A Proposed Framework for Loan Default Prediction Using Machine Learning Techniques

Mona Aly SharafEldin[1], Amira M. Idrees[2], Shimaa Ouf[3]
Business Information Systems, Helwan University, Egypt[1]
College of Business, King Khalid University, Saudi Arabia[2]
Information Systems Department, Helwan University, Egypt[3]

***Abstract*—The accurate prediction of loan defaults is critical for the risk management strategies of financial institutions. Traditional credit assessment approaches have often relied on subjective judgment, leading to inconsistent decisions and heightened financial risk. This study investigates the application of machine learning techniques—namely Random Forest, Decision Tree, and Gradient Boosting—to predict loan defaults using customer data from the Agricultural Bank of Egypt. The research emphasizes the role of feature selection in enhancing model performance, utilizing both embedded and recursive methods to isolate key predictive attributes. Among the evaluated features, loan balance, due amount, and delinquency history emerged as the most influential, while demographic variables like gender and employment status were found to be less significant. The Decision Tree model demonstrated superior performance with an overall accuracy of 88%, a recall of 53%, and a specificity of 89%, making it the most effective among the tested classifiers. The findings highlight the importance of combining robust feature selection with interpretable models to support informed decision-making in banking.**

***Keywords*—Random forest; decision trees; gradient boosting machines; feature selection; feature importance; loan default**

I. Introduction

In recent decades, the financial sector has increasingly adopted credit lending as a central business activity. However, many of these institutions are still facing serious problems with respect to the proper assessment of credit risk, mainly because there is a lack in traditional evaluating models that do not count on the real behavior of the borrower. Accurate estimation of credit risk is a key factor in maintaining financial stability and improving the performance of institutions in the data-driven economy of today [1].

shimaaouf@commerce.helwan.edu.egWith the proliferation of data and computing capabilities, machine learning has become available for mainstream risk management tasks. The latest technologies provide enhanced prediction on loan repayment and default identification. One of the biggest problems in creating accurate predictive models focusing on relevant features is the process of eliminating noise and irrelevant signals and zeroing in on what really matters: input variables to those most impactful on prediction results. Removing irrelevant or repetitive features not only reduces the risk of overfitting but also simplifies the model and enhances interpretability [2].

Mona.Aly21@commerce.helwan.edu.eg[1]
aidrees@kku.edu.sa[2]
shimaaouf@commerce.helwan.edu.eg[3]

To address these challenges, various feature selection approaches, such as filtering, wrapping, and embedded methods, are employed to refine datasets and ensure that learning algorithms can generalize effectively. When assessing the likelihood of loan default, classification algorithms, like decision trees, support vector machines, and random forests, are commonly used. Furthermore, ensemble learning methods, such as AdaBoost [3], have proven to improve performance by combining multiple predictive models.

These algorithms analyze borrower-specific financial data, including income levels, employment details, credit scores, and historical payment patterns. The integration of feature selection with these advanced models enables financial institutions to build more reliable predictive systems that not only improve forecast accuracy but also provide clear insights into the variables that influence repayment behavior [4].

Essentially, a loan is a financial arrangement in which a borrower receives funds from a lender under specific repayment terms. This agreement includes the repayment of the principal amount, interest, and other applicable charges. To approve a loan, lenders typically examine a borrower's financial history and current obligations to assess their capacity to fulfill the repayment terms. If the borrower fails to meet payment deadlines beyond a predetermined grace period, the loan may be categorized as default [5].

As the financial sector continues to evolve, early identification of potential loan defaults has become increasingly important for minimizing institutional risk. By leveraging machine learning methods and carefully selected data inputs, lenders can make more informed, accurate, and strategic credit decisions.

The primary objective of this study is to identify the key factors that influence decision-making in the context of loan defaults. By applying machine learning methods, the study aims to explore which attributes significantly affect default risk and to evaluate the effectiveness of different machine learning methods in this domain. Additionally, the study seeks to determine the most suitable technique for accurately predicting loan default outcomes, thereby supporting more informed and data-driven financial decision-making.

The contributions were made to the field of financial risk modelling by focusing on loan default prediction in a specific context: livestock-based microfinance within Egypt's agricultural sector. The study applies machine learning

techniques tailored to the particular characteristics and challenges of this domain. The contributions can be summarized as follows:

A. Loan Default Prediction in Livestock-Based Lending

The study addresses a relatively under-researched area: livestock-based lending using real-world data obtained from the Agricultural Bank of Egypt, which serves rural communities across more than 1,190 branches. Unlike commonly studied domains such as credit cards or personal loans, this dataset reflects borrowing behavior in a specialized agricultural finance context.

B. Application of Machine Learning Techniques

Various supervised machine learning algorithms, including Random Forest, Decision Tree, and Gradient Boosting, were utilized to evaluate feature significance and create predictive models. Important variables, such as balance (BAL), due amount (DUE), and delinquency score (DELI), emerged as strong predictors of loan defaults. Conversely, attributes like gender and occupation showed minimal predictive power and were omitted to streamline the models and enhance clarity.

C. Decision Tree Analysis for Interpretability

The Decision Tree model, which recorded the highest accuracy (88%) among the models tested, was chosen for deeper examination due to its interpretability. Analysis of the root node indicated that a balance threshold of ≤ 2140 was a key risk indicator. Additional branches included factors like region (ZONE_NAME), DELI, and DUE, allowing the model to recognize interactions within borrower attributes.

D. Regional Influence on Risk Classification

Borrowers from the "East Delta" region exhibited more uncertain classification outcomes, suggesting that regional factors may influence loan repayment behavior. This highlights the value of incorporating geographic context in risk assessment models, especially in rural or sector-specific lending scenarios.

E. Gini Index as a Measure of Classification Certainty

The Gini index was employed to assess the certainty of classification outcomes across various decision tree paths. Lower Gini values indicated greater confidence in predictions, whereas moderate values signaled more variability in borrower characteristics. This measure facilitated the identification of segments where the model's forecasts were more or less dependable, aiding in practical decision-making.

F. Comprehensive Performance Assessment

In addition to accuracy, the study included further evaluation metrics like precision, recall, and F1-score to offer a more thorough insight into model performance. This methodology is particularly important in financial contexts, where both false positives and false negatives have significant operational ramifications.

The research is structured into eight sections: Introduction is given in Section I. Related Work in Section II. Proposed Framework in Section III. Data Integration in Section IV. Data preprocessing in Section V. Feature Selection in Section VI. Model Building in Section VII. Model Evaluation in Section VIII. Results in Section IX. Discussion in Section X. Finally, Conclusion and Future Work in Section XI and XII.

II. Related Work

This section provides an overview of existing studies related to the prediction of default loans using machine learning methods. In recent years, this area has been the focus of numerous studies.

Puli (2024) [6] utilized various machine learning methods, including random forests, naïve Bayes, gradient boosting, support vector machines, neural networks, k-nearest neighbors, and decision trees, along with statistical approaches, such as logistic regression. The findings indicate that neural networks and random forest models demonstrated notable efficacy in predicting banking crises in India. Additionally, the research evaluated the performance of multiple algorithms, including neural networks, naïve Bayes, and ensemble techniques, such as random forest bagging, boosted decision trees, stacking of logistic regression, support vector machines, and neural networks, for forecasting loan defaults. The ensemble models consistently outperformed individual classifiers. Specifically, boosted decision trees achieved the highest accuracy (84.9 %), followed by random forest (83.1 %). Neural networks recorded an accuracy of 80.3%, surpassing other individual classifiers, whereas naïve Bayes exhibited significantly lower performance at 46.5%. These results underscore the superior predictive capability of the ensemble methods in this domain.

Saini (2023) [7] explored the application of various machine learning algorithms to predict loan approval decisions. The study implemented Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression (LR) to evaluate their predictive capabilities. The comparative analysis revealed that the Random Forest model delivered superior performance, achieving an accuracy of 98.04%, thus indicating its robustness and effectiveness in loan approval prediction tasks.

Jovanne (2023) [8] employed various supervised machine learning algorithms, k-nearest neighbors (k-NN), decision trees (J48), naïve Bayes, and logistic regression to predict loan default using a dataset of 1,000 instances. Preprocessing steps included missing value imputation, normalization, and SMOTE for class balancing. Feature selection was conducted using correlation, information gain, and wrapper methods. The models were evaluated using classification accuracy, F-measure, and kappa statistics. Results showed that k-NN (k=3) and logistic regression achieved the highest accuracy rates of 78.38% and 77.31%, respectively. These findings highlight the potential of data mining techniques in enhancing credit risk assessment in financial institutions.

