UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Comparative Evaluation of Supervised Machine Learning Models for Credit Score Prediction using CACPECO Data

Proyecto de Titulación

Ángel David Llerena Camacho

Israel Pineda, Ph.D.

Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magíster en Ciencia de Datos

Quito, 01 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Comparative Evaluation of Supervised Machine Learning Models for Credit Score Prediction using CACPECO Data

Angel David Llerena Camacho

Nombre del Director del Programa:	Felipe Grijalva
Título académico:	Ph.D. en Ingeniería Eléctrica
Director del programa de:	Ciencia de Datos

Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor en Ciencias Matemáticas
Decano del Colegio:	Ciencias e Ingenierías

Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor en Física

Quito, diciembre 2024

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombre del estudiante:	Ángel David Llerena Camacho
Código de estudiante:	00338870
C.I.:	1804551305
Lugar y fecha:	Quito, 01 de diciembre de 2024.

DEDICATORIA

A mis padres, por su incalculable respaldo, guía y entrega que han sido la base esencial en mi crecimiento profesional y personal, quiero destacar que este éxito refleja los valores y principios que me inculcaron, y con sincera gratitud dedico este trabajo a ustedes.

AGRADECIMIENTOS

Expreso mi profundo agradecimiento a CACPECO, institución que ha jugado un papel muy importante en el desarrollo de mi carrera profesional; su apoyo incondicional para cursar esta maestría y al permitirme utilizar su base de datos para este proyecto han sido fundamentales. Su creencia en mi potencial me ha impulsado profesionalmente. En igual medida, extiendo mi más sincero agradecimiento a los profesores de la Universidad San Francisco de Quito, quienes, con su saber, dedicación y vocación académica, enaltecen el valor de mi proceso de aprendizaje de manera incalculable. En particular, agradezco al PhD. Israel Pineda, mi tutor, por su paciencia, experta guía y aportes invaluables que han sido determinantes para la concreción de este proyecto. A todos ustedes, mi más alto reconocimiento y profunda gratitud por haber sido parte integral de este importante recorrido.

RESUMEN

La calificación crediticia es un instrumento crucial para que los asesores financieros determinen la elegibilidad de un cliente para un préstamo. Tradicionalmente, estas calificaciones dependen de agencias de crédito externas que consolidan datos financieros globales, incluidas las deudas totales y los historiales crediticios. Sin embargo, las instituciones financieras más pequeñas enfrentan desafíos para acceder a esta información restringida. Este estudio presenta un enfoque basado en el aprendizaje supervisado para predecir de forma independiente las calificaciones crediticias, aprovechando los datos internos de los clientes, como activos, pasivos, edad, género, dependientes familiares y otros factores socioeconómicos. Se desarrolló una sólida línea de trabajo, que incluye pasos de preprocesamiento como codificación de etiquetas, sobremuestreo aleatorio, escalamiento de características y selección de características. Se entrenaron y ajustaron siete modelos de aprendizaje automático (Linear Regression, Random Forest Classifier, Support Vector Classifier, K-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naive Bayes, and Decision Tree Classifier) utilizando GridSearchCV. La metodología se probó en dos conjuntos de datos: el conjunto de datos de Lending Club para evaluación comparativa y los datos del mundo real de CACPECO. La evaluación del desempeño incluyó métricas de regresión (MAE, MSE, RMSE, R^2) y métricas de clasificación (accuracy, precision, recall, F1-score). Los resultados mostraron que Random Forest Regressor, Gaussian Naive Bayes y Linear Discriminant Analysis superaron a otros modelos, lo que demuestra su capacidad para manejar datos complejos y desequilibrados de manera efectiva. Esta investigación proporciona un marco reproducible para que las instituciones financieras predigan de manera independiente las calificaciones crediticias, lo que reduce la dependencia de las agencias de crédito externas. Al utilizar datos internos, este enfoque mejora los procesos de toma de decisiones y proporciona a las instituciones con recursos limitados las herramientas necesarias para evaluar de manera confiable la solvencia crediticia.

Palabras clave: score crediticio, aprendizaje supervisado, GridSearchCV, linear regression, random forest classifier, linear discriminant analysis, k-neighbors classifier, gaussian naive bayes, support vector classifier, decision tree classifier.

ABSTRACT

Credit scoring serves as a crucial instrument for financial advisors in determining a client's loan eligibility. Traditionally, these scores rely on external credit bureaus that consolidate global financial data, including total debts and credit histories. However, smaller financial institutions face challenges accessing this restricted information. This study introduces a supervised learning-based approach to independently predict credit scores, leveraging internal client data such as assets, liabilities, age, gender, family dependents, and other socio-economic factors. A robust pipeline was developed, including preprocessing steps like label encoding, random oversampling, feature scaling, and feature selection. Seven machine learning models—Linear Regression, Random Forest Classifier, Support Vector Classifier, K-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naive Bayes, and Decision Tree Classifier—were trained and fine-tuned using GridSearchCV. The methodology was tested on two datasets: the Lending Club dataset for benchmarking and CACPECO's real-world data. Performance evaluation included regression metrics (MAE, MSE, RMSE, R^2) and classification metrics (accuracy, precision, recall, F1-score). Results showed that Random Forest Regressor, Gaussian Naive Bayes and Linear Discriminant Analysis outperformed other models, demonstrating their ability to handle imbalanced and complex data effectively. This research provides a reproducible framework for financial institutions to independently predict credit scores, reducing dependence on external credit bureaus. By utilizing internal data, this approach improves decision-making processes and equips resource-constrained institutions with the tools needed to reliably assess creditworthiness.

