Comparative Predictive Performance of Logistic Regression, Naive Bayes, and Support Vector Machine Models in Loan Default Classification among Microfinance Institution Clients in Makueni County, Kenya

Jackson Musau[1*], Dr. Martin Kasina[2] & Dr. Ayubu Anapapa[3]

[1]Department of Mathematics and Statistics, Machakos University, Machakos, Kenya

[2]Department of Mathematics and Statistics, Machakos University, Machakos, Kenya

[3]Department of Mathematics and Actuarial Science, Murang'a University of Technology, Murang'a, Kenya

***Correspondence Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100600627

Received: 07 June 2026; Accepted: 12 June 2026; Published: 30 June 2026

ABSTRACT

Microfinance institutions (MFIs) play a critical role in improving financial inclusion in Kenya; however, high loan default rates continue to threaten their financial sustainability. This study aimed to develop and compare the performance of Logistic Regression, Naïve Bayes and Support Vector Machine (SVM) models in predicting loan default among clients of MFIs in Makueni County, Kenya. The study adopted a quantitative research design and used secondary data comprising 4,592 borrower records obtained from selected MFIs operating within the county. Data analysis was done using python programming. Borrower socio-economic and financial attributes were extracted from loan records and preprocessed through data cleaning, normalization an. encoding procedures. To ensure robust and unbiased evaluation, Stratified 5-fold cross-validation combined with Grid Search CV hyperparameter tuning was applied across all models. Model performance was assessed using accuracy, precision, recall, F1-score, specificity, and Area Under the Curve (AUC). The results showed that the SVM model achieved the highest predictive performance (accuracy = 0.857, AUC = 0.914), followed by Logistic Regression (accuracy = 0.830, AUC = 0.904), while Naïve Bayes performed least effectively (accuracy = 0.740, AUC = 0.769). The findings demonstrate that SVM provides superior classification ability in capturing complex borrower patterns, while Logistic Regression remains a strong and interpretable baseline model. The study concludes that machine-learning models, particularly SVM, significantly improve credit risk prediction in microfinance environments and can support data-driven lending decisions in rural financial institutions.

Key Word: Loan default, Microfinance institutions, Logistic regression, Naïve Bayes, Support Vector Machine.

INTRODUCTION

Microfinance institutions (MFIs) are the key instruments for increasing financial inclusion in developing economies. This is because they provide seamless access to credit services to the poor households and micro-enterprises. [10] argues that entrepreneurship and equitable financial growth become possible when there is access to credit facilities. However, sustainability of MFIs is based on their capacity to manage credit risk, which is a quantitative issue of estimating the likelihood of a loan defaulter, a problem that lies at the core of applied statistical modelling.

In Kenya, microfinance industry has increased financial accessibility but there is growing pressure on credit-risk. In 2024, the Central Bank of Kenya [3] reported a non-performing loan (NPL) ratio of 16.4%. In 2025, the State of the Banking Industry Report [4] noted a slow credit growth and poor transmission of the monetary

policy. According to an article by [6], the overall MFI assets decreased by almost ten percent in 2024 with the gross NPLs increasing 5 times between 2014 and 2024 taking the ratio to over 35 percent. These statistics measure systemic vulnerability requiring statistical research. This quantitative view is supported by empirical research by scholars in Kenya. Through multiple regression, [7] demonstrated that the capital adequacy is significantly inversely related to Non-Performing Loans (NPLs) among microfinance banks. Using logistic regression to analyze digital-loan data of commercial banks, [8] verified the model strength in estimating probabilities of default, the income-to-loan ratio and credit score were found to be determinants of loan default. [12] evaluated the performance of logistic regression, XG Boost and random-forest algorithms to predict credit risk using 6,771 loan accounts of Machakos County MFIs. The study found out that XG Boosting was the best performer with 83.3% accuracy. Findings from these studies indicate the necessity of strict comparative analysis of statistical models of classification in Kenyan credit-risk environments such as in MFIs.

In Makueni County, the main credit providers among rural households and smallholders are MFIs, but repayment problems are very acute. [11] found that in Nzaui/Kilili/Kalamba Ward the default rate was 61 percent attributed to the literacy of the borrowers, off-farm earnings, frequency and level of supervision by the loan officers. Such socio-economic variables are quantifiable and easy to model statistically and are often seldom included in forecasting models. Reliance on rain-based agriculture and a lack of access to banking services increases the exposure to repayment risk thereby undermining objectivity through subjective lending practices.

Although the study by [11] offers valuable descriptive information on the loan repayment issues in Nzaui/Kilili/Kalamba Ward, the study methodology and the level of analysis of the current study are of a different nature. The main methods applied in the work of Mwaka were the traditional methods of inferential identification of socio-economic factors that affect the repayment behavior with the focus on the borrower literacy, off-farm earnings and loan supervision. The analysis was not aimed at creating predictive models and its results were much context-based and descriptive.

Conversely, the proposed research used a predictive and comparative modelling method whereby the Logistic Regression, Naive Bayes and Support Vector Machines (SVM) are used to determine the likelihood of loan defaults among clients of MFIs in Makueni County quantitatively. In contrast to [11], this study merges socio-economic variables into formal classification algorithms in a systematic manner and allows objective and data-driven decision-making. Also, the model performance was compared and assessed with the commonly used measures of accuracy, precision, recall, F1-score and ROC-AUC, thereby filling methodological and operational gaps of previous studies.

The fact that the loan default rates are still on the increase in Makueni County is an indication that there may be some structural and methodological flaws in the credit risk assessment practice. This has been mostly due to the over dependence on the use of single models, mostly logistic regression, without a systematic comparison with other classification algorithms that may be able to capture nonlinear relationship and complex behavior of borrowers. The lack of localized and empirically validated comparative modeling frameworks has thus limited the timeliness, precision and objectivity in credit decisions made by MFIs leading to ongoing increases in non-performing loans and endangering the sustainability of the institutions.

This study therefore mainly sought to address the deficit of comparative statistical analysis of the major classification models i.e Logistic Regression, Naive Bayes and Support Vector Machines, in predicting loan default among clients of microfinance institutions in Makueni County. It aimed at closing the methodological and contextual gaps by using applied statistical models with a focus on the accuracy of the model, its interpretability and its practical value in making decisions.

The specific aim of this study was to compare the predictive performance of logistic regression, Naive Bayes and SVM models in classifying loan default among microfinance institution clients in Makueni County, Kenya. This study may improve credit risk assessment, reduce non-performing losses and become more sustainable by making the credit decisions based on the objective and data-driven lending decisions.

MATERIAL AND METHODS

This study adopted a quantitative research design with a comparative approach to evaluate and compare the predictive accuracy and effectiveness of the selected models. The design was appropriate because the study aimed at developing and comparing statistical classification models in predicting the probability of loan default. There were 9 MFIs with 22 branches operating within Makueni County [2]. This study adopted a purposive sampling technique to select nine microfinance institutions (MFIs) for inclusion in the analysis. Each MFI was represented by one branch to form a sample of nine institutions. The institutions were chosen based on the availability of complete and relevant loan-level records necessary for developing and evaluating predictive models for loan default. However, the use of purposive sampling was appreciated as a limitation for this study since it might have introduced selection bias, as only MFIs with complete datasets were included. Data analysis was done using python programming.

Preprocessing

Before data was subjected to analysis, it was first prepared to accommodate the predictive algorithms. Data was cleaned to address any missing entries by imputation. Before fitting the models, the dataset was prepared by handling outliers through capping, eliminating duplicates and scaling continuous variables to a 0 to 1 range through normalization. Categorical variables were encoded as shown in Table 1.

