are selected and compared, including KNN and XGBoost. The dataset is divided into training and test sets with stratified sampling to maintain the class distribution. The model is trained and evaluated using performance metrics, and a confusion matrix is displayed to analyze the accuracy of the model.

3 Methodology

This section details the datasets used for training and the proposed methodology, which encompasses various techniques to achieve the research objective of developing predictive credit risk models based on boosting techniques.

3.1 Credit datasets description

After reviewing the research provided by various authors such as [7, 31], the datasets “Statlog German Credit Data”, “Statlog Australian Credit Approval”, “Bank Marketing”, “Credit Approval”, and “South German Credit Data” were selected for the development of predictive credit risk models. These datasets are described below.

3.1.1 Statlog German Credit Data

The German dataset contains 1000 instances, out of which 700 are non-debtor clients and 300 are debtor clients [19]. Each instance contains 20 features such as *age* and *loan amount*, with the variable *highlighted* as the main objective for analysis. Seven of its attributes are integers, two are binary, and the remaining are categorical. Some of the included variables are: *duration*, *credit history, and credit amount*.

3.1.2 Statlog Australian Credit Approval

The Australian dataset contains 690 instances, of which 307 are non-debtor clients and 383 are debtor clients [33]. Each instance is composed by 14 features, 6 of which are continuous, and the remaining 8 are categorical. To preserve data security, attributes of the Australian entity were replaced with random identifiers, and some features contain missing values.

3.1.3 Bank Marketing

The banking marketing dataset contains 45211 instances, out of which 39922 are positive and 5289 are negative [28]. This dataset is crucial for direct marketing strategies implemented by banking institutions in Portugal, where marketing campaigns are conducted via telephone calls. It is common to require multiple contacts with the same client before achieving product subscription. Some of the variables included are: *campaign, duration, loan.*

3.1.4 Credit Approval

The credit approval dataset contains a total of 690 instances. Each instance is described by 15 attributes and a class label [32]. The attributes include categorical, integer, and real variables. To preserve data security, the names and values of the features have been encoded with meaningless identifiers. Additionally, some instances have multiple missing values.

3.1.5 South German Credit Data

The German credit dataset contains 1000 instances, with 700 classified as non-debtors and 300 as debtors. Unlike the previously described "Statlog German Credit Data" dataset, this dataset provides corrections to some instances and backgrounds, based on LMU Open Data [13]. Each instance contains twenty features, distributed among seven numeric variables and thirteen categorical variables. Some of these variables include *bank account balance* and *debt amount*.

Table 1 summarizes the selected datasets, including the number of variables, the number of instances, and the distribution between positive and negative instances.

3.2 Methodology

Due to the emerging growth of credit risk as one of the main issues in the financial sector [34], we used predictive models based on boosting techniques in the datasets selected in the previous subsection. Figure 1 illustrates the methodology approach proposed.

According to Figure 1, the proposed methodology consists of various subsections developed using the datasets mentioned earlier. Within the framework of exploratory data analysis, techniques include univariate and multivariate analysis, data imputation using a KNN strategy, encoding categorical variables using *LabelEncoder*, *One Hot Encoding*, and *Binary Encoding* techniques, and scaling using *RobustScaler* and *MinMax* methods.

3.2.1 Enhancing Models Performance with Clustering

Furthermore, we propose a new strategy that involves adding a column based on an unsupervised algorithm such as *Kmeans*, *Kmedoids*, *K-prototype*, and *Aggloremative Clustering*. This strategy offers several advantages. First, it segments the dataset into groups with similar characteristics, allowing the identification of underlying patterns and relationships that might not be evident with the original variables. Second, the generated cluster column represents the assignment of each observation to a specific group, enriching the dataset by providing an additional dimension of information for the predictive model. Finally, including the cluster column in the training data allowed the model to learn significant differences between clusters, which improved the model's performance, especially in cases where the data exhibited complex structures not fully captured by the original variables. Grouping data into clusters also reduced variance within each group, as observations within a cluster were more similar to each other than to those in other clusters, helping the model learn more consistent patterns and generalize better to new data.

Within the framework of machine learning based on boosting techniques, models such as XGBoost, LightGBM, and Boosted Random Forest will be applied. Subsequently, 5-fold and 10-fold

Table 1. Overview of dataset sizes and attributes

Dataset	Attributes	Instances	Positive Instances	Negative Instances
Statlog German Credit Data	20	1000	700	300
Statlog Australian Credit Approval	14	690	383	307
Bank Marketing	16	45211	39922	5289
Credit Approval	15	690	383	307
South German Credit Data	21	1000	700	300

cross-validation and hyperparameter tuning will be conducted. Finally, results will be compared using metrics such as Area Under the Curve and F-Score.

4 Results and Discussion

In this section, the results obtained according to the proposed methodology are presented. Starting with a descriptive analysis, the experimental protocols, hyperparameter tuning, and the experimental results obtained are discussed.

4.1 Exploratory Data Analysis

Firstly, we present the results of univariate and multivariate analysis, highlighting the most relevant graphs for the current research in the different datasets.

For Statlog German Credit Data, as shown in Figure 2(a), the data reveals that clients aged between 25 and 40 years acquire the highest number of credits, indicating that this age group requires special attention. Otherwise, Figure 2(b) illustrates a general positive trend, where a longer duration of credit corresponds to a higher amount. However, the distribution of both types of risk across the chart does not show a clear separation by duration or amount, suggesting that there is no distinct pattern in terms of credit risk. The regression line and its confidence interval indicate a positive linear relationship.

In the case of South German Credit Data, as depicted in Figure 3(a), we observed that the categories of credit purpose with the highest percentages are cars (used), recycling, and furniture/equipment. We also see that clients do not invest their funds in vacations, repairs, and radio/television.