Key words: credit scoring, supervised learning, GridSearchCV, linear regression, random forest classifier, linear discriminant analysis, k-neighbors classifier, gaussian naive bayes, support vector classifier, decision tree classifier.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

Comparative Evaluation of Supervised Machine Learning Models for Credit Score Prediction using CACPECO Data

Ángel Llerena, *Master Degree in Data Science, USFQ, (dllerena@estud.usfq.edu.ec)*

***Abstract*—Credit scoring serves as a crucial instrument for financial advisors in determining a client's loan eligibility. Traditionally, these scores rely on external credit bureaus that consolidate global financial data, including total debts and credit histories. However, smaller financial institutions face challenges accessing this restricted information. This study introduces a supervised learning-based approach to independently predict credit scores, leveraging internal client data such as assets, liabilities, age, gender, family dependents, and other socio-economic factors. A robust pipeline was developed, including preprocessing steps like label encoding, random oversampling, feature scaling, and feature selection. Seven machine learning models—Linear Regression, Random Forest Classifier, Support Vector Classifier, K-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naive Bayes, and Decision Tree Classifier—were trained and fine-tuned using GridSearchCV. The methodology was tested on two datasets: the Lending Club dataset for benchmarking and CACPECO's real-world data. Performance evaluation included regression metrics (MAE, MSE, RMSE, R^2) and classification metrics (accuracy, precision, recall, F1-score). Results showed that Random Forest Regressor, Gaussian Naive Bayes and Linear Discriminant Analysis outperformed other models, demonstrating their ability to handle imbalanced and complex data effectively. This research provides a reproducible framework for financial institutions to independently predict credit scores, reducing dependence on external credit bureaus. By utilizing internal data, this approach improves decision-making processes and equips resource-constrained institutions with the tools needed to reliably assess creditworthiness.**

***Index Terms*—credit scoring, supervised learning, GridSearchCV, linear regression, random forest classifier, linear discriminant analysis, k-neighbors classifier, gaussian naive bayes, support vector classifier, decision tree classifier.**

I. Introduction

CREDIT scoring is an essential metric in the financial sector, as it allows institutions to assess the creditworthiness of credit applicants. Typically, this score is calculated by analyzing aggregated financial data from various sources. However, access to such information is often restricted to certain entities, which can limit the assessment capacity of many financial institutions. CACEPCO is a financial institution that offers various products, with loans being the ones that generate the most profits for the institution. Through a loan application process, clients are qualified to determine if they are potential candidates to access this product. A key to reducing the level of default and properly selecting good payers is to determine the credit score of clients. The use of supervised learning methods in predicting credit scores has demonstrated significant effectiveness.

For example, in [1] they proposed an advanced model that combines LightGBM, XGBoost and TabNet, together with data balancing techniques such as SMOTEENN, achieving significant improvements in the accuracy of credit risk prediction.

Likewise, in [2] they introduced a classifier based on Bagging and supervised Autoencoders for credit scoring, obtaining promising results in the classification of credit applicants.

Conversely, [3] investigated the application of interpretable ensemble learning methods for credit scoring prediction, emphasizing the critical role of interpretability in financial risk modeling.

These studies, along with others discussed later, highlight the significance and efficacy of machine learning techniques in credit risk assessment, providing more precise tools tailored to the evolving demands of the financial sector.

The primary goal of this study is to conduct a comparative analysis of various supervised learning models for credit score prediction, utilizing well-established algorithms such as Linear Regression, Random Forest Classifier, Linear Discriminant Analysis, K-Neighbors Classifier, Gaussian Naive Bayes, Support Vector Classifier (SVC) and Decision Tree Classifier. These models were selected due to their ability to address classification problems in various contexts, offering a broad representation of linear, non-linear, proximity-based and decision tree approaches.

Linear Regression is a supervised learning model designed to capture the relationship between a dependent variable and one or more independent variables through a linear equation.

Represented as:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

where β are the estimated coefficients and ϵ represents the random error.

This model is widely used in predictive analytics and statistical inference [4]. In financial applications, such

Á. Llerena is with Universidad San Francisco de Quito USFQ as credit score prediction, linear regression allows for assessing how factors such as income, age, and debt affect an individual's financial behavior. Its simplicity and ease of interpretation make it ideal for cases where relationships between variables are approximately linear and data meet the assumptions of independence of errors and homoscedasticity [5].

Despite being a basic model, its relevance persists in initial exploratory analysis and as a baseline in more complex problems. However, its performance may be affected by the presence of nonlinear relationships or high multicollinearity between independent variables. Advanced methods such as regularized regression (Ridge or Lasso) can address these limitations [6].

Random Forest Classifier is a supervised learning algorithm that uses multiple randomly constructed decision trees to perform classification and regression tasks. Introduced by Breiman in 2001, this model combines the results of individual trees using ensemble techniques such as bagging, reducing the risk of overfitting and improving generalization [7].

In the financial sector, Random Forest has demonstrated its effectiveness in tasks like credit risk evaluation and fraud detection. Its capability to manage high-dimensional datasets and accommodate both categorical and continuous variables makes it highly adaptable. Additionally, hyperparameters such as the number of trees and maximum depth enable precise model tuning to achieve a balance between accuracy and computational efficiency [8].

The main challenge of this model lies in its higher computational cost compared to simpler models, especially on large data sets. However, its robustness against noisy and imbalanced data positions it as a reliable option for complex financial problems [9].

Linear Discriminant Analysis is a supervised statistical model designed to enhance the separation between two or more classes through a linear combination of features. It operates under the assumption that the classes adhere to a Gaussian distribution and share identical covariance matrices, enabling the determination of optimal decision boundaries for classification tasks [10].

In applications such as credit default prediction, LDA could handle small data sets and provide an interpretable projection of the variables into lower-dimensional spaces. This model has proven to be efficient and effective on problems where the features are well linearly separated [11]. Despite its simplicity, LDA faces limitations in the presence of non-linear data or when Gaussian assumptions do not hold. However, its ease of implementation and robustness against small samples make it a popular choice for initial classification tasks in financial problems [12].

The K-Neighbors Classifier is a supervised learning algorithm that assigns a class to a sample based on the majority vote of its nearest neighbors within the feature space. As a non-parametric model, KNN does not rely on any specific assumptions regarding the underlying data distribution. The most relevant hyperparameter in this model is kkk, the number of neighbors considered in the classification decision. The algorithm uses distance metrics, such as Euclidean or Manhattan, to identify the nearest neighbors of a given sample [4].