Table 1: Coding of Categorical Variables

Variable Name	Description	Coding
Gender	Sex of borrower	1 = Male, 2 = Female
Marital Status	Marital status	1 = Single, 2 = Married, 3 = Others
Education Level	Highest education attained	1 = Primary, 2 = Secondary, 3 = Tertiary, 4 = University
Loan Purpose	Purpose of loan	1 = Business, 2 = Education, 3 = Personal, 4 = Other
Collateral	Loan security status	1 = Yes, 0 = No
Employment Status	Employment status	1 = Employed, 2 = Self-employed, 3 = Unemployed
New Repeat	Loan history status	1 = New, 2 = Repeat
Loan Status (Target)	Indicates loan default status	1 = Defaulted, 0 = Not Defaulted

To address class imbalance, class weighting was applied using the class weight=’balanced’ parameter in Python’s scikit-learn library. This automatically adjusted penalties such that misclassification of the minority class carries higher weight, effectively balancing model sensitivity without altering the original dataset.

Model Validation

To ensure robust and unbiased evaluation, the study employed Stratified 5-Fold Cross-Validation, which partitions the dataset into five subsets while preserving class distribution in each fold. Each model was trained on four folds and tested on the remaining fold, and this process was repeated five times. Final performance metrics were obtained by averaging results across all folds.

Hyperparameter Optimization

Model performance was improved using Grid Search CV-based hyperparameter tuning. For Logistic Regression, the regularization parameter (C) and solver type were optimized. For SVM, the kernel type (linear, RBF), penalty

parameter (C), and gamma values were tuned. Naïve Bayes was used without hyperparameter tuning due to its non-parametric nature.

Models

Logistic Regression Model

Logistic Regression is a classification model that is used for classification tasks. In this study, binary logistic regression was used to predict the probability of a client defaulting on a loan, which is a binary outcome, i.e., default (1) or no default (0). The choice of binary logistic regression was motivated by several considerations specific to this study. Binary logistic regression directly models the probability of a binary outcome (default vs. non-default) using the sigmoid function which produces default likelihoods. As opposed to advanced machine learning algorithms, logistic regression provides transparent coefficient estimates and odds ratios, which are easily interpretable, enabling loan officers in Makueni County MFIs to understand and explain which borrower characteristics drive default risk. In addition, logistic regression handles class imbalance through class weighting, which was applied in this study to ensure the model does not bias predictions toward the majority (non-default) class. These characteristics establish logistic regression as both an appropriate baseline model and a practically useful tool for MFI credit risk assessment in Makueni County.

Naïve Bayes Model

Naive Bayes is a classification model based on Bayes theorem that predicts the category of a data point using probability. It assumes that all features are independent of each other hence the name Naive. In our study, we used this model to predict the category of a loan client, to determine whether the client is a defaulter or a non-defaulter. The inclusion of Naïve Bayes in this study was based on its simplicity and strong performance in classification problems, particularly in situations involving high-dimensional data and relatively small datasets. Despite the independence (Naive) assumption being a simplification of real-world conditions, the model has demonstrated competitive performance in credit risk prediction studies, making it a relevant benchmark model for comparison with Logistic Regression and Support Vector Machine (SVM) in this study.

Support Vector Machine Model

Support Vector Machine (SVM) is a model that can be used for classification and regression tasks. In this study, we used SVM to classify loan clients by predicting whether a loan client defaults or not. The selection of SVM in this study was based on its ability to handle both linear and non-linear relationships, making it suitable for complex credit risk patterns in loan default prediction. It is also effective in high-dimensional datasets and demonstrates strong classification performance even when class boundaries are not clearly separable. This makes SVM a suitable comparative model alongside Logistic Regression and Naïve Bayes in this study.

Consider a dataset made of loan defaulters and non-defaulters. We seek to define the best decision boundary, a plane that separates the two classes of loan clients. This boundary is called the hyperplane and the closest data points to the hyperplane on both sides of the line are the support vectors. The distance between the hyperplane and the support vectors is called the margin. This is illustrated in Figure 1 below:

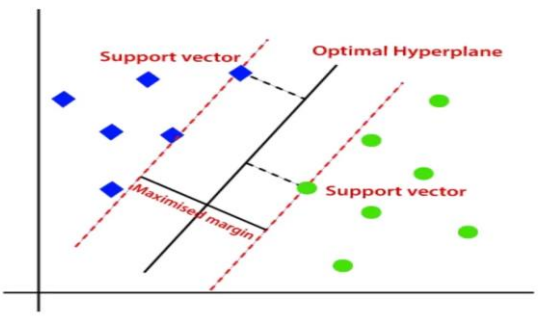

Figure 1. Support Vector Machine Diagram

THEORETICAL RESULTS

Logistic Regression

Logistic regression model uses sigmoid function to transform linear combination of input features into a probability. Suppose we take the input borrower's features, such as income, age, employment status and other relevant predictor variables to be represented by a vector $X = x_1, x_2, \ldots, x_n$ and the dependent variable Y with binary values 0 and 1 such that:

$$y = \begin{cases} 0 & \text{if No Default} \\ 1 & \text{if Default} \end{cases} \tag{1}$$

We first compute a linear combination z of input features such that:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

where β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients corresponding to the predictor variables $x_1, x_2, \ldots, x_n$. We now take the log odds of a defaulter such that:

$$\log\left(\frac{p}{1-p}\right) = z \tag{3}$$

where p is the likelihood of a loan applicant defaulting. The value z is a continuous value from a linear model. To transform z into a probability function, we use the sigmoid function:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4}$$

where x is the input value to the sigmoid function such that;

$$\lim_{x\to\infty} \sigma(x) = 1, \qquad \lim_{x\to-\infty} \sigma(x) = 0. \tag{5}$$

resulting in the characteristic S-shaped curve shown in Figure 2.

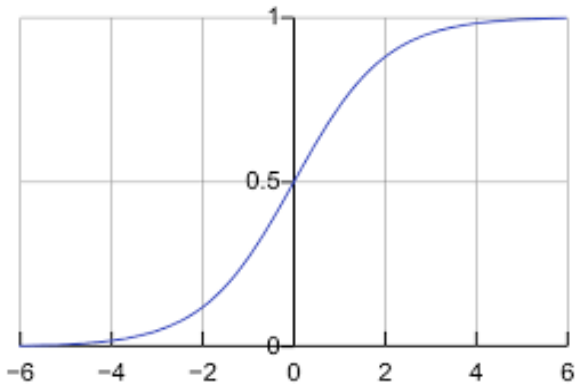

Figure 2. Sigmoid function curve.

In our case, x is the value of z. The sigmoid function then becomes as follows:

$$\sigma(z) = p(y = 1|X) = \frac{1}{1+e^{-z}} \tag{6}$$

Taking the value of z from equation (2), we obtain a probability function:

$$p(y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}} \tag{7}$$

where $p(y = 1|X)$ represents the probability that a client defaults. The output is a probability, hence lies between 0 and 1. To classify a new loan client we compute the probability of default $\hat{p} = P(y = 1 \mid X)$ and classify the client such that:

$$\hat{y} = \begin{cases} 1, & \text{if } \hat{p} \geq 0.5 \\ 0, & \text{if } \hat{p} < 0.5 \end{cases} \tag{8}$$

Naïve Bayes Model

Consider our response variable Y defined as:

$$Y = \begin{cases} 1, & \text{if the borrower defaults} \\ 0, & \text{if the borrower does not default} \end{cases} \tag{9}$$

and let $X = x_1, x_2, \dots, x_n$ be a vector of the borrower's features, such as income, age, employment status and other relevant predictor variables. To estimate the probability of default:

$$P(Y = 1 \mid x_1, x_2, \dots, x_n) \tag{10}$$

We apply Bayes' theorem:

$$P(Y \mid X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \tag{11}$$

where:

$P(Y \mid X)$: Posterior probability — the probability that a loan client defaults ($Y = 1$) or does not default ($Y = 0$) given their features X.