In contrast, as shown in Figure 3(b), the category “200 DM or more” exhibits a wider range of loan amounts and longer durations, indicating that individuals with higher current account balances generally have longer loan terms. In contrast, the categories “less than 0 DM” and “0 to 200 DM” are associated with shorter durations and focus on lower values. The “no checking account” category demonstrates a moderate distribution of loan durations.

4.2 Experimental Protocols

Secondly, we conduct our experiments using two schemes: Experimental Protocol I (PR.I) and Experimental Protocol II (PR.II).

— **PR.I.** This first protocol was proposed by [7]. The author used the “Statlog German Credit Data” and “Statlog Australian Credit Approval” datasets. The experiments were carried out using 10-fold cross-validation with mutually exclusive folds. For comparison, the area under the curve (AUC) was used as a metric, since it represents the probability that a credit applicant with a good rating will score higher than an applicant with a bad rating. This means that a model with a higher AUC is better at distinguishing between good and bad credit applicants. The AUC is defined by the Equation 1:

$$AUC = \frac{1 + TPR - FPR}{2} \times 100\%, \quad (1)$$

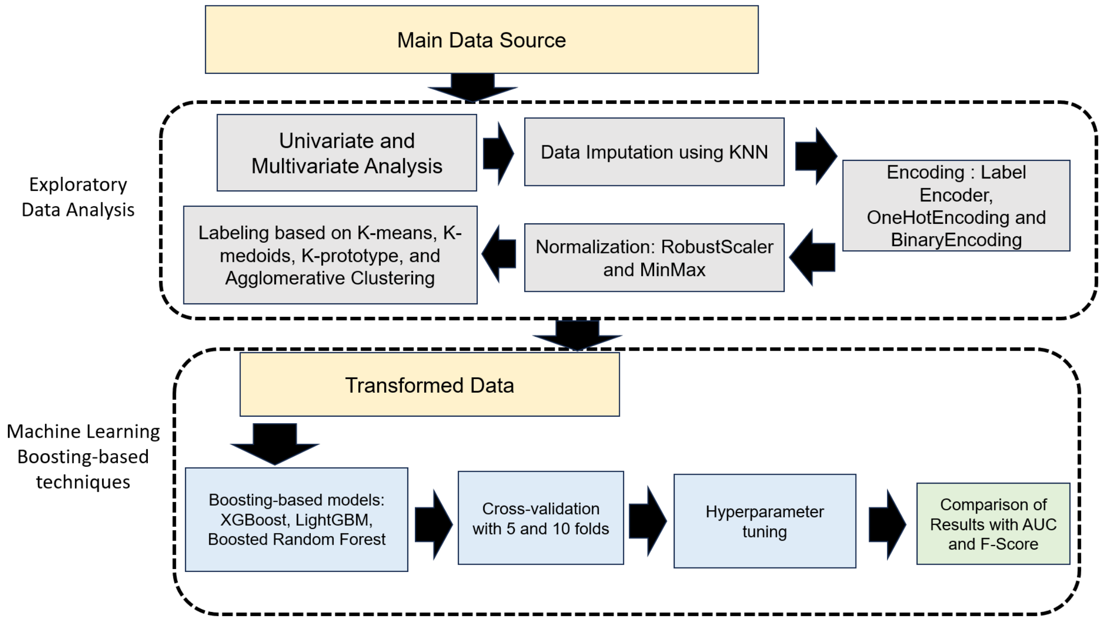


Fig. 1. Methodology applied to the development of predictive models to forecast credit risk based on boosting techniques such as LightGBM, XGBoost, and Boosted Random Forest.

where TP rate represents the true positive rate and FP rate represents the false positive rate. The best results obtained by [7] using Boosted Decision Trees were an AUC of 81.08% for “Statlog German Credit Data” and 94.03% for “Statlog Australian Credit Approval”. In this research, before data splitting and training, techniques such as *Label Encoder* were applied for encoding categorical variables to maintain variable simplicity. Subsequently, the *K-means* technique was applied to obtain labels. The elbow method and silhouette method were used to determine that the optimal number of clusters is three, and models were trained with an additional column. Following this, the training strategy applied by [7] mentioned earlier was employed. Finally, we employed hyperparameter optimization using 10-fold cross-validation to enhance the performance of the models. The results will be compared using AUC, and ROC curves for each fold as well as the average will be presented. Additionally, new metrics such as precision, recall, and F-Score will be introduced to reduce the bias.

— **PR.II.** This second protocol is proposed by [31], who used the “Statlog Australian Credit Approval”, “Bank Marketing”, “Credit Approval”, and “South German Credit Data” datasets. The experimentation was conducted using techniques such as *OneHotEncoding* for categorical variable encoding, a 70-30 training-test split, and 5-fold cross-validation. As a comparative metric, the authors applied the F-Score defined by the Equation 1:

$$\text{F-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \times 100\%. \tag{1}$$

The best results obtained by [31] using “Factorization Machine” were an F-Score of 55.35% for “Bank Marketing,” 96.47% for “Credit Approval,” 83.99% for “South German Credit Data,” and 87.80% for “Statlog Australian Credit Approval.”