In financial applications, such as credit risk classification, KNN is useful due to its simplicity and ease of implementation. This model is particularly effective when classes are well separated in the feature space. However, it can become computationally intensive on large datasets, as it requires computing the distances of each sample to all others during inference. Furthermore, KNN can be sensitive to the scale of the features, so normalizing or standardizing the data is crucial for its performance [10].

KNN has an intuitive interpretation that makes it popular in early machine learning applications, but it is susceptible to overfitting when kkk is small or underfitting if kkk is too large. Its simplicity makes it a useful tool for rapid prototyping and for problems where the relationship between the features and the target variable remains relatively simple [13].

Gaussian Naive Bayes is a probabilistic classification model based on Bayes' theorem, which assumes conditional independence among features. It presumes that the features within each class follow a Gaussian (normal) distribution. GNB computes the posterior probabilities for all classes and assigns the sample to the class with the highest posterior probability [9].

GNB is frequently used in the classification of high-dimensional data, as it can handle large and complex data sets efficiently. Its computational efficiency enables the processing of large datasets within a short time frame, making it particularly well-suited for applications like financial fraud detection. However, its performance can suffer if features do not meet the assumption of conditional independence, although it is usually surprisingly robust against moderate violations of this assumption [6].

This model is also used to establish fast baselines in classification problems, due to its simplicity and speed. Although it can be outperformed by more advanced algorithms, it remains a popular choice in problems where interpretability and efficiency are more important than absolute accuracy [7].

Support Vector Classifier is a supervised learning algorithm aimed at identifying the optimal hyperplane that maximizes the margin of separation between classes within the feature space. SVC uses a kernel function-based approach, such as linear, polynomial, or radial (RBF), to transform the data and make it linearly separable in a higher dimensional space. This model is particularly effective on complex and non-linear data sets [8].

In the financial domain, SVC has been used in applications such as credit default prediction and fraud pattern detection, where relationships between variables may not be linear. SVC's ability to handle high-dimensional and complex problems makes it a powerful tool in challenging problems. However, its computational cost can be high on large data sets, and its performance depends heavily on the proper selection of hyperparameters such as the CCC penalty coefficient and the type of kernel used [14].

Despite these limitations, SVC offers a balance between flexibility and robustness, making it a suitable choice for tasks where accuracy and generalization ability are crucial [15].

Decision Tree Classifier is a hierarchical model that makes decisions through simple rules based on feature values. Each internal node corresponds to a decision about a specific feature, and the terminal leaves represent the target classes. Decision trees are highly interpretable and allow the decision process to be visualized, making them useful for problems where interpretability is crucial [16].

In credit assessment, decision trees are used to classify customers into risk categories, based on characteristics such as income, credit history, and age. Their ability to handle categorical and continuous data, along with their ease of implementation, makes them a versatile tool. However, individual decision trees are prone to overfitting, especially on small data sets, which can be mitigated by techniques such as pruning or their integration into ensemble methods such as Random Forest or Gradient Boosting [11].

Although not as accurate as other advanced models, decision trees offer a fast and understandable solution to classification problems. They are ideal for initial applications or when a clear interpretation of the results is needed [12].

The primary aim of this study is to evaluate the performance of these models through a detailed analysis of key metrics, such as:

- **Mean Absolute Error (MAE)**: to assess the average magnitude of errors without considering their sign.
- **Mean Squared Error (MSE)**: which weights larger errors more significantly by squaring them.
- **Root Mean Squared Error (RMSE)**: as a derived metric that facilitates interpretation in the same units of the credit score.
- **$\mathbf{R^2}$ (Coefficient of Determination)**: to measure the proportion of variation explained by the model.
- **Cross-Validated MSE**: which ensures the robustness of the models by evaluating their performance in different partitions of the dataset.
- **Accuracy, Precision, Recall and F1-Score**: classification-specific metrics that are fundamental to understanding the balance between positive and negative predictions.

These metrics are essential because they allow the performance of models to be evaluated from multiple perspectives. For example, error metrics (MAE, MSE, RMSE) are critical to quantify the overall accuracy in credit score prediction as a continuous value, while metrics such as accuracy, precision, recall and F1-score offer detailed insight into the behavior of models in correctly categorizing relevant cases, a crucial aspect in financial risk management.

The main contribution of this work lies in providing a systematic evaluation that not only validates previous findings, but also establishes a comparative framework to identify which models best fit the nature of the data used. This not only helps to explore new perspectives in credit score prediction, but also offers practical recommendations based on the balance between accuracy and model complexity.

Furthermore, this work has a practical approach for financial institutions with limited resources, providing a reproducible framework to select models that optimize credit score prediction.

II. Related work

Credit score analysis is essential for financial institutions, as it allows identifying clients with a high probability of default, thus optimizing capital management and minimizing losses due to late payments. Numerous studies have explored strategies to address this challenge by using supervised learning models and advanced preprocessing and parameter tuning techniques, some of which are explained below.

The work presented in [1] examines the use of LightGBM together with PCA and SMOTEENN, obtaining outstanding results such as an F1-score of 0.9989 and an AUC-ROC of 0.9999. The combination of dimensionality reduction and data balancing proved to be critical to improve the performance of the models. This study underlines the relevance of techniques such as Random Oversampling, Feature Selection and Standard Scaler, which play a key role in the pipeline adopted in this work.

On the other hand, [2] highlights the effectiveness of supervised algorithms such as Random Forest, Logistic Regression and SVM in classifying loan applicants. It also emphasizes the value of ensemble techniques and supervised autoencoders to address the class imbalance problem. These strategies justify the inclusion of Linear Regression and RandomForestClassifier in this study, due to their ability to handle large volumes of financial data, also offering interpretability, necessary in the case that a partner/client requests an explanation of the decision made using AI.

