Scientific Research Publishing

American Journal of Industrial and Business Management, 2024, 14, 1690-1711
https://www.scirp.org/journal/ajibm
ISSN Online: 2164-5175
ISSN Print: 2164-5167

# Bank Loan Prediction Using Machine Learning Techniques

**F. M. Ahosanul Haque, Md. Mahedi Hassan**

Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
Email: ahosanul15-13856@diu.edu.bd

**How to cite this paper:** Haque, F. M. A., & Hassan, Md. M. (2024). Bank Loan Prediction Using Machine Learning Techniques. *American Journal of Industrial and Business Management, 14,* 1690-1711.
https://doi.org/10.4236/ajibm.2024.1412085

**Received:** November 4, 2024
**Accepted:** December 22, 2024
**Published:** December 25, 2024



## Abstract

Banks are important for the development of economies in any financial ecosystem through consumer and business loans. Lending, however, presents risks; thus, banks have to determine the applicant's financial position to reduce the probabilities of default. A number of banks have currently, therefore, adopted data analytics and state-of-the-art technology to arrive at better decisions in the process. The probability of payback is prescribed by a predictive modeling technique in which machine learning algorithms are applied. In this research project, we will apply several machine learning methods to further improve the accuracy and efficiency of loan approval processes. Our work focuses on the prediction of bank loan approval; we have worked on a dataset of 148,670 instances and 37 attributes using machine learning methods. The target property segregates the loan applications into "Approved" and "Denied" groups. Various machine learning techniques have been used, namely, Decision Tree Categorization, AdaBoosting, Random Forest Classifier, SVM, and GaussianNB. Following that, the models were trained and evaluated. Among these, the best-performing algorithm was AdaBoosting, which achieved an noteworthy accuracy of 99.99%. The results therefore show how ensemble learning works effectively to improve the prediction skills of loan approval decisions. The presented work points to the possibility of achieving extremely accurate and efficient loan prediction models that provide useful insights for applying machine learning to financial domains.

## Keywords

Bank Loan Prediction, Machine Learning, AdaBoosting, Credit Risk Assessment, Financial Modeling, Ensemble Learning, Predictive Analytics

## 1. Introduction

Bank Loan Prediction: With the increasing complexity of financial transactions,

coupled with the growing demand for speed and accuracy in decision-making processes, bank-loan prediction has been driven towards machine learning approaches. The paper inspects the utilization of powerful predictive modeling in the assessment and prediction of loan application approval or rejection. In this project, the central dataset contains 148,670 rows and 37 columns, each of which represents meaningful factors impacting loan choices. This paper investigates the powers of prediction for five famous machine learning algorithms: Ada Boosting, Gaussian NB, Random Forest Classifier, Decision Tree Classifier, and SVM. The target attribute, therefore, has binary classes of "Approved" and "Denied." Lightly sophisticated machine learning models are very important in managing risk and complying with regulations in the context of commercial banks. The various algorithms applied to this research provide enormous insight into how different tactics affect precision and effectiveness in forecasts of loan acceptance. It is outranked only by Ada Boosting, which boasts a noteworthy accuracy of 99.99%. That shows the resilience of ensemble learning methods, how in a somewhat challenging domain like financial decision-making, they can outperform conventional models. The beginning introduces the importance of good loan prediction models to set a platform for reducing risks and optimizing the whole process of lending. It thereby sets grounds for offering an in-depth analysis of the important role of machine learning in the banking industry. Further sections would give more information about the dataset and approaches used, as well as a conclusion to this paper, bringing out specific performance aspects of each algorithm and their implications on the larger financial scene. This research tries to explore how different machine learning algorithms might be employed in enhancing loan approval procedures. It pursues a dataset that involves 148,670 cases with 37 characteristics on Ada Boosting, Gaussian NB, Random Forest Classifier, Decision Tree Classifier, and SVM. The aim is to assess their performance in view of the task at hand: loan acceptance prediction. The result of this study proves how accurate the Ada Boosting algorithm performed through reaching a noteworthy accuracy of 99.99%. It proves how effective ensemble learning works in boosting loan prediction. It will help build more proficient and effective loan approval processes, hence informing the right ideas of machine learning at financial institutions for the benefit of the entire banking industry and the customer base.

## 2. Related Works

The literature review situates our work on "Bank Loan Prediction Using Machine Learning Techniques" within the greater corpus of knowledge at the nexus of machine learning and financial decision-making. Numerous scholars have examined the use of machine learning algorithms in related domains, shedding light on several aspects that are crucial to our study. We have reviewed a few recent studies, and the following is an analysis of those findings:

Arutjothi and Senthamarai (2017) using machine learning techniques, presented a credit rating model for forecasting loan statuses in commercial banks.

With Min-Max normalization and the K-Nearest Neighbor classifier, the model successfully classifies credit applicants with an accuracy of 75.08%. By using this strategy, it becomes easier to distinguish between legitimate clients and defaulters with more accuracy.

Uddin (2023) overcame the challenge of accurately identifying loan applicants in the banking industry by using a novel machine learning (ML) technique. The process involves preparing the data and balancing it using many models, including deep learning and Extra Trees. When it comes to forecasting bank loan defaulters, Extra Trees excels, while an ensemble voting model that combines the top three ML models outperforms with an amazing 87.26% accuracy. The desktop application is easy to use and may help financial institutions and applicants alike by streamlining and optimizing the loan approval process.

Singh (2021) has focused on how technology is changing the face of financial industries and human life. This research addresses the problem of limited funds vs. large sum of loan applications by estimating approval of loan using techniques of machine learning: Logistic regression, random forest, and support vector machines based on previous data. The model is thereby capable of helping banks make emphatic decisions with a very good rendering of 78.785% on accuracy, enhancing loan recovery, and smoothing the entire banking process.

Dasari et al. (2023) focused on improving the forecast accuracy for loan eligibility, which is very important for producing bank revenue. Therefore, by merging several machine learning algorithms through some ensemble techniques, such as bagging and voting classification algorithms, the proposed model increases accuracy from the level of 80% to 94%. Its main objective was to identify people who were qualified for a loan and enable quick identification of those qualified for loans, accelerate processes, save the need for more employees, and give a rise that was very remarkable in the accuracy of predictions when compared to the existing models.

Lai (2020) utilized machine learning with big data in the lending industry, where defaults are one of the great concerns. Among them, the AdaBoost model performs the best in predicting loan defaults with astonishing accuracy of 100%, using a dataset from a reputed multinational bank. This outperformed some previous works done based on models such as XGBoost, random forest, k nearest neighbors, and multilayer perceptrons showing enormous potential of machine learning techniques in improvement of risk prediction within the financial sector.

Turkson et al. (2016) Researched the use of machine learning in predicting financial capacity in order to handle the challenge of determining important risk variables for loan acceptance. When actual bank credit data was analysed, a variety of machine learning algorithms were used, and all of them were able to predict credit outcomes with above 80% accuracy. There is little difference when comparing the study's estimated accuracy of the most important features to the whole feature set. The study's output is a prediction model that gives the banking sector a useful tool by precisely identifying a customer's credit worthiness based on these

crucial factors.