Table 2. Hyperparameter Tuning

			S.G.D	S.A.D	B.K.	C.A.	S.G.C.D
N° Folds			10	10	5	5	5
Hiperparameter	**Standard Terminology**	**Values Set**	**Obtained Value**	**Obtained Value**	**Obtained Value**	**Obtained Value**	**Obtained Value**
Boosted Random Forest (B.R.F)							
n_estimators	Number of trees	[50, 100, 200, 300, 400]	300	300	300	300	300
max_depth	Maximum depth	[1, 2, 3, 5, 8, 11, 13, 15]	15	15	15	15	15
min_samples_leaf	Min samples per leaf	[1, 2, 5, 8]	2	2	2	2	2
min_samples_split	Min samples to split	[5, 10, 15, 20]	15	15	15	15	15
XGBoost (XGB)							
n_estimators	Number of trees	[50, 100, 200, 300, 400]	200	100	100	100	100
max_depth	Maximum depth	[1, 2, 3, 5, 8, 11, 13, 15]	5	8	1	1	2
learning_rate	Learning rate	[0.05, 0.1, 0.2, 0.3, 0.5]	0.1	0.1	0.1	0.2	0.3
subsample	Subsample ratio	[0.3, 0.4, 0.5, 0.8, 1]	0.8	0.4	1	0.3	0.3
gamma	Complexity penalty	[1, 2, 3, 5, 6]	2	3	6	3	1
min_child_weight	Min child weight	[1, 2, 3, 4, 5]	5	2	3	3	3
LightGBM (L.GBM)							
n_estimators	Number of trees	[50, 100, 200, 300, 400]	100	100	100	100	100
colsample_bytree	Subsample features	[0.1, 0.5, 0.8]	0.8	0.8	0.8	0.8	0.8
num_leaves	Number of leaves	[10, 20, 30, 31]	31	31	31	31	31
learning_rate	Learning rate	[0.05, 0.1, 0.3, 0.5]	0.05	0.05	0.05	0.05	0.05
max_depth	Maximum depth	[1, 2, 3, 5, 8, 11, 13, 15]	3	3	3	3	3
min_child_samples	Min samples per leaf	[10, 20, 30, 50]	20	20	20	20	20
subsample	Subsample ratio	[0.3, 0.4, 0.5, 0.8, 1]	0.8	0.8	0.8	0.8	0.8

In this research, we conducted an exploratory data analysis before splitting the data into training and testing sets. Subsequently, we applied numerical variable encoding techniques, such as *RobustScaler*, and conducted data imputation using the *KNN* strategy. Outlier treatment was handled by the interquartile range method, and we used the *KMeans* algorithm for label assignment. Then, the training strategy applied by [31] was implemented. Finally, we conducted hyperparameter tuning with 5 folds to improve the model performance. The results will be evaluated and compared using metrics such as F-Score, Matthews Correlation, Accuracy, Precision, Recall, and AUC.

4.3 Hyperparameter Calibration

The combination of optimal hyperparameters allows us to achieve an improvement in accuracy and a reduction in overfitting in the models [38]. Table 2 shows the range of values explored, and the best values obtained using the *Grid Search* technique with 5-fold and 10-fold cross-validation for each algorithm (Boosted Random Forest, XGBoost, and LightGBM). In addition, we present the hyperparameter settings with their standard terminology for each dataset: “Statlog German Credit Data” (S.G.C.D.), “Statlog Australian Credit Approval” (S.A.D.), “Bank Marketing” (B.K.), “Credit Approval” (C.A.), and “South German Credit Data” (S.G.D.).

4.4 Results

The performance of credit scoring models is determined in terms of their ability to differentiate between non-debtor clients and debtor clients. This evaluation was performed using AUC for the first experimental protocol and F-Score for the second experimental protocol, detailed in the previous subsection 4.2.

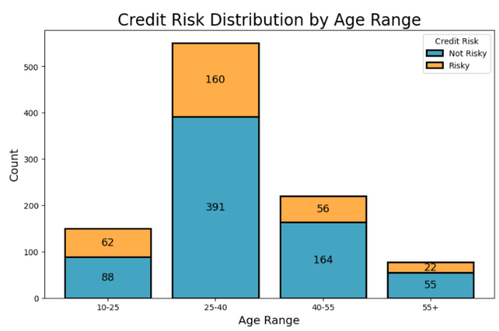
4.4.1 Results for PR.I

Tables 3 and 4 provide the results of the metrics AUC, precision, recall, and F-Score for the models XGBoost (XGB), LightGBM (LGM), and Boosted Random Forest (BRF). These comparisons were made under the first experimental protocol, utilizing categorical variable encoding methods such as *One Hot Encoding* (O.H.E.), Label Encoder (L.E.), and *Binary Encoding* (B.E.), as well as clustering methods like *K-Means*, *K-Medoids*, *Agglomerative Clustering*, and *K-Prototype*, in comparison to the original Boosted Decision Trees (B.D.T) model.

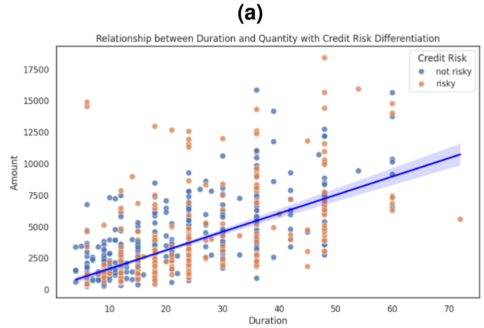
ISSN 2007-9737

Table 3. Comparison of AUC, precision, recall, and F-Score for the implemented machine learning models in the Statlog German Credit Data.

