University of Ibadan
Journal of Science and Logics in ICT Research

# Evaluation of Machine Learning-Based Algorithm to Predicting Loan Default in Nigeria

[1] Efekodo K. O., [2]Akinola O. S., and [3]Waheed A. A.

*[1, 3]Department of Computer Sciences, Lead City University, Ibadan, Oyo State, Nigeria*
*[2]Department of Computer Science, University of Ibadan, Ibadan, Oyo State, Nigeria*
*Emails: keoghenekaro@yahoo.com, solom202@yahoo.co.uk, waheed.azeez@lcu.edu.ng*

*Abstract*

Accurately predicting loan defaults is critical in the financial sector to minimize losses and optimize credit risk management. Traditional creditworthiness assessment methods often fail to capture the complex, dynamic interactions in financial data, leading to inaccurate predictions. This study harnesses advanced machine learning techniques to enhance the prediction of loan defaults, aiming to outperform traditional statistical models. A dataset containing 50,000 borrower records with diverse characteristics, including demographic, financial, and loan-specific features, was utilized. The data was split into training (70%) and test (30%) sets for model development and evaluation. Various machine learning algorithms were tested, including Logistic Regression, Decision Trees, Gradient Boosting Classifiers, Random Forest, and Gaussian Naive Bayes. The Gaussian Naive Bayes (GaussianNB) model demonstrated superior performance, achieving an accuracy of 78.8% on the test set. This model effectively captured complex patterns in the high-dimensional data, significantly reducing false positives and false negatives compared to other models. The findings suggest that machine learning models, particularly GaussianNB, offer substantial improvements in predictive accuracy for loan default risk assessments. This findings can enhance lenders' decision-making processes by improving risk stratification and resource allocation. Future research should explore integrating non-traditional data sources, such as behavioral and macroeconomic variables, and employing deep learning techniques to further refine predictive accuracy.

***Keywords:*** *Accuracy, Classifier, Decision Trees, Gaussian Naive Bayes, Gradient Boosting Classifiers, Logistic Regression, Random Forest*

## 1. Introduction

Consumer expenditure plays a critical role in shaping the overall economic landscape and contributing to financial risks. A significant aspect of this spending often involves consumer credit, where individuals utilize loans to meet their consumption needs. As of 2023, the Consumer Marketplace Lending sector is projected to achieve a transaction value of $78.57 million, with an expected compound annual growth rate (CAGR) of 5.29% between 2023 and 2027. By 2027, this sector is estimated to grow to $96.57 million, with a per-user transaction value of $48.75 million forecasted for 2023. The United States is expected to lead the international market, recording a transaction value of $26.18 billion in the same year [1].

Efekodo K. O., Akinola O. S., and Waheed A. A. (2025). Evaluation of Machine Learning-Based Algorithm to Predicting Loan Default in Nigeria, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 13 No. 1, pp. 118 – 129



In Nigeria, the credit market is primarily regulated by the Central Bank of Nigeria (CBN), which oversees Deposit Money Banks (DMBs) [2]. However, the market also comprises credit lenders not directly governed by the CBN, such as Primary Mortgage Institutions, which report to the Federal Mortgage Bank, and leasing firms regulated by the Equipment Leasing Association of Nigeria [3] Loans, in their essence, represent agreements in which lenders provide funds or assets to borrowers under the assurance that the borrowers will repay, often with interest. This practice is a core operation of many banking institutions, as interest earned from loans is a key revenue source [4].

Despite the importance of loans, there exists a risk of default by borrowers, which can occur during the tenure of the loan if they fail to meet repayment obligations. Evaluating the probability of default is crucial for effective risk management. Traditionally, credit officers have assessed borrower creditworthiness manually through historical credit data. However, recent advancements in technology have led to the adoption of more sophisticated tools, including automated systems and machine learning models, to analyze credit risk [5]. Credit risk is defined as the potential loss lenders may face if borrowers fail to meet their obligations, remains a fundamental concern in lending. This risk influences the terms of credit approvals and the interest rates charged by financial institutions. Accurately predicting defaults is a major challenge, necessitating robust models to minimize human error in consumer credit assessments [6].

The advent of big data and machine learning has transformed credit risk analysis by enabling more accurate and scalable predictive models. Unlike traditional methods relying on statistical regression, machine learning algorithms analyze datasets to predict outcomes using sophisticated techniques [7]. Automated loan default models have gained popularity among lending institutions, offering efficient and accurate risk assessments.

Machine learning algorithms, such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees (DT), and Bayesian classifiers, are commonly employed to predict loan defaults [8]. These models were selected due to their ability to handle high-dimensional data (Gradient Boosting Classifiers and Random Forest), interpretability (Logistic Regression), and performance in classification tasks with limited data points (Gaussian NB). Additionally, Decision Trees provide a clear decision-making process that aligns with creditworthiness assessment frameworks [9]. However, loan prediction models often encounter errors, notably false positives (incorrect rejection of creditworthy applicants) and false negatives (approval of non-creditworthy applicants), which reduce prediction accuracy and overall efficiency [6].

Class imbalance, where non-defaulting borrowers significantly outnumber defaulters, exacerbates prediction errors by skewing model outputs. Techniques such as SMOTE and cost-sensitive learning have been employed to address these imbalances, improving prediction reliability [10]. Ensemble methods that combine multiple machine learning algorithms have also been explored to address overestimation of defaults. However, identifying variables that effectively reduce the incidence of misclassification remains a challenge [11]. This complexity underlines the need for continuous innovation in model development, using diverse datasets to achieve optimal performance.