$P(X \mid Y)$: Likelihood — the probability of observing the loan client's features X given their default status Y.

$P(Y)$: Prior probability — the overall probability of loan default or non-default in the population of loan borrowers.

$P(X)$: Marginal likelihood — the probability of observing these features X across all loan borrowers.

Taking the naive assumption that all features are independent given the class, that is, for a borrower with features $X = (x_1, x_2, \dots, x_n)$, we obtain:

$$P(x_1, x_2, \dots, x_n \mid Y) = \prod_{i=1}^{n} P\,(x_i \mid Y) \tag{12}$$

We then substitute equation (12) to Bayes' theorem to get:

$$P(Y \mid x_1, x_2, \dots, x_n) = \frac{P(Y) \cdot \prod_{i=1}^{n} P(x_i|Y)}{P(x_1)P(x_2)\dots P(x_n)} \tag{13}$$

The denominator $P(x_1)P(x_2) \dots P(x_n)$ does not depend on Y hence is a constant for a given input x_i. We can then simplify the expression to get:

$$P(Y \mid x_1, x_2, \dots, x_n) \propto P(Y) \cdot \prod_{i=1}^{n} P\,(x_i \mid Y) \tag{14}$$

To classify a new loan client, we compute the posterior probability for each class Y ,i.e. $P(Y \mid x_1, x_2, \dots, x_n)$ and then use $argmax$ to choose the class with the highest probability such that:

$$\hat{Y} = \underset{Y}{\text{argmax}}\ P(Y) \cdot \prod_{i=1}^{n} P\,(x_i \mid Y) \tag{15}$$

This expression defines the best Naive Bayes classifier that we can use to classify a new loan client.

Support Vector Machine Model

We consider a feature vector $X = x_1, x_2, \ldots, x_n$ where $x_1, x_2, \ldots, x_n$ include borrower features like age, income, loan amount, employment status and other appropriate features. Our response variable $y_i \in \{+1, -1\}$ represents:

$$y_i = \begin{cases} +1, & \text{if the borrower defaults} \\ -1, & \text{if the borrower does not default} \end{cases} \tag{16}$$

A hyperplane that separates borrowers who default from those who do not is defined as:

$$w \cdot x + b = 0 \tag{17}$$

where:

w is the weight vector, the coefficients for the borrower's features.

x is a borrower's feature vector.

b is the bias or intercept, shifting the hyperplane away from the origin without changing its orientation.

A new loan applicant $\boldsymbol{x}_{\text{new}}$ is classified based on which side of the hyperplane they fall given that:

$$\hat{y} = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x}_{\text{new}} + b) \tag{18}$$

such that

$$\hat{y} = \begin{cases} +1, & \text{if } w \cdot x + b \geq 0 \\ -1, & \text{if } w \cdot x + b < 0 \end{cases} \tag{19}$$

If $\boldsymbol{w} \cdot \boldsymbol{x}_{\text{new}} + b > 0$, we predict class $+1$ while if $\boldsymbol{w} \cdot \boldsymbol{x}_{\text{new}} + b < 0$, we predict class -1. To obtain the optimal classifier, we seek to maximize the margin, i.e., the distance between the support vectors and the decision boundary such that:

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = +1 \quad \text{and} \quad \boldsymbol{w} \cdot \boldsymbol{x} + b = -1 \tag{20}$$

The margin M between these planes is given by:

$$M = \frac{2}{\|\boldsymbol{w}\|} \tag{21}$$

where $\|\boldsymbol{w}\|$ is the magnitude of the weight vector $\boldsymbol{w}$. In practice, to maximize the margin M, we minimize $\| \boldsymbol{w} \|$. For computation convenience, we use $\frac{1}{2}\|\boldsymbol{w}\|^2$ which is equivalent to$\|\boldsymbol{w}\|$. For perfectly separable borrowers' data:

$$\text{yi}(\text{w} \cdot \text{xi} + \text{b}) \geq 1, \forall \text{i} \tag{22}$$

We seek to optimize the classifier by finding the $\boldsymbol{w}$ and b that minimize $\frac{1}{2}\|\boldsymbol{w}\|^2$ such that every data point is correctly classified:

$$\min_{\text{w,b}} \frac{1}{2}|\text{w}|^2 \text{ Subject to } \text{y}_\text{i}(\text{w} \cdot \text{x}_\text{i} + \text{b}) \geq 1, \forall \text{i} \tag{23}$$

In practice however, borrowers' data may not be perfectly separable in real life. We hence introduce a slack variable $\xi_i \geq 0$ that allows some misclassifications such that:

$$\text{yi}(\text{w} \cdot \text{xi} + \text{b}) \geq 1 - \xi_i, \ \xi_i \geq 0, \forall \text{i} \tag{24}$$

If $\xi_i = 0$, the point is correctly classified outside the margin. If $0 < \xi_i \leq 1$, the point is inside the margin but still on the correct side. If $\xi_i > 1$, the point is misclassified. To balance the effect of the slack variable, we plug in a penalty factor$C$. The optimized classifier now with both maximized margin and penalized misclassifications becomes:

$$\min_{w,b,\xi} \frac{1}{2}|w|^2 + C\sum_{i=1}^{N}\xi_i \quad \text{subject to:} \quad \xi i \geq 0, yi(w \cdot xi + b) \geq 1 - \xi i, \forall i \qquad (25)$$

where C is a hyperparameter called the cost or regularization parameter controlling the balance between margin width and classification errors. A high C penalizes misclassifications heavily resulting in a narrow margin, hence the model can overfit while a low C allows some misclassifications resulting to a wider margin hence the model is more robust to outliers and may underfit. In practice, the model selects the optimal C that yields the highest accuracy, producing the best model that can be used to classify a new loan client.

RESULTS

The sample dataset was made up of 4,592 records drawn from the 9 selected MFIs. The adequacy of the dataset size (n = 4,592) was evaluated using learning curves generated from Logistic Regression, Support Vector Machine and Naïve Bayes models as shown in Figure 3