Alaradi and Hilal (2020) [9] utilized a predictive model for assessing loan approval using a variety of decision tree-based algorithms, ranging from simple decision trees to more complex ensemble methods like random forests. Their findings indicated that basic decision trees struggled to deliver strong performance, likely due to their limited capacity to capture the intricate and highly correlated relationships among key loan-related features. Despite this, the decision tree (DT) model demonstrated a strong balance between accuracy, interpretability, and practical

relevance. It achieved a test accuracy of 97.25%, making it a viable option for automating and accelerating the evaluation of loan applications based on applicant characteristics. The study recommended that using the DT-based prediction model can support more transparent and efficient decision-making in loan processing systems.

Güder and Köse (2024) [10] conducted a study to develop predictive models for home loan approval decisions using machine learning algorithms. The primary goal was to enhance the efficiency and accuracy of loan processing in the banking sector, minimizing human error and improving decision-making speed. The researchers employed four popular supervised learning techniques: k-nearest neighbors (KNN), random forest (RF), support vector machines (SVM), and logistic regression (LR). Two separate datasets were used to assess the performance of these models across varying data conditions. Each dataset was split into 90% for training and 10% for testing. The performance metrics included accuracy, precision, recall, F1-score, specificity, and others. The results indicated that SVM achieved the best performance on the first dataset with an accuracy of 88.7%, while Random Forest performed best on the second dataset, reaching an accuracy of 98.8%. Additionally, Random Forest achieved the highest precision (98.9%) and recall (99.3%) in the second dataset, highlighting its strong predictive capability in more homogeneous data environments. In contrast, the performance of algorithms on the smaller, more variable dataset was slightly lower, illustrating the impact of data quality and size on model accuracy.

Sampurna and Vidya (2023) [11] discuss how the XGBoost and Random Forest algorithms outperformed decision trees, logistic regression, and regularized logistic regression in terms of classification performance. The research highlights the importance of evaluating multiple algorithms to identify the most effective one.

Tabassum, Namita, and Prachi (2025) [12] conducted a study to evaluate the effectiveness of various machine learning models in forecasting financial distress among publicly listed companies in Vietnam. The results indicate that the Extreme Gradient Boosting (XGBoost) model achieved the highest predictive accuracy at 95.66%, while the Artificial Neural Network (ANN) yielded the lowest accuracy at 91.68%. To enhance the interpretability of the models, SHAP (Shapley Additive Explanations) values were employed to identify the most influential financial indicators. Key variables, such as the long-term debt-to-equity ratio, enterprise value-to-sales ratio, accounts payable-to-equity ratio, and diluted earnings per share (EPS), were found to play a critical role in predicting financial instability. The study not only offers a novel analytical tool for credit rating agencies to assess default risks but also contributes to the broader effort to make complex machine learning models more transparent and explainable.

Hussain (2024) [13] conducted a study employing supervised machine learning algorithms, including Decision Tree, Naive Bayes, Multilayer Perceptron, and a Stacking Ensemble Model, to predict loan approval outcomes in the banking sector. The dataset used comprised borrower-related features, such as income, credit history, loan amount, and employment status. Data preprocessing was enhanced using K-Nearest Neighbors (KNN) for missing value imputation, alongside clustering techniques (K-Means) to segment applicants based on income and dependent attributes.

The study applied classification algorithms to both general and cluster-specific data. The Stacking Model demonstrated the highest testing accuracy (83.24%), outperforming Naive Bayes (82.16%), Decision Tree (80%), and Multilayer Perceptron (73.51%). Additionally, clustering models based on applicant income and marital/dependent status yielded varied accuracies across clusters, ranging from 33.33% to 81%, indicating the effectiveness of group-specific modeling.

Chen (2023) [14] proposed a framework for detecting financial statement fraud in publicly listed companies by utilizing several machine learning models, including LightGBM, XGBoost, Gradient Boosting Decision Trees (GBDT), and Random Forest. To enhance model performance, the study also introduced an integrated feature selection approach. Furthermore, the use of the Synthetic Minority Over-sampling Technique (SMOTE) effectively addressed the issue of class imbalance, leading to a notable improvement in fraud detection accuracy. Among the evaluated models, GBDT demonstrated superior performance, achieving the highest Area Under the Curve (AUC) and sensitivity metrics.

Ali (2023) [15] introduced a fraud detection model based on the XGBoost algorithm, aimed at identifying fraudulent financial activities within companies across the Middle East and North Africa (MENA) region. To address the issue of class imbalance in the dataset, the study employed the Synthetic Minority Over-sampling Technique (SMOTE). Various machine learning algorithms were implemented using Python to predict instances of financial statement fraud. The experimental results demonstrated that XGBoost achieved the highest accuracy, reaching 96.05%, outperforming other models such as Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM).

Gupta and Mehta (2021) [16] conducted a study to examine the application of statistical and machine learning methodologies in detecting financial statement fraud. The findings indicate that the fuzzy technique, neural network, decision tree, SVM, logistic regression, and probit regression achieved the highest precision rates of 89.5%, 71.7%, 73.6%, 90.4%, and 86.8%, respectively. Consequently, machine learning techniques demonstrated superior performance compared with statistical methods in predicting the likelihood of fraud in businesses, particularly when the sample had limited data access, achieving a precision rate of 96.05%.

Venkatesan, A. Kumar and S. Sabni, (2020) [17] used a machine learning algorithm to detect credit card fraud based on the customer's transaction, and the accuracy of the system was predicted using machine classification algorithms, such as logistic regression and KNN, based on credit card data. The KNN algorithm delivered the best results in terms of statistical measures, such as precision 0.95), recall (0.72), and f1-score (0.82) for fraud.

Vahid (2024) [18] conducted research employing machine learning models, Logistic Regression, KNN, SVM, Decision Tree, and Random Forest, to predict loan approval. Using a

Kaggle dataset, the study tested the impact of two-feature selection methods: K-Best and Recursive Feature Elimination (RFE). The Random Forest model achieved the highest accuracy (97.71%) when paired with RFE and cross-validation. Results showed that selecting key features significantly improves model performance, and Random Forest is especially effective in loan prediction tasks.

Kozinaa, (2023) [19] proposed automated models for predicting defaults in leasing companies using machine-learning techniques. These models use a combination of internal data from company systems and external sources from financial supervisory institutions, incorporating features related to leasing contracts and asset specifications. The experimental results showed that the Random Forest algorithm achieved the highest precision for default prediction, reaching up to 75%, whereas deep neural networks recorded the highest recall at 81.6%. For non-default predictions, the deep learning methods demonstrated the highest precision, whereas AdaBoost achieved the best recall. Although these models were developed in the context of leasing firms, the authors suggest that their flexibility allows them to be applied across various financial institutions to support credit risk management. However, a notable limitation is the substantial amount of detailed input data required, which could present practical challenges for real-world deployment. The authors recommend that future studies explore statistical forecasting techniques to benchmark and possibly enhance current models.

Jumaa (2023) [20] introduced a deep-learning model to predict loan default risks among UAE bank customers. Using data from 1000 respondents, the model translated demographic and financial factors, such as income, debt ratio, and credit card use, into predictive inputs. After pre-processing, a neural network was developed with 25 input features and tested using Keras and TensorFlow. The model achieved 97.6% training accuracy and 95.2% training accuracy on the test data, showing a strong potential in supporting lending decisions. The authors emphasized the model's practical value and called for further research using regression and variance analyses.

Viswanatha (2023) [21] addressed the ongoing difficulties faced by banks in accurately identifying suitable loan applicants amid rising demand. To enhance the selection process, the study suggested integrating various machine learning techniques with ensemble learning methods. The research applied algorithms, such as Random Forest, Naive Bayes, Decision Tree, and K-Nearest Neighbors (KNN), achieving a maximum accuracy of 83.73%, with Naive Bayes outperforming the others. This approach not only improves selection precision but also significantly reduces loan processing time, benefiting both applicants and banking staff by streamlining the approval procedure.

Btoush (2025) [22] proposed a hybrid ensemble framework combining classical machine learning techniques—such as decision tree (DT), random forest (RF), support vector machine (SVM), XGBoost, CatBoost, and logistic regression (LR)—with deep learning models, including convolutional neural networks (CNN) and bidirectional LSTM (BiLSTM) enhanced with attention mechanisms. The model employed a stacking approach to integrate predictions from all classifiers, using Random Forest as the final estimator. To address the challenge of data imbalance in detecting credit card fraud, the authors applied robust pre-processing and evaluation strategies. The hybrid model achieved outstanding performance, with an accuracy of 99.97%, precision of 97.62%, recall of 83.67%, and an F1-score of 90.11%. These results demonstrated its effectiveness in minimizing both false positives and undetected fraudulent transactions. The study highlighted the model's scalability, interpretability, and suitability for real-time fraud detection.