Machine learning is a subset of Artificial Intelligence (AI), encompasses data analysis and model building based on structured data. ML algorithms enable computers to learn from data and make informed predictions or classifications [12]. The two primary methodologies in ML are supervised and unsupervised learning. Supervised learning involves training algorithms with labeled data to predict outcomes accurately, while unsupervised learning identifies patterns within unlabeled datasets [13]. Supervised techniques, including binary classification models, are widely used in credit scoring and loan prediction, emphasizing data quality and appropriate feature selection for model optimization [14].

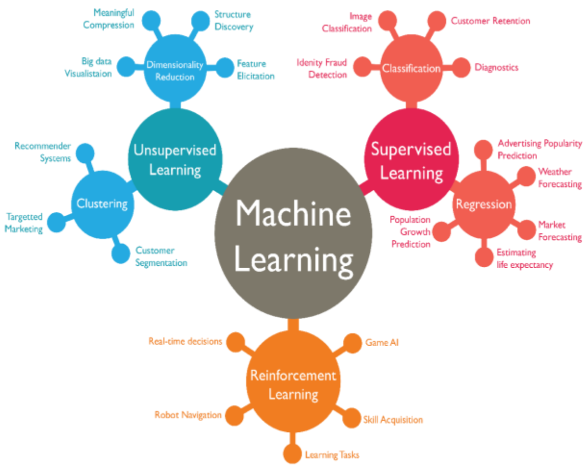


**Figure 1**: **Simple Illustration of Machine Learning Techniques [15]**.

The development of effective machine learning models depends on the nature and quality of data inputs. As such, data-driven approaches are pivotal in improving predictive accuracy and minimizing risk in loan default predictions. The ultimate goal is to create models that enhance decision-making capabilities within lending institutions, thereby improving risk management, reducing bad debt, and fostering financial ]stability in the credit market [16]. Leveraging machine learning for credit risk assessment offers significant potential to enhance predictive accuracy, reduce default rates, and streamline credit decision processes. This approach is particularly relevant in Nigeria's evolving credit landscape, where a balance between technological innovation and regulatory oversight is essential. Thus, this study proposes evaluating machine learning-based algorithms to improve loan eligibility predictions based on factors like education, employment, and loan history, as some studies [17][18] had shown limited accuracy and precision with alternative algorithms. Hence, this study conducts a comparative analysis of five algorithms Logistic Regression, Decision Trees, Gradient Boosting Classifiers, Random Forest, and Gaussian NB and other performance metrics, offering a detailed performance assessment in the context of predicting the creditworthiness of borrowers.

## 2. Related Works

Machine learning algorithms have been extensively applied to loan prediction, with studies showcasing varying degrees of success and limitations. Sharma *et al.*, [19] developed a model integrating multiple machine learning algorithms and ensemble methods like bagging and voting classifiers, achieving a prediction accuracy of 94%. While their model reduced human intervention and optimized decision-making processes, it did not address class imbalance or challenges related to ensemble model interpretability key factors for practical deployment in lending institutions.

Similarly, Nandipati & Boddala [20] used machine learning algorithms such as decision trees, random forests, support vector machines, and K-nearest neighbors to evaluate loan eligibility. The ensemble decision tree with AdaBoost achieved superior accuracy. However, their study highlighted concerns about overfitting in ensemble models, which were not fully resolved. In contrast, Thakar *et al.*, [21] employed an Artificial Neural Network (ANN) to create a loan prediction system with 92% accuracy, demonstrating its ability to predict repayment probabilities. Despite the high accuracy, the complexity and lack of transparency in ANN models limit their interpretability and scalability.

Saini *et al.*, [22] compared Random Forest, Logistic Regression, K-nearest Neighbors, and Support Vector Machines for loan approval prediction, with Random Forest achieving the highest accuracy (98.04%). However, their reliance on accuracy as the sole performance metric ignored precision and recall essential for assessing performance in imbalanced datasets. Akça & Sevli [23] used Logistic Regression with sensitivity and specificity metrics, incorporating demographic features like age and credit history. Their focus on interpretability provided a practical approach, though their evaluation was limited to one algorithm, restricting broader insights.

Several studies have addressed the challenge of imbalanced datasets. Babo & Beyene [24] used Support Vector Machines with various kernels, achieving 97.2% accuracy with the poly kernel. However, they relied heavily on data preprocessing without exploring cost-sensitive learning or other advanced techniques for imbalanced data. Patel & Bhavsar [25] combined MSMOTE with ensemble classifiers to balance datasets, achieving 99% accuracy and precision with Random Forest and Bagging models. This approach underscores the importance of preprocessing in mitigating imbalance but highlights the need for comparison with alternative methods.

Ensemble and deep learning methods have shown promise in recent studies. Archana & Divyalakshmi [26] highlighted the superior performance of ensemble methods like bagging and deep learning models for loan eligibility prediction. However, their findings raised concerns about the computational costs of deep learning models, which may limit scalability for financial institutions. Sujatha *et al.*, [27] provided a comprehensive review of advancements in loan prediction, identifying key trends but offering limited empirical evidence to support the implementation of specific techniques.

Other studies have emphasized the comparative performance of traditional algorithms. Similarly, **Abdullah *et al.*, [28]** found Logistic Regression to be the most effective model for predicting loan approvals using demographic data. Both studies emphasized the practicality of Logistic Regression but did not address its limitations in handling complex, nonlinear relationships. Several commonalities emerge across these studies. Ensemble methods such as AdaBoost and bagging consistently demonstrate high accuracy in loan prediction, as shown by Sharma *et al.*, [19] and Nandipati & Boddala [20]. However, their computational intensity and potential overfitting require careful consideration. Logistic Regression remains a reliable baseline for interpretable models, as supported by Akça & Sevli [23] **and** Abdullah *et al.,* [28]**,** though its limitations with complex data warrant supplementation with ensemble or advanced machine learning techniques.