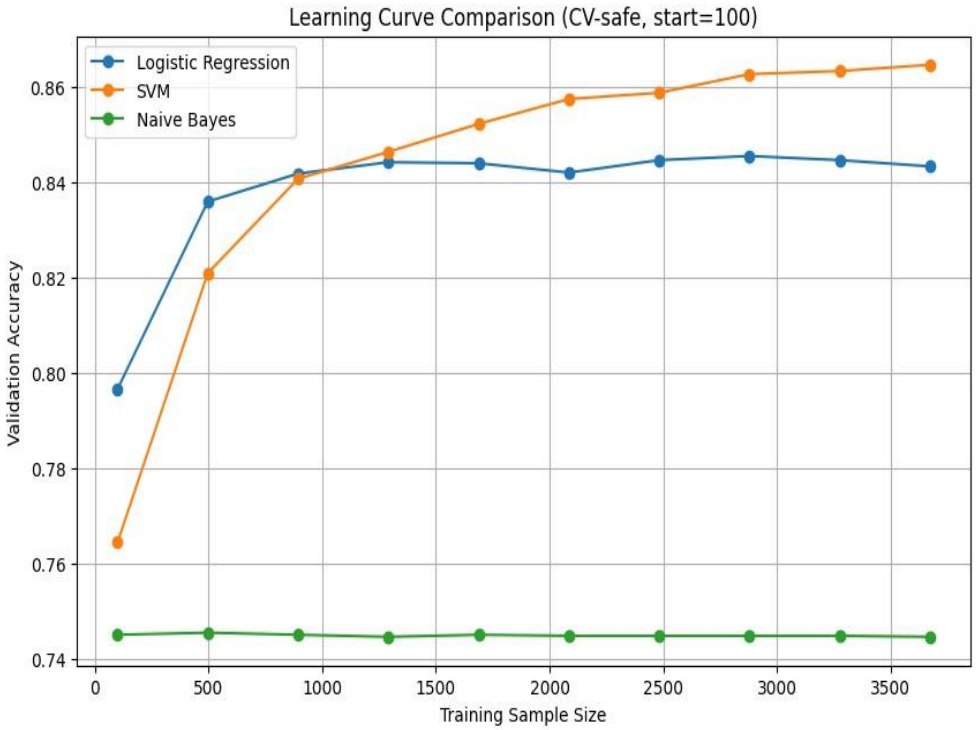

Figure 3. Learning Curves

The results of the learning curve analysis across sample sizes of 500 to 3,500 show distinct optimal sample size requirements for the three classifiers. Logistic Regression plateaued at 1,500-sample size, SVM at 3,000 while Naive Bayes showed no improvement beyond 500-sample size. These findings establish 3,000-sample size as the optimal training size for maximum accuracy. This indicates that the current dataset size (n=4,592) was sufficient for stable model performance.

Exploratory Data Analysis

Target Distribution

An analysis of the distribution of the target variable was conducted to evaluate the proportion of defaulted and non-defaulted loans. The results are presented in Figure 4.

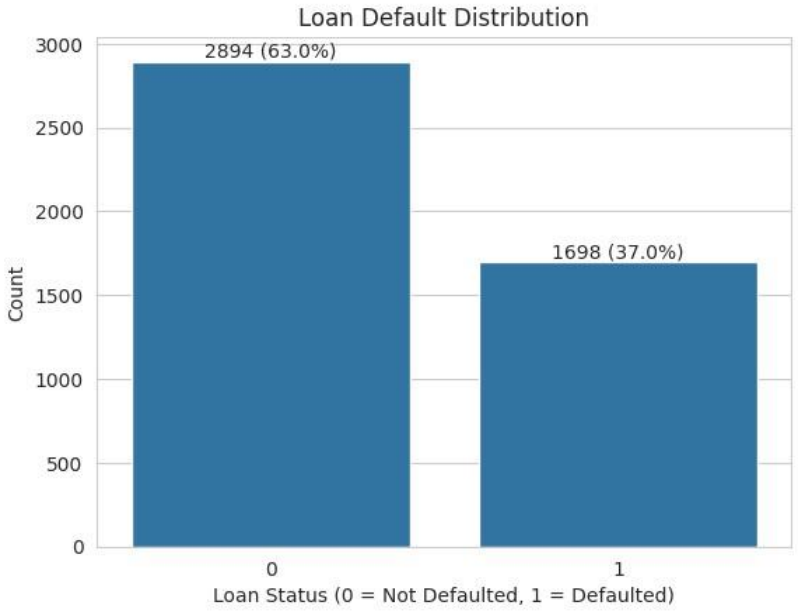

Figure 4. Distribution of the Target Variable

The analysis shows that 2894 (63.0%) of borrowers did not default, while 1698 (37.0%) defaulted. This indicates that although the majority of borrowers repay their loans successfully, a significant proportion still default. This finding are similar to the study by [11], which reported that loan datasets are often characterized by class imbalance with a higher proportion of non-defaulters. This imbalanced dataset hence call for application of classification methods that are robust to class imbalance.

Descriptive Analysis of Continuous Variables

A descriptive analysis for continuous variables was done to summarize their key feature characteristics, thereby providing an overview of the dataset structure. The results are presented in Table 2.

Table 2. Descriptive Statistics of the Continuous Variables

Variable	***N***	**Mean**	**SD**	**Min**	**Max**
Age	4592	36.97	8.78	18.00	73.00
Loan Amount	4592	22150.70	12089.75	5000	40000
Interest Amount	4592	6255.16	5040.37	1000	26000
Loan Term	4592	15.41	9.89	4	35
Interest Rate	4592	31.62	21.09	8	60
OLB	4592	11134.26	9492.17	0	36000
Monthly Income	4592	27776.79	12293.02	3000	50000
Serviced Loans	4592	3.29	2.04	0	10

The dataset contains 4,592 observations across all continuous variables. Loan amounts average Ksh 22,150, with a wide dispersion (SD = 12,089), indicating variation in borrowing capacity among clients. Interest rates average was at 31.62%, with a maximum of 60%, suggesting risk-based pricing across borrowers. Monthly income averages Ksh 27,776, indicating that most borrowers fall within the low-to-middle income category typical of

microfinance clients. The outstanding loan balance (OLB) shows a mean of Ksh 11,134, but ranges up to Ksh 36,000, indicating that some borrowers still carried significant repayment burdens.

Descriptive analysis of Categorical Variables

A visualization of how the distribution of loan default varied across different categorical variables was done and results presented in Figure 5.

