MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 13 Issue 02 (2026) | Pages: 370-377

Comparative Analysis of Machine Learning Models in Credit Risk Evaluation

Shweta Meshram[1], S.S. Bavlatti[2], A.R. Bensekar[3], A.V. Binorkar[4], S.V. Dafal[5], T.A. Kharat[6]

[1,2,3,4,5,6]Department of MCA, MES'IMCC, Pune

[1]*sdm.imcc@mespune.in*, [2]*samarth.bavlatti@gmail.com*, [3]*aamirbensekar325@gmail.com*, [4]*adityabinorkar@gmail.com*, [5]*dafalsanika30@gmail.com*, [6]*tejaskharat9999@gmail.com*

Peer Review Information

Type: *Article*
Received: *20 March 2026*
Revised: *03 April 2026*
Accepted: *21 May 2026*
Published: *03 June 2026*

Abstract

Loan approval is an important process in the financial sector, influencing both lenders and applicants. Traditional methods often rely on manual evaluation, which tend to be slow, inconsistent, and prone to bias. To do away with various such limitations, this paper explores employing machine learning techniques for predicting loan approval outcomes. The research compares three widely used classification algorithms: Random Forest, Logistic Regression, Decision Tree . A structured dataset containing applicant details such as financial status, employment information, and credit history was used. Data preprocessing techniques which include: handling absent values, encoding categorical variables, and feature scaling were applied to ensure data quality. The models had to be trained and examined using performance metrics like precision, recall, F1-score, and ROC-AUC. Among the models, Random Forest demonstrated the most reliable performance given its capacity to capture complex patterns and reduce overfitting. The findings showcase the underlying capacity of machine learning in improving decision-making efficiency and accuracy in loan approval systems.

Keywords: Machine Learning; Loan Prediction; Credit Risk Evaluation; Random Forest; Logistic Regression; Decision Tree Classification; Data Preprocessing.

How to Cite This Article

Meshram, S., Bavlatti, S. S., Bensekar, A. R., Binorkar, A. V., Dafal, S. V., & Kharat, T. A. (2026). Comparative analysis of machine learning models in credit risk evaluation. *Multidisciplinary Journal of Research in Engineering and Technology*, 13(2), 370–377.

Introduction

In today's financial environment, loan approval plays a necessary role in enabling individuals to achieve goals such as education, housing, or business development. Financial institutions must undertake careful evaluation of applications to minimize risk while ensuring fair access to credit. Traditionally, this process depends on manual assessment of factors such as income, credit history, and assets, which can sometimes prove time-consuming and sometimes inconsistent.

With the advancement of data-driven technologies, machine learning has emerged as a powerful tool to support decision-making in finance. These models have ability to process large amounts of data, identify hidden patterns, and generate predictions with higher consistency compared to manual methods.

This study focuses on analyzing and comparing the performance of three distinct machine learning algorithms-Decision Tree, Logistic Regression and Random Forest for predicting loan approval. Each model offers unique advantages: Decision Trees handle non-linear relationships effectively, Logistic Regression provides simplicity and interpretability, also Random Forest can improve prediction accuracy via combining multiple trees.

The prime objective of this study is to identify the most suitable model for real-world loan approval systems based on accuracy, reliability, and scalability.

Literature Review

Recent studies show increasing interest in leveraging machine learning to automate loan approval processes. Compared to traditional approaches, these methods provide faster processing, improved accuracy, and reduced dependency on manual intervention. A number of researchers have evaluated unique machine learning models for this task. For example, studies have tested and compared algorithms such as Logistic Regression, Random Forest, SVM(Support Vector Machine), Decision Tree, and Gradient Boosting. Results from such studies depict how ensemble methods like Random Forest often achieve better performance attributing to their capacity to undertake complex relationships within data.