Class imbalance is a recurring challenge in loan prediction studies. Techniques like MSMOTE and data preprocessing employed by Babo & Beyene [24]**,** have proven effective in mitigating this issue. Despite these efforts, alternative solutions, including cost-sensitive learning and threshold adjustment, remain underexplored. While accuracy is the most commonly reported metric, other performance measures like precision, recall, and F1-score are critical for imbalanced datasets but often overlooked. Few studies, such as those by Akça & Sevli [23]**,** have emphasized sensitivity and specificity, signaling a need for more comprehensive evaluation metrics in this domain.

### *2.1 Research Gap*

The reviewed studies demonstrate the efficacy of various machine learning approaches but often lack a balanced consideration of interpretability, scalability, and real-world applicability. Additionally, the integration of demographic and behavioral attributes, such as educational attainment, employment status, and repayment history, is limited. These attributes are essential for accurately assessing creditworthiness and tailoring predictions to diverse borrower profiles. Furthermore, while some studies address class imbalance, few evaluate the comparative performance of techniques like MSMOTE against alternative approaches.

This study uniquely addresses these gaps by incorporating borrower attributes and employing advanced machine learning models, including Logistic Regression, Decision Trees, Gradient Boosting Classifiers, Random Forest, and Gaussian NB. By evaluating these models across multiple metrics, such as precision, recall, and F1-score, the research aims to provide a nuanced understanding of their effectiveness in loan eligibility prediction.

## 3. Methodology

### *3.1 Research Approach*

The research employs a supervised learning approach, focusing on training a classification model to predict loan default. Supervised learning is well-suited for this task, as it relies on labeled datasets where the input features are used to predict specific target outcomes. The dataset used in this research is publicly accessible, sourced from [29]. It includes detailed demographic, performance, and previous loan data to support the study's objectives. The dataset contains 3 different datasets for both train and test; Demographic data (traindemographics.csv), Performance data (trainperf.csv), Previous loans data (trainprevloans.csv) making it suitable for the computational capabilities of the selected hardware.

Cross-validation techniques, such as k-fold cross-validation, were applied to ensure the robustness of the models. This approach involves splitting the dataset into k subsets (folds), iteratively training the model on k-1 folds and testing it on the remaining fold, thereby minimizing the risk of overfitting and enhancing generalization. Four classification algorithms are utilized in the study: Decision Tree, Gradient Boosting Classifier, Random Forest, and Gaussian Naive Bayes (NB) Classifier. The study's approach includes dividing the dataset into training and testing subsets, ensuring that the model is exposed to different data during training and evaluation. Additionally, cross-validation techniques are employed to further ensure the robustness of the model. Cross-validation involves splitting the data into several folds, training on some folds while testing on others, which helps in minimizing overfitting and obtaining a more generalizable model.

### *3.2 Requirement Specification*

**Hardware:** The study utilizes a personal computer with an Intel Core i5 processor (2.2 GHz) and 8GB RAM, suitable for processing and training moderately-sized datasets without significant delays.

**Software requirements:** IDEs such as PyCharm, Jupyter Notebook, and Visual Studio Code for coding and debugging. Data manipulation is performed using Pandas and NumPy, while data visualization employs Matplotlib and Seaborn. Scikit-Learn is used for implementing and evaluating machine learning models, and Anaconda ensures effective dependency management and smooth installation of necessary libraries, facilitating seamless development and integration of project components.

### *3.3 Research Design*

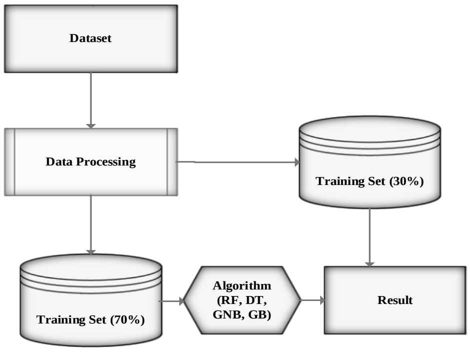


**Figure 2: Conceptual Model of the Proposed Design**

**Data Collection:** The dataset was downloaded from [29] a publicly accessible repository. It includes three segments: demographic data, performance data, and historical loan data. These segments were selected for their relevance and completeness in addressing the study's objectives.

**Dataset Details:** The dataset is composed of three key segments: demographic data, performance data, and previous loan data, each of which provides critical insights into customer behavior. The training and test datasets are structured to facilitate machine learning model development, with the target variable labeled as "good" (1) or "bad" (0), indicating whether a loan is likely to default. This binary classification allows the models to distinguish between high-risk and low-risk loan applicants effectively.

**Dataset Description:**

- *Demographic Data:* This dataset includes information like customer ID, birthdate, type of bank account, geographical coordinates (longitude and latitude), bank name and branch, employment status, and the highest level of education attained.
- *Performance Data:* This subset focuses on the repeat loans taken by customers and the likelihood of their repayment based on historical performance. It helps in assessing whether a customer who has taken a previous loan is likely to default again.
- *Previous Loans Data:* This dataset records all past loans associated with each customer, with unique identifiers for each loan.

**Data Preprocessing and Balancing:** In this study, missing data was addressed through imputation, using mean values for numerical variables and mode values for categorical variables to ensure that all data points are filled appropriately.

The data was then normalized and transformed to ensure that all features are on a similar scale, which is important for algorithms sensitive to feature scales. Scikit-learn's pipeline was used for seamless implementation of these preprocessing steps, allowing the data transformation to be carried out in a streamlined manner. Additionally, categorical variables were converted into numerical formats using dummy variables, making them suitable for model input while retaining their original information.