Several previous studies have explored the use of machine learning methods for classification tasks. Nevertheless, many of these studies have primarily relied on publicly accessible or benchmark datasets and have not thoroughly tackled significant challenges such as class imbalance or the relevance of findings in real-world scenarios. Although models such as XGBoost, Random Forest, and SVM have shown promising outcomes in certain contexts, the degree to which these results translate to operational environments remains unclear.

This study aims to fill two identified gaps in the existing literature. First, regarding the choice of algorithms, it employs established machine learning techniques—specifically, Decision Tree, Random Forest, and XGBoost—selected for their effective combination of predictive accuracy, interpretability, and resilience across various data distributions. Second, and more importantly, the research utilizes actual operational banking data from the Agricultural Bank of Egypt, as opposed to depending on conventional or synthetic datasets. This choice offers a more realistic basis for developing and assessing models.

In Table I, the results of the machine learning techniques utilized in the previous studies have been presented.

TABLE I. RESULTS OF THE EMPLOYED TECHNIQUES IN PREVIOUS STUDIES

Reference	Techniques	Accuracy	Recall	Precision	F1-Score
Puli (2024) [6]	DT	84.9%9	98.98%	100%	99.49%
	RF	83.10%	-	-	-
	NN	80.30%	-	-	-
	Naïve Bayes	46.50%	-	-	-
Saini (2023) [7]	RF	98.04%	-	-	-
Jovanne (2023) [8]	DT	78.38%	-	-	-
	k-NN	78.58%			
	Naïve Bayes	76.54%			

	Logistic	77.31%			
Alaradi &Hilal(2020) [9]	LR	79.50%	-	-	-
	KNN	78.41%	-	-	-
	SVM	93.78%	-	-	-
	DT	87%	-	-	-
	RF	97.68%	-	-	-
Güder a& Köse (2024) [10]	Data set1 KNN	87%	95.70%	88.20%	91.80%
	Data set1RF	87.00%	95.70%	88.20%	91.80%
	Data set1LR	87.00%	97.90%	95.70%	92%
	Data set1SVM	88.70%	97.90%	87%	93.10%
	Data set2 KNN	92.90%	92.50%	96.10%	94.30%
	Data set2RF	98.80%	92.50%	98.90%	99.10%
	Data set2LR,	90.80%	94%	92.90%	92.70%
	Data set2SVM	93.90%	94%%	96.20%	95.10%
Sampurna & Vidya (2023) [12]	XGBoost, RF, LR				
Tabassum (2025) [13]	XGBoost	95.66%,	-	-	-
	ANN	91.68%.	-	-	-
Hussain (2024) [14]	DT	80%	-	-	-
	Naïve Bayes	82.16%	-	-	-
	Multilayer Perceptron	73.51%	-	-	-
	Stacking Model	83.24%	-	-	-
Chen (2023) [15]	RF+SMOTE	60.55%	-	-	-
	GBDT+SMOTE	72.24%	-	-	-
	XGBoost + SMOTE	66.46%	-	-	-
	LGB+SMOTE	69.47%	-	-	-
Ali (2023) [16]	LR	73.88%	83.44%	80.55%	81.96%
	DT	82.22%	50%	41.11%	45.13%
	SVM	88.88%	84.11%	80.34%	82.18%
	RF	80.55%	50%	40.27%	0.4461%
	AdaBoost	83.33%	50%	41.66%	45.45%
	XGBoost	93.66%	86.37%	79.38%	82.72%
Gupta & Mehta (2021) [17]	SVM	90.4%	-	-	-
	LR	86.8%	-	-	-
	DT	73.6%	-	-	-
Vengatesan (2020) [18]	LR, KNN		72%	95%	82%
Vahid (2024) [19]	LR	79.50%	-	-	-
	KNN	78.41%	-	-	-
	SVM	93.78%	-	-	-
	DT	87.02%	-	-	-
	RF	97.68%	-	-	-
Kozina & Michaland (2023) [20]	RF	-	81.6 %	75.0 %	-
Jumaa (2023) [21]	NN	95.2%	-	-	-
Viswanatha (2023) [22]	RF,	77.23%			
	NB	83.73%,	-	-	-
	DT	63.41%	-	-	-
	KNN	77.23%	-	-	-
Btoush (2025) [23]	DT	99.93%	81.63%	89.89%	-
	RF	99.96%	76.53%	97.40%	-
	SVM	99.94%	66.33%	97.02%	-
	XGBoost,	99.96%	77.55%	95.00%	-
	LR	99.92%	60.20%	88.06%	-

III. Proposed Framework

Fig. 1 illustrates the proposed framework for predicting loan defaults using machine-learning techniques. The model began with data integration and selection, combining historical customer records and geographic information from the Egyptian Agricultural Bank. Following this, a comprehensive data preprocessing phase ensures data quality through steps such as handling missing values, normalization, encoding, and outlier detection. Feature selection is then conducted to extract the most informative attributes and uncover significant variable interactions. The refined dataset was subsequently used to train predictive models using machine learning algorithms. Finally, the different models are evaluated using robust performance metrics and visualized through interpretive tools to facilitate informed decision-making in credit risk assessment.

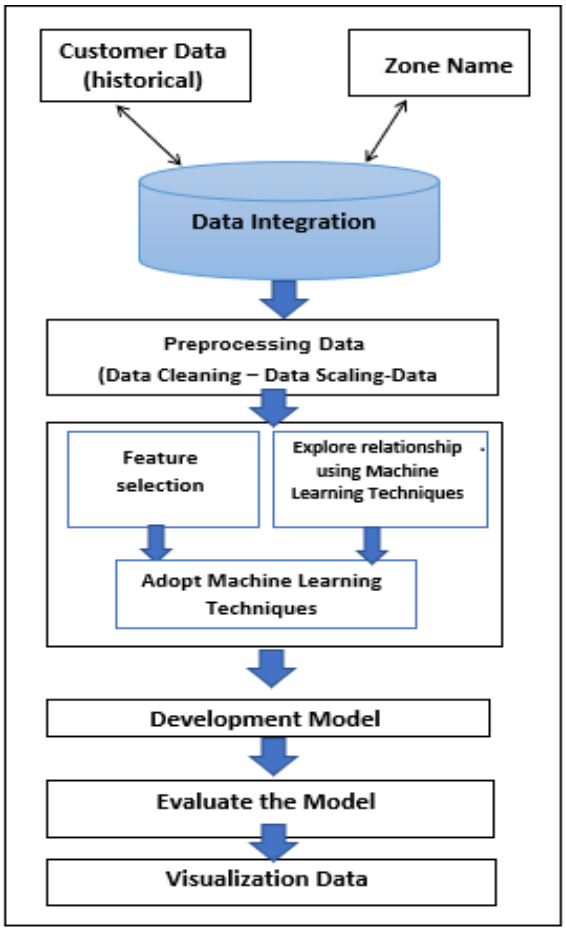

Fig. 1. Proposed framework.

IV. Data Integration

The first stage of the proposed framework centers on merging two key data sources: historical customer information and geographic zone data sourced from the Egyptian Agricultural Bank. The combined dataset consists of 168,95100 entries, each featuring attributes that represent the demographic, financial, and geographic characteristics of individual borrowers. This integration process guarantees the data's consistency, completeness, and relevance, thus setting it up for subsequent analytical tasks. As illustrated in Table II, the chosen features were directly sourced from customer records. These attributes were selected based on their potential to predict loan defaults and include variables such as gender, age category, employment status, loan amount, history of delinquency, loan due date, and the corresponding geographic zone. The final compilation of selected features is as follows:

TABLE II. The List of Attributes of ABE Data

Feature	Data	Column name
Name of zone	Each zone name included the branch it follows.	ZONE_NAME
Gender	Gender and Sex of borrowers.	GENDER
Age of the borrower	age range	AGE_RANGE
job of the borrower	yes\no	JOB
Balance of loan	Loan amount requested by the borrower	BAL
Delinquency	The number of times the borrower had been delinquent for 30+ days in the past 3 years	DELI
Due date for loan	DATE of due loan	LOAN_DUE_DATE
Classification	target for prediction (loan status according to table (0\1)	CLASSIFICATION

To get ready for model training, the target variable (CLASSIFICATION) was established by transforming the original delinquency status into a binary format (Yes/No), based on the days past due (DPD). As explained in the corresponding delinquency classification in Table III, records with DPD values of 91 days or more (i.e., buckets 4 to 7) were categorized as "defaulted (Yes=1)", while those with values below this threshold (Current to Bucket-3) were designated as "non-defaulted (No=0)".