		Statlog German Credit Data											
		K-Means			**K-Medoids**			**Agglomerative Clustering**			**K-Prototype**		
		O.H.E	**L.E.**	**B.E.**	**O.H.E**	**L.E.**	**B.E.**	**O.H.E**	**L.E.**	**B.E.**	**O.H.E**	**L.E.**	**B.E.**
XGB	**AUC**	79.60	**80.94**	76.20	79.17	80.28	75.87	79.59	80.49	75.87	78.18	**80.84**	78.18
	Precision	65.46	69.13	60.15	65.07	68.08	59.34	65.69	68.88	61.19	61.97	69.46	61.97
	Recall	47.00	51.00	37.33	45.67	51.33	40.00	46.67	51.67	40.67	43.00	51.00	43.00
	F-Score	54.44	58.35	45.79	53.48	58.16	47.56	54.30	58.75	48.63	50.52	58.40	50.52
LGM	**AUC**	78.62	79.36	75.89	78.04	79.08	76.18	78.32	79.30	75.97	77.30	79.11	77.07
	Precision	62.76	66.68	62.31	62.19	65.47	64.41	63.48	65.80	66.50	64.81	68.27	65.37
	Recall	39.00	44.00	31.67	36.67	42.33	31.33	39.33	43.67	32.01	31.00	44.33	33.33
	F-Score	47.81	52.67	41.79	45.87	50.93	42.07	48.22	52.15	42.81	41.52	53.30	43.87
BRF	**AUC**	**80.59**	**81.12**	76.66	80.24	**80.58**	76.20	80.18	**80.99**	76.26	76.00	**80.86**	77.27
	Precision	71.38	69.12	69.45	67.96	68.19	69.64	69.95	68.95	73.78	58.70	68.65	71.89
	Recall	42.00	43.67	21.67	39.33	43.33	21.00	41.01	44.01	25.01	40.67	44.35	24.33
	F-Score	52.41	53.13	32.28	49.42	52.64	32.01	51.28	53.29	37.03	47.68	53.46	36.20
Bastos (2022)	**AUC**	**81.08**	**81.08**	**81.08**	**81.08**	**81.08**	**81.08**	**81.08**	**81.08**	**81.08**	**81.08**	**81.08**	**81.08**



(a)



(b)

Fig. 2. (a) Bar chart of the target variable based on *age* and (b) Scatter plot with regression line between *duration, quantity* and *objective* in the dataset “Statlog German Credit Data”.

We can determine that the Boosted Random Forest and XGBoost models demonstrate greater predictive performance on the “Statlog German Credit Data” and “Statlog Australian Credit Approval” datasets, outperforming the results initially reported by [7]. In addition, Table 5 presents the cluster profiling, which details the means of each variable within each cluster. This is a fundamental part of the labeling process, as it allows identifying and describing the distinctive features of each group in the dataset.

According to Table 5, there are notable differences in the mean values of different variables within the “Statlog German Credit Data” and “Statlog Australian Credit Approval” datasets. For the “Statlog German Credit Data,” variables such as *duration*, *loan purpose*, *credit amount*, and *installment rate* vary across clusters. Specifically, cluster 0 comprises clients with average credit amounts and loan durations of approximately 19 months; cluster 1 includes clients with lower credit amounts and shorter loan durations of around 11 months; and cluster 2 features clients with higher credit amounts and longer loan durations of about 23 months. Cluster 1 predominantly contains non-debtors, whereas cluster 2 includes a higher proportion of debtors.

In line with Figure 4, two ROC curves are presented. Figure 4 (a) shows the ROC curve of each of the 10 folds, achieving the best performance with the ROC curve of fold 7 at 85.90%, and the lowest performance with fold 6 at 75.55%. On the other hand, Figure 4 (b) shows the average ROC curve, achieving an area under

Table 4. Comparison of AUC, precision, recall, and F-Score for the implemented machine learning models in the Statlog Australian Credit Approval Data.

		Statlog Australian Credit Approval											
		K-Means			**K-Medoids**			**Agglomerative Clustering**			**K-Prototype**		
		O.H.E	**L.E.**	**B.E.**	**O.H.E**	**L.E.**	**B.E.**	**O.H.E**	**L.E.**	**B.E.**	**O.H.E**	**L.E.**	**B.E.**
XGB	**AUC**	93.97	**94.47**	93.78	93.97	**94.47**	93.79	93.97	**94.47**	93.93	92.35	**94.25**	**94.11**
	Precision	85.28	86.41	87.26	85.28	86.41	85.98	85.28	86.41	86.12	82.09	86.59	85.86
	Recall	85.98	87.94	87.32	85.98	87.94	87.32	85.98	87.94	87.00	89.59	86.65	87.32
	F-Score	85.45	86.95	87.16	85.45	86.95	86.52	85.45	86.95	86.43	85.51	86.42	86.46
LGM	**AUC**	93.75	93.70	**94.29**	93.70	93.64	**94.48**	93.75	93.70	**94.39**	92.60	93.76	**94.10**
	Precision	84.11	83.81	86.48	84.22	84.22	86.51	84.11	83.81	86.80	82.62	84.73	86.76
	Recall	84.37	85.98	87.32	84.37	85.67	87.63	84.37	85.98	87.96	87.61	85.35	87.30
	F-Score	84.06	84.64	86.76	84.10	84.71	86.94	84.06	84.64	87.27	84.87	84.80	86.94
BRF	**AUC**	93.88	93.80	**94.45**	93.91	93.69	**94.45**	93.88	93.80	**94.51**	92.81	**94.03**	**94.64**
	Precision	87.41	87.92	87.10	87.86	87.98	86.99	87.41	87.92	86.82	85.57	88.55	87.64
	Recall	86.31	85.67	87.96	85.99	85.32	87.63	86.31	85.67	87.96	85.37	85.99	88.29
	F-Score	86.59	86.52	87.42	86.51	86.34	87.22	86.59	86.52	87.27	85.27	86.92	87.83
Bastos (2022)	**AUC**	**94.03**	**94.03**	**94.03**	**94.03**	**94.03**	**94.03**	**94.03**	**94.03**	**94.03**	**94.03**	**94.03**	**94.03**

the curve of 81.12%, suggesting that the model performs well in distinguishing between risk and no risk classes. Apart from that, in the Statlog Australian Credit Approval dataset, significant differences are observed in the obscured attributes A1 and A2. Cluster 2 shows a difference of approximately 180 in A1, and cluster 1 has a difference of around 80 in A2 compared to the other clusters.