**Correlation Analysis:** Correlation analysis was conducted to evaluate relationships between independent variables and the target variable (loan default). Strongly correlated features were prioritized for inclusion in model training. A heatmap was generated to visually represent the strength and direction of these correlations.

**Data Splitting:** The pre-processed loan dataset is randomly split into two parts: 70% of the data is used for training the model, while the remaining 30% is reserved for testing. By separating the training and testing sets, the study aims to prevent overfitting and ensure that the model performs well on new data.

## 4.0 Results

### *4.1 Result on Dataset Processing*

The dataset has been segmented into three distinct categories: demographic information, performance metrics, and historical borrowing records. To ensure data integrity, columns containing null entries were rigorously examined and cleansed if they failed to satisfy an established threshold for validity. This process was critical to ascertain the proportion of data points falling short of accuracy within each column. Subsequent to this refinement, Figure 5 presents a detailed classification of the remaining data types within the cleaned dataset, specifically pertaining to demographic details, performance statistics, and antecedent loan transactions, all of which are pivotal for the model's application.

```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   customerid                 3269 non-null   int32
 1   systemloanid               3269 non-null   int64
 2   loannumber                 3269 non-null   int64
 3   approveddate               3269 non-null   datetime64[ns]
 4   creationdate               3269 non-null   datetime64[ns]
 5   loanamount                 3269 non-null   float64
 6   totaldue                   3269 non-null   float64
 7   termdays                   3269 non-null   int64
 8   referredby                 3269 non-null   int32
 9   good_bad_flag              3269 non-null   int32
 10  birthdate                  3269 non-null   int32
 11  bank_account_type          3269 non-null   int32
 12  longitude_gps              3269 non-null   float64
 13  latitude_gps               3269 non-null   float64
 14  bank_name_clients          3269 non-null   int32
 15  bank_branch_clients        3269 non-null   int32
 16  employment_status_clients  3269 non-null   int32
 17  level_of_education_clients 3269 non-null   int32
 18  approved_year              3269 non-null   int64
 19  approved_month             3269 non-null   int64
 20  approved_day               3269 non-null   int64
 21  approved_dayofweek         3269 non-null   int64
 22  approved_weekofyear        3269 non-null   int64
 23  creation_year              3269 non-null   int64
 24  creation_month             3269 non-null   int64
 25  creation_day               3269 non-null   int64
 26  creation_dayofweek         3269 non-null   int64
 27  creation_weekofyear        3269 non-null   int64
 28  amount_due_ratio           3269 non-null   float64
 29  avg_loan_amount            3269 non-null   float64
 30  avg_loan_term              3269 non-null   float64
dtypes: datetime64[ns](2), float64(7), int32(9), int64(13)
memory usage: 702.3 KB
```

**Figure 3: Columns Remaining after Data Cleaning**

From the figure, the dataset has 30 distinct columns, each representing a different attribute or feature related to loans and customers. Each column has a 'non-null' count of 3269, which suggests that there are no missing values across the entire dataframe for the columns displayed. This is a positive sign, indicating that the dataset is complete and may not require further cleaning for missing values. The absence of null values implies that there won't be a need for imputation strategies typically required to handle missing data. The columns represent potential features for predictive modeling. For instance, factors such as loan amount, interest rates, and customer bank account flags might be used to predict loan default (good_bad_flag). With no null values and proper data types, the dataset exhibits high data integrity, which is conducive to reliable outcomes from data analysis or machine learning models.

***Loan Amount Distribution***

Loan amount by good or bad and Educational level by good/bad for better understanding and analysis of the data was plotted. Which shows that in terms of loan amount and educational level, the good out performed the bad

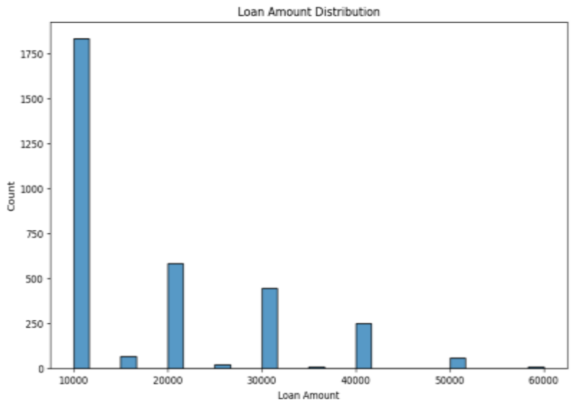


**Figure 4: Plots Showing Loan Amount Distribution**

Figure 4 shows the bar chart which plots the frequency of loans at various loan amount levels. From the chart, it is evident that the most common loan amount is in the lowest bracket shown, 10,000, which has a count significantly higher than any other amount, with over 1750 occurrences. The frequency of loans decreases as the loan amount increases, showing fewer loans distributed in the higher amounts of 20,000, 30,000, 40,000, and 50,000.