TABLE III. Show Retail Loans Classification

Delinquency buckets	DPD as per Due Date	Retail Classification ORR	Non-accrual status	Classification based on past due days count
Current	0	Performing	no	
bucket-1	Jan-30		no	
bucket-2	31-60	Sub-standard	no	
bucket-3	61-90		no	
bucket-4	91-120	Doubtful	yes	
bucket-5	121-150	bad/non-performing	yes	
bucket-6	151-180		yes	
bucket-7	181+		yes	

A. Descriptive Analysis

1) Sample distribution by geographic zone. Table IV and associated graph presented in Fig. 2 below offer a descriptive summary of the dataset, emphasizing the geographic distribution of loan applicants. This analysis indicates that the West Delta region is the primary source of observations, representing the largest share of the sample. Conversely, the Branches Sector makes up the smallest portion.

This distribution highlights the regional concentration of applicants, suggesting a potential geographic influence on lending patterns that may be explored further in subsequent analyses.

TABLE IV. BRIEF RESULT OF THE DISTRIBUTION OF THE SAMPLE ACCORDING TO ZONE

Category	Frequency	Percentage
"West Delta"	1729335	35.93%
"East Delta"	1079627	22.43%
"South Upper Egypt"	1044724	21.70%
"North Upper Egypt"	932144	19.37%
"Branches sector"	27686	0.58%

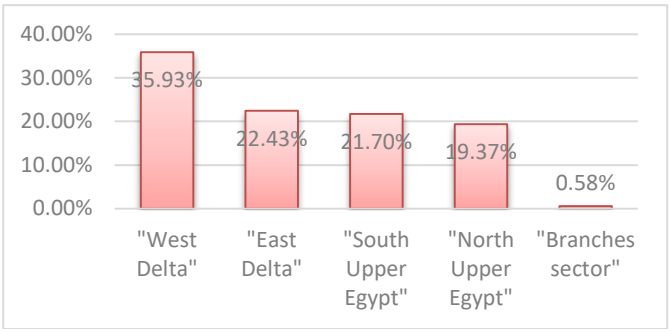

Fig. 2. Distribution of the sample according to zone.

2) Loan default status. The following table summarizes the dataset concerning the loan default status of applicants. It is clear that a significant majority of the instances are classified in the "No Fault" category, signifying that most borrowers have not defaulted on their loans. A mere fraction of the sample is linked to default (see Table V).

TABLE V. BRIEF RESULT OF THE DISTRIBUTION OF THE SAMPLE ACCORDING TO LOAN DEFAULT

Fault	Frequency	Percentage
No	4638395	4%
Yes	173299	96%

3) Age distribution. Table VI illustrates the age distribution of the sample. It is clear that the highest percentage of individuals is in the 56 to 60 age bracket, comprising 37.9% of the overall sample. In contrast, the 21 to 25 age cohort constitutes the smallest portion, with merely 0.8% of the cases. This indicates that the dataset is predominantly made up of older individuals, whereas younger age groups are significantly underrepresented. Fig. 3 shows the distribution of the sample according to age range.

4) Gender distribution. The distribution of the dataset by gender indicates a significant imbalance, with male clients comprising the majority at approximately 79%, while female clients represent only 21%, as shown in Table VII. Disparity may reflect broader demographic or socioeconomic trends within the bank's customer base, and highlights the importance of examining whether gender has a substantial impact on loan default prediction

TABLE VI. BRIEF RESULT OF THE DISTRIBUTION OF THE SAMPLE ACCORDING TO AGE RANGE

AGE_RANGE	Frequency	Percentage
21-25	39602	0.8%
26-30	211996	4.4%
31-35	298947	6.2%
36-40	400692	8.3%
41-45	428711	8.9%
46-50	447490	9.3%
51-55	483336	10.0%
56-60	1825130	37.9%
61-65	417645	8.7%
66-70	259967	5.4%

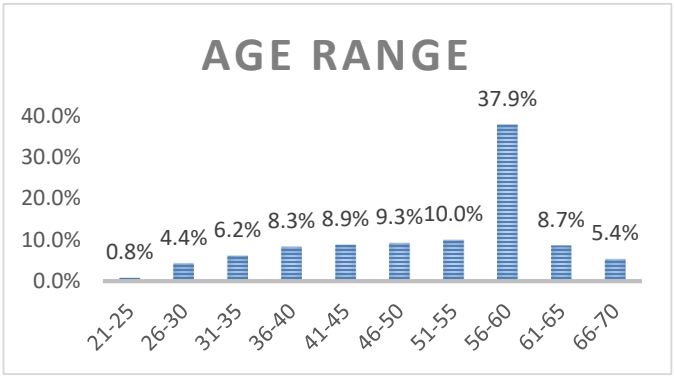

Fig. 3. Display of the distribution of the sample according to age range.

TABLE VII. BRIEF RESULT OF THE DISTRIBUTION OF THE SAMPLE ACCORDING TO GENDER

Gender	Frequency	Percentage
M	3799435	79%
F	1014081	21%

5) Job distribution. Table VIII illustrates the distribution of the sample according to employment status. It is clear that the overwhelming majority of individuals in the dataset are employed, representing 99.7% of the total sample, whereas a negligible portion (0.3%) are not currently employed.

TABLE VIII. BRIEF RESULT OF DISTRIBUTION OF THE SAMPLE ACCORDING TO JOB

Job	Frequency	Percentage
Work	4798416	99.7%
Not Work	15100	0.3%

After presenting how the data are integrated and presenting descriptive analysis of the data, now the second step is to present how the data is preprocessed.

V. Data Preprocessing

Before conducting the analysis, the integrated dataset underwent several preprocessing steps to ensure data quality and prepare it for modeling. These steps are essential, particularly in financial applications such as bank loan risk assessment, where the reliability of data directly impacts model performance.

A. Data Cleaning

Handling Missing Values: in this study, an available case method was used, which is that only complete records were retained for analysis. Although this approach may lead to information loss especially if the proportion of missing values is high it was considered appropriate given the large dataset size. Retaining complete cases helped maintain data consistency without significantly affecting the representativeness of the sample.

Outlier Detection: The dataset was reviewed for outliers using the Z-score test, but no significant extreme values were found. While outlier treatment is important for models sensitive to such values (e.g., linear regression), it is less critical for more robust methods like decision trees and ensemble algorithms.

B. Data Scaling

To prepare the dataset for machine learning models and maintain uniformity among various variable types, a range of data transformation methods was utilized. These preprocessing procedures were driven by the organization and meaning of the accessible data, which comprises a combination of categorical and numerical attributes, including ZONE_NAME, GENDER, AGE_RANGE, JOB, YEAR, MONTH, BAL, DUE, DELI, and LOAN_DUE_DATE.

1) Normalization of numerical features. Continuous variables, particularly balance (BAL), amount due (DUE), and delinquency (DELI), were standardized to align their scales and enhance computational efficiency. For example, the balance values, which frequently attained high magnitudes, were converted to thousands to facilitate interpretation and minimize computational demands during training. This normalization guaranteed that individual features would not unduly impact the learning process because of variations in unit scales.

2) Removing duplicates and noise. The dataset underwent a thorough examination to spot and remove duplicate entries, along with irrelevant or inconsistent data that might skew the analysis, which accounted for no more than 4 per cent of the total. This process was essential to uphold data integrity and avoid misleading trends from affecting model training. However infrequent, entries with unlikely balance or delinquency figures were assessed in relation to their corresponding loan due dates and financial behavior patterns prior to deciding whether to keep or discard them.

C. Data Splitting

In this study, the dataset was divided into two parts: a training set and a testing set, utilizing an unconventional split ratio. After trying out several partitioning methods, a 60% training and 40% testing split was chosen due to its superior performance in initial assessments. This selected ratio guaranteed that the training subset was adequate for effective model learning, while the somewhat larger test set facilitated a more thorough and reliable evaluation of the model's generalization abilities. This division was also reinforced by implementing K-fold cross-validation, which improved the robustness and credibility of the performance evaluation process.