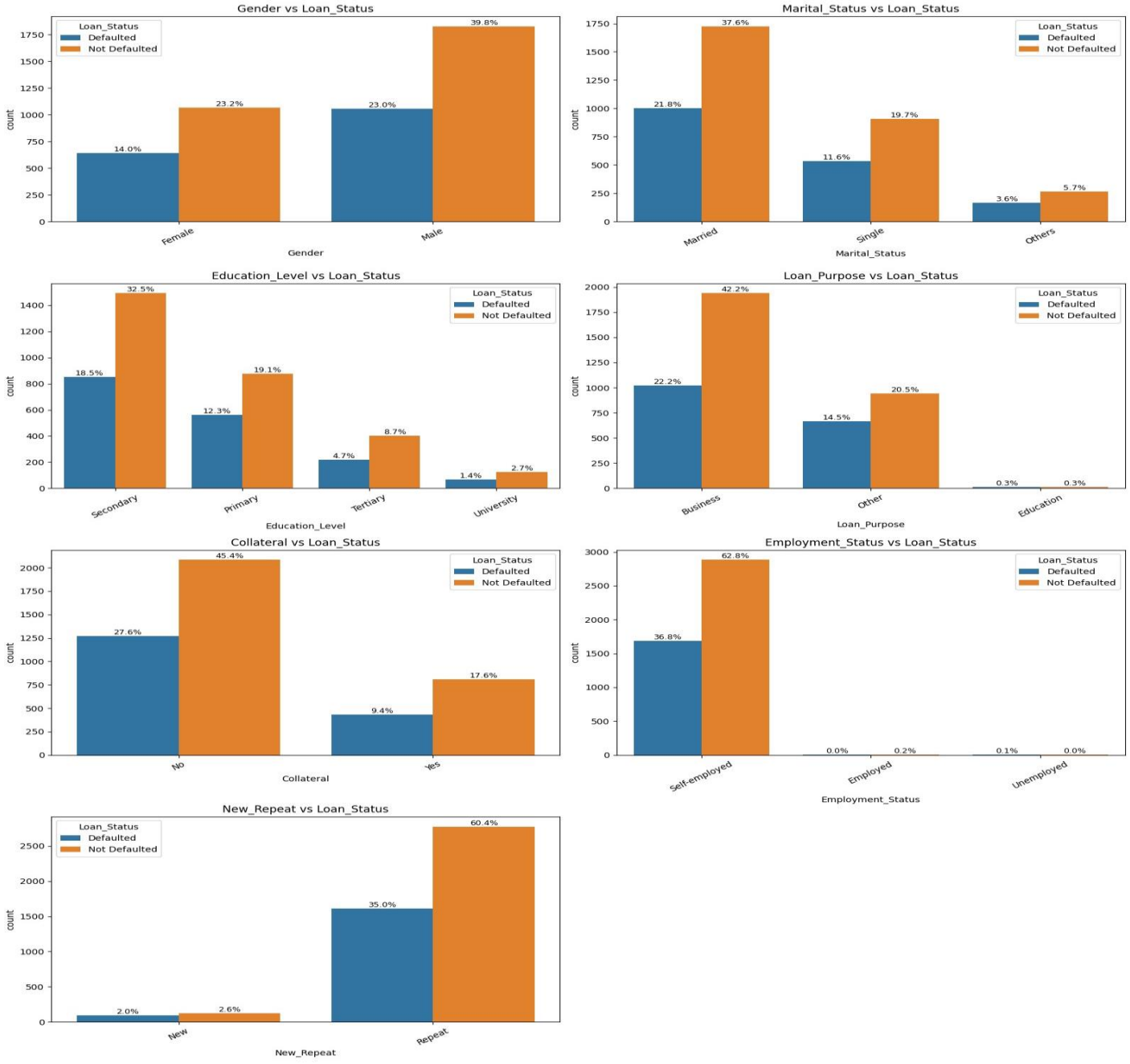

Figure 5. Categorical Variables and Loan Default Status

The series of bar charts shows relationship between borrower characteristics and loan default status. The findings are summarized below.

Gender and Loan Status

Male borrowers exhibited higher frequencies in both defaulted (23.0%) and non-defaulted loans (39.8%) compared to female borrowers (14.0% defaulted; 23.2% not defaulted). This suggests that males constitute a larger proportion of borrowers overall, though default patterns appear proportional across gender.

Marital Status and Loan Status

Married individuals accounted for the highest proportion of both defaulted (21.8%) and non-defaulted loans (37.6%), followed by single borrowers (11.6% defaulted; 19.7% not defaulted). Borrowers classified as "others" represented the smallest share (3.6% defaulted; 5.7% not defaulted). This indicates that marital status may be associated with loan participation but not necessarily with disproportionate default risk.

Education Level and Loan Status

Borrowers with secondary education formed the largest group (18.5% defaulted; 32.5% not defaulted), followed by primary education (12.3% defaulted; 19.1% not defaulted). Tertiary and university-educated borrowers had considerably lower representation in both categories. This pattern suggests that individuals with lower education levels dominate the borrower population.

Loan Purpose and Loan Status

Loans taken for business purposes had the highest proportions (22.2% defaulted; 42.2% not defaulted), followed by other purposes (14.5% defaulted; 20.5% not defaulted). Loans for education purposes were negligible (0.3% for both categories). This indicates that business loans are the most common and may carry relatively higher exposure to default simply due to volume.

Collateral and Loan Status

Borrowers without collateral showed higher default rates (27.6%) compared to those with collateral (9.4%). Similarly, non-default rates were higher among those without collateral (45.4%) than those with collateral (17.6%). These findings suggest that lack of collateral is associated with increased loan uptake and higher default counts.

Employment Status and Loan Status

Self-employed individuals constituted the majority of borrowers (36.8% defaulted; 62.8% not defaulted), while employed and unemployed categories had negligible representation. This indicates that the dataset is heavily skewed toward self-employed individuals.

Loan History and Loan Status

Repeat borrowers accounted for the majority of both defaulted (35.0%) and non-defaulted loans (60.4%), whereas new borrowers represented a very small proportion (2.0% defaulted; 2.6% not defaulted). This suggests that repeat borrowing is common and may be associated with sustained loan access regardless of default status.

These findings are consistent with [5], who observed that borrower characteristics such as business engagement, collateral status and loan usage patterns play an important role in determining the repayment outcomes in microfinance settings. The dominance of business-related loans and self-employed borrowers reveals the typical characteristics of microfinance client base, where access to credit is closely tied to informal economic activities.

Correlation Analysis

A correlation analysis was computed to assess relationships and multicollinearity among the independent variables before model fitting was done. The results were visualized using a correlation heatmap as shown on Figure 6

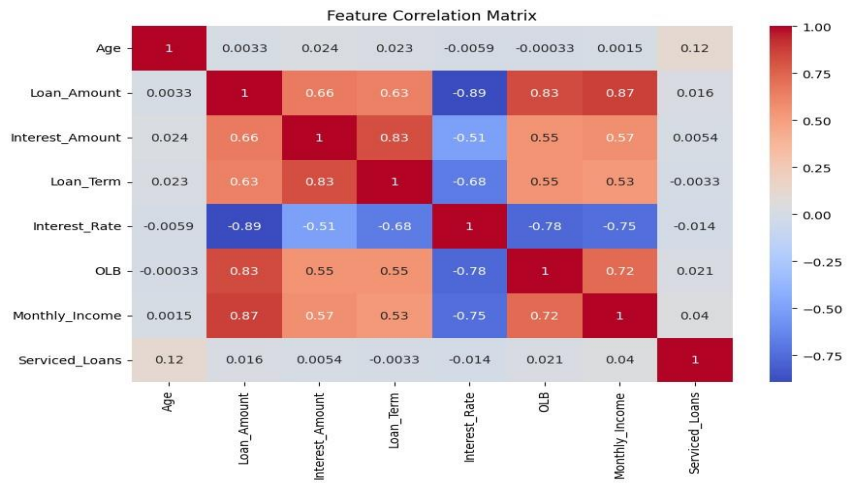

Figure 6. Correlation Heatmap

The correlation matrix revealed several strong relationships among the financial variables. Loan amount was strongly positively correlated with outstanding loan balance (OLB) (r = .83) and monthly income (r = .87), indicating that borrowers with higher incomes tend to receive larger loans and maintain higher outstanding balances. Loan amount was also moderately correlated with interest amount (r = .66) and loan term (r = .63).

Interest amount showed a strong positive correlation with loan term (r = .83), suggesting that longer loan durations are associated with higher interest charges. Additionally, OLB was strongly correlated with monthly income (r = .72), indicating a relationship between income levels and remaining loan obligations.

Notably, interest rate exhibited strong negative correlations with loan amount ($r = -.89$), OLB (r = −.78) and monthly income (r = −.75), suggesting that borrowers with higher loan amounts and incomes tend to receive lower interest rates. Age and serviced loans showed weak or negligible correlations with most variables, indicating minimal linear association with the other predictors.

It is worth noting that, the presence of several strong correlations (|r|) among key predictor variables suggests potential multicollinearity. Since Logistic Regression, one of our models in this research, is highly sensitive to multicollinearity, it was therefore necessary to further assess multicollinearity using the Variance Inflation Factor (VIF) prior to model fitting.