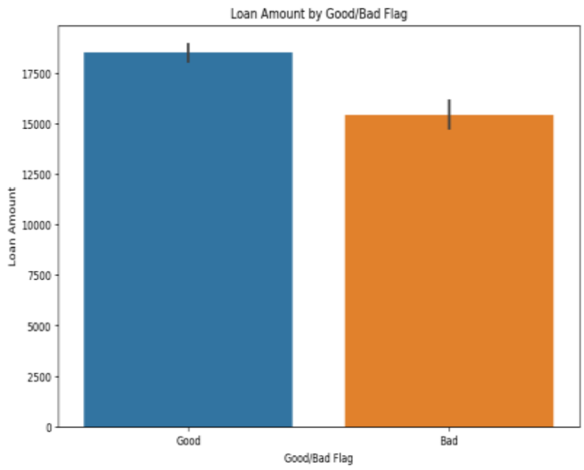


**Figure 5 Plots Showing Loan Amount by Good/Bad**

Figure 5 shows a bar chart comparing the total loan amounts categorized by the "Good/Bad Flag," which represents the creditworthiness or repayment history of the borrowers. 'Good' indicating reliable borrowers and 'Bad' indicating those who may have defaulted or are at risk. The blue bar, representing 'Good' borrowers, shows a higher total loan amount compared to the orange bar for 'Bad' borrowers. The higher total loan amount for 'Good' borrowers suggests that the lender's strategy may favour extending more credit to individuals with a positive repayment history. The chart also imply that 'Bad' borrowers are less likely to be approved for larger loans, reflecting a risk-averse lending approach. The borrowers who are classified as 'Good' may generally be more financially stable, allowing them to take out larger loans, while 'Bad' borrowers may either apply for smaller loans or be approved for less due to their credit history.

***Education Level by Good/bad***

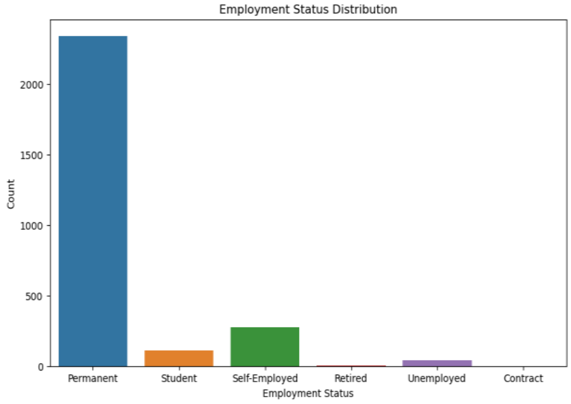


**Figure 6**: Plots Showing Educational Level By Good/Bad

Figure 6 depicts a bar chart titled which presents the count of individuals across various employment categories. These categories include Permanent, Student, Self-Employed, Retired, Unemployed, and Contract. From the chart, the 'Permanent' category has the highest count by a significant margin, indicating that the majority of individuals in this dataset are permanently employed. The counts for 'Student', 'Self-Employed', 'Retired', 'Unemployed', and 'Contract' are substantially lower, with 'Students' and 'Self-Employed' being slightly more than the other categories, but still much less compared to 'Permanent'. The high count of permanently employed individuals shows a lower credit risk for lenders, as these individuals potentially have a stable income source. Also, it reflects the lender's target market, indicating a focus on individuals with permanent employment.

***Model Building***

**1. Decision Tree**

**Table 1: Classification Report of Decision Tree**

| | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| **0** | 0.25 | 0.26 | 0.26 | 152 |
| **1** | 0.77 | 0.77 | 0.77 | 502 |
| **Accuracy** | | | 0.65 | 654 |
| **Macro avg** | 0.51 | 0.52 | 0.51 | 654 |
| **Weighed avg** | 0.65 | 0.65 | 0.65 | 654 |

The model correctly predicted 65% of the total instances. While not outstanding, this suggests moderate performance. Macro Average (Precision: 0.51, Recall: 0.52, F1-score: 0.51). Weighted Average (Precision: 0.65, Recall: 0.65, F1-score: 0.65). The decision tree model performed well with the majority class (class 1) but poorly with the minority class (class 0). The significant class imbalance (with class 1 having more than three times the instances of class 0) is likely affecting the model's ability to predict class 0 accurately.

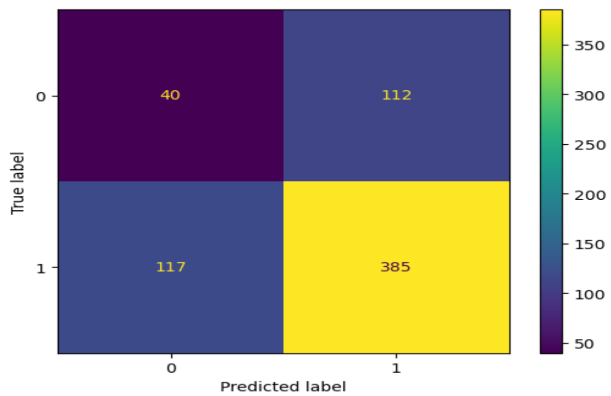


**Figure 7: Confusion Matrix of Decision Tree**

Figure 7 shows the confusion matrix of decision tree classifier, to measure the performance of a classification model. The matrix displays the actual versus predicted classifications that a model has made. True Positive (TP) shows the number 385, indicating that the model correctly predicted the positive class ('1') 385 times. True Negative (TN) with the number 40 shows that the model correctly predicted the negative class ('0') 40 times.

False Positive (FP) with the number 112, shows the instances where the model incorrectly predicted the positive class ('1') when it was actually the negative class ('0'). False Negative (FN) showing the number 117, represents the instances where the model incorrectly predicted the negative class ('0') when it was actually the positive class ('1'). The model has a higher number of true positives and true negatives than false positives and false negatives, which generally indicates a model that is performing reasonably well. However, the number of false negatives is close to the number of true negatives, which could be a concern depending on the cost or risk associated with a false negative in the specific application for this model. The relatively high number of false positives suggests that the model may be over-predicting the positive class.