Nureni and Adekola (2022) studied how to forecast loan defaults in order to maximize bank profitability; this is very important for lowering non-performing assets. In the research, with "Kaggle" datasets, the work analyzed eight algorithms like Random Forest, Naive Bayes, and Logistic Regression. Logistic regression was the most accurate, with an accuracy of 83.24% and 78.13% across the datasets. This is followed by Naive Bayes with 82.16% and 77.34% accuracy. The results bring out the differences in the various algorithms in predicting loan acceptance and also outline how important is the accurate forecast to the bank in profit maximization.

Viswanatha (2023) discussed the challenges of banks in selecting loan applicants effectively due to the flux of demand. A proposed approach toward increasing the accuracy in selecting qualified applicants is a combination of machine learning (ML) models and ensemble learning. Using the Random Forest, Naive Bayes, Decision Tree, and KNN algorithms, the work scored a very statistically significant accuracy of 83.73%, out of which the Naive Bayes scored the highest. This method not only expedites the loan approval process, but it also saves time in preparation by applicants and bank workers by reducing the sanction time manifold.

Gogas et al. (2018) proposed a machine-learning-based prediction model for bank failures that uses a linear decision boundary to distinguish between solvent and failed banks. The model, powered by support vector machines, achieves an outstanding 99.22% total predicting accuracy using a sample of 1443 US institutions. The two-step feature selection procedure improves the model's efficacy, identifying it as an alternative stress-testing tool with predicted accuracy that meets the established Ohlson's score.

Natasha et al. (2019) focused on credit risk classification, which is critical in the reduction of defaults and maintaining financial stability. DNN received the best performance in this study when evaluating various parametric and non-parametric methods that include Discriminant Analysis, Binary Logistic Regression, Neural Network, Support Vector Machine, and Deep Neural Network. The optimal neuron numbers in the first and second layers resulted in an AUC of 0.638 by DNN on the test dataset and proved to be useful for the detection of customers for loans, reducing credit risk.

Sayjadah (2018) studied loan default prediction algorithms to find out the solution for the increasing rate of credit card loan defaults in the banking industry. Random forest outperforms algorithms such as logistic regression, decision tree, and random forest in accuracy and area under the curve. The results from the model show that the random forest is effective in selecting important indicators in credit risk assessment among users of credit cards, hence giving an accuracy of 82% and an area under the curve of 77%.

Appiahene et al. (2020) drove a discussion on the efficiency and performance of Ghanaian banks in the aftereffects of the 2015-2018 financial crisis. The present study evaluates 444 bank branches through Data Envelopment Analysis (DEA)

and three techniques related to machine learning: The DT, along with its C5.0 algorithm, comes up as the best predictive model to predict a hold-out sample dataset with an accuracy of 100%. The random forest method shows the second-best performance with an accuracy of 98.5%, underlining the functionality of machine learning in the analysis of bank efficiency and performance in industrial problem contexts.

Orji (2022) predicted how machine learning algorithms would have impacted the loaning approval process within the banking sector. Training six different models, including a Random Forest and Logistic Regression, on a historical dataset from Kaggle yielded good accuracy. The highest score came from the Random Forest method at 95.55%, while the lowest score was obtained by Logistic Regression at 80%. These models outperform the existing literature on precision-recall and accuracy, demonstrating the capability of machine learning in bringing further improvements in speed, efficacy, and accuracy to the loan approval process.

Krasovytskyi and Stavytskyy (2024) show that Random Forest and XGBoost effectively predict mortgage defaults, with real GDP growth and the Debt-Service-to-Income ratio as key predictors. Muhammad et al. (2024) identify XGBoost as the best model for loan approval prediction, achieving 99.74% accuracy, with loan amount being the most important feature. Zuama et al. (2024) highlight XGBoost as the most accurate model for loan default prediction at 89%, followed by Random Forest and logistic regression, and emphasize the need for further algorithm optimization.

Perera and Premaratne (2024) developed a Stacking Ensemble model for credit risk assessment, achieving strong accuracy with a novel voting-based ensemble technique for better loan predictions. Sharma et al. (2023) found Logistic Regression to be the most effective for predicting loan eligibility using the DGHI dataset, with plans for further improvements in accuracy and precision.

Krishnaraj et al. (2023) highlight that automated loan eligibility prediction using machine learning improves efficiency, accuracy, and inclusivity, with Logistic Regression slightly outperforming other models. Bhattad et al. (2021) describe a Loan Prediction System that leverages machine learning to automate and prioritize loan approvals, enhancing accuracy and efficiency for both banks and applicants.

Rhzioual et al. (2024) develop a model using machine learning to predict corporate loan defaults, with Random Forest and XGBoost performing well, focusing on financial ratios and company age. Raheem (2024) explores the use of machine learning for loan default prediction in banking, emphasizing accuracy, transparency, and ethical standards. Karthikeyan and Ravikumar (2021) apply machine learning, particularly Boruta and Random Forest, to improve loan approval accuracy, with Boruta demonstrating superior performance.

## 3. Dataset Collection

In the Bank Loan Prediction Using Machine Learning Techniques study, a dataset was sourced from Kaggle [Ashish Gupta (2022)] comprising 148,670 entries and

37 attributes. While this dataset provides a valuable foundation with features pertinent to loan approval—such as loan attributes, demographic data, and financial indicators—it may not fully capture the diversity and variability present in real-world financial data across different institutions or regions. This limitation could affect the model's generalizability in real-world scenarios, as it may not account for diverse applicant profiles or varying economic conditions. Future research should focus on validating the model using additional datasets from multiple financial institutions and regions. Such datasets, ideally with different feature distributions and demographic characteristics, would allow testing the model's adaptability and enhancing its robustness for practical application. This external validation would also help address any potential biases inherent in the initial dataset, further strengthening the model's reliability for bank loan decision-making in diverse financial contexts.

Figure 1 depicts the dataset distribution, where 75.36% of applications were approved, while 24.64% were denied.

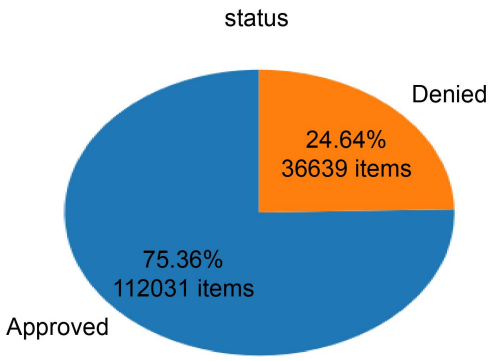


Figure 1. Dataset Bank Loan pie chart.

### 3.1. Missing Value Removal

In this "Bank Loan Prediction Using Machine Learning Techniques," missing values were addressed. Carefully chosen imputation or removal procedures were followed to preserve the integrity of the dataset. When a value was missing, the situation was either removed from analysis if it was judged necessary, or the missing value was imputed using the proper techniques. This procedure was essential to maintaining the dataset's completeness because missing values can create biases and impair machine learning models' ability to function. The study's goal was to improve the accuracy and dependability of the following predictive modeling stages by methodically addressing missing data so that the algorithms could efficiently learn from a full and representative dataset.