VI. Feature Selection

Feature selection is a critical step in building effective machine learning models, particularly in domains such as bank loan risk assessment. Broadly, feature selection techniques can be categorized into three types: filter, wrapper, and embedded methods. Filter methods evaluate the statistical relevance of features independently of any learning algorithm. Examples include the correlation coefficient, which measures linear relationships between features and the target variable; the chi-square test, which assesses statistical independence in categorical data; and mutual information, which quantifies how much information a feature provides about the target. These methods are computationally efficient and help in quickly identifying potentially useful features [23, 24].

Wrapper methods, on the other hand, assess the performance of feature subsets by training models and selecting features based on predictive accuracy. Techniques like Recursive Feature Elimination (RFE) iteratively remove the least important features to find an optimal subset. Forward selection and backward elimination also follow similar principles but vary in their direction of feature inclusion or exclusion. Embedded methods integrate feature selection directly into model training. Notable examples include Lasso regression, which shrinks less relevant feature coefficients to zero through L1 regularization, and tree-based models like Decision Trees, Random Forests, and Gradient Boosting, which inherently perform feature selection by evaluating feature importance during the learning process [25]. These methods not only improve model performance but also enhance interpretability, a critical factor for institutions like banks.

When working with datasets that contain a mix of categorical and numerical features common in banking applications certain techniques stand out for their effectiveness. Both RFE and tree-based models are suitable, especially when categorical variables are appropriately encoded. However, tree-based models offer several advantages over RFE: they naturally handle mixed data types, capture non-linear relationships, are robust to multicollinearity, and provide computational efficiency [26]. Furthermore, tree-based models deliver built-in feature importance scores, simplifying the selection process, and advancements have been made to reduce biases in these scores. Their interpretability and versatility across classification and regression tasks make them particularly valuable in the context of loan risk modeling. Consequently, this research prioritizes tree-based feature selection methods to enhance both model performance and practical applicability. Accordingly, in this research, tree-based models were adopted.

In this section, we will highlight the most important features using tree-based methods: Decision Trees, Random Forests, and Gradient Boosting. These methods are applied to both dependent variables. The first one is to use classification in its original

format, and the other one is to categorize the dependent variable into two categories only. The first one is not default, and the second category is loan default. The following are the results of the three methods: Random Forest, decision tree, and gradient boosting.

A. Classification (Four Categories)

In this section, feature importance was evaluated using tree-based ensemble techniques—specifically, Decision Trees, Random Forest, and Gradient Boosting—applied to the original multi-class classification target variable. This classification represents the internal credit grading standards followed by the bank. The objective was to identify the most influential predictors affecting loan status as classified into four distinct risk categories as outlined in the preceding credit rating table.

1) Random forest feature importance. The Random Forest algorithm was employed to assess the input features based on their significance in contributing to classification accuracy. As shown in Table IX and depicted in Fig. 4, the findings reveal that loan balance (BAL) emerged as the most significant feature by a considerable margin, trailed by delinquency history (DELI) and amount due (DUE). In contrast, gender and job status showed little importance, indicating limited predictive capability in the realm of loan classification.

TABLE IX. BRIEF RESULT OF THE RANDOM FOREST FEATURE IMPORTANCE

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.009437	GENDER	0.002049
AGE_RANGE	0.015941	JOB	0.000363
YEAR	0.030530	BAL	0.672653
DUE	0.130977	DELI	0.138049

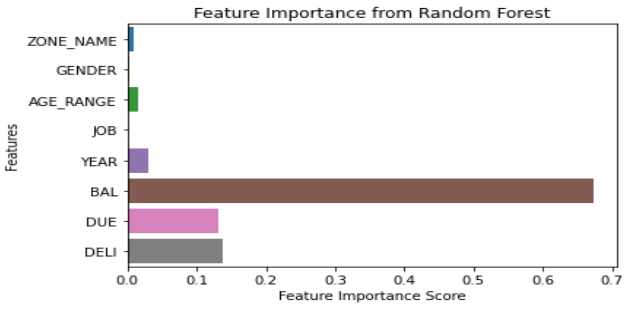

Fig. 4. Feature importance from Random Forest.

2) Decision tree feature importance. The Decision Tree was employed to compute feature importance scores. This method provides a straightforward interpretation by evaluating how each feature contributes to reducing impurity at decision nodes. As shown in Table X and visualized in Fig. 5, the most dominant variable was the loan balance (BAL), followed by delinquency history (DELI) and year of the record (YEAR).

3) Gradient boosting feature importance. From the following Table XI and Fig. 6, the least two important factors are gender and job.

TABLE X. BRIEF RESULT OF THE DECISION TREE FEATURE IMPORTANCE

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.012394	GENDER	0.006294
AGE_RANGE	0.028994	JOB	0.000518
YEAR	0.048181	BAL	0.644940
DUE	0.042377	DELI	0.216300

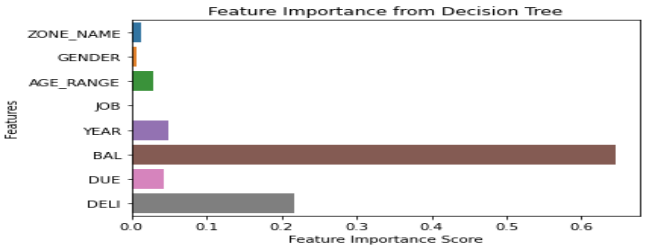

Fig. 5. Feature importance from Decision Tree.

TABLE XI. BRIEF RESULT OF THE GRADIENT BOOSTING FEATURE IMPORTANCE

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.017195	GENDER	0.000568
AGE_RANGE	0.003375	JOB	0.000873
YEAR	0.151289	BAL	0.074997
DUE	0.461394	DELI	0.290309

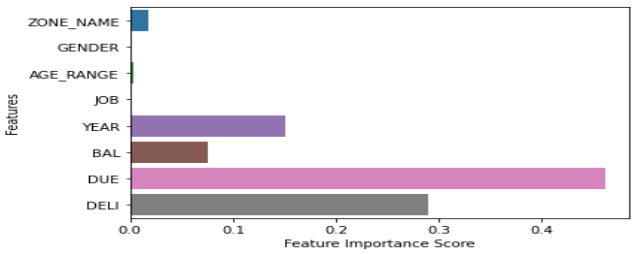

Fig. 6. Feature importance score from Gradient Boosting.

Three tree-based machine learning methods—Decision Tree, Random Forest, and Gradient Boosting (Table XI)—were utilized to evaluate the importance of various features in predicting loan defaults. The analysis revealed a consistent trend in the ranking of variables, although some differences in their significance were noted among the models. The most critical features identified across all models were Loan Balance (BAL), Amount Due (DUE), Delinquency (DELI), and Year (YEAR). Conversely, the least significant features in all three models were, distinctly, Gender (GENDER) Job status (JOB). These two variables consistently received the lowest importance ratings, reflecting their limited influence on default prediction. Consequently, gender and job status were removed from the final modeling stage to improve efficiency and decrease noise, without affecting the framework performance. Fig. 6 shows feature importance score from Gradient boosting.

B. Loan Default (Binary Classification)

During this stage, the target variable was transformed into a binary format (either defaulted or non-defaulted). The findings

reveal that most of the loan records within the sample are categorized as non-defaulted.

1) Random forest feature importance. As shown in the Table XII, the Random Forest model was used to identify the most influential features in predicting loan defaults. The analysis reveals that gender and job status had the lowest importance scores, suggesting they contributed minimally to the predictive performance of the model.

TABLE XII. BRIEF RESULT OF THE RANDOM FOREST FEATURE IMPORTANCE IN LOAN DEFAULT

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.006080	GENDER	0.001318
AGE_RANGE	0.011007	JOB	0.000294
YEAR	0.033104	BAL	0.655241
DUE	0.168674	DELI	0.124281

2) Decision tree feature importance. In the analysis of binary classification utilizing the Decision Tree algorithm, the model assessed the significance of each variable in forecasting loan default results. As shown in Table XIII, gender and employment status displayed the least importance scores compared to all other features.