Other research emphasises that machine learning models might significantly cut down processing time and improve consistency in decision-making. Ensemble approaches, in particular, combine the power of multiple models, resulting in more stable and accurate predictions. Along with performance, contemporary work has also emphasised on the importance of fairness and transparency in automated systems. Researchers are inclined to an opinion that while machine learning improves efficiency, it is necessary to make sure these systems do not introduce bias into financial decisions.

Overall, the literature indicates that machine learning, specifically ensemble techniques, provides an effective and practical solution for modern loan approval systems.

Methodology

The step-by-step implementation of our approach is clearly outlined in Fig. 1., which provides a visual guide to the entire workflow.

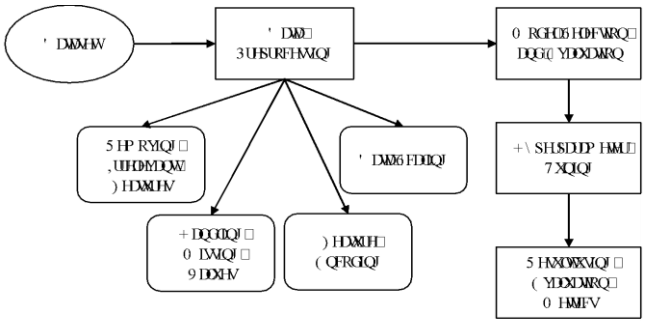


Fig. 1. *Flowchart of Methodology*

The overall workflow of this study is illustrated in Fig. 1, which outlines the complete process followed to develop and asses the loan approval prediction models. The methodology is differentiated into various stages, also including data preprocessing, exploratory analysis, model development, tuning, and evaluation.

Data Acquisition and Preprocessing

The dataset acquired for this research was resourced from Kaggle, a popular platform that provides open datasets for data analysis and machine learning. It contains detailed information about loan applicants such as demographic details, employment status, annual income, loan amount, credit score (CIBIL), and asset-related information.

Before applying any kind of machine learning techniques, the dataset was carefully prepared to ensure quality and consistency. Data preprocessing becomes an important part in improving model accuracy and reliability.

- Removing Irrelevant Features: A column named loan_id was dropped as it only serves as a unique identifier and does not contribute to prediction, helping reduce noise in the data.
- Handling Missing Data: Missing data was identified in several attributes. The number of missing values for each feature were as follows: number of dependents (13), education (12), self-employment status (21), annual income (12), loan amount (14), loan term (9), CIBIL score (18), residential assets value (15), commercial assets value (22), luxury assets value (11), bank asset value (14), and loan status (20).To maintain data quality and avoid biased results, the dropna() method was applied to eliminate rows containing missing values. Although this procedure resulted in a reduction of the dataset size, it ensured that the remaining records were complete and suitable for accurate and consistent analysis.
- Encoding Categorial Variables: Since machine learning algorithms need numerical input, categorical features columns like education and self_employed were converted using label encoding. The target column loan_status, which indicates whether a loan was approved or rejected, was encoded as 1 (approved) and 0 (not approved).
- Feature Scaling: To bring all features onto a similar scale and enhance model performance, specially for gradient-based algorithms, standardization was applied using StandardScaler. This technique transforms numerical features to have a mean of zero and a standard deviation of one.
- Train-Test Split: The data set was categorised into training and testing sets in an 80-20 ratio, ensuring the class distribution remains consistent through stratified sampling. A random seed (random_state=42) was used for reproducibility.

Exploratory Data Analysis

Understanding the patterns in the dataset is a vital step before developing predictive models. This section presents an exploratory data analysis (EDA) aimed at uncovering insights related to loan approval decisions. The analysis considers both categorical and numerical features, including education, self-employment status, and CIBIL score, and examines their association with loan outcomes.

- Distribution of Education and Self-Employment: The dataset revealed an almost equal distribution between graduates (2,127 applicants) and non-graduates (2,144 applicants), as well as between self-employed (2,152) and non-self-employed (2,119) individuals. This balanced distribution eliminates concerns of class imbalance for these attributes, supporting fair and representative model training. Furthermore, it suggests that loan outcomes cannot be inferred directly from these features, encouraging further multivariate analysis.