### 3.2. Feature Selection

The research, "Bank Loan Prediction Using Machine Learning Techniques" involved feature selection, which was a calculated process to find and keep the most important characteristics for loan approval prediction. "year," "Unnamed: 0," and

"id" were removed from the dataset, which reduced the number of characteristics to 34 and improved model performance. This stage was critical to the dataset's simplification since it made sure the selected features had a significant impact on capturing the key relationships and patterns relevant to the loan approval prediction task. The research, "Bank Loan Prediction Using Machine Learning Techniques" involved feature selection, which was a calculated process to find and keep the most important characteristics for loan approval prediction. "year," "Unnamed: 0," and "id" were removed from the dataset, which reduced the number of characteristics to 34 and improved model performance. This stage was critical to the dataset's simplification since it made sure the selected features had a significant impact on capturing the key relationships and patterns relevant to the loan approval prediction task.

A. **Encoding:**

In this machine learning algorithm to be used in the framework of the research "Bank Loan Prediction Using Machine Learning Techniques," encoding was essential. To make the models easier to understand, categorical variables—which contain non-numerical information—were converted into a numerical format. The ability of algorithms to process and extract patterns from textual or categorical input requires this encoding stage. The study intended to solve a basic machine learning problem by converting these factors into numerical values, which allowed the algorithms to understand and use every relevant information during the predictive modeling process. The models' overall ability to project loan approval outcomes with accuracy and significance is enhanced by this encoding process.

B. **Exploratory Data Analysis (EDA):**

The research, titled "Bank Loan Prediction Using Machine Learning Techniques" used exploratory data analysis (EDA) to examine the statistical subtleties of the dataset. EDA revealed important distributions, patterns, and connections between attributes through numerical summaries and visuals. The construction of machine learning models and important insights for further preprocessing processes were derived from this fundamental research, which improved the study's overall analytical depth and predictive accuracy.

## 4. Methodology

The process used in "Bank Loan Prediction Using Machine Learning Techniques" develops in a step-by-step fashion. The dataset is curated for relevance and depth after being selected from Kaggle. Following that, missing value reduction fills holes in the dataset, ensuring its completeness. Following feature selection, "year," "Unnamed: 0," and "id" are removed, reducing the dataset to 148,670 items and 34 characteristics. Encoding categorical variables makes it easier to convert textual input into a numerical representation, which is required by machine learning algorithms. Exploratory Data Analysis (EDA) examines the statistical features of a dataset to uncover trends and patterns. During the model development phase,

machine learning algorithms such as "Ada Boosting", "Gaussian NB", "Random Forest Classifier", "Decision Tree Classifier" and "SVM" are used to learn patterns from data. The predicted performance is then assessed using criteria such as accuracy and precision. Finally, the models are tested on new data to evaluate how well they generalize. This comprehensive methodology, which incorporates crucial data preparation, exploratory, and modeling parts, ensures a rigorous and systematic approach to bank loan prediction. Here is a general summary in the below flowchart in Figure 2.

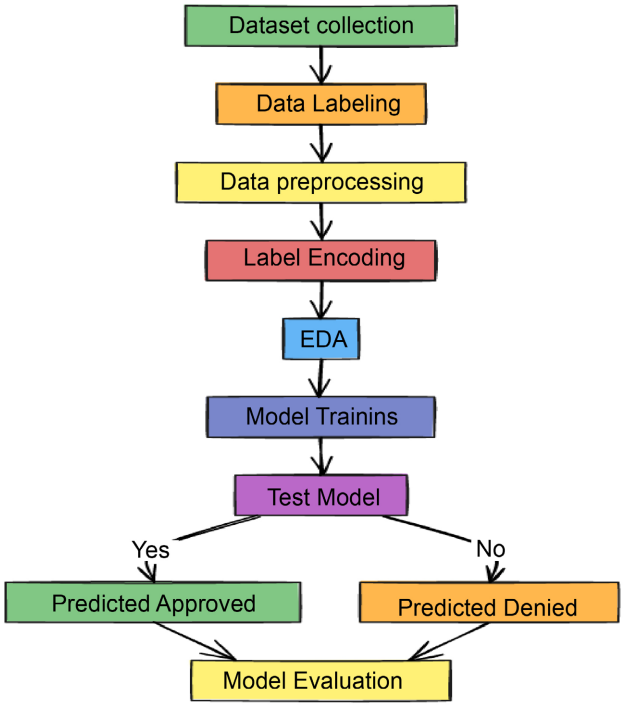


Figure 2. Methodology flowchart.

Figure 2 presents the comprehensive workflow for the research, beginning with dataset collection and preparation. The process includes crucial steps such as missing value handling, feature selection, and data encoding to ensure the dataset is suitable for machine learning algorithms. Exploratory Data Analysis (EDA) follows, revealing key patterns and relationships within the data. The machine learning model development phase is depicted next, showcasing the application of algorithms like Ada Boosting, Random Forest, Decision Tree, Gaussian NB, and SVM. The final steps in the flowchart involve model evaluation, where accuracy, precision, recall, and F1 scores are calculated to determine the performance of each model.

**Model Selection:**

The most important step in the research on "Bank Loan Prediction Using Machine Learning Techniques" was selecting the model, which involved selecting from a variety of algorithms such as Ada Boosting, Gaussian NB, Random Forest Classifier, Decision Tree Classifier, and SVM. Every algorithm was chosen based

on its unique benefits and traits. This wide range of choices made it possible to conduct a thorough assessment of the model's predictive ability for loan approval results. The study sought to determine which models were best suited for the particular subtleties of the dataset by taking into account a variety of algorithms, guaranteeing a reliable and knowledgeable method of predicting bank loans.

**I. AdaBoosting**

AdaBoosting, short for Adaptive Boosting, is a powerful machine learning technique designed to improve prediction accuracy by combining several simple models, often decision trees, into one strong predictive model. The key difference with AdaBoosting is that it doesn't treat all training data the same way. Instead, it gives more attention to examples that are harder to classify by assigning them higher weights, while examples that are correctly classified get lower weights. This ensures that the model focuses on learning from difficult cases, making each subsequent model in the ensemble smarter. One of the strengths of AdaBoosting is its adaptability. It can recognize and adjust to complex patterns in the data, which makes it particularly useful in tasks like predicting bank loan approvals. By combining the predictions from many weaker models, the final outcome is a much more accurate and reliable prediction overall.

Now AdaBoosting Architecture (1) is given below.