TABLE XIII. BRIEF RESULT OF THE DECISION TREE FEATURE IMPORTANCE IN LOAN DEFAULT

Variable name	Importance	Variable name	Importance
ZONE_NAME	0.011449	GENDER	0.004770
AGE_RANGE	0.025480	JOB	0.000544
YEAR	0.048742	BAL	0.621828
DUE	0.266804	DELI	0.020383

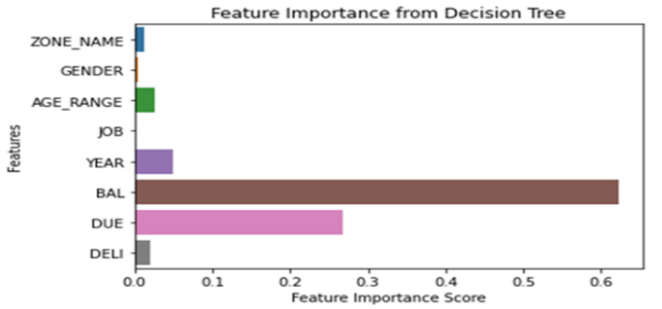

Fig. 7. Feature importance score from Decision Tree (loan default results).

In the analysis of binary classification utilizing the Decision Tree algorithm, the model assessed the significance of each variable in forecasting loan default results. As shown in Table XIV, gender and employment status displayed the least importance scores compared to all other features (see Fig. 7).

3) Gradient boosting feature importance. In the context of the binary classification task, the Gradient Boosting algorithm was utilized to assess the significance of each input feature in predicting loan default. The findings are presented in Table XIV.

It reveals that gender and job status possess the least importance scores, indicating a limited effect on the results of the model (Fig. 8).

TABLE XIV. BRIEF RESULT OF THE GRADIENT BOOSTING FEATURE IMPORTANCE IN LOAN DEFAULT

Variable Name	Importance	Variable Name	Importance
ZONE_NAME	0.004796	GENDER	0.000187
AGE_RANGE	0.001554	JOB	0.000184
YEAR	0.133705	BAL	0.042997
DUE	0.680054	DELI	0.136524

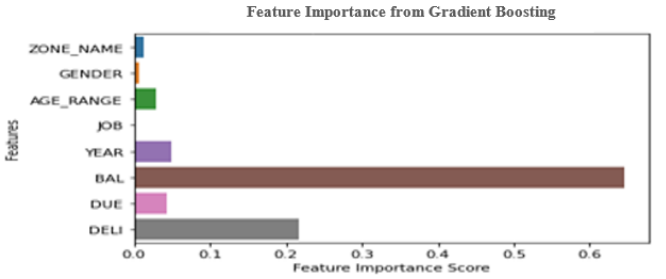

Fig. 8. Feature importance score from Gradient Boosting (loan default results).

VII. MODEL BUILDING

Following the selection of relevant features and the identification of key relationships among variables, suitable Machine learning algorithms are applied to develop predictive models. Techniques such as Random Forest, Decision Trees, and Gradient Boosting Machines are employed due to their proven effectiveness in handling complex, high-dimensional financial data. These algorithms are chosen for their ability to capture non-linear patterns and interactions within the dataset, thereby enhancing the accuracy and reliability of loan default predictions.

Before deciding, we have run more than one model: KNN, DT, RF, and Gradient Boost, and the following table shows that the best one is decision tree (Table XV).

The performance of the developed predictive model is assessed using a set of standard evaluation Metrics:

1) Accuracy. This metric measures the overall correctness of the model by calculating the proportion of correctly predicted instances out of the total number of cases.

2) Precision and recall. Precision reflects the proportion of correctly identified positive cases among all cases predicted as positive, while recall (also known as sensitivity) measures the model's ability to identify actual positive cases. Together, these metrics help evaluate the trade-off between false positives and false negatives [27].

3) Specificity and sensitivity. Sensitivity indicates the model's effectiveness in detecting loan defaults (positive cases), whereas specificity measures its ability to correctly identify non-defaults (negative cases). These metrics provide a comprehensive understanding of the model's capability to distinguish between default and non-default scenarios [27].

So, we are going to run a decision tree model for the second target variable that is categorized into two categories. A decision tree is a popular machine learning algorithm used for classification and regression tasks. It is a flowchart-like structure where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents a class label or a continuous value.

TABLE XV. THE EVALUATION METRICS OF VARIOUS ML MODELS

Model	F1-score	Accuracy	Recall	Precision	ROC curve
DT	66%	88%	53%	16%	80%
KNN	61%	82%	49%	11%	68%
RF	63%	84%	51%	12%	74%
Gradient boost	53%	70%	42%	10%	72%

A decision tree is a structured model used for both classification and regression tasks, consisting of several key components. The root node sits at the top and represents the entire dataset, initiating the first decision split. Internal nodes follow, each representing a decision based on a particular feature. Branches connect the nodes and illustrate the outcomes of those decisions, guiding the data down various paths. Finally, leaf nodes represent the terminal points, where a specific outcome or class label is assigned. This hierarchical structure enables the decision tree to map inputs to outputs through a series of straightforward rules.

One of the main advantages of decision trees is their interpretability. They are easy to understand and visualize, making them accessible to non-technical stakeholders. Additionally, they work well with both numerical and categorical data and are capable of modeling non-linear relationships. Unlike many other algorithms, decision trees do not require data normalization or scaling. They also inherently provide feature importance scores, offering insights into which variables most significantly impact predictions. However, a potential drawback is their tendency to overfit the training data, especially in the presence of noise, which can reduce their generalization performance. Despite this, their simplicity and explanatory power make decision trees a valuable tool in many machine learning applications.

The selected algorithms are implemented in Python. The model is trained on the cleaned dataset, with optimization techniques such as cross-validation applied to improve accuracy and prevent overfitting.

VIII. MODEL EVALUATION

The performance of the developed predictive model is assessed using a set of standard evaluation metrics:

Accuracy: This metric measures the overall correctness of the model by calculating the proportion of correctly predicted instances out of the total number of cases [28] [29].

Confusion Matrix: This provides a detailed view of the model's classification performance by showing the counts of true positives, true negatives, false positives, and false negatives [30]. Fig. 9 shows the confusion matrix of the model.

The overall accuracy is 88%, which means that the model could predict 88% of the non-fault and Fault loans correctly predicted, and this is a very high percentage. The sensitivity of the model 53%, which means that 53% of the Fault loans correctly specified. While specificity is 89%, which means that 89% of the non-Fault loans correctly specified.

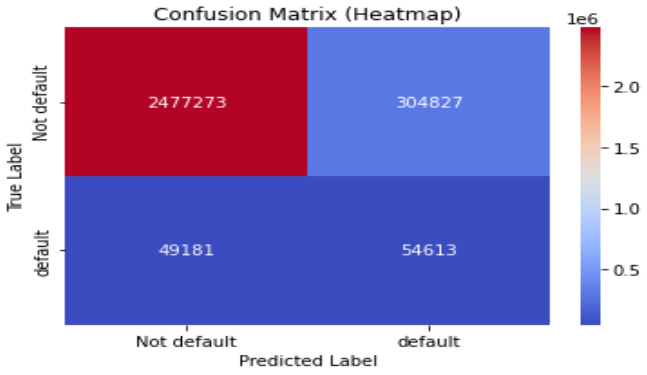

Fig. 9. Confusion matrix for model.

Another measure for the Goodness of fit is the ROC curve. The ROC Curve (Receiver Operating Characteristic Curve) is a graphical representation used to evaluate the performance of a binary classification model. It shows the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate) [29] at different classification thresholds. From Table XV, it is clear that the area under the curve is 80%, which indicates that the model is a good fit [30]. Fig. 10 shows the ROC curve.

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$FPR = \frac{False\ Positives}{False\ Positives + TrueNegatives}$$

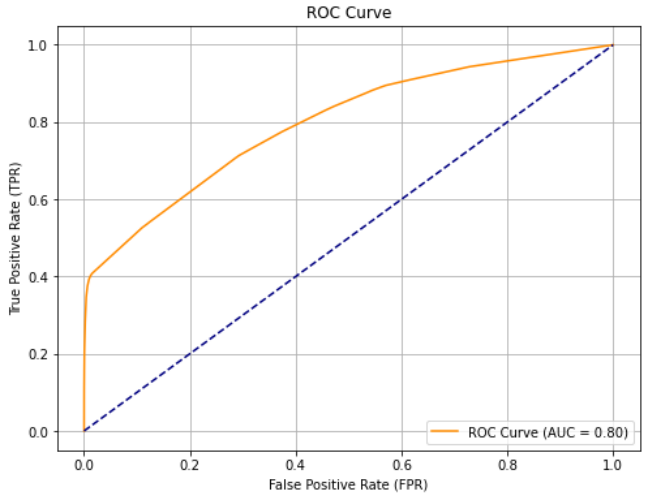

Fig. 10. ROC Curve.