### 2. Gradient Boosting Classifier

**Table 2: Classification Report of Gradient Boosting Classifier**

| | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| **0** | 0.41 | 0.06 | 0.10 | 152 |
| **1** | 0.77 | 0.97 | 0.85 | 502 |
| **Accuracy** | | | 0.76 | 654 |
| **Macro avg** | 0.59 | 0.52 | 0.48 | 654 |
| **Weighed avg** | 0.69 | 0.76 | 0.69 | 654 |

From the table, across both classes, the model accurately predicts 76% of the instances. Macro Average (Precision: 0.59, Recall: 0.52, F1-score: 0.48), Weighted Average (Precision: 0.69, Recall: 0.76, F1-score: 0.69)

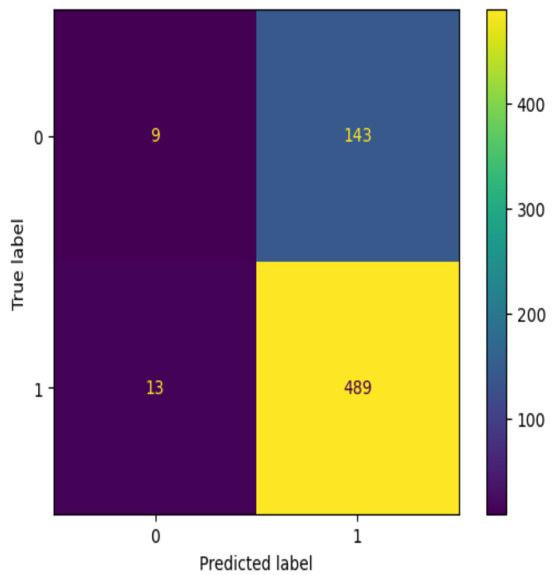


**Figure 8: Confusion Matrix of Gradient Boosting Classifier**

Figure 8 shows the confusion matrix for the Gradient Boosting Classifier. The model predicted the negative class '0' correctly 9 times (True Negatives), but it incorrectly predicted the positive class '1' as '0' on 13 occasions (False Negatives). For the positive class '1', the model predicted correctly 489 times (True Positives) and incorrectly predicted the negative class '0' as '1' 143 times (False Positives). The diagonal from the top left to the bottom right shows the correct predictions by the model, with the larger numbers indicating the model's tendency to predict class '1' correctly more often than class '0'. The small number of True Negatives compared to False Negatives suggests that the model has difficulty identifying the negative class. The high number of True Positives and low number of False Negatives for class '1' indicates that the model is much better at predicting the positive class.

### 3. Random Forest Classifier

**Table 3: Classification Report of Random Forest Classifier**

| | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| **0** | 0.41 | 0.06 | 0.10 | 152 |
| **1** | 0.77 | 0.97 | 0.85 | 502 |
| **Accuracy** | | | 0.76 | 654 |
| **Macro avg** | 0.59 | 0.52 | 0.48 | 654 |
| **Weighed avg** | 0.69 | 0.76 | 0.69 | 654 |

Table 3 shows the classification report for the Random Forest Classifier. The overall accuracy of the model is 76%, meaning it correctly predicts 76% of the time when considering both classes. Precision (0.59), Recall (0.52), F1-score (0.48):

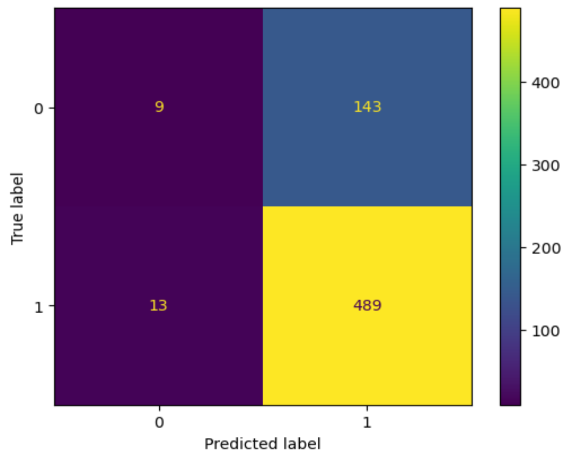


**Figure 9: Confusion Matrix of Random Forest Classifier.**

Figure 9 shows the confusion matrix for the Random Forest Classifier with the number of correct and incorrect predictions made by the model. There are 9 true negatives, indicating that the model correctly predicted the negative class '0' nine times. There are 489 true positives, where the model correctly predicted the positive class '1'. However, there are 13 false negatives, meaning the model incorrectly predicted the negative class when it was actually the positive class, and 143 false positives, where the model incorrectly predicted the positive class when it was actually the negative class. The model is substantially better at predicting the positive class than the negative class.

### 4. Gaussian Naive Beyes

**Table 4 : Classification Report of Gaussian N B Classifier**

| | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| **0** | 0.41 | 0.06 | 0.10 | 152 |
| **1** | 0.77 | 0.97 | 0.85 | 502 |
| **Accuracy** | | | 0.76 | 654 |
| **Macro avg** | 0.59 | 0.52 | 0.48 | 654 |
| **Weighed avg** | 0.69 | 0.76 | 0.69 | 654 |

**Research Design, 2024**

Table 4 presents the classification report for the Gaussian Naive Bayes (NB). The model has an overall accuracy of 76%, which indicates that it correctly predicts the class for 76% of the instances across both classes. Precision (0.59), Recall (0.52), F1-score (0.48).