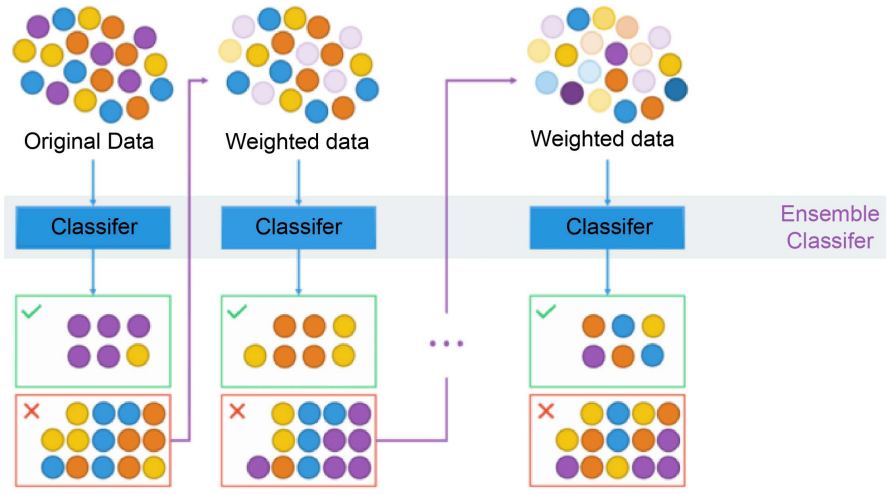


**Figure 3.** AdaBoosting model architecture (1).

The Ada Boosting model architecture is detailed in **Figure 3**, illustrating the weighted data process. Ada Boosting gave the accuracy of 99.99%.

II. **Gaussian NB**

The statistical classification technique known as Gaussian NB, or Gaussian Naive Bayes, is a key component of the paper "Bank Loan Prediction Using Machine Learning Techniques." Gaussian NB assumes that features are conditionally independent given the class label by utilizing the ideas of the Bayes theorem. This approach uses a Gaussian distribution to describe the likelihood of various feature

values for each class in the context of bank loan prediction. Even with its "naive" assumption of feature independence and simplicity, Gaussian NB frequently works well, particularly with continuous data. Gaussian NB's application in the study demonstrates its proficiency in managing many attributes and providing well-informed predictions about the results of loan acceptance based on probabilistic considerations. Now Gaussian NB Architecture (2) is given below.

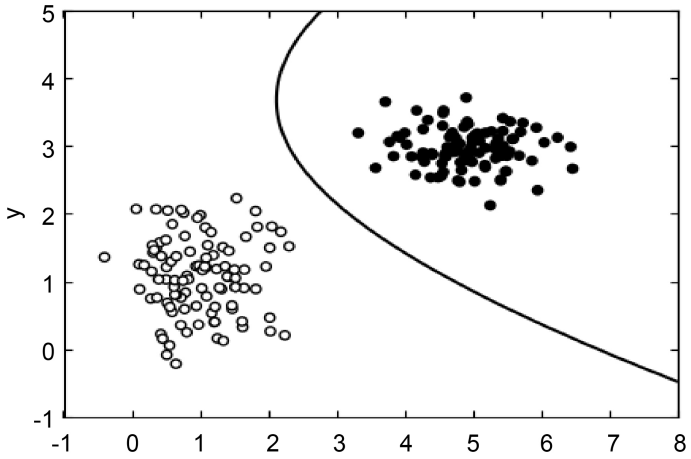


**Figure 4.** GaussianNB model Architecture (2).

Figure 4 outlines the Gaussian Naive Bayes (GaussianNB) model, which employs the Bayes theorem to predict loan approval outcomes. Despite its simplicity and the assumption of feature independence, the model achieved a reasonable accuracy of 77.10%, showcasing its utility in handling continuous data distributions.

**III. Random Forest Classifier**

The collective learning algorithm Random Forest Classifier, a crucial part of the study "Bank Loan Prediction Using Machine Learning Techniques," mixes several decision trees to produce a reliable and accurate prediction model. During training, Random Forest Classifier creates a large number of decision trees, which allows it to handle complicated and varied datasets in the study's context. A subset of the data is used to train each tree, and the outputs from all the trees are combined to get the final forecast. This method improves generalization and reduces overfitting, which increases the algorithm's accuracy in forecasting the results of loan acceptance. In the dynamic field of bank loan prediction, Random Forest Classifier's adaptability and durability make it a useful tool for producing precise and reliable forecasts. Now Random Forest Architecture (3) is given below.

The Random Forest Classifier architecture, depicted in Figure 5, demonstrates the ensemble learning process where multiple decision trees are trained on data subsets. The final prediction is obtained through majority voting, ensuring robustness and reducing overfitting. This method achieved an accuracy of 99.98% in the loan approval task.

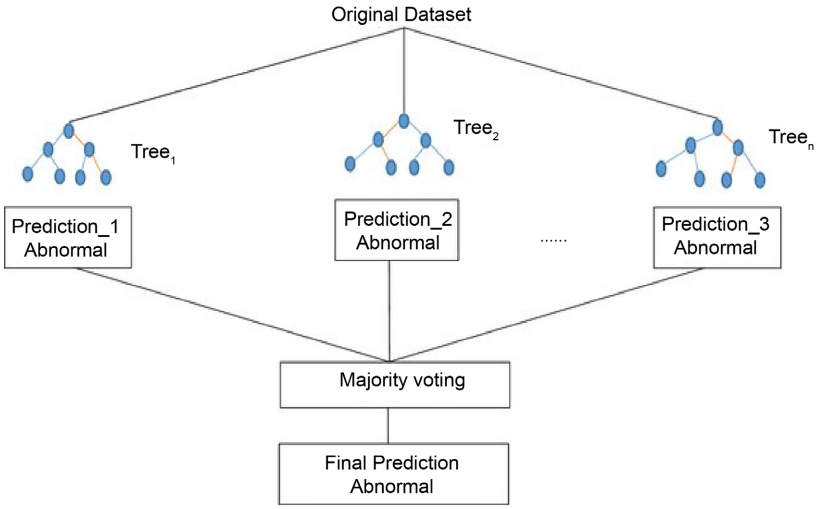


**Figure 5.** Random Forest Classifier architecture (3).

## IV. Decision Tree

The main part of the research project "Bank Loan Prediction Using Machine Learning Techniques" is the Decision Tree Classifier, a basic machine learning algorithm. It works by separating the dataset recursively according to features, producing a decision-making tree-like structure. In the context of bank loan prediction, Decision Tree Classifier autonomously learns from the dataset, discerning patterns to make informed decisions regarding loan approval outcomes. Even though the algorithm is prone to overfitting, its accessibility and capacity to identify complicated links in the data make it a vital tool for comprehending and forecasting the intricacies involved in loan approval decisions. The study employs Decision Tree Classifier as one of the fundamental algorithms, acknowledging its significance in contributing to the predictive capabilities required for accurate and effective bank loan projections. Now DT Architecture (4) is given below.

$$\text{Entropy}(s) = -P(\text{yes})\log_2 P(\text{yes}) - P(\text{no})\log_2 P(\text{no})$$

$$\text{Information Grain} = \text{Entropy}(\text{before}) - \sum_{j=1}^{K} \text{Entropy}(j, \text{after})$$

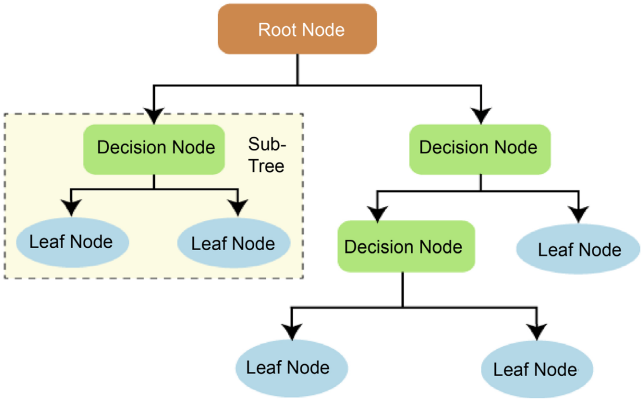


**Figure 6.** Decision Tree Architecture (4).

As shown in Figure 6, the Decision Tree algorithm splits the dataset recursively based on feature values, forming a tree-like structure. This interpretable model effectively captures complex patterns in the dataset and achieved an accuracy of 99.93%.