The research in [17] focuses on decision trees and shows how algorithms such as C5.0 can transform credit risk assessment into a more dynamic and quantitative process. This reinforces the selection of models such as DecisionTreeClassifier, which bring transparency to decisions, something crucial for financial institutions that, as mentioned above, require interpretable results.

The study in [18] addresses the use of Decision Trees and Random Forest to diagnose heart diseases in imbalanced datasets, using SMOTE and hyperparameter optimization techniques. This approach highlights the importance of tools such as GridSearchCV and Label Encoding to tune models on imbalanced data, allowing to maximize their performance and robustness.

In [3], ensemble techniques are explored using CatBoost combined with SHAP, demonstrating how interpretability and ensembles can significantly improve credit score prediction. Although CatBoost is not part of this work, this approach inspires the use of models such as LinearDiscriminantAnalysis and Logistic Regression, known for their ability to explain classification decisions in a transparent manner.

On the other hand, [19] evaluates the impact of feature selection and normalization on the performance of models such as SVM, GaussianNB and Random Forest. This analysis shows that proper data preparation substantially improves the accuracy of the models, reinforcing the use of tools such as Standard Scaler and Feature Selection in this study.

The study presented in [20] introduces a deep learning model designed to predict high-risk behaviors in financial traders, achieving an accuracy of 99%. The implementation of techniques such as unsupervised pre-training improves the extraction of distributed representations and allows for accurate classification in complex datasets. This approach highlights the applicability of deep learning in the context of credit scoring, especially for identifying patterns in voluminous and unstructured financial data.

In [17], the effectiveness of the C5.0 algorithm is evaluated in transforming credit risk assessment from a qualitative to a quantitative process. The results showed that incorporating key variables, such as current account balances and loan duration, significantly improves predictive capability. This approach reinforces the importance of explainable techniques such as DecisionTreeClassifier, which facilitate the interpretation of results and are essential to ensure transparency in financial prediction systems.

In [21], a dynamic theory of credit scoring is presented, considering how individuals past actions affect their reputation and credit terms through Bayesian updates. Although linear regression is not explicitly mentioned, the importance of simple and transparent models in credit risk assessment is highlighted. This highlights how predictive models, including linear variants, can be integrated into a theoretical framework to better understand credit markets.

[16] explores the use of LightGBM and other advanced models for user credit assessment, comparing various models including linear regression. Although LightGBM proved superior due to its ability to handle large and diverse data sets, linear regression was used as a base model to identify its utility compared to more complex algorithms. This shows that linear regression, although effective in simple tasks, may not be suitable for data with non-linear or highly dimensional features.

[9] uses linear regression in conjunction with KNN and LDA to detect credit card fraud, emphasizing model sensitivity (recall). Linear regression, although less sophisticated than other methods, contributed significantly to the improvement of recall, especially when combined with other algorithms. This demonstrates how simple linear models can play a critical role in applications where minimizing false negatives is key.

[22] does not directly use linear regression but highlights the limitations of basic models such as logistic regression for imbalanced data scenarios in credit assessment. The proposed model, "Logistic-BWE," addresses this problem through an ensemble approach that improves predictive ability in minority samples (default payments). This method highlights interpretability and robustness, although it does not address linear regression specifically, underlining the relevance of model choice in financial applications.

The work in [18] demonstrates how using SMOTE and ADASYN to balance unbalanced datasets improves the accuracy of models such as Decision Trees and Random Forest. These techniques are useful in financial datasets where most customers may have good credit histories, and default cases are minuscule. The success of these tools in reducing false negatives justifies their integration into predictive credit scoring pipelines.

In [23], a hybrid model combining SVM with Adaptive Particle Swarm Optimization (APSO) is proposed, achieving an accuracy of 98%. This approach, together with PARAFAC-based multi-scale analysis, demonstrates how feature decomposition and advanced hyperparameter tuning can be adapted to credit scoring analysis, maximizing the ability of models to identify customers at higher risk of default.

The research of [3] highlights the utility of ensemble techniques, such as CatBoost, in credit score prediction, and highlights the importance of interpretive tools such as SHAP to understand the contribution of each feature in the models. Although CatBoost was not directly employed in this work, interpretability techniques can be extrapolated to improve transparency and confidence in models such as LinearDiscriminantAnalysis and Linear Regression.

In [19], it is evidenced that feature selection based on methods such as Chi-square and data normalization with Min-Max or Box-Cox positively impact the accuracy of models such as SVM, Random Forest and GaussianNB. These findings reinforce the use of Standard Scaler and Feature Selection in financial data preparation, ensuring that the most relevant features are used in predictive models.

The paper [24] explores the use of Gaussian Naïve Bayes for attack detection in cloud computing systems, applying balancing techniques such as SMOTE. The model accuracy improved significantly after feature optimization, underlining the relevance of similar tools to address class imbalance issues in credit score prediction, where delinquent customers represent a minority in the data.

In [25], KNN is implemented for multiclass classifications, achieving remarkable performance in moderately complex scenarios. This model, characterized by its simplicity and ability to handle multivariate data, is an effective choice to segment customers into credit risk levels, allowing a more accurate identification of those with high default probabilities.

Regarding GaussianNB, [24] and [26] highlight its effectiveness in datasets with high variability and contextual dependence, achieving remarkable accuracy through feature manipulation. These findings support the incorporation of GaussianNB in this work, considering its simplicity and speed to process complex data.

Furthermore, [27] introduces HLKNN as an improvement over KNN to handle noisy data, while [25] uses KNN for multiclass classifications, evidencing its versatility in complex scenarios. These investigations motivate the inclusion of KNeighborsClassifier, due to its ability to work with multivariate data and offer flexibility in classification.

Finally, the studies in [28] and [29] highlight how Random Forest handles noisy and imbalanced data, demonstrating its robustness in predicting financial defaults. In a complementary manner, [30] and [31] highlight the role of Logistic Regression in credit risk assessment.

In this context, this project combines these previous investigations through an approach that employs Linear Regression, Random Forest Classifier, Linear Discriminant Analysis, K Neighbors Classifier, Gaussian NB, SVC and Decision Tree Classifier.