Foremost, Figure 5(a) presents the ROC curve of each of the 10 folds, with the best performance from folds 5 and 10 achieving 1.000, and the lowest performance from fold 8 at 87.20%. On the other hand, Figure 5(b) shows the average ROC curve, achieving an area under the curve of 94.47%, indicating a high level of model accuracy.

4.4.2 Results for PR.II

The Table 6 presents the results for the second experimental protocol, focusing on the F-Score as a comparison metric. We found that the XGBoost model consistently outperformed the Factorization Machine model implemented by [31], effectively classifying debtors and non-debtors across all four scenarios in terms of F-Score.

4.5 Discussion

Based on the results obtained in Tables 3, 4 and 6, we observed that both experimental protocols significantly enhanced their predictive capability with the inclusion of clustering-based labeling. The additional cluster column (0, 1, 2, 3) provided valuable insights into the natural segmentation of the data, allowing the model to capture underlying patterns and relationships that were not apparent from the original features. This enrichment of the feature space enabled the model to better differentiate between distinct data groups, reduce noise, and ultimately improve prediction metrics.

Furthermore, we evaluated three encoding methods: *Label Encoding*, *One Hot Encoding*, and *Binary Encoding*. Our findings indicate that *Label Encoding* is the most effective approach for encoding categorical variables. This is because *One Hot Encoding* and *Binary Encoding* significantly increase data dimensionality, which can introduce noise and reduce the model's predictive performance. In contrast, *Label Encoding* provides a more compact and efficient representation, leading to improved model performance. These results are consistent with research by [30], which also found that *Label Encoding* enhances predictive accuracy by avoiding the dimensionality increase associated with *One Hot Encoding* and *Helmert encoding*.

Two methodologies presented in the research by [7, 31] offer contrasting approaches. The first methodology applies 10-fold cross-validation with mutually exclusive folds, using the average AUC as the sole validation metric. However, relying on AUC alone can lead to biased and

Table 5. Cluster profiling of the attributes with the greatest variation in each cluster applied to the Statlog German Credit Data and Statlog Australian Credit Approval datasets

	Statlog German Credit Data			**Statlog Australian Credit Approval Data**			
Cluster	**0**	**1**	**2**	**Cluster**	**0**	**1**	**2**
Duration (months)	19.48	11.21	23.98	**A1**	228.00	238.00	48.91
Loan purpose	3.39	3.28	2.77	**A2**	45.29	139.67	57.75
Credit amount	758.68	320.94	892.50	**A3**	0.00	0.33	0.18
Installment rate	1.63	2.13	1.29	**A4**	5.18	2.33	2.23
Telephone (Y/N)	0.53	0.33	0.82	**A7**	64.76	45.00	35.38

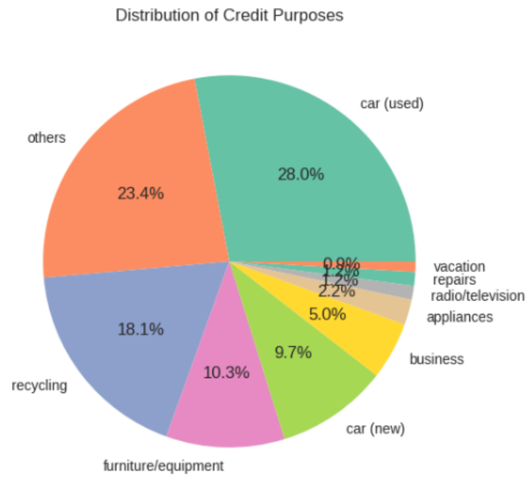
Table 6. Comparison of F-Score, Matthews Correlation (M.C), Accuracy, Precision, Recall, and AUC for the implemented machine learning models in relation to the model presented by [31]

	F-Score	**Accuracy**	**M.C.**	**Precision**	**Recall**	**AUC**
Bank Marketing						
XGB	**55.71**	**90.67**	**51.02**	63.17	49.82	72.97
LGM	35.65	89.41	34.99	62.68	24.91	61.46
BRF	51.46	90.59	47.88	**65.54**	42.37	69.70
Quan (2024)	55.35	90.21	49.22	53.48	**57.36**	**73.43**
Statlog Australian Credit Approval						
XGB	**88.34**	**90.82**	**80.77**	87.80	**88.89**	**90.45**
LGM	82.21	86.00	70.66	81.71	82.72	85.41
BRF	82.58	86.96	72.38	86.49	79.01	85.54
Quan (2024)	87.80	88.44	76.78	**88.52**	86.63	89.28
Credit Approval						
XGB	84.72	84.54	69.32	80.95	87.63	84.16
LGM	84.42	85.02	70.10	82.35	86.60	85.12
BRF	86.32	87.44	74.77	88.17	84.54	87.27
Quan (2024)	**96.47**	**94.64**	**85.46**	**95.34**	**97.61**	**90.53**
South German Credit Data						
XGB	**85.25**	**78.67**	**47.47**	81.86	88.94	72.19
LGM	83.68	76.33	41.47	80.18	87.50	69.29
BRF	84.30	76.67	41.04	79.00	**90.38**	68.02
Quan (2024)	83.99	76.96	47.25	**93.29**	76.37	**81.65**

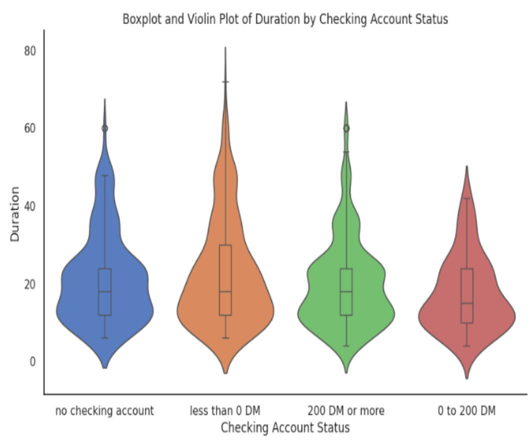
potentially inaccurate results. Furthermore, this methodology lacks preprocessing techniques such as encoding, normalization, or the labeling method applied in our research. The second methodology employs *One Hot Encoding* and cross-validation but does not specify the number of folds, a crucial detail for reproducibility. Additionally, it omits normalization techniques like *RobustScaler* and imputation techniques such as KNN, both of which are applied in our research to ensure more robust data preparation and processing for optimal model performance.