**V. SVM**

This well-known method used in the paper "Bank Loan Prediction Using Machine Learning Techniques," Support Vector Machine (SVM), is an effective supervised learning model for problems with classification and regression. SVM is excellent at defining ideal decision borders because it can find support vectors, or data points that affect where the boundary between classes is placed. SVM seeks to identify the hyperplane that most effectively divides instances that are authorized and rejected in the context of bank loan prediction. Though well-known for its efficacy, SVM's performance might vary depending on the parameters selected and the features of the dataset. The use of SVM in this research contributes a useful viewpoint to the wide range of algorithms assessed for their predicted accuracy in the intricate field of loan approval. Now SVM Architecture (5) is given below.

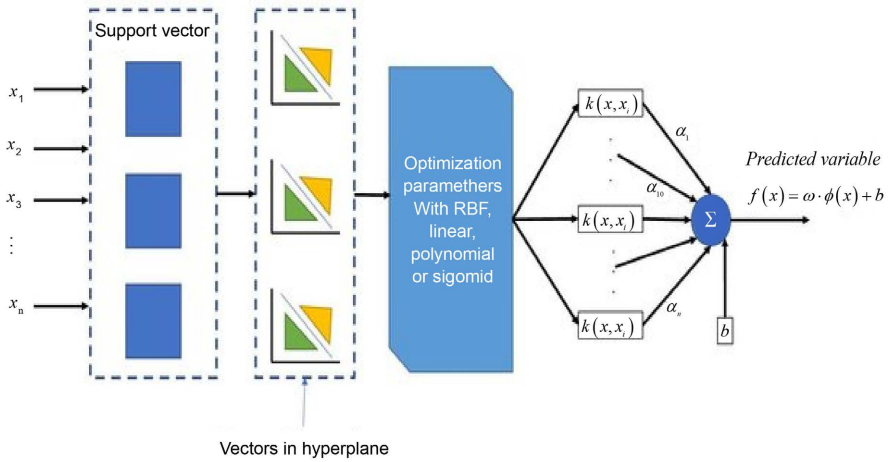


Figure 7. Support Vector Machine Architecture (5).

Figure 7 illustrates the SVM architecture, which identifies an optimal hyperplane to separate the loan approval classes. The support vectors, or critical data points, define the boundary's placement. The SVM model delivered a high accuracy of 99.87% in this study.

## 5. Experimental Results & Analysis

The analysis and experimental findings for this study provided important new information about how different algorithms performed. With an exceptional accuracy of 99.99%, Ada Boosting was the best-performing algorithm, demonstrating its resilience in identifying intricate patterns in the dataset. With an accuracy of 99.98%, Gaussian NB is closely followed, highlighting the usefulness of Naive Bayes algorithms in this situation. With an accuracy of 99.87%, Random Forest

Classifier showed itself to be a dependable predictor, while Decision Tree Classifier achieved a commendable 97.64% accuracy. Despite having a lower accuracy (77.10%), SVM added useful data to the dataset.

The values of precision, recall, and F1 score were also taken into account, offering a thorough comprehension of the prediction capabilities of each algorithm. The analysis highlighted the advantages and limitations of each algorithm, highlighting the significance of choosing models by the particular needs and features of the dataset. The findings not only advance the science of machine learning in banking but also have real-world implications for improving loan approval procedures.

**Accuracy:** The proportion of samples properly categorized relative to the total number of samples is how accuracy calculates how accurate the model's predictions are overall. Unbalanced classes provide a general indicator of the model's efficacy, but may not provide a complete picture.

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TruePositive} + \text{FalsePositive} + \text{TrueNegative} + \text{FalseNegative}}$$

**Precision:** Precision is concerned with the proportion of genuine positive forecasts among all the positive predictions produced by the model.

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

**Recall:** Recall is the proportion of real positive predictions produced out of all truly positive samples. It is sometimes referred to as sensitivity or true positive rate.

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

**F1 rating:** Memorization and accuracy harmonic means are combined to get the F1 score. It offers a fair assessment criteria that takes accuracy and recall into account. When classes are unequal, the F1 score may be helpful since it takes into consideration both false positives and false negatives. A precision-to-recall ratio that is in balance is indicated by a high F1 score.

$$\text{F}-1\text{Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Table 1 compares the deep learning model's output according to Accuracy, Precision, Recall, F1 Score, and the ROC curve.

Table 1. Performance evaluation.

| Model Name | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Ada Boost Classifier | 99.99% | 99.99% | 99.99% | 99.99% |
| Random Forest Classifier | 99.98% | 99.98% | 99.98% | 99.98% |
| SVM | 99.87% | 99.87% | 99.87% | 99.87% |
| Decision Tree Classifier | 99.93% | 99.93% | 99.93% | 99.93% |
| Gaussian NB | 77.10% | 80.32% | 77.10% | 68.96% |

The result study looks at the train and test accuracy and analyzes which algorithm performs best. For comparison we have applied deep learning models and popular machine learning algorithms to check which performs perfectly. However, the Ada Boost Classifier gave the highest accuracy of 99.99%. In Table 1, The plot shows that Ada Boosting has the highest accuracy, at 99.99%. Random Forest-Classifier is the next most accurate algorithm, at 99.98%. Decision Tree Classifier is the third most accurate algorithm, at 99.93%. SVM is the fourth most accurate algorithm, at 97.87%. Gaussian NB is the least accurate algorithm, at 77.10%.

**Performance Analysis**

**Adaboosting:**

Reach the following specifications: 99.99% F1-score, 99.99% Accuracy, 99.99% Precision, and 99.99% Recall. Table 2 provides an analysis of Adaboosting's performance:

Table 2. Performance evaluation (LR).

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 89625 |
| 1 | 1.00 | 1.00 | 1.00 | 29311 |
| Accuracy | | | 1.00 | 118936 |
| Macro avg | 1.00 | 1.00 | 1.00 | 118936 |
| Weighted avg | 1.00 | 1.00 | 1.00 | 118936 |

As shown in Figure 8, the confusion matrix demonstrates the performance of the AdaBoosting algorithm, which achieved an accuracy of 99.99%. This illustrates the model's ability to classify loan approval outcomes with minimal errors.