From the heatmap results in Figure 6, Interest Rate shows high correlation across all other variable. The researcher then drew attention to this particular variable. A Variance Inflation Factor (VIF) analysis was hence conducted before and after dropping Interest Rate variable. A VIF value greater than 10 indicates severe multicollinearity, while values between 5 and 10 suggest moderate multicollinearity. The results are presented in Table 3.

Table 3. Comparison of VIF Values Before and After Removing Interest Rate

Variable	VIF (Before)	VIF (After)
Loan Amount	17.97	7.42
Interest Rate	12.48	–
Loan Term	7.74	3.40
Interest Amount	7.61	3.57
Monthly Income	4.30	4.23

OLB	3.33	3.30
Serviced Loans	1.02	1.02
Age	1.02	1.02

This results indicate that prior to removal of Interest Rate variable, Loan Amount exhibited severe multicollinearity (VIF = 17.97), while Interest Rate also showed high collinearity (VIF = 12.48). Loan Term (VIF = 7.74) and Interest Amount (VIF = 7.61) indicated moderate to high multicollinearity. In contrast, Monthly Income (VIF = 4.30), OLB (VIF = 3.33), Serviced Loans (VIF = 1.02) and Age (VIF = 1.02) showed acceptable levels of multicollinearity.

After removing Interest Rate, the VIF values significantly decreased. Loan Amount reduced to 7.42, Loan Term to 3.40 and Interest Amount to 3.57. Monthly Income (VIF = 4.23), OLB (VIF = 3.30), Serviced Loans (VIF = 1.02) and Age (VIF = 1.02) remained largely unchanged. Overall, the results confirm that multicollinearity was sufficiently reduced and the dataset was suitable for subsequent model fitting.

Feature Importance Analysis

To assess feature influence on the loan default, a feature analysis was carried out for each model and results presented as follows.

Logistic Regression Model

A logistic regression model was fitted on the data and the following results were obtained. Table 4 presents the estimated coefficients of the logistic regression model used to identify the determinants of loan default among microfinance institution clients. The results include parameter estimates, standard errors, z-values and p-values, which are used to assess the statistical significance and direction of influence of each predictor variable.

Table 4. Weighted Binary Logistic Regression Results

Variable	**Coef.**	**Std. Err**	**z**	**p-value**	**2.5%**	**97.5%**
Constant	-1.4459	0.086	-16.753	< .001	-1.615	-1.277
Age	0.0863	0.049	1.768	.077	-0.009	0.182
Loan Amount	-6.3741	0.260	-24.531	< .001*	-6.883	-5.865
Interest Amount	0.7533	0.112	6.737	< .001*	0.534	0.972
Loan Term	-0.9904	0.105	-9.434	< .001*	-1.196	-0.785
OLB	5.1987	0.204	25.515	< .001*	4.799	5.598
Monthly Income	-0.1435	0.102	-1.413	.158	-0.343	0.056
Serviced Loans	0.4099	0.059	7.005	< .001*	0.295	0.525
Gender (2)	0.0140	0.048	0.293	.770	-0.080	0.108
Marital Status (2)	-0.0060	0.053	-0.114	.910	-0.110	0.098
Marital Status (3)	-0.0009	0.051	-0.018	.986	-0.101	0.100
Education Level (2)	-0.0706	0.055	-1.287	.198	-0.178	0.037

Education Level (3)	-0.0179	0.053	-0.336	.737	-0.122	0.087
Education Level (4)	-0.0060	0.053	-0.112	.911	-0.111	0.099
Loan Purpose (2)	-0.0159	0.053	-0.299	.765	-0.120	0.088
Loan Purpose (4)	0.3209	0.054	5.927	< .001*	0.215	0.427
Collateral (1)	-0.0290	0.049	-0.590	.555	-0.125	0.067
Employment Status (2)	0.1178	0.057	2.064	.039*	0.006	0.230
Employment Status (3)	0.3240	0.064	5.027	< .001*	0.198	0.450
New Repeat (2)	-0.2036	0.055	-3.692	< .001*	-0.312	-0.096

Note. * indicates statistically significant variables at $p < .05$.

As shown in Table 4, a unit increase in age increased the log odds of defaulting by 0.0863, meaning that older borrowers were slightly more likely to default compared to younger borrowers, holding other variables constant. For every additional unit increase in loan amount, the log odds of defaulting decreased by 6.3741, indicating that higher loan amounts were associated with a lower likelihood of default, holding other factors constant. For very additional unit increase in Interest Amount, the log odds of defaulting increased by 0.7533 suggesting that an increase in the interest amount increased the chances of defaulting. For every additional unit increase in loan term, the log odds of defaulting decreased by 0.9904, suggesting that longer repayment periods reduce the probability of default. A unit increase in outstanding loan balance (OLB) increased the log odds of defaulting by 5.1987, indicating that borrowers with higher outstanding balances were significantly more likely to default, holding all other variables constant. For every additional unit increase in monthly income, the log odds of defaulting decreased by 0.1435, although this effect was not statistically significant at $\alpha = 0.05$. For every additional unit increase in loans serviced, the log odds of defaulting increased by 0.4099, indicating that borrowers with more prior loan engagements were more likely to default. A male applicant changed the log odds of defaulting by 0.0140, indicating a very small increase in the likelihood of default compared to female applicants, holding all other variables constant. Applicants marital status and education level did not have a significant effect ($p > 0.05$) to loan defaulting at $\alpha = 0.05$. Borrowers whose loan purpose category was 4(other loan uses) had their log odds of defaulting increased by 0.3209 compared to the reference category. Collateral category 1(loan with collateral) did not have a significant effect to the loan defaulting. Being in employment status category 3(unemployed) increased the log odds of defaulting by 0.3240, while employment status category 2(self employed) increased the log odds of defaulting by 0.1178, although this effect was marginally significant. Being a repeat borrower (New/Repeat = 2) reduced the log odds of defaulting by 0.2036, indicating that repeat clients were less likely to default compared to new borrowers.

Results show that Loan Amount, Interest Amount, Loan Term, OLB, Serviced Loans, Loan Purpose, Employment Status and New/Repeat (loan history) predictors were statistically significant ($p < 0.05$) at $\alpha = 0.05$. This means that they significantly affect the applicants' ability to repay their loans. Using these statistically significant factors, the researcher constructed the parsimonious model as follows.

From Equation 3, the logistic regression model is expressed in terms of probability as:

$$p(y = 1|X) = \frac{1}{1+e^{-z}} \tag{26}$$

where z is the linear predictor given by:

$$z = 3.0403 - 0.0005(LoanAmount) + 0.0001(InterestAmount) - 0.0992(LoanTerm) + 0.0005(OLB) + 0.1975(ServicedLoans) + 0.6951(LoanPurpose\_4) + 1.7667(EmploymentStatus\_2) + 7.6876(EmploymentStatus\_3) - 0.9194(New/Repeat\_2) \quad (27)$$

Thus, the model estimates the probability that the outcome variable equals one given the explanatory variables included in the model ($p(y = 1|X)$). If $p(y = 1|X) > 0.5$, the client is classified as a defaulter otherwise a non-defaulter. This is the best binary logistic model that can be used to predict the probability of default.

Naïve Bayes Model

To asses which features influenced on the target variable (defualt status), a Naïve Bayes classifier was fitted, a feature influence analysis was done and results present in Table 5.