Fig. 2. and Fig. 3. illustrates the count distribution of applicants based on educational background and employment status.

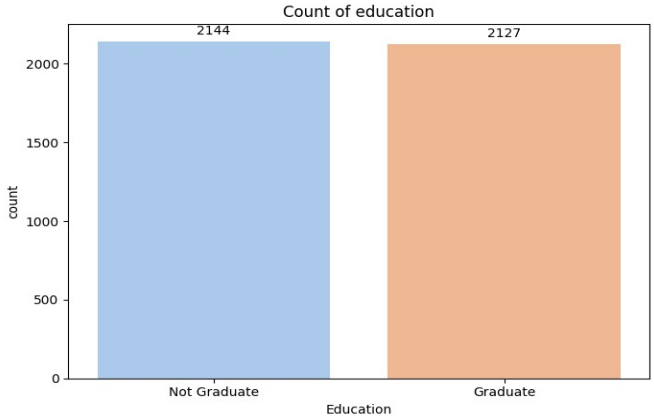


Fig. 2. *Count of Applicants by Education*

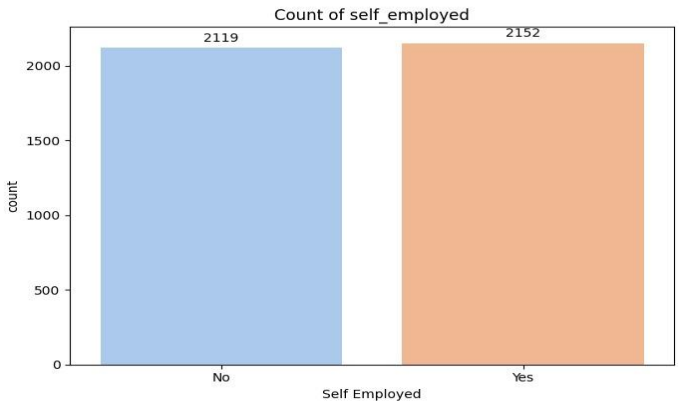


Fig. 3. *Count of Applicants by Employment*

- Loan Approval by Self-Employment Status: The loan approval distribution was examined with respect to self-employment status. Among non-self-employed individuals, approximately 1,320 loans were approved and 800 were rejected, while among the self-employed, about 1,340 were approved and 820 were not, as shown in Fig. 4. This nearly symmetrical distribution indicates that employment type has minimal influence on approval decisions. Contrary to assumptions that self-employed individuals face higher rejection rates due to perceived income instability, the data suggests equitable treatment across employment categories.

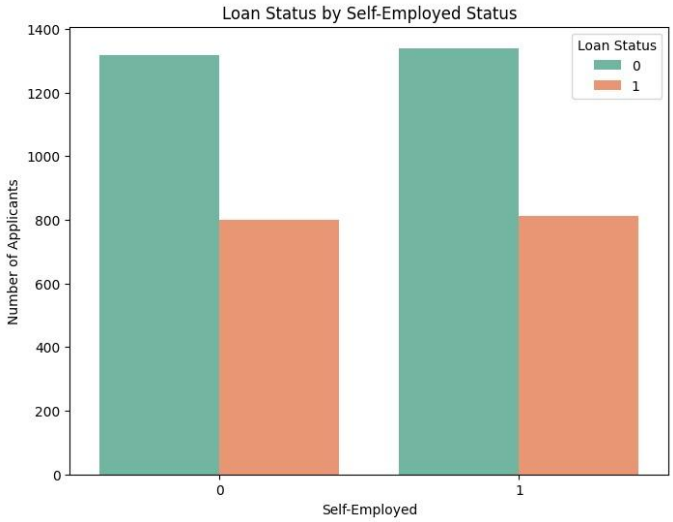


Fig. 4. *Loan status distribution by self-employment status*

- CIBIL Score vs Loan Status: A notable pattern was observed between CIBIL scores and loan approval outcomes. Approved loans were predominantly clustered in the 600–900 range, whereas rejected applications were mainly in the 300–600 band, as shown in Fig. 5. Very few applicants with high scores were denied loans, likely due to additional financial risk factors. These findings confirm that CIBIL score is a significant determinant in loan evaluation, which reflects its importance as an indicator of an applicant's creditworthiness in financial decisions.