Moreover, it has been observed that credit scoring models tend to perform better with

ISSN 2007-9737

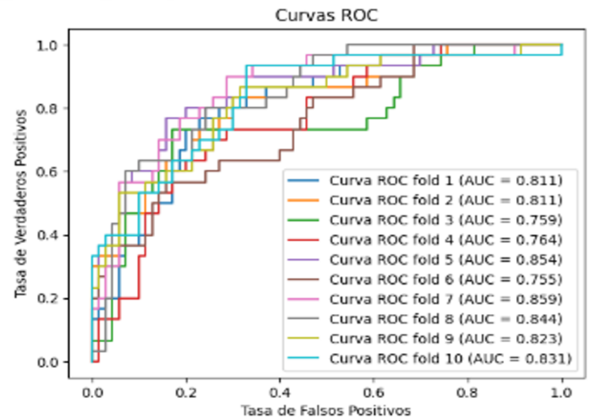


(a)

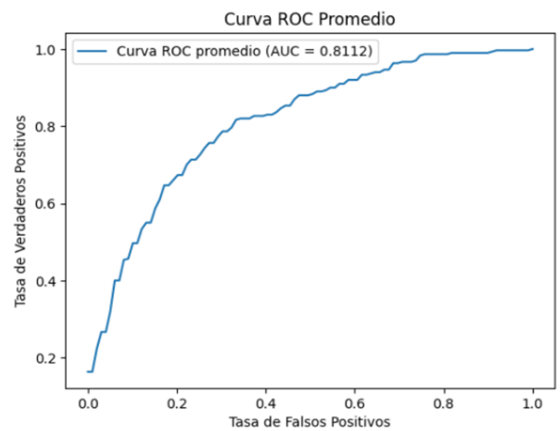


(b)

Fig. 3. (a) Pie chart in relation to the purpose of the credit and (b) Boxplot combined with violin of duration by checking account status in the dataset "South German Credit Data"



(a)



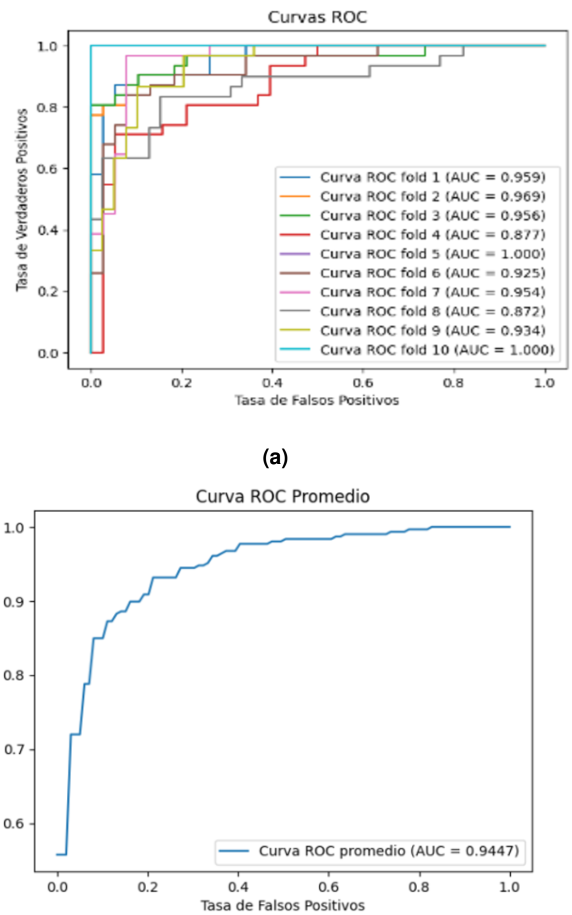
(b)

Fig. 4. (a) AUC per fold and (b) Average AUC of the Boosted Random Forest model for Statlog German Credit Data

tree-based methods, such as Boosted Random Forest, XGBoost, and LightGBM. The most notable results reported by [7] were achieved using the Boosted Decision Trees model, highlighting its effectiveness in accurately predicting credit approval. This finding aligns with the research by [1], which demonstrated that tree-based models like random forest and gradient boosting provide superior performance and stability in credit risk classification compared to neural network models with two or three hidden layers.

Finally, in the study by [31], there is a discrepancy in the reported dataset size for the "Credit Approval" dataset. While their research discusses results based on 690 instances, the original research mentions only 300 instances, raising concerns about the accuracy and consistency of the data used in their experiments.

ISSN 2007-9737



(a)

(b)

Fig. 5. (a) AUC per fold and (b) Average AUC of the XGBoost model for Statlog Australian Credit Approval.

5 Conclusion and Future Work

Effective credit risk management plays a crucial role in maintaining the stability and profitability of financial institutions. By adopting advanced machine learning boosting techniques, such as LightGBM, XGBoost, and Boosted Random Forest, institutions can significantly enhance the accuracy of credit risk predictions. These models enable more precise borrower assessments, leading to improved metrics like F-Score and AUC, and optimizing the decision-making process in credit approvals. As financial institutions continue to embrace these technologies, they stand to benefit from reduced non-payment rates, more efficient resource allocation, and a more resilient financial system overall.