A. Visualization and Analysis

In the final stage, model results are visualized to enhance interpretability and support effective communication with stakeholders. These visual representations enhance transparency, aid in validating the model's behavior, and ensure

that findings can be clearly communicated to both technical and non-technical audiences. Fig. 11 illustrates the structure of a decision tree.

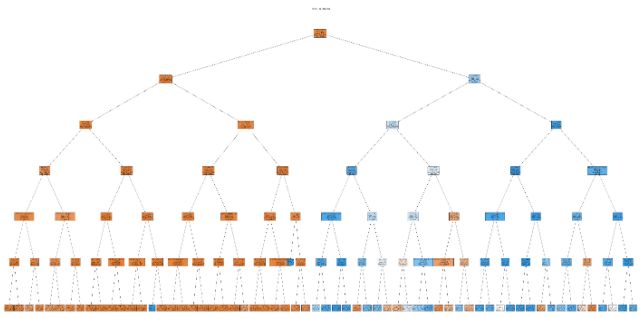

Fig. 11. Illustrates the structure of a decision tree.

The root node analysis of the decision tree highlights how the feature BAL (balance) serves as the most critical determinant in the classification of loan default risk. The initial split occurs at BAL ≤ 2140.0, making this threshold the most influential decision point. Lower values typically indicate a "Not Fault" classification, while higher balances often correspond with an increased likelihood of "Fault." This foundational split sets the tone for the downstream branches, guiding how further conditions refine the classification outcome.

- Path 1 reveals a scenario of high confidence for "Not Fault" classification. When BAL is further reduced to ≤ 2097.6 and the borrower's zone is not "East Delta," the Gini index is extremely low, signaling strong consensus in the data. This indicates that such a combination of low BAL and non-"East Delta" region consistently leads to non-default outcomes. The model is highly reliable in this path, showing minimal misclassification and suggesting that borrowers in this group present low risk.
- Path 2 follows a similar route but introduces moderate uncertainty. Here, while BAL remains ≤ 2097.6, the zone is identified as "East Delta." This regional factor slightly increases the Gini index, indicating some ambiguity. Although the majority of predictions still point to "Not Fault," the presence of the "East Delta" zone introduces variability, suggesting that geographical factors may contribute to nuanced risk assessments, especially in marginal financial conditions.
- In Path 3, the model addresses a more complex situation leading to a "Fault" classification. This occurs when BAL exceeds 2140.0, DUE is also high (above 0.68), but DELI (delinquency) remains low (≤ 0.15). Despite the low DELI, the combination of high BAL and DUE tips the classification toward "Fault." The Gini value here is moderate, suggesting a less decisive split. It reflects a group where most instances are defaults, but a few are not, highlighting the complexity of high-risk profiles, where even a favorable feature like low DELI cannot fully offset the effect of high balances and dues.
- Path 4 introduces a highly uncertain classification scenario. When BAL is within a borderline range (≤ 2140.0 but > 2097.6), DELI is above 0.05, and DUE varies, the model shows significant fluctuation in its output. With moderate DUE (≤ 11364.07), the classification remains "Not Fault," but once DUE exceeds this threshold, the likelihood of a "Fault" classification rises and the Gini index drops, indicating greater certainty. This path illustrates the dominance of the DUE variable in edge cases and how it can override other indicators in determining default risk.
- Finally, Path 5 emphasizes the strong influence of DELI in reducing risk. When BAL is low and the borrower is not in the "East Delta" zone, a DELI ≤ 0.05 leads to a highly confident "Not Fault" prediction, despite other variables suggesting moderate risk. The Gini index is low in this scenario, showing that low delinquency plays a powerful protective role. This finding underscores the importance of monitoring delinquency closely, as it can significantly alter the risk profile even when other conditions are less favorable.

IX. Results

The predictive model showcases a thoughtful and structured decision-making process shaped by the relative importance of selected customer attributes. Rather than relying on a single indicator, the model navigates through combinations of factors to reach decisions that are both data-driven and sensitive to nuanced behavioral patterns.

A. Balance (BAL): Entry Point for Risk Assessment

At the initial stage, the customer's balance acts as a key determinant: When the balance is less than or equal to 2140.0, the likelihood of loan default is low, and the model typically moves toward a "no fault" decision.

Balances exceeding this threshold trigger additional evaluation steps, as they may signal heightened financial pressure or exposure, leading to a higher probability of default.

This distinction reflects the model's cautious approach: it does not generalize but instead digs deeper when early indicators suggest potential risk.

B. Geographic Factor: ZONE_NAME

Regional context also plays an essential role: customers outside the "East Delta" region generally present lower risk, and the model tends to assign them to the "non-fault" category.

However, being from "East Delta" introduces a degree of uncertainty. This does not imply automatic risk but encourages the model to be more attentive, possibly reflecting local economic or demographic conditions.

Such consideration reflects how risk may be influenced, not just by individuals but also by broader regional dynamics.

C. Delinquency Score (DELI): Strong Behavioral Signal

This variable carries significant weight in the model's decisions: scores at or below 0.05 offer strong reassurance, often guiding the model to a "not fault" outcome, even in the presence of moderate financial concerns.

In contrast, scores above 0.15 often signal concern, sometimes shifting the model's outcome toward a "fault" decision—even if other variables seem relatively stable.

This behavior highlights the model's ability to recognize patterns in payment discipline as a reliable reflection of future risk.

D. DUE (Outstanding Amount): Key in Borderline Decisions

In more complex or uncertain cases, the outstanding amount becomes a tipping factor: values below 11,364.07 often help stabilize decisions in favor of a non-default classification.

Higher amounts, however, tend to increase the model's confidence that default risk may be present.

Here, the model demonstrates sensitivity to financial load, adjusting its judgment when individuals may be reaching unsustainable repayment levels.

E. Gini Index: Evaluating Predictive Fairness

The model's ability to separate risky from safe cases is validated through the Gini coefficient, a well-regarded metric in credit scoring. A higher Gini score indicates that the model can reliably distinguish between borrowers who are likely to default and those who are not.

This ensures that the decisions made are not only technically sound but also aligned with ethical considerations of fairness and accountability.

X. Discussion

This study offers a unique and novel contribution by concentrating on a very specific and little-studied domain: livestock-based loans within the Agricultural Bank of Egypt. This is in contrast to previous studies that have extensively examined the use of machine learning algorithms to predict loan approval and default risks, with a primary focus on general loan products or credit scoring mechanisms in various banking sectors. Just as Puli (2024), Saini (2023), which employed algorithms like Random Forest, SVM, and Neural Networks on public or generalized datasets, the present research leverages actual, real-world customer data from one of Egypt's largest rural-focused banking institutions, comprising 1,190 branches nationwide. The dataset reflects true client behaviors in repaying livestock-oriented loans—an area that remains largely unexplored in prior academic research.

XI. Conclusion

Machine learning has become an increasingly valuable tool in assessing loan risk within the banking sector. By leveraging these techniques, it is possible to identify patterns associated with loan default and to determine the most relevant predictive features. The present study emphasizes the role of appropriate feature selection in improving both model accuracy and interpretability. Approaches such as filter methods, wrapper strategies, and embedded techniques were considered to identify key variables.

Among the models implemented, algorithms such as Random Forest, Decision Tree, and Gradient Boosting demonstrated favorable performance, with Random Forest achieving an accuracy of approximately 88%. These models also offer a reasonable balance between predictive power and interpretability, making them suitable for practical use.

The feature importance analysis revealed that financial indicators had the greatest influence on prediction outcomes. Notably:

Balance (BAL) showed a strong association with default risk.

Do Amounts (DUE) were identified as critical in flagging potential non-repayment.

Delinquency Scores (DELI) provide insight into historical repayment behavior.

On the other hand, demographic attributes, such as gender and occupation, were found to have limited impact on model performance, suggesting that behavioral and financial factors may play a more central role in loan risk evaluation.

In conclusion, machine learning, when applied with thoughtful feature selection, can contribute meaningfully to improving credit risk assessment. While no single model can guarantee perfect accuracy, decision trees and ensemble methods have shown promise in balancing accuracy, transparency, and real-world applicability.

XII. Future Work

While the current study demonstrates the potential of traditional machine learning models such as Decision Tree and Random Forest in predicting loan defaults, future research can benefit from exploring more advanced algorithms and broader datasets. Incorporating modern ensemble and deep learning techniques, such as Extreme Gradient Boosting (XGBoost) Artificial Neural Networks (ANNs), may offer improved predictive accuracy, especially in handling complex, nonlinear relationships within financial data.