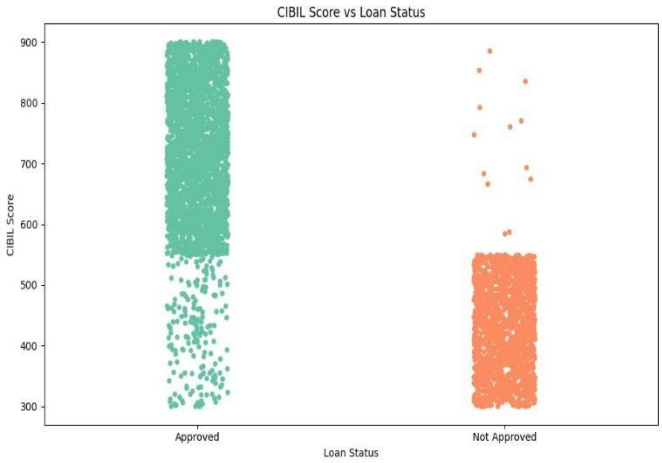


Fig. 5. *CIBIL Score Distribution by Loan Approval Status*

The exploratory analysis highlights several important trends: education level is balanced but not a strong standalone predictor of loan approval; self-employment status shows minimal influence, suggesting equitable evaluation across employment types; and CIBIL score is a critical variable, with higher scores being strongly linked to loan approvals. These insights validate the value of comprehensive feature analysis beforehand.

Model Selection and Training

To identify the primary appropriate algorithm for loan approval prediction, three commonly employed supervised learning models were chosen for comparison.:

- Logistic Regression: This is a basic linear model more often used for binary classification tasks. It estimates probabilities using a logistic function and is valued for its interpretability.
- Decision Tree Classifier: This model works by segregating the data into smaller sets based on feature values. While easy to understand and capable of handling complex data, it may overfit, especially on tiny datasets

- Random Forest Classifier: An ensemble method that brings together multiple decision trees and uses majority voting for predictions. It generally offers improved accuracy and reduced overfitting compared to single trees.

Each of these models was trained on the same training dataset for a fair comparison.

Hyperparameter Tuning

To improve the overall efficiency of the models hyperparameter tuning was performed to find most effective parameter combination for each algorithm. This process is critical for improving model generalizability and achieving higher predictive accuracy.

Hyperparameter grids were defined for three machine learning models: Random Forest, Logistic Regression, and Decision Tree, and Each grid defined a set of possible values for important variables that affect in turn how the model learns and its overall complexity. The configurations included:

These parameters govern how each model learns from the data. Systematically testing various combinations allows for the selection of the most effective settings for optimal performance.

Table 1. Hyperparameter Grids Used for Model Tuning with Gridsearchcv

Model	**Hyperparameter**	**Values Explored**
Logistic Regression	C	[0.01, 0.1, 1, 10, 100]
	Solver	["liblinear", "lbfgs"]
Decision Tree	max_depth	[3, 5, 10, None]
	min_samples_split	[2, 5, 10]
	criterion	["gini", "entropy"]
Random Forest	n_estimators	[50, 100, 200]
	max_depth	[None, 10, 20]
	min_samples_split	[2, 5]
	criterion	["gini", "entropy"]

To automate this process, GridSearchCV was used to perform an exhaustive search covering all parameters combinations within the defined grids. This method included 5-fold cross-validation to evaluate performance and prevent overfitting. The scoring metric was set to **F1-score**, providing a balanced assessment of both precision and recall-particularly useful for imbalanced datasets.