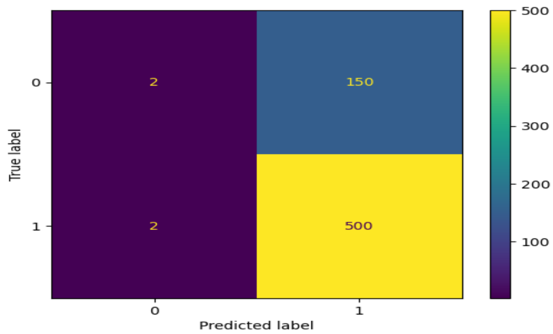


**Figure 10: Confusion Matrix of Gaussian NB Classifier**

The confusion matrix for the Gaussian Naive Bayes Classifier shows that the model has correctly predicted class 0 (True Negatives) only 2 times and class 1 (True Positives) 500 times. It has incorrectly predicted class 0 as class 1 (False Positives) 150 times, and class 1 as class 0 (False Negatives) 2 times. This suggests the model is highly effective at identifying class 1 instances but struggles significantly with class 0, failing to identify the majority of actual class 0 instances correctly. The disproportionally small number of correct predictions for class 0 indicates a possible bias towards class 1.

**Table 5 : Classification Report of the Models**

| Model | Score |
|---|---|
| GaussianNB | 0.787584 |
| Random Forest Classifier | 0.762997 |
| Gradient Boosting Classifier | 0.762997 |
| Decision Tree Classifier | 0.647615 |

The classification report in Table 5 present the overall accuracy, of four different models. The Gaussian Naive Bayes (NB) Classifier with a score of approximately 0.788, the Gaussian NB model has the highest accuracy among the four models. This suggests that, despite its simplicity, the Gaussian NB classifier is best at generalizing from the training data to the test data for this particular dataset. Random Forest Classifier has an accuracy score of around 0.763. This ensemble method, which typically performs well on a wide range of classification tasks due to its capacity for reducing overfitting, is slightly less accurate than the Gaussian NB model for this dataset.

Gradient Boosting Classifier also with a score of 0.763, it performs equivalently to the Random Forest model. Gradient Boosting is another ensemble method that focuses on learning from the errors of previous trees. Its performance being similar to the Random Forest suggests that both ensemble methods are benefiting similarly from the dataset's characteristics. Decision Tree Classifier with a score of approximately 0.648 is significantly lower than the other models. As a single decision tree, it's more prone to overfitting and generally less accurate on unseen data compared to ensemble methods. The simplest model, Gaussian NB, outperforms the more complex ensemble models in this case. This could suggest that the dataset's features have a relationship that aligns with the conditional independence assumption of Naive Bayes.

### *Discussion*

The Decision Tree Classifier demonstrated a moderate performance with an accuracy of 65%. It showed a stronger predictive capability for the majority class but struggled with the minority class, indicating a potential issue with class imbalance and overfitting [30]. Ensemble methods like Gradient Boosting and Random Forest achieved better performance, with an accuracy of 76%, but similarly faced difficulties in generalizing across classes due to class imbalance. Gaussian Naive Bayes achieved the highest accuracy at 78.8%, suggesting that its simplicity and assumption of feature independence suited this dataset well, although it also struggled with predicting the minority class. The confusion matrices provided detailed insights into actual versus predicted classifications, revealing specific areas where models faltered, particularly in terms of false positives and false negatives. This information underscored the practical implications of deploying these models, emphasizing the potential risks of misclassifications, especially in the context of loan approvals. Comparative studies have shown varying performances of machine learning models.

In one study, Decision Tree outperformed Logistic Regression and Random Forest for loan approval predictions [30]. Conversely, other research indicated Random Forest as the best-performing model among Random Forest, SVM, KNN, and Logistic Regression for loan prediction accuracy (Tumuluru et al., 2022). Another study found Logistic Regression to have the highest accuracy (83.24%) compared to Random Forest, XGBoost, and Decision Tree [31]. Similarly, a broader comparison involving eight models showed Logistic Regression achieving the highest accuracy, followed closely by Naive Bayes and Random Forest [30].

These findings highlight that while ensemble methods and more complex models offer better predictive capabilities in some cases, simpler models like Logistic Regression and Gaussian Naive Bayes can still perform robustly, depending on the dataset characteristics and preprocessing steps. The results emphasize the importance of addressing class imbalance and selecting appropriate models based on specific data requirements and operational goals.

The findings have practical implications for financial institutions aiming to deploy machine learning models for credit risk assessment. Gaussian Naive Bayes, as the top-performing model, offers a simple and computationally efficient tool for initial credit screening. Its ease of implementation means it can be integrated into existing systems with minimal overhead. However, its limitations in handling class imbalance necessitate complementary measures such as advanced preprocessing (e.g., SMOTE) or cost-sensitive learning approaches to mitigate the risk of misclassifications. These steps are crucial to reduce false negatives, which could lead to approving high-risk loans, thereby increasing financial liabilities.

Ensemble methods, such as Gradient Boosting and Random Forest, while slightly less accurate, offer advantages in handling more complex datasets. These models are particularly suitable for institutions with access to significant computational resources. When paired with explainability tools like SHAP (Shapley Additive Explanations), they can provide insights into loan decisions, aligning with regulatory and transparency requirements. For instance, explaining why a loan was rejected or approved can help build trust with applicants and regulators alike.

## 5. Conclusion

This study successfully utilised Machine Learning based approach to predict loan default. The study's comprehensive approach to data preparation and model evaluation has yielded a deep understanding of the performance characteristics of various machine learning models when applied to a rigorously cleansed dataset. This loan default prediction data, was segmented into demographic information, performance metrics, and historical borrowing records, was shown to have high integrity, which is vital for reliable machine learning applications.

The performance analysis of the models Decision Tree, Random Forest, Gradient Boosting, and Gaussian Naive Bayes revealed that while each model has strengths, they also exhibit significant limitations, particularly in handling class imbalance. The Decision Tree model demonstrated moderate accuracy and highlighted potential overfitting issues, as it performed significantly better on the majority class. Similarly, both ensemble methods, Random Forest and Gradient Boosting, although robust with a higher accuracy of 76%, struggled with the minority class, suggesting that even sophisticated models can falter without strategies to address class imbalance. The Gaussian Naive Bayes model emerged as the top performer with the highest accuracy, suggesting that its underlying assumptions might be particularly well-suited to the dataset’s features. However, like the other models, it also showed a bias towards the majority class, indicating a common challenge across the board.