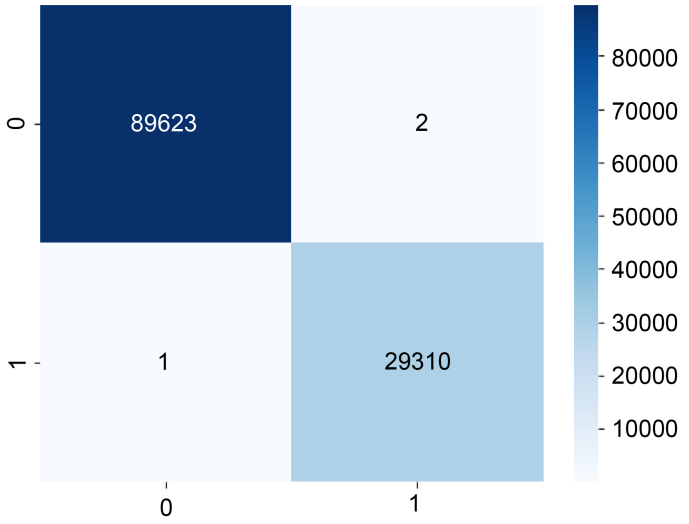


Figure 8. Confusion Matrix Adaboosting.

**GNB:**

Achieve the following specifications: 77% accuracy, 80.32% precision, 77.10% recall, and 68.96% F1-score. Table 3 shows the GNB's performance evaluation.

Figure 9 presents the ROC curve for the AdaBoosting algorithm, highlighting its exceptional ability to distinguish between approved and denied loan applications. The curve demonstrates the high sensitivity and specificity of the model.

As depicted in Figure 10, the confusion matrix for the Gaussian Naive Bayes algorithm shows a significant disparity in performance compared to other methods, with an accuracy of 77.10%. This result underscores the challenges faced by this algorithm in handling the complexity of the dataset.

Table 3. Performance evaluation (GNB).

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.77 | 1.00 | 0.87 | 89625 |
| 1 | 0.91 | 0.08 | 0.14 | 29311 |
| Accuracy | | | 0.77 | 118936 |
| Macro avg | 0.84 | 0.54 | 0.51 | 118936 |
| Weighted avg | 0.80 | 0.77 | 0.69 | 118936 |

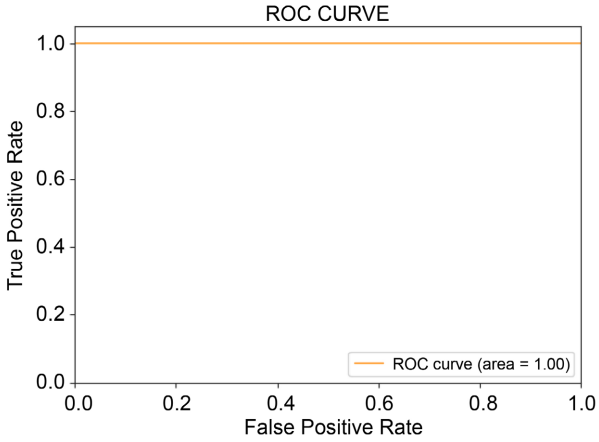


Figure 9. ROC CURVE Adaboosting.

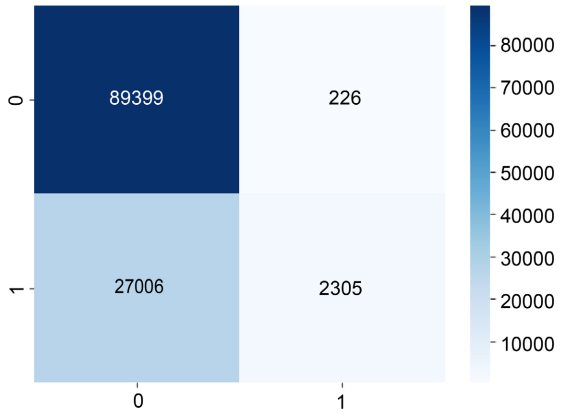


Figure 10. Confusion Matrix GNB.

**Random Forest Classifier:**

Achieve a 99.98% accuracy rate, a 99.98% precision rate, a 99.98% recall rate, and a 99.98% F1 score. Table 4 shows the performance assessment of DT.

See Figure 11 for the ROC curve of the Gaussian Naive Bayes algorithm. The curve reflects its moderate performance, suggesting room for improvement in predicting loan approval outcomes accurately.

Figure 12 illustrates the confusion matrix for the Random Forest Classifier, which achieved an impressive accuracy of 99.98%. The matrix reveals the model's robustness in correctly classifying the majority of loan applications.

Table 4. Performance Evaluation (Random Forest Classifier).

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 1.0 | 1.0 | 1.0 | 89,625 |
| 1 | 1.0 | 1.0 | 1.0 | 29,311 |
| Accuracy | | | 1.0 | 118,936 |
| Macro avg | 1.00 | 1.00 | 1.00 | 118,936 |
| Weighted avg | 1.00 | 1.00 | 1.00 | 118,936 |

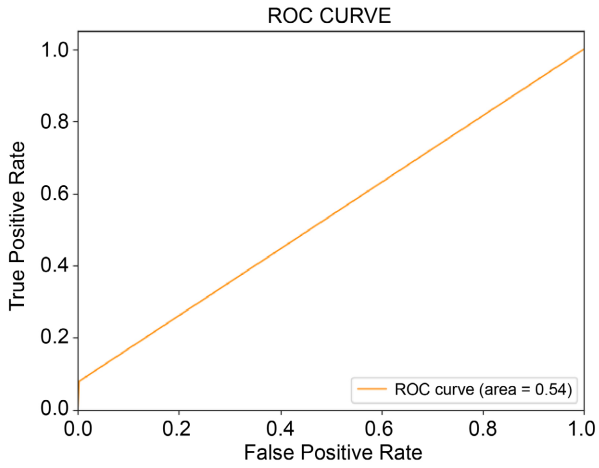


Figure 11. ROC CURVE GNB.

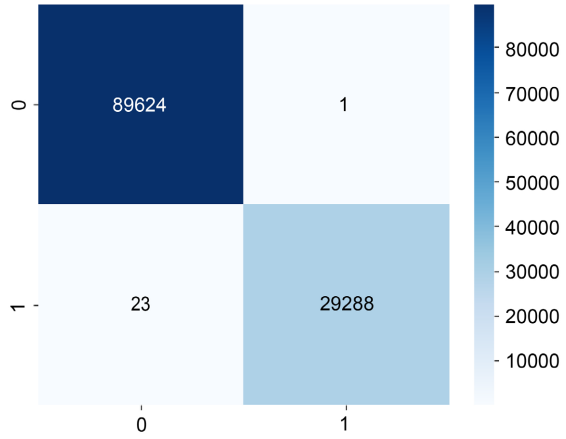


Figure 12. Confusion matrix random forest classifier.

**Decision Tree:**

Attained a 99.93% accuracy rate, a 99.93% precision rate, a 99.93% recall rate, and a 99.93% F1 score. Table 5 shows the results of DT's performance evaluation.

As seen in Figure 13, the ROC curve for the Random Forest Classifier highlights its excellent predictive performance, confirming its capability to handle diverse data patterns effectively.

As shown in Figure 14, the ROC curve for the Decision Tree Classifier provides insight into its performance, which achieved a high accuracy of 99.93%. This reflects the model's ability to balance interpretability and predictive power.

Table 5. Performance Evaluation (DT).