The grid search identified the best-performing hyperparameters for every model based on the highest F1-score achieved during cross-validation. These optimal configurations were retained for final model training and evaluation.

Performance Evaluation

Post training, model efficiency was measured with the help of several key metrics to understand their strengths and weaknesses:

- Precision: Indicates how many of the predicted approved cases were actually correct. It helps reduce the chances of approving applicants who may not be eligible.
- Recall (Sensitivity): Measures how many of the truly approved applications were accurately identified by the model, ensuring that eligible applicants are not getting overlooked.
- F1-Score: A combined measure of recall & precision that provides a balanced evaluation, especially useful when the data is unevenly distributed.
- ROC AUC Score: Represents the area trapped under the ROC curve, which indicates the model's overall ability to distinguish between approved and rejected applications. A higher AUC reflects better performance.

Using multiple evaluation metrics provides a information of how each model performs, especially in contexts where false positives or false negatives carry high risk.

Visualization and Comparative Analysis

To make the results easier to interpret and compare, two different types of visualizations were created:

- Pie Chart: Showing the proportion of approved versus not-approved loans.
- Bar Graphs: Used to compare the precision, recall, and F1-score of all models' side-by-side. These visuals clearly highlight the comparative strengths and limitations of each model.

- ROC Curves: All models' ROC curves were plotted on the single plot to compare their performance along different classification thresholds. A model with a curve closer to the top-left corner (and a higher AUC) is considered more accurate

Through this structured approach, the study aimed to identify the classification model that delivers the most reliable, accurate, and generalizable predictions for loan approval decisions.

Result And Discussion

This study analyzed the performance of three widely used machine learning models - Random Forest, Logistic Regression, and Decision Tree to predict loan approval outcomes using applicant data. The dataset was carefully preprocessed, including the encoding of categorical variables and normalization of numerical features. All models were trained and tested using an 80-20 split to ensure an unbiased evaluation.

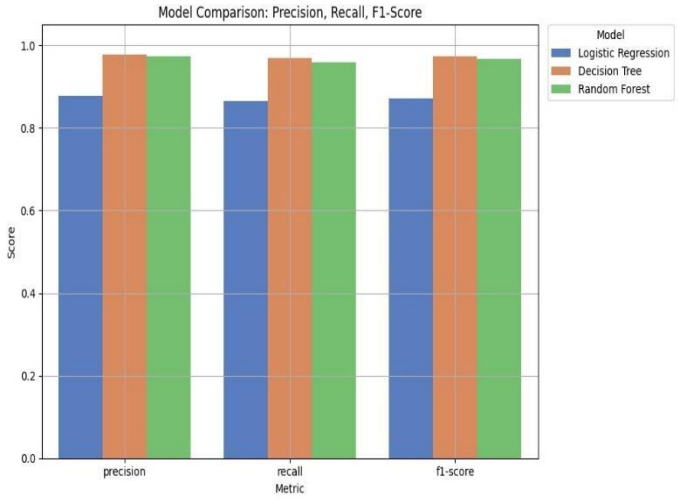
Table 2. Performance Evaluation Of Different Classification Models For Loan Prediction

Classifier	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.88	0.90	0.89	0.97
Decision Tree	0.97	0.97	0.97	0.98
Random Forest	0.98	0.96	0.97	1.00

Model Performance Metrics

Each model was assessed using key classification metrics: recall, precision, F1-score, and ROC-AUC. These metrics helped measure how accurately each model predicted loan approvals and rejections.

The results indicate that all three models achieved good performance, particularly when identifying approved loans:



Random Forest consistently delivered the best results across all metrics. Its recall and high precision show that it was excellent at both making accurate predictions and capturing the most of true positive cases.

Decision Tree and Logistic Regression performed strongly, with scores just slightly behind Random Forest. Their ability to make correct predictions was commendable, though not quite as refined.

The ROC Curve analysis offered deeper insight. Random Forest obtained a perfect AUC score of 1.00, meaning it could flawlessly distinguish between approved and non-approved loans across all thresholds. Decision