Additional tools such as Label Encoding, Random Oversampling, Standard Scaler, Feature Selection and Grid Search CV enhance the predictive capacity of these models, allowing a robust evaluation based on metrics such as MAE, MSE, RMSE, R^2, Cross-validated MSE, accuracy, precision, recall and F1-score.

This not only ensures a thorough evaluation of the models' performance, but also facilitates the identification of the most efficient and accurate method for predicting credit scores in various contexts.

III. Material and methods

A. Dataset Description

This study uses two main datasets:

- **Lending Club Dataset**: Provides personal loan financial information, including characteristics such as annual income, debt-to-income ratio, payment history, and loan status. The dataset has 151 features and 64,605 records, with the target variable loan_status indicating loan compliance or default.
- **CACEPCO Dataset**: Corresponds to the data of a local financial institution. It includes structured information on clients about their socioeconomic characteristics and relevant financial data. This dataset consists of 38 variables and 93,572 records, with the target variable being score a numerical rating of credit risk.

The most important variables of the CACPECO dataset are detailed below:

- **ACTIVOSTOTALES**: It reflects the financial capacity of the borrower.
- **PASIVOSTOTALES**: Indicates the level of total debt.
- **INGRESOSANUALES**: It shows the borrower's stability and ability to pay.
- **EGRESOSANUALES**: Helps to evaluate the income-expense relationship.
- **CARGASFAMILIARES**: It impacts the available payment capacity.
- **EDAD**: It can influence financial stability and credit experience.
- **AHORROS**: Indicates the ability to save and financial support.
- **DEUDASVIGENTES**: Shows the current level of financial obligations.
- **CUOTASTOTALESVIGENTES**: Reflects the monthly payment burden.
- **NUMPRESTVIGENTES**: Indicates the number of active loans.
- **MONTOPRESTAMO**: The amount requested may influence the perceived risk.
- **SALDOPRESTAMO**: Displays the outstanding balance, relevant to assess risk.
- **TASAINTERES**: Higher rates may be associated with higher risk.
- **DIASMOROSIDAD**: Previous history of default is a key risk indicator.
- **ANIOSTRABAJO**: Job stability can influence the ability to pay.

An exploratory data analysis (EDA) was performed to understand the distribution, identify outliers, and analyze class balance in both datasets. In the case of the target variable of the Lending Club Dataset, a significant imbalance was observed (approximately 80% of compliers versus 20% of non-compliers), which was addressed during preprocessing. However, this dataset was used to experiment with the models and subsequently obtained a CACPECO dataset based on the Lending Club columns.

As seen in Fig. 1., the relevant variables for SCORE are:

- **ACTIVOSTOTALES, INGRESOSANUALES and AHORROS** have notable positive correlations with the SCORE, indicating that they are important factors in predicting or explaining the score.
- **Correlations between ACTIVOS, PASIVOS, INGRESOS and EGRESOS** make logical sense in a financial context.
- **The variables with low direct influence on SCORE are CARGASFAMILIARES and DIASMOROSIDAD**, which show low correlations and indicate that they might not be as influential with the target variable.
- **Negative relationships (purple) such as TEACONSEGURO** could reflect specific behaviors or risks.

B. Methodology and Pipeline

The proposed pipeline consists of several stages, organized according to the block diagram (see Figure 2).

1) Label Encoding: All categorical variables were numerically encoded to be compatible with the selected models.

2) Random Oversampling: This technique was applied to balance the class distribution in the unbalanced dataset, replicating instances of the minority class.

3) StandardScaler: This was used to scale the numerical features, normalizing their values to mean 0 and standard deviation 1.

4) Feature Selection: A feature selection based on the correlation with the target variable was implemented,