Looking ahead, we plan to expand the scope of datasets to include diverse geographical regions and financial sectors, which will help assess the robustness and global applicability of these models. Further research should explore the combination of boosting techniques with other machine learning and deep learning algorithms to improve accuracy and efficiency. Employing advanced hyperparameter optimization methods, such as Bayesian search or evolutionary algorithms, could also yield significant performance improvements. Moreover, evaluating the impact of data quality on model performance is crucial, including how inaccuracies in input data and data cleaning methods affect outcomes. Developing and testing real-time credit risk prediction systems that integrate directly with current financial platforms presents a promising area for innovation.

Lastly, we plan to investigate interpretability and explainability techniques for boosting models to enhance users' understanding of model decisions and consider applying these models in other sectors like insurance and telecommunications for broader impact.

References

1. **Addo, P. M., Guegan, D., Hassani, B. (2018).** Credit Risk Analysis Using Machine and Deep Learning Models. Risks, Vol. 6, No. 2, pp. 1–20. DOI: `10.3390/risks6020038`.
2. **AL-Shatnwai, A. M., Faris, M. (2020).** Predicting Customer Retention Using XG-Boost and Balancing Methods. International Journal of Advanced Computer Science and Applications, Vol. 11, No. 7, pp. 1–7. DOI: `10.14569/IJACSA.2020.0110785`.
3. **Al-Shiam, S. A., Hasan, M. M., Pantho, M. J., Shochona, S. A., Nayeem, M. B., Choudhury, M. T. H., Nguyen, T. N. (2024).** Credit Risk Prediction Using Explainable AI. Journal of Business and Management Studies,

Vol. 6, No. 2, pp. 61–66. DOI: 10.32996/jbms.2024.6.2.6.

4. **Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., Selmanovic, E. (2024).** Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data. Machine Learning and Knowledge Extraction, Vol. 6, No. 1, pp. 53–77. DOI: 10.3390/make6010004.

5. **Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., Segovia-Vargas, M.-J. (2020).** Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. IEEE Access, Vol. 8, pp. 64873–64890. DOI: 10.1109/ACCESS.2020.2984412.

6. **Asencios, R., Asencios, C., Ramos, E. (2023).** Profit Scoring for Credit Unions Using the Multilayer Perceptron, XGBoost and TabNet Algorithms: Evidence From Peru. Expert Systems with Applications, Vol. 213, pp. 1–10. DOI: 10.1016/j.eswa.2022.119201.

7. **Bastos, J. A. (2022).** Predicting Credit Scores With Boosted Decision Trees. Forecasting, Vol. 4, No. 4, pp. 925–935. DOI: 10.3390/forecast4040050.

8. **Bazzana, F., Bee, M., Hussin-Adam-Khatir, A. A. (2024).** Machine Learning Techniques for Default Prediction: An Application to Small Italian Companies. Risk Management, Vol. 26, No. 1, pp. 1–23. DOI: 10.1057/s41283-023-00132-2.

9. **Bhatt, T. K., Ahmed, N., Iqbal, M. B., Ullah, M. (2023).** Examining the Determinants of Credit Risk Management and Their Relationship With the Performance of Commercial Banks in Nepal. Journal of Risk and Financial Management, Vol. 16, No. 4, pp. 1–23. DOI: 10.3390/jrfm16040235.

10. **Biswas, N., Mondal, A. S., Kusumastuti, A., Saha, S., Mondal, K. C. (2022).** Automated Credit Assessment Framework Using ETL Process and Machine Learning. Innovations in Systems and Software Engineering, Vol. 21, pp. 257–270. DOI: 10.1007/s11334-022-00522-x.

11. **Bitetto, A., Cerchiello, P., Filomeni, S., Tanda, A., Tarantino, B. (2023).** Machine Learning and Credit Risk: Empirical Evidence From Small-and Mid-Sized Businesses. Socio-Economic Planning Sciences, Vol. 90, pp. 1–9. DOI: 10.1016/j.seps.2023.101746.

12. **Chang, V., Xu, Q. A., Akinloye, S. H., Benson, V., Hall, K. (2024).** Prediction of Bank Credit Worthiness Through Credit Risk Analysis: An Explainable Machine Learning Study. Annals of Operations Research, pp. 1–25. DOI: 10.1007/s10479-024-06134-x.

13. **Credit, S. G. (2019).** South German Credit. UCI Machine Learning Repository. DOI: 10.24432/C5X89F.

14. **Doko, F., Kalajdziski, S., Mishkovski, I. (2021).** Credit Risk Model Based on Central Bank Credit Registry Data. Journal of Risk and Financial Management, Vol. 14, No. 3, pp. 1–17. DOI: 10.3390/jrfm14030138.

15. **Erickson, B. J., Kitamura, F. (2021).** Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiology: Artificial Intelligence, Vol. 3, No. 3, pp. 1–7. DOI: 10.1148/ryai.2021200126.

16. **Feng, Y. (2022).** Bank Green Credit Risk Assessment and Management by Mobile Computing and Machine Learning Neural Network under the Efficient Wireless Communication. Wireless Communications and Mobile Computing, Vol. 2022, No. 1, pp. 1–11. DOI: 10.1155/2022/3444317.

17. **Golbayani, P., Florescu, I., Chatterjee, R. (2020).** A Comparative Study of Forecasting Corporate Credit Ratings Using Neural Networks, Support Vector Machines, and Decision Trees. The North American Journal of Economics and Finance, Vol. 54, pp. 1–16. DOI: 10.1016/j.najef.2020.101251.

18. **Hasni, M., Aguir, M. S., Babai, M. Z., Jemai, Z. (2024).** A Machine Learning Model for Accurate Credit Risk Forecasting in Banking Systems: An Empirical Investigation. International Journal of Supply & Operations Management, Vol. 11, No. 2, pp. 168–187. DOI: 10.22034/ijsom.2023.109898.2722.