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 89,625 |
| 1 | 1.00 | 1.00 | 1.00 | 29,311 |
| Accuracy | | | 1.00 | 118,936 |
| Macro avg | 1.00 | 1.00 | 1.00 | 118,936 |
| Weighted avg | 1.00 | 1.00 | 1.00 | 118,936 |

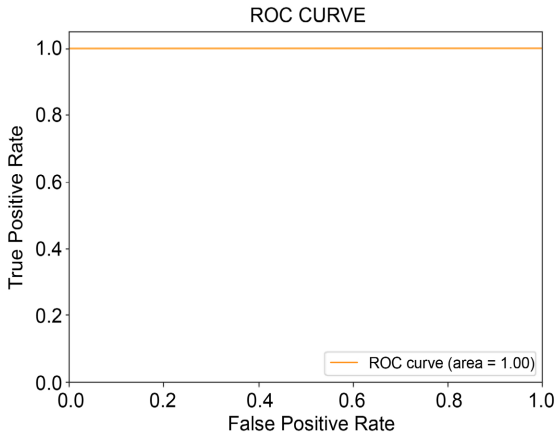


Figure 13. ROC CURVE random forest classifier.

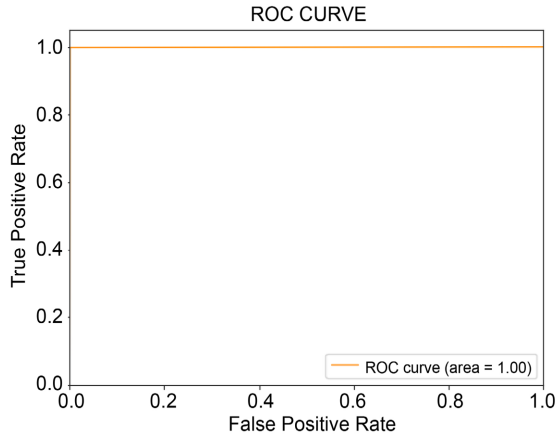


Figure 14. ROC CURVE DT.

Figure 15 presents the confusion matrix for the Decision Tree Classifier. The matrix visually demonstrates the model's effectiveness in classifying loan applications accurately.

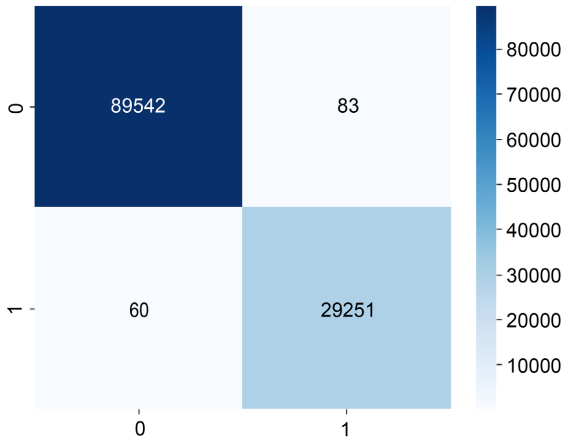


Figure 15. Confusion Matrix DT.

**Support Vector Machine:**

Attain the highest possible scores for accuracy, precision, recall, and F1 99.87%, 99.87%, and 99.87%, respectively. Table 6 shows the performance assessment of RF.

See Figure 16 for the confusion matrix of the Support Vector Machine (SVM) model. The matrix highlights the model's accuracy of 99.87% and its strength in delineating approved and denied loans.

Table 6. Performance Evaluation (SVM).

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 89,625 |
| 1 | 1.00 | 1.00 | 1.00 | 29,311 |
| Accuracy | | | 1.00 | 118,936 |
| Macro avg | 1.00 | 1.00 | 1.00 | 118,936 |
| Weighted avg | 1.00 | 1.00 | 1.00 | 118,936 |

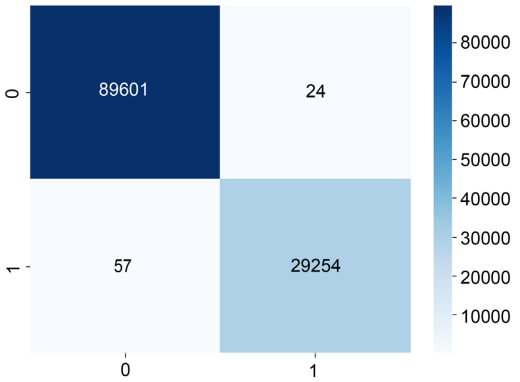


Figure 16. Confusion Matrix SVM.

As illustrated in Figure 17, the ROC curve for the Support Vector Machine demonstrates its high accuracy and effective separation of loan approval categories. This showcases the algorithm's performance in the loan prediction task.

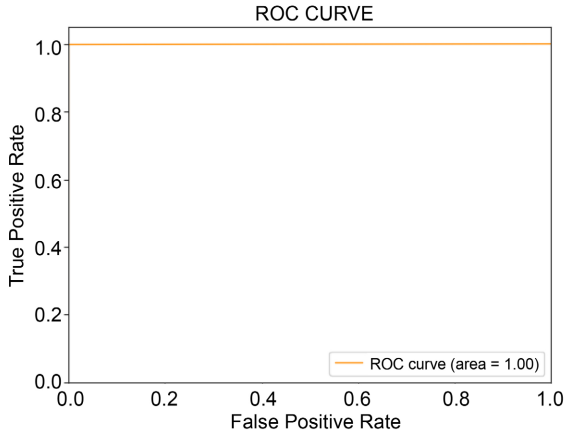


Figure 17. ROC CURVE SVM.

## 6. Discussion

This article deals with the performance of various machine learning algorithms regarding the prediction of bank loan approvals. Thus, the obtained accuracies that turn out to be statistically significant horrifyingly indicate the potential that these models hold regarding applicant eligibility. Algorithm selection is an essential part of machine learning and significantly influences the reliability and effectiveness of predictive models. In the present study on bank loan prediction, a few commonly used algorithms were assessed for their efficiency. The respective accuracies achieved by each were as follows: Ada Boosting at 99.99%, SVM at 99.87%, Decision Tree at 99.93%, Random Forest Classifier at 99.98%, and Gaussian NB at 77.10%. Random Forest Classifier and Ada Boosting, being ensemble methods, have performed incomparably well; this is understandable as several such ensemble methods are usually famous for their good generalization and avoidance of overfitting. Their accuracies, 99.98% and 99.99%, respectively, are a testimony to the strength of ensemble learning in complex predictive tasks such as bank loan approval. Others that did an excellent performance were the Support Vector Machine and Decision Tree, which had accuracies of 99.87% and 99.93%, respectively. It can be seen here that the robustness of SVM in high-dimensional space is very evident, while the decision tree, although slightly higher in accuracy, does offer interpretability and simplicity we'll look into later. The performance was rather poorer by Gaussian NB, reaching only an accuracy of 77.10%, probably due to the independence of features in its mathematics, which may be less relevant in such a complex dataset as bank loan prediction.