Further research could explore creating hybrid ensemble models by combining different machine learning algorithms, such as decision trees with neural networks or boosting methods. Investigating how these models can complement each other in handling imbalanced datasets and optimizing performance could yield significant improvements in predictive accuracy and reliability.

## References


[1] Hency, J. A., & Padmajavalli, R. (2016). Comparative Analysis of Classification Algorithms on Endometrial Cancer Data. *Indian Journal of Science and Technology, 9*(28), 205-211. doi: 10.17485/ijst/2016/v9i28/93846

[2] Aniceto, M.C., Barboza, F., & Kimura, H. (2020). Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal*, 6(1), 1-4.

[3] Egwa, A.A., Bello, H., Ahmad, A.A., & Bizi, M.S. (2022). Default prediction for loan lenders using machine learning algorithms. *SLU Journal of Science and Technology*, 5(1&2), 1-2.

[4] Isa, F., & Isa, R. (2021). Treatment of toxic asset by deposit money banks in Nigeria: A review of literature. *TSU-International Journal of Accounting and Finance*, 1(1), 42-50.

[5] Ahmed, A.I., & Rajaleximi, P.R. (2019). An empirical study on credit scoring and credit scorecard for financial institutions. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 8(7), 2278-1323

[6] Moradi, S., & Rafiei, F.M. (2019). A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. *Financial Innovation*, 5(1), 1-27

[7] Nilsson, M., & Shan, Q. (2018). Credit risk analysis with machine learning techniques in peer-to-peer lending market. *Stockholm Business School Master's Degree Thesis Master's Programme in Banking and Finance.*

[8] Alaka, H.A., Oyedele, L.O., Owolabi, H.A., Kumar, V., Ajayi, S.O., Akinade, O.O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, 94, 164-184

[9] Leal, A. A. (2024). Algorithms, Creditworthiness, and Lending Decisions Check for updates. *Legal Aspects of Autonomous Systems: A Comparative Approach*, *4*, 321

[10] Mienye, I. D., & Sun, Y. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, *25*, 100690.

[11] Begum, C., & Deniz, U. (2019). Comparison of data mining classification algorithms: Determining the default risk. *Research Article, Hindawi Scientific Programming.*

[12] Varshney, K.R. (2019). Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3), 26-29

[13] Jamali, A.A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R., & Ebrahimie, E. (2016). DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug Discovery Today*, 21(5), 718-724

[14] Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, 273-282

[15] https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

[16] Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17, 16-2

[17] Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2023). Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*, *330*(1), 609-637.

[18] Bello, O. A. (2023). Machine learning algorithms for credit risk assessment: an economic and financial analysis. *International Journal of Management*, *10*(1), 109-133.

[19] Sheikh, M. A., Goel, A. K., & Kumar, T. (2020, July). An approach for prediction of loan approval using machine learning algorithm. In *2020 international conference on electronics and sustainable communication systems (ICESC)* (pp. 490-494). IEEE

[20] Nandipati, V. S. S., & Boddala, L. V. (2024). Credit Card Approval Prediction: A comparative analysis between Logistic Regression, KNN, Decision Trees, Random Forest, XGBoost.

[21] Thakar, S., Patel, D., Gandhi, V. C., & Trivedi, D. (2024, October). Financial Analytics with Artificial Neural Networks: Predicting Loan Repayment. In *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 1666-1073). IEEE.

[22] Saini, P. S., Bhatnagar, A., & Rani, L. (2023, May). Loan approval prediction using machine learning: A comparative analysis of classification algorithms. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1821-1826). IEEE.

[23] Akça, M. F., & Sevli, O. (2022). Predicting acceptance of the bank loan offers by using support vector machines. *International Advanced Researches and Engineering Journal*, *6*(2), 142-147.

[24] Babo, S. B., & Beyene, A. M. (2023). Bank Loan Classification of Imbalanced Dataset Using Machine Learning Approach.

[25] Patel, V., & Bhavsar, H. (2023, March). Detailed Performance Study of Data Balancing Techniques for Skew Dataset Classification. In *Congress on Control, Robotics, and Mechatronics* (pp. 187-201). Singapore: Springer Nature Singapore.

[26] Archana, S., & Divyalakshmi, K. S. A. (2023). Comparison of Various Machine Learning Algorithms and Deep Learning Algorithms for Prediction of Loan Eligibility. *Ijraset Journal For Research in Applied Science and Engineering Technology*, *11*(VI).

[27] Sujatha, C. N., Gudipalli, A., Pushyami, B., Karthik, N., & Sanjana, B. N. (2021, November). Loan prediction using machine learning and its deployement on web application. In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-7). IEEE.

[28] Abdullah, S., Othman, Z., & Mohamad, R. (2023). Predicting the Risk of SME Loan Repayment using AI Technology-Machine Learning Techniques: A Perspective of Malaysian Financing Institutions. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *31*(2), 320-326

[29] https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data

[30] Tejaswini, J., Kavya, T.M., Ramya, R.D., Triveni, P.S., & Maddumala, V.R. (2020). Accurate loan approval prediction based on machine learning approach. *Journal of Engineering Science*, 11(4), 523-32

[31] Prathap, K.R., & Bhavani, R. (2023). Study comparing classification algorithms for loan approval predictability (Logistic Regression, XG boost, Random Forest, Decision Tree). *Journal of Survey in Fisheries Sciences*, 10(1S), 2438-47