Table 5. Feature Importance Based on Mean Difference

Feature	Mean Difference
Loan Amount	10275.592
Monthly Income	8722.168
Interest Amount	2878.710
OLB	2078.392
Loan Term	6.468
Age	0.528
Serviced Loans	0.227
Loan Purpose	0.220
Education Level	0.040
Collateral	0.033
Gender	0.011
Marital Status	0.010
New Repeat	0.010
Employment Status	0.004

Mean differences between classes (Class 0 minus Class 1) indicated that the top three features separating classes were Loan Amount ($M_{\text{diff}} = 10275.59$), Monthly Income ($M_{\text{diff}} = 8722.17$) and Interest Amount ($M_{\text{diff}} = 2878.71$), with positive differences suggesting these values were higher on average for Class 0. In contrast, features such as Gender ($M_{\text{diff}} = 0.01$), Marital Status ($M_{\text{diff}} = 0.01$) and Employment Status ($M_{\text{diff}} = 0.0036$) showed negligible class separation hence minimal influence.

Support Vector Machine Model

To evaluate the influence of each feature on the target variable (loan default status), SVM feature importance analysis was conducted. The results are presented in Table 6.

Table 6. SVM Feature Importance

Feature	Importance
Loan Amount	0.301
OLB	0.157
Serviced Loans	0.033
Loan Purpose	0.029
Loan Term	0.022
Monthly Income	0.020
Interest Amount	0.006
New Repeat	0.004
Employment Status	0.003
Collateral	0.001
Age	-0.000
Marital Status	-0.001
Gender	-0.003
Education Level	-0.006

The feature importance results indicate that the most influential predictors of loan default were Loan Amount (0.236), OLB (0.158) and Interest Rate (0.144). These variables had the highest impact on model performance, suggesting they play a major role in distinguishing between defaulters and non-defaulters. On the other hand, variables such as Education Level, Collateral and Age displayed negligible importance, indicating minimal contribution to classification accuracy

Metric Evaluations Analysis

A comparison of the three classification models, i.e., Logistic Regression, Naive Bayes and Support Vector Machine (SVM), was conducted using Confusion Matrix, accuracy, precision, recall, F1-score and area under the curve (AUC) as evaluation metrics. Table 7 presents the results of the confusion matrices obtained which summarizes the number of correct and incorrect predictions.

Table 7. Confusion Matrix Results

Model	**TN**	**FP**	**FN**	**TP**
Logistic Regression	455	124	28	312
Naive Bayes	482	97	119	221
SVM (RBF)	460	119	8	332

Note: TN = True Negatives; FP = False Positives; FN = False Negatives; TP = True Positives.

From this results, SVM model achieved the highest classification accuracy, correctly identifying 460 non-defaulters (true negatives) and 332 defaulters (true positives), with relatively few misclassifications (119 false positives and 8 false negatives). This indicates strong predictive performance and balanced classification ability. The logistic regression model also performed well, correctly classifying 455 non-defaulters and 312 defaulters, with 124 false positives and 28 false negatives. This was slightly less accurate than SVM, but still demonstrated good classification capability. In contrast, the Naive Bayes model showed comparatively weaker performance, correctly identifying 482 non-defaulters and 221 defaulters, but with higher misclassification rates (97 false positives and 119 false negatives), particularly in failing to correctly identify defaulters. The confusion matrices confirm that SVM model provides the most accurate and reliable classification, followed by logistic regression, while Naive Bayes performs least effectively.

The predictive performance of the three classification models was assessed using Stratified 5-fold cross-validation. The evaluation was based on multiple performance metrics, including accuracy, precision, recall, F1-score, specificity, and the area under the receiver operating characteristic curve (AUC), as summarized in Table 8.

Table 8. Performance Comparison of Classification Models

Model	Accuracy	Precision	Recall	F1	Specificity	AUC
SVM (RBF)	0.857	0.739	0.949	0.831	0.803	0.914
Logistic Regression	0.830	0.715	0.900	0.797	0.789	0.904
Naive Bayes	0.740	0.652	0.641	0.646	0.799	0.769

Further analysis confirmed that SVM demonstrated the best overall performance, achieving the highest accuracy (0.857), precision (0.739), recall (0.949), F1-score (0.831), specificity (0.803) and AUC (0.914). This indicates superior classification ability and discriminative power compared to the other models. The logistic regression model also performed well, with an accuracy of 0.830, precision of 0.715, recall of 0.900, F1-score of 0.804, specificity of 0.797 and AUC of 0.904. Although slightly lower than SVM, it still demonstrated strong predictive capability. In contrast, Naive Bayes model showed the lowest performance across all metrics, with an accuracy of 0.740, precision of 0.652, recall of 0.641, F1-score of 0.646, specificity of 0.799 and AUC of 0.769, suggesting comparatively weaker classification performance. Overall, this results indicate that the SVM model outperformed both logistic regression and Naive Bayes, making it the most suitable model for predicting loan default in this study.

A visualization of the Receiver Operating Characteristic (ROC) curve was conducted to evaluate the discriminative ability of the three classification models and results present in Figure 7.

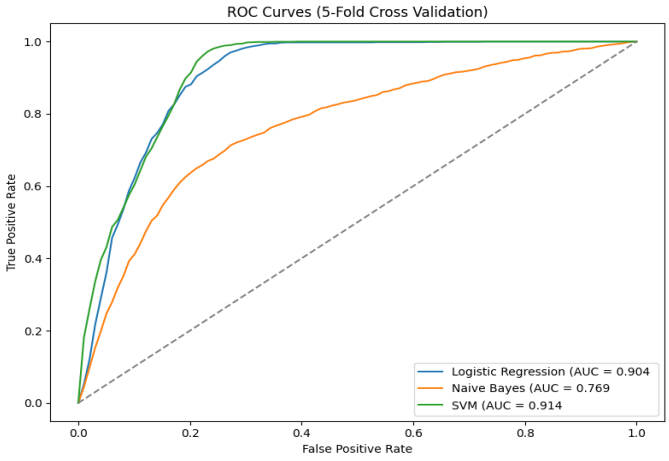

Figure 7. Area Under the Curve (AUC) for Classification Models

From the visualization results, SVM exhibited the best performance, with an Area Under the Curve (AUC) of 0.914. Its ROC curve lies closest to the top-left corner of the plot, indicating a strong ability to correctly distinguish between defaulters and non-defaulters across various threshold levels. The logistic regression model also demonstrated high discriminative power, with an AUC of 0.904. Its curve closely follows that of the SVM, suggesting that it performed well but was slightly less effective in classification compared to SVM. In contrast, the Naive Bayes model recorded a lower AUC of 0.769. Its ROC curve is further away from the top-left corner and closer to the diagonal reference line, indicating comparatively weaker classification performance. The diagonal dashed line represents a random classifier (AUC = 0.5) and all three models perform above this baseline, confirming their predictive usefulness. Overall, the ROC analysis reinforced that the SVM model provided the best classification performance among the three models.

Model Significance Test

A DeLong's test for paired AUC comparisons was conducted to determine whether the observed differences in discriminative performance among Logistic Regression, Naive Bayes and Support Vector Machine (SVM) models were statistically significant. The results were presented in Table 9.

Table 9. DeLong's test

Comparison	Difference	z	p	Significant (α = 0.05)
Logistic Regression vs. Naive Bayes	0.1380	19.540	< .001	Yes
Logistic Regression vs. SVM	-0.0098	-3.067	.002	Yes
Naive Bayes vs. SVM	-0.1478	-23.571	< .001	Yes

The results of DeLong's test for paired AUC comparisons revealed statistically significant differences across all three models. SVM achieved the highest AUC (0.914), significantly outperforming Logistic Regression ($z = -3.067$, $p = 0.002$) and Naive Bayes ($z = -23.571$, $p < 0.001$). Logistic Regression also significantly outperformed Naive Bayes ($z = 19.540$, $p < 0.001$). These results provide rigorous statistical evidence supporting SVM's superior discriminative performance in loan default classification.