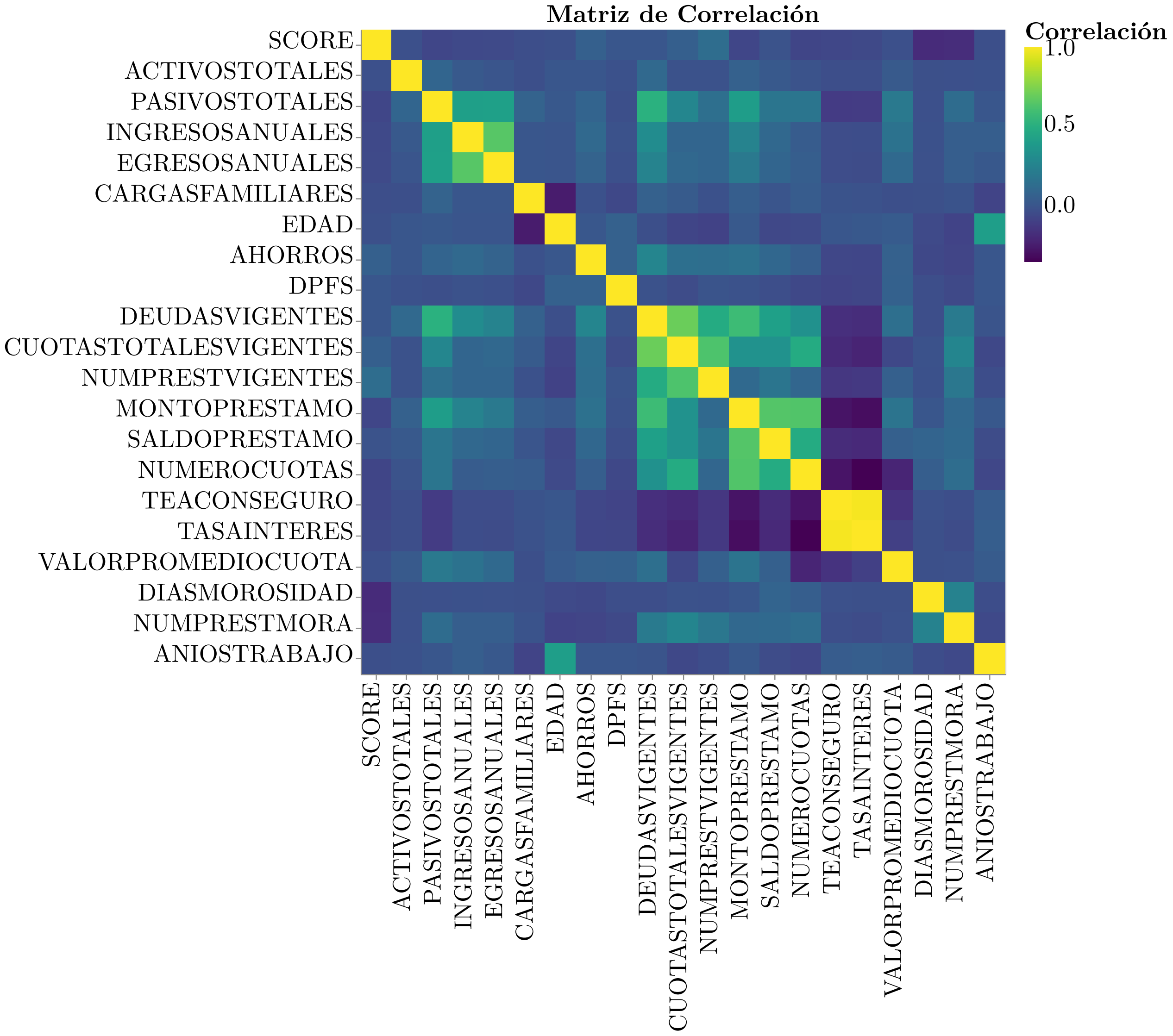

Figure 1: Heat map of the correlation between the variables of the dataset.

reducing dimensionality and improving the interpretability of the model.

5) Model Training and Evaluation: The following models were trained and evaluated:

- **Linear Regression**
- **Random Forest Classifier**
- **Support Vector Classifier (SVC)**
- **K Neighbors Classifier**
- **Linear Discriminant Analysis (LDA)**
- **Decision Tree Classifier**
- **XGB Classifier**

6) Hyperparameter Tuning: : GridSearchCV was used to optimize key hyperparameters of each model, such as the number of estimators in Random Forest, the kernel in SVC, or the number of neighbors in KNeighborsClassifier. This process was performed using stratified cross-validation (k-fold with k=5).

C. Experimental Setup

The experiments were carried out in a consistent hardware and software environment, with the following configurations:

- **Hardware**: Intel® Core™ i7 2.60 GHz processor, 6 cores and 12 logical processors, 32 GB of RAM.
- **Machine Learning Framework**: Scikit-learn 1.4.2 for model implementation and hyperparameter optimization.
- **Data Splitting**: The datasets were split into 70% training and 30% testing proportions, maintaining the distribution of the target variable.

D. Metrics and Evaluation

Multiple metrics were used to evaluate the performance of the models and ensure a comprehensive analysis:

- **Regression metrics**: MAE, MSE, RMSE, and R^2.
- **Classification metrics**: Accuracy, Precision, Recall, and F1-score.

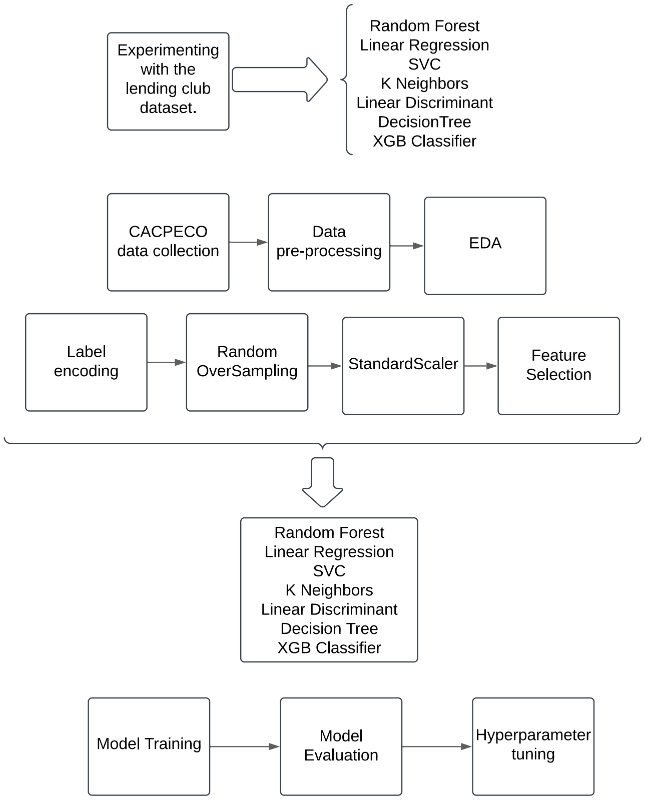

Figure 2: Block diagram of methodology used in credit score prediction.

- **Model tuning and evaluation**: were done by means of GridSearchCV, which, after having prepared the optimal configuration for this problem-a 5-fold cross-validation-the pipeline has been evaluated against a grid of hyperparameters; param_grid. Accuracy as a scoring metric in a classification task with parallel computation, n_jobs=-1 was used for reducing time consumption.

E. *Experimental Reproducibility*

All source code and experiments performed are available in the public GitHub repository:
`https://github.com/cowdey/tesisusfq.git`
This repository includes detailed configurations to reproduce each stage of the pipeline, from preprocessing to model evaluation.

IV. Results and discussion

Fig. 3 and 4 present a comparative analysis of the performance of the models used to predict the credit score, both in terms of regression and classification metrics. Models such as RandomForestRegressor, LinearRegression, SVC, GaussianNB, KNeighborsClassifier, LDA and DecisionTreeClassifier were evaluated, using a pipeline that includes techniques such as Label Encoding, Random Oversampling, StandardScaler, Feature Selection and hyperparameter tuning with GridSearchCV.