In addition, expanding the dataset to include behavioral patterns, real-time transactional data, and external credit factors may also lead to deeper insights and more generalized models.

Acknowledgment

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the General Research Project under grant number GRP/6/46.

References

[1] S. Mestiri and S. Mestiri, "Credit scoring using machine learning and deep Learning-Based models," Data Science in Finance and Economics 2024 2:236, vol. 4, no. 2, pp. 236–248, 2024, doi: 10.3934/DSFE.2024009.

[2] M. Ranjan, K. Barot, V. Khairnar, V. Rawal, A. Pimpalgaonkar, S. Saxena and A. Sattar, "Python: Empowering Data Science Applications and Research," Journal of Operating Systems Development & Trends, 2023, Volume 10, Issue 1, 2023,pp27-33

[3] J. Quan and X. Sun, "Credit risk assessment using the factorization machine model with feature interactions," Humanities and Social Sciences Communications 2024 11:1, vol. 11, no. 1, pp. 1–10, Feb. 2024, doi: 10.1057/s41599-024-02700-7.

[4] N. Gulo, E. Nurninawati, R. A. R. S, and D. P. Kristiadi, "Decision Support System for Submitting Credit using Analytical Hierarchy Process (AHP) Method Based on Android on Save and Loan Cooperatives Cubg Pasar Kemis Tangerang," IJISTECH (International Journal of

Information System and Technology), vol. 6, no. 4, pp. 441–448, Dec. 2022, Accessed: Feb. 20, 2025. [Online]. Available: https://ijistech.org/ijistech/index.php/ijistech/article/view/259

[5] Tran, Phuong ,and , Nga Phan " Digital Transformation of the Banking Industry in Developing Countries " International Journal of Professional Business Review (2023), Vol. 8 , 5 pp. 8-0.

[6] Vahid sinap "AComparative Study of Loan Approval Predicayion Using Machine Learning Methods",Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Ankara, Türkiye,Apirl 2024,pp644-663

[7] P. S. Saini, A. Bhatnagar, and L. Rani, "Loan approval prediction using machine learning: A comparative analysis of classification algorithms," in 2023 3rd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE), May 2 023, pp. 1821-1826.

[8] Jovanne C. Alejandrino1 , Jovito Jr. P. Bolacoy1 , John Vianne B. Murcia, " Supervised and unsupervised data mining approaches in loan default prediction," International Journal of Electrical and Computer Engineering (IJECE) , April 2023, Vol. 13, pp. 1837-1847

[9] M. Alaradi and S. Hilal, "Tree-based methods for loan approval," in 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Oct. 2020, pp. 1-6.

[10] Güder, G., & Köse, U. (2024). Prediction of home loan approval with machine learning. *Advances in Artificial Intelligence Research (AAIR), 4*(2), pp.87–95.

[11] Sampurna , Sahil and Vidya " Predicting Credit Risk in European P2P Lending: A Case Study of "Bondora" Using Supervised Machine Learning Techniques" December 2023, 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2023, pp. 1-6.

[12] Tabassum, Namita, Vaishnavi and Prachi ,Aditya and" A Machine Learning Based Framework For Bankruptcy Prediction In Corporate Finances Using Explainable AI Techniques" Vol 49, No 15 (2025) ,pp 15–26.

[13] Hussain, M. Z., Ejaz, S., Batool, E., Hasan, M. Z., Mustafa, M., Khalid, A., Hussain, U., Khan, Z., Javaid, A., Ashraf, M. F., Awan, R., & Yaqub, M. A" Bank Loan Prediction System Using Machine Learning Models" 2024 IEEE 9th International Conference for Convergence in Technology No I2 ,pp. 1-7

[14] chen, Y. (2023). Financial Statement Fraud Detection based on Integrated Feature Selection and Imbalance Learning. Frontiers in Business, Economics and Management,VOL 8(3), pp46-48.

[15] Ali, A. A., Khedr, A. M., El-Bannany, M., & Kanakkayil, S. (2023). A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. Applied Sciences, VOL 13(4), 2272

[16] A. Gupta, V. Pant, S. Kumar, and P. K. Bansal, "Bank Loan Prediction System using Machine Learning," in 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), Dec. 2020, pp. 423-426

[17] .K. Vengatesan , A. Kumar and S. Sabni,. Credit Card Fraud Detection Using Data Analytic Techniques. Advances in Mathematics: Scientific Journal 9, (2020)Vol3, pp1185–1196.

[18] Vahid SİNAP."A Comparative Study of Loan Approval Prediction Using Machine Learning Methods" Gazi University Journal of Science, Part C: Design and Technology, (2024),VOL 12(2), P.P644– 663.

[19] Agata Kozinaa , Łukasz Kuźmińskia , Michał Nadolnya , Karolina Miałkowskaa , Piotr Tutaka , Jakub Janusa , Filip Płotnickia , Ewa Walaszczy (2023)"The default of leasing contracts prediction using machine learning", 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systemsp,p424–433

[20] Muhamad, Muhamad JUMA and Arif (2023) Improving Credit Risk Assessment through Deep Learning-based Consumer Loan Default Prediction Model, INTERNATIONAL JOURNAL OF FINANCE & BANKING STUDIES VOL 12,p.p85-92.

[21] Viswanatha, V. (2023)." Prediction of loan approval in banks using machine learning approach", International, ournal of Engineering and Management Research, Vol 13(4), P.P7–19

[22] Btoush, W., Alsharari, N. M., & Al-Dhamari, R. A. (2025). A hybrid deep learning and machine learning stacking model for credit card fraud detection. *Applied Sciences, 15*(6), 1–20. https://doi.org/10.3390/app15063320

[23] Hiba, A. (2024). A review of feature selection methods in big data. *International Journal of Intelligent Systems and Applications in Engineering*, VOL*12*(4), P.P4347–4366.

[24] Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data, 11*(44),P.P1-16.

[25] Pödör, Z., & Hekfusz, M. (2024). Comparing feature selection methods on metagenomic data using r andom forest classifier. *Transactions on Engineering and Computing Sciences, 12*(1), P.P175–187

[26] S. Rajora, D. Li, C. Jha, N. Bharill, O. Patel, S. Joshi, and M. Prasad(2018), "A comparative study of machine learning techniques for credit card fraud detection based on time variance," In 2018 IEEE symposium series on computational intelligence (SSCI), IEEE, pp. 1958-1963.

[27] Jasmina, Goran and Drago (2020)New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers, Advanced Engineering Informatics, Vol 45, p.p

[28] Silvia ,Federico and Emanuele Giovannini (2017),Solvency prediction for small and medium enterprises in banking, Decision Support Systems, Vo 102, P.P 91-97

[29] Cabot, J. H., & Ross, E. G. (2023). Evaluating Prediction Model Performance. *Surgery*, Elsevier VOL174,NO(3), P.P 723–726.

[30] Tariq Saeed (2020)The Application of Data Mining Techniques for Financial Risk Management: A classification framework, IJCSNS International Journal of Computer Science and Network Security, VOL.20 No.8,P.P84-93.

AUTHORS' PROFILE

Mona Sharafeldin received her master degree of computer science from the Sadat Academy for Management Sciences in 2014. She is currently PhD student in Business Information System, Helwan University.

Prof. Amira M. Idrees: I'm currently a professor in College of Business, King Khalid University. I have been the head of scientific departments and the vice dean of the community services and environmental development, Faculty of Computers and Information, Fayoum University. I have also been a professor in the Faculty of Computers and Information Technology in Future University, the head of IS department and the head of University Requirements Unit. My research interests include Knowledge Discovery, Text Mining, Opinion Mining, Cloud Computing, E-Learning, Software Engineering, Data Science, and Data warehousing.

Assoc.Prof/Shimaa Ouf: I am currently an Associate Professor at the Information Systems Department, Faculty of Commerce and Business Administration, Helwan University. I was born in Cairo, Egypt. She received an M.Sc. degree in Information Systems from the Faculty of Computers and Information, Helwan University, Egypt. Ph.D. degree in Information Systems from the Faculty of Computers and Information, Helwan University, Egypt. I have published many Scientific articles in International journals and conferences in the area of E-learning Ecosystems, Web 2.0 Technologies, Cloud Computing, Intelligent learning environments, Personalized E-learning Ecosystem using Ontology and Semantic Web Rule Language, Business Intelligence, Big Data, Artificial Intelligence, and Blockchain. I have deep experience in creating intelligent systems using the semantic web and Artificial Intelligence technologies. I am an active reviewer for numerous international journals.