DISCUSSION

Feature Importance

Loan amount emerged as one of the most influential predictors across the three model feature importance analysis. This finding is consistent with a study by [12] who found that loan size significantly influenced credit risk among microfinance institutions in Kenya. Large loan amounts may increase repayment burden and financial exposure, thereby increasing the likelihood of borrower default. Outstanding loan balance (OLB) also demonstrated strong predictive influence in the present study. This finding supports the observations of [7], who established that loan portfolio characteristics significantly affect non-performing loans in Kenyan microfinance banks. High outstanding balances may indicate over-indebtedness and repayment difficulties among borrowers, which consequently increases the default risk.

Interest amount and loan term were also found to significantly influence loan default. These findings agree with [5], who reported that interest rates and loan repayment conditions significantly affect borrower repayment behavior among small business owners in Rwanda. Higher interest obligations and unfavorable repayment periods may place financial pressure on borrowers and increase the probability of default. The number of serviced loans was another important predictor retained in the final model. This suggests that previous borrowing experience and loan management history significantly influence repayment performance. These findings are consistent with [11], who established that borrower monitoring and repayment behavior significantly affect loan repayment among microfinance clients in Makueni County.

Employment status and borrower type (new or repeat client) were also statistically significant predictors in the final model. Repeat borrowers were less likely to default compared to new borrowers, suggesting that borrower familiarity and repayment history contribute positively to creditworthiness assessment. These findings confirm that loan default among microfinance institution clients is influenced by a combination of borrower financial characteristics, loan conditions and behavioral factors.

Model Comparison

The comparative analysis of logistic regression, naive Bayes and support vector machine revealed significant differences in the predictive performance of the three classification models. Among the three models evaluated, SVM achieved the best overall performance with an accuracy of 85.7% and AUC of 0.914. Logistic Regression followed closely with an accuracy of 83.0% and AUC of 0.904, while Naïve Bayes recorded the weakest performance with an accuracy of 74.0% and AUC of 0.769.

The superior performance of SVM observed in this study is consistent with findings by [9], who demonstrated that SVM outperformed Logistic Regression in predicting loan default using Moroccan banking data. Their study reported a ROC value of 98% for SVM compared to 71.7% for Logistic Regression. Similarly, [13] reported that SVM and Logistic Regression significantly improved classification performance in Kenyan microfinance institutions. The strong performance of SVM in the present study suggests that robust nonlinear machine learning classification algorithms are highly effective in capturing the complex interactions and patterns among borrower characteristics and loan repayment behavior.

The results also align with [1], who emphasized that predictive performance varies depending on dataset characteristics and model selection. The findings of the current study demonstrate that although Logistic Regression remains effective and interpretable, SVM provides better discriminative power and classification accuracy for loan default prediction. This reinforces the argument that advanced machine learning techniques can improve predictive performance in financial risk assessment.

CONCLUSION

These results indicate that loan default is primarily driven by financial exposure (loan amount, outstanding balance, repayment term) and borrower characteristics (employment status and repayment history). Notably, repeat borrowers were less likely to default, suggesting that credit history is a strong protective factor. Employment status emerged as the most influential categorical predictor, showing that income stability significantly affects repayment behaviour.

The comparative analysis showed clear differences in predictive performance across the three models. SVM outperformed all models with an accuracy of 86%, precision of 0.739, recall of 0.949, F1-score of 0.831, specificity of 0.7803 and AUC of 0.914. Logistic Regression followed closely with an accuracy of 83% and AUC of 0.904, demonstrating strong interpretability and competitive predictive power. Naive Bayes performed least effectively, with an accuracy of 74% and AUC of 0.769, indicating weaker assumptions of feature independence limited its performance. Overall, SVM demonstrated superior discriminatory ability in separating defaulters from non-defaulters, making it the most suitable model for deployment in similar credit scoring environments.

Future research should explore ensemble-learning methods to further improve classification performance.

REFERENCES

1. R. Kandula et al., “Comparative Analysis for Loan Approval Prediction System Using Machine Learning Algorithms,” in Proceedings of Fifth International Conference on Computer and Communication Technologies (IC3T 2023), B. R. Devi, K. Kumar, M. Raju, K. S. Raju, and M. Sellathurai, Eds. Springer, 2024, vol. 897. doi: 10.1007/978-981-99-9704-6_18.
2. Association of Micro-finance Institutions–Kenya, Microfinance Sector Report 2025. AMFI-K, 2025. [Online]. Available: https://www.amfikenya.com.

3. Central Bank of Kenya, Bank Supervision Annual Report 2024. Central Bank of Kenya, 2024.
4. Central Bank of Kenya, State of the Banking Industry Report 2025. Central Bank of Kenya, 2025.
5. D. Twesige, A. Uwamahoro, P. Ndikubwimana, F. Gasheja, I. K. Misago, and U. Hategikimana, "Causes of Loan Defaults Within Microfinance Institutions: Learning from Micro and Small Business Owners in Rwanda: A Case of MSEs in Kigali," Rwanda Journal of Social Sciences, Humanities and Business, vol. 2, no. 1, pp. 27–49, 2021. doi: 10.4314/rjsshb.v2i1.3.
6. H. Njuguna, "Kenya's Microfinance Banks: A Decade of Declining Fortunes," 2025. [Online]. Available: https://kenyanwallstreet.com/kenyas-microfinance-banks-a-decade-of-declining-fortunes.
7. H. R. Boye and F. Mithi, "Firm Characteristics and Non-Performing Loans of Microfinance Banks in Kenya," The Strategic Journal of Business & Change Management, vol. 10, no. 4, pp. 368–381, 2023. doi: 10.61426/sjbcm.v10i4.2762.
8. I. N. Barasa, S. W. Wanyonyi, and M. M. Kololi, "Application of Logistic Regression in Enhancing Digital Credit Risk Management in Commercial Banks," Asian Journal of Probability and Statistics, vol. 27, no. 2, pp. 13–26, 2025. doi: 10.9734/ajpas/2025/v27i2710.
9. K. Amzile and H. Mohamed, "Assessment of Support Vector Machine performance for default prediction and credit rating," Banks and Bank Systems, vol. 17, no. 1, pp. 161–175, 2022. doi: 10.21511/bbs.17(1).2022.14.
10. M. A. Rahman, "Financial Inclusion and Economic Development," International Journal of Economics and Management Intellectuals (IJEMI), 2024. doi: 10.63665/ijemi-y1f1a002. [Ahead of print].
11. M. K. Mwaka, "Factors Influencing Repayment Among Microfinance Loan Consumers in Makueni County: A Case of Nzaui/Kilili/Kalamba Ward, Makueni County, Kenya," Master's thesis, University of Nairobi, 2017.
12. M. M. Kasina, J. M. Kihoro, and A. Kibet, "Performance of Different Machine Learning Algorithms for Credit Risk Classification," Journal of Data Analysis and Information Processing, vol. 13, pp. 504–519, 2025b. doi: 10.4236/jdaip.2025.134029.
13. M. Raymond, "The Impact of Machine Learning on Reducing Credit Risk in Microfinance Institutions," 2025. [Online]. Available: https://www.researchgate.net/publication/394414373.