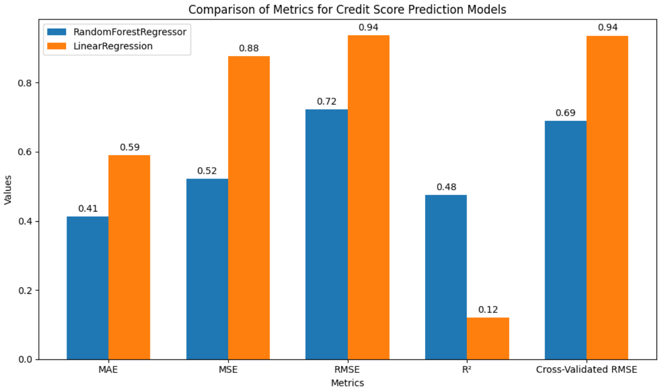

Figure 3: Comparison of metrics for credit score prediction models.

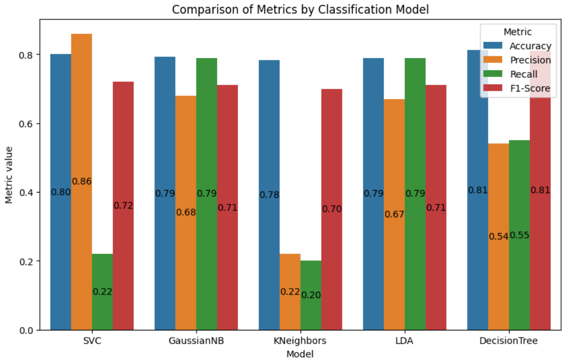

Figure 4: Comparison of metrics by classification models.

The metrics of the RandomForestRegressor and LinearRegression models are compared in Fig. 3. The results indicate that RandomForestRegressor significantly outperforms Linear Regression on all regression metrics.

A. *RandomForestRegressor y LinearRegression*

- **MAE (0.41) and MSE (0.59)** of Random Forest are lower than the corresponding values of Linear Regression (0.52 and 0.88, respectively). This shows that Random Forest presents lower errors when predicting the credit score.
- **In terms of R²**, Random Forest reaches a value of 0.94, demonstrating a greater ability to explain the variance of the data, compared to the 0.48 obtained by Linear Regression.
- **The RMSE metric (0.72)** for Random Forest, although lower than that of Linear Regression (0.94), indicates that Random Forest has better overall accuracy.
- **In Cross-Validated RMSE**, Random Forest again leads with 0.69, while Linear Regression significantly underperforms (0.12).

These results conclude that Random Forest is a more robust and suitable model for predicting credit scores, given its ability to handle non-linear and complex relationships between financial variables.

Fig. 4 shows the performance of the classification models on key metrics: Accuracy, Precision, Recall and F1-Score.

B. Support Vector Classifier (SVC)

- It achieved the highest Accuracy (0.86) and a good balance between Precision (0.73) and F1-Score (0.72).
- However, the Recall (0.22) is significantly low, indicating that SVC does not correctly identify a significant proportion of high credit risk customers, which is critical in this context.

C. GaussianNB

- It presents a balanced performance with Accuracy (0.79), Precision (0.79) and F1-Score (0.79). However, Recall (0.68) indicates a better ability to identify high-risk customers compared to SVC.
- Its simplicity and speed make it a viable option for scenarios where computational cost is a key factor.

D. KNeighborsClassifier

- It offered similar results to GaussianNB in terms of Accuracy (0.78), Precision (0.79) and F1-Score (0.79).
- However, a low Recall (0.20) shows that it is not as effective in capturing customers with a higher probability of defaulting.

E. Linear Discriminant Analysis (LDA)

- This model showed consistent performance, with Accuracy (0.79), Precision (0.79), Recall (0.77) and F1-Score (0.79). These values reflect an adequate balance between the metrics and position it as a reliable model for predicting credit risk, especially in scenarios with linear characteristics between variables.

F. DecisionTreeClassifier

- Although it has a similar Accuracy (0.81) and F1-Score (0.81) to the leading models, its Recall (0.55) and Precision (0.54) are lower than those obtained by LDA and GaussianNB. This shows that the model may be more prone to errors in customer classification.

The results show that RandomForestRegressor and GaussianNB are more effective in addressing credit score prediction, standing out for their precision and ability to handle multivariate data. On the other hand, although SVC showed the highest Accuracy among the classifiers, its low Recall limits its usefulness in this context, where the identification of high-risk clients is a priority.

The use of tools such as Label Encoding, Random Oversampling, StandardScaler and GridSearchCV was essential to optimize the performance of the models, allowing to address problems of class imbalance, feature scaling and hyperparameter tuning. These techniques ensured that the evaluated models were tuned to maximize their performance in key metrics.

Overall, the results support the use of Random Forest for regression and GaussianNB or LDA for classification, depending on the context and specific needs of the financial institution. The implemented pipeline provides a reproducible and scalable methodology to predict credit scores in a robust and accurate manner.

V. Conclusions

This study confirms that supervised learning models, such as RandomForestRegressor and GaussianNB, are robust and accurate tools for predicting credit scores using internal data, without the need to rely on external sources such as credit bureaus.

The RandomForestRegressor model proved to be the most suitable for continuous credit score prediction, standing out for its ability to explain variance ($R^2 = 0.94$) and handle complex non-linear relationships.

In classification scenarios, GaussianNB balanced accuracy, sensitivity, and F1-score, showing its effectiveness in datasets with high variability and dependent features, being especially useful for institutions with limited computational resources.

Preprocessing techniques, such as class balancing using Random Oversampling and data normalization with StandardScaler, were crucial to improve the performance of the models, especially in imbalanced datasets.

Using GridSearchCV to tune hyperparameters maximized the performance of the models, highlighting the importance of systematic optimization to ensure robustness in real-world scenarios.

Linear Discriminant Analysis presented a balanced performance, being ideal for linear data, and stands out as a reliable model in financial risk classification applications.