19. **Hofmann, H. (1994).** Statlog (German Credit Data). UCI Machine Learning Repository. DOI: 10.24432/C5NC77.

20. **Jain, H., Khunteta, A., Srivastava, S. (2020).** Churn Prediction in Telecommunication Using Logistic Regression and Logit Boost. Procedia Computer Science, Vol. 167, pp. 101–112. DOI: 10.1016/j.procs.2020.03.187.

21. **Kadam, A., Namde, P., Shirke, S., Nandgaonkar, S., Ingle, D. (2021).** Loan Credibility Prediction System using Data Mining Techniques. International Research Journal of Engineering and Technology, Vol. 8, No. 5, pp. 1354–1357.

22. **Li, Y., Stasinakis, C., Yeo, W. M. (2022).** A Hybrid XGBoost-MLP Model for Credit Risk Assessment on Digital Supply Chain Finance. Forecasting, Vol. 4, No. 1, pp. 184–207. DOI: 10.3390/forecast4010011.

23. **Liu, J., Zhang, S., Fan, H. (2022).** A Two-Stage Hybrid Credit Risk Prediction Model Based on XGBoost and Graph-Based Deep Neural Network. Expert Systems with Applications, Vol. 195, pp. 1–13. DOI: 10.1016/j.eswa.2022.116624.

24. **Luo, C. (2022).** A Comparison Analysis for Credit Scoring Using Bagging Ensembles. Expert Systems, Vol. 39, No. 2, pp. 1–7. DOI: 10.1111/exsy.12297.

25. **Madeira, C. (2023).** Adverse Selection, Loan Access and Default Behavior in the Chilean Consumer Debt Market. Financial Innovation, Vol. 9, No. 49, pp. 1–29. DOI: 10.1186/s40854-023-00458-6.

26. **Markov, A., Seleznyova, Z., Lapshin, V. (2022).** Credit Scoring Methods: Latest Trends and Points to Consider. The Journal of Finance and Data Science, Vol. 8, pp. 180–201. DOI: 10.1016/j.jfds.2022.07.002.

27. **Mitra, R., Goswami, A., Tiwari, M. K. (2022).** Financial Supply Chain Analysis With Borrower Identification in Smart Lending Platform. Expert Systems with Applications, Vol. 208, pp. 1–13. DOI: 10.1016/j.eswa.2022.118026.

28. **Moro, S., Rita, P., Cortez, P. (2014).** Bank Marketing. UCI Machine Learning Repository. DOI: 10.24432/C5K306.

29. **Noriega, J. P., Rivera, L. A., Herrera, J. A. (2023).** Machine Learning for Credit Risk Prediction: A Systematic Literature Review. Data, Vol. 8, No. 11, pp. 1–17. DOI: 10.3390/data8110169.

30. **Poslavskaya, E., Korolev, A. (2023).** Encoding Categorical Data: Is There Yet Anything 'Hotter' Than One-Hot Encoding?. arXiv, pp. 1–11. DOI: 10.48550/arXiv.2312.16930.

31. **Quan, J., Sun, X. (2024).** Credit Risk Assessment Using the Factorization Machine Model With Feature Interactions. Humanities and Social Sciences Communications, Vol. 11, No. 234, pp. 1–10. DOI: 10.1057/s41599-024-02700-7.

32. **Quinlan, J. R. (1987).** Credit Approval. UCI Machine Learning Repository. DOI: 10.24432/C5FS30.

33. **Quinlan, R. (1987).** Statlog (Australian Credit Approval). UCI Machine Learning Repository. DOI: 10.24432/C59012.

34. **Sabbeh, S. F. (2018).** Machine-Learning Techniques for Customer Retention: A Comparative Study. International Journal of advanced computer Science and applications, Vol. 9, No. 2, pp. 1–9. DOI: 10.14569/IJACSA.2018.090238.

35. **Talaat, F. M., Aljadani, A., Badawy, M., Elhosseini, M. (2024).** Toward Interpretable Credit Scoring: Integrating Explainable Artificial Intelligence With Deep Learning for Credit Card Default Prediction. Neural Computing and Applications, Vol. 36, No. 9, pp. 4847–4865. DOI: 10.1007/s00521-023-09232-2.

36. **Wang, J., Liu, G., Xu, X., Xing, X. (2024).** Credit Risk Prediction for Small and Medium Enterprises Utilizing Adjacent Enterprise Data and a Relational Graph Attention Network. Journal of Management Science and Engineering, Vol. 9, No. 2, pp. 177–192. DOI: 10.1016/j.jmse.2023.11.005.

37. **Wattanakitrungroj, N., Wijitkajee, P., Jaiyen, S., Sathapornvajana, S., Tongman, S. (2024).** Enhancing Supervised Model Performance in Credit Risk Classification Using Sampling Strategies and Feature Ranking. Big Data and Cognitive Computing, Vol. 8, No. 3, pp. 1–22. DOI: 10.3390/bdcc8030028.

38. **Weerts, H. J. P., Mueller, A. C., Vanschoren, J. (2020).** Importance of Tuning Hyperparameters of Machine Learning Algorithms. arXiv, pp. 1–17. DOI: 10.48550/arXiv.2007.07588.

39. **Xiao, H.-H., Yang, W.-K., Hu, J., Zhang, Y.-P., Jing, L.-J., Chen, Z.-Y. (2022).** Significance and Methodology: Preprocessing the Big Data for Machine Learning on TBM Performance. Underground Space, Vol. 7, No. 4, pp. 680–701. DOI: 10.1016/j.undsp.2021.12.003.

Article received on 17/09/2024; accepted on 13/01/2025.
**Corresponding author is Renzo Orlando Villanueva Mora.*