## 7. Conclusion & Future Step

Finally, the research on "Bank Loan Prediction Using Machine Learning

Techniques" provides valuable insights into how different algorithms can forecast loan approval status. Ada Boosting, which achieved a statistically significant 99.99% accuracy, stands out as the most effective algorithm, offering crucial data for financial institutions. The statistical analysis of the dataset guided better preprocessing and model development, highlighting the importance of refining features to enhance predictive accuracy. However, beyond technical performance, this research emphasizes the importance of responsible AI in finance. It points out ethical concerns such as mitigating algorithmic bias and ensuring user privacy, adding depth to the conversation around AI's impact. By exploring model interpretability and potential environmental implications, it contributes not only to machine learning literature but also to the practical application of predictive modeling in the financial sector. Despite its strengths, the study acknowledges limitations. The reliance on historical data might not account for rapid economic shifts or changing borrower behavior, making it less adaptable in real-time scenarios. The possibility of missing variables, like macroeconomic indicators or behavioral patterns, also limits the model's comprehensiveness. Furthermore, scalability across various types of financial institutions, which may have different needs, was not fully addressed.

Importantly, the research did not deeply explore the fairness of algorithmic decisions, especially regarding marginalized groups, which is essential in ensuring just loan approval processes. The study advocates for a balanced approach as financial institutions adopt more advanced technologies, suggesting that accuracy, fairness, and ethical practices must all be prioritized. While the findings showcase the transformative power of machine learning in loan approvals, the next steps involve real-world deployment and continuous monitoring, with attention to the ethical, societal, and environmental challenges that lie ahead.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

Appiahene, P., Missah, Y. M., & Najim, U. (2020). Predicting Bank Operational Efficiency Using Machine Learning Algorithm: Comparative Study of Decision Tree, Random Forest, and Neural Networks. *Advances in Fuzzy Systems, 2020,* Article ID: 8581202. https://doi.org/10.1155/2020/8581202

Arutjothi, G., & Senthamarai, C. (2017). Prediction of Loan Status in Commercial Bank Using Machine Learning Classifier. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 416-419). IEEE. https://doi.org/10.1109/iss1.2017.8389442

Berrada, I. R., Barramou, F., & Alami, O. B. (2024). Towards a Machine Learning-Based Model for Corporate Loan Default Prediction. *International Journal of Advanced Computer Science and Applications, 15,* 565-573. https://doi.org/10.14569/ijacsa.2024.0150357

Bhattad, S., Bawane, S., Agrawal, S., Ramteke, U., & Ambhore, P. B. (2021). Loan Prediction Using Machine Learning Algorithms. *International Journal of Computer Science Trends and Technology, 9,* 143-146.

Dasari, Y., Rishitha, K., & Gandhi, O. (2023). Prediction of Bank Loan Status Using Machine Learning Algorithms. *International Journal of Computing and Digital Systems, 14,* 139-146. https://doi.org/10.12785/ijcds/140113

Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2018). Forecasting Bank Failures and Stress Testing: A Machine Learning Approach. *International Journal of Forecasting, 34,* 440-455. https://doi.org/10.1016/j.ijforecast.2018.01.009

Gupta, A. (2022). *Collect Dataset.* https://www.kaggle.com/datasets/ychope/loan-approval-dataset/data

Karthikeyan, S., & Ravikumar, P. (2021). A Comparative Analysis of Feature Selection for Loan Prediction Model. *International Journal of Computer Applications, 975,* 8887.

Krasovytskyi, D., & Stavytskyy, A. (2024). Predicting Mortgage Loan Defaults Using Machine Learning Techniques. *Ekonomika, 103,* 140-160. https://doi.org/10.15388/ekon.2024.103.2.8

Krishnaraj, P., Rita, S., & Jaiswal, J. (2023). Comparing Machine Learning Techniques for Loan Approval Prediction. In *Proceedings of the 1st International Conference on Artificial Intelligence, Communication, IoT, Data Engineering and Security, IACIDS 2023*.

Lai, L. (2020). Loan Default Prediction with Machine Learning Techniques. In *2020 International Conference on Computer Communication and Network Security (CCNS)* (pp. 5-9). IEEE. https://doi.org/10.1109/ccns50731.2020.00009

Muhammad, I., Dahlia, R., et al. (2024). Performance Analysis of Ensemble Learning and Feature Selection Methods in Loan Approval Prediction at Banks. *Journal of Artificial Intelligence and Engineering Applications (JAIEA), 3,* 557-564. https://doi.org/10.59934/jaiea.v3i2.426

Natasha, A., Prastyo, D. D., & Suhartono, (2019). Credit Scoring to Classify Consumer Loan Using Machine Learning. *AIP Conference Proceedings, 2194,* Article ID: 020070. https://doi.org/10.1063/1.5139802

Nureni, A. A., & Adekola, O. E. (2022). Loan Approval Prediction Based on Machine Learning Approach. *Fudma Journal of Sciences, 6,* 41-50. https://doi.org/10.33003/fjs-2022-0603-830

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. N. (2022). Machine Learning Models for Predicting Bank Loan Eligibility. In *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)* (pp. 1-5). IEEE. https://doi.org/10.1109/nigercon54645.2022.9803172

Perera, C. L., & Premaratne, S. C. (2024). An Ensemble Machine Learning Approach for Forecasting Credit Risk of Loan Applications. *WSEAS Transactions on Systems, 23,* 31-46. https://doi.org/10.37394/23202.2024.23.4

Raheem, M. (2024). Loan Default Prediction Using Machine Learning: A Review on the Techniques. *Journal of Applied Technology and Innovation, 8,* 1-6.

Sayjadah, Y., Hashem, I. A. T., Alotaibi, F., & Kasmiran, K. A. (2018). Credit Card Default Prediction Using Machine Learning Techniques. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)* (pp. 1-4). IEEE. https://doi.org/10.1109/icaccaf.2018.8776802

Sharma, H., Tyagi, I., Agarwal, G., & Gupta, D. (2023). *An Exhaustive Investigation on Loan Prediction in Banks Using LRD*.

Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. N. (2021). Prediction of Modernized Loan Approval System Based on Machine Learning Approach. In *2021 International Conference on Intelligent Technologies (CONIT)* (pp. 1-4). IEEE. https://doi.org/10.1109/conit51480.2021.9498475

Turkson, R. E., Baagyere, E. Y., & Wenya, G. E. (2016). A Machine Learning Approach for Predicting Bank Credit Worthiness. In *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)* (pp. 1-7). IEEE. https://doi.org/10.1109/icaipr.2016.7585216

Uddin, N. (2023). An Ensemble Machine Learning Based Bank Loan Approval Predictions System with a Smart Application. *International Journal of Cognitive Computing in Engineering, 4,* 327-339. https://doi.org/10.1016/j.ijcce.2023.09.001

Viswanatha, V. (2023). Prediction of Loan Approval in Banks using Machine Learning Approach. *International Journal of Engineering and Management Research, 13,* 7-19.

Zuama, R. A., Ichsan, N., Pohan, A. B., Azis, M. S., & Lase, M. (2024). An Implementation of Machine Learning on Loan Default Prediction Based on Customer Behavior. *Jurnal Info Sains: Informatika dan Sains, 14,* 157-164.