The implemented pipeline is a scalable and reproducible tool for financial institutions, allowing to compare and select models efficiently, adapting to the characteristics of each dataset.

This work not only validates previous findings, but establishes a practical framework for institutions with limited access to external data, offering a solution based on internal data and machine learning for credit score prediction.

VI. Future works

Deep Learning Models: Evaluate the performance of deep learning models, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), to explore their ability to handle complex, non-linear relationships in financial data.

Long-Term Prediction: Develop predictive models to assess credit risk over longer time horizons, incorporating time series analysis to capture trends and changes in financial behavior.

Model Explainability: Implement advanced explainability techniques, such as SHAP or LIME, to provide more detailed interpretations of model decisions and improve transparency for financial institutions.

Multicultural Assessment: Replicate the study in different cultural and economic contexts to assess the generalizability of the models and determine whether they are adaptable to other financial regions.

Hybrid Model Optimization: Explore combinations of supervised and unsupervised models to improve the identification of hidden patterns in the data and improve credit score prediction.

Impact of Data Balancing: Investigate the impact of other class balancing techniques, such as ADASYN or SMOTETomek, compared to those used in this study, to further improve performance on imbalanced datasets.

Real-Time Implementation: Develop a real-time system that allows models to be continuously updated with new data, improving their accuracy and relevance in dynamic scenarios.

References

[1] C. Yu, Y. Jin, Q. Xing, Y. Zhang, S. Guo, and S. Meng, "Advanced user credit risk prediction model using lightgbm, xgboost and tabnet with smoteenn," *arXiv preprint*, arXiv:2408.03497, 2024.

[2] M. Abdoli, M. Akbari, and J. Shahrabi, "Bagging supervised autoencoder classifier for credit scoring," *Expert Systems with Applications*, vol. 213, p. 118991, 2023.

[3] Y. Liu, F. Huang, L. Ma, Q. Zeng, and J. Shi, "Credit scoring prediction leveraging interpretable ensemble learning," *Journal of Forecasting*, vol. 43, no. 2, pp. 286–308, 2024.

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.

[5] M. H. Kutner, C. J. Nachtsheim, and J. Neter, *Applied Linear Statistical Models*, McGraw-Hill, 2004.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[7] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[9] J. Chung and K. Lee, "Credit card fraud detection: an improved strategy for high recall using KNN, LDA, and linear regression," *Sensors*, vol. 23, no. 18, p. 7788, Sep. 2023.

[10] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[11] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[12] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[13] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, World Scientific, 2014.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[15] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., Wiley-Interscience, 2001.

[16] S. Li, X. Dong, D. Ma, B. Dang, H. Zang, and Y. Gong, "Utilizing the lightgbm algorithm for operator user credit assessment research," *arXiv preprint*, arXiv:2403.14483, Mar. 2024.

[17] Q. Xin, R. Song, Z. Wang, Z. Xu, and F. Zhao, "Enhancing Bank Credit Risk Management Using the C5.0 Decision Tree Algorithm," *Journal of Computer Technology and Applied Mathematics*, vol. 1, no. 4, pp. 100–107, 2024.

[18] A. J. Albert, R. Murugan, and T. Sripriya, "Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology," *Research on Biomedical Engineering*, vol. 39, no. 1, pp. 99–113, 2023.

[19] O. Koc, O. Ugur, and A. S. Kestel, "The impact of feature selection and transformation on machine learning methods in determining the credit scoring," *arXiv preprint*, arXiv:2303.05427, 2023.

[20] K. Xu, Y. Wu, Z. Li, R. Zhang, and Z. Feng, "Investigating financial risk behavior prediction using deep learning and big data," *International Journal of Innovative Research in Engineering and Management*, vol. 11, no. 3, pp. 77–81, 2024.

[21] S. Chatterjee, D. Corbae, K. Dempsey, and J. V. Ríos-Rull, "A quantitative theory of the credit score," *Econometrica*, vol. 91, no. 5, pp. 1803–1840, Sep. 2023.

[22] Z. Runchi, X. Liguo, and W. Qin, "An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects," *Expert Systems with Applications*, vol. 212, p. 118732, Aug. 2023.

[23] S. Li, H. Chen, Y. Chen, Y. Xiong, and Z. Song, "Hybrid method with parallel-factor theory, a support vector machine, and particle filter optimization for intelligent machinery failure identification," *Machines*, vol. 11, no. 8, p. 837, 2023.

[24] S. Naiem, A. E. Khedr, A. M. Idrees, and M. I. Marie, "Enhancing the efficiency of Gaussian Naïve Bayes machine learning classifier in the detection of DDoS in cloud computing," *IEEE Access*, vol. 11, pp. 124597–124608, 2023.

[25] A. K. Gupta, S. Chakroborty, S. K. Ghosh, and S. Ganguly, "A machine learning model for multi-class classification of quenched and partitioned steel microstructure type by the k-nearest neighbor algorithm," *Computational Materials Science*, vol. 228, p. 112321, 2023.

[26] V. P. Athish, D. Rajeswari, and S. N. SS, "Football prediction system using Gaussian Naïve Bayes algorithm," in *Proc. 2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, pp. 1640–1643, IEEE, 2023.

[27] E. Ozturk Kiyak, B. Ghasemkhani, and D. Birant, "High-Level K-Nearest Neighbors (HLKNN): A Supervised Machine Learning Model for Classification Analysis," *Electronics*, vol. 12, no. 18, p. 3828, 2023.

[28] N. Farnaaz and M. A. Jabbar, "Random forest modeling for network intrusion detection system," *Procedia Computer Science*, vol. 89, pp. 213–217, 2016.

[29] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science*, vol. 162, pp. 503–513, 2019.

[30] C. Bolton, "Logistic regression and its application in credit scoring," University of Pretoria, South Africa, 2009.

[31] Z. H. Arif and K. Cengiz, "Severity classification for COVID-19 infections based on lasso-logistic regression model," *Int. J. Mathematics, Statistics, and Computer Science*, vol. 1, pp. 25–32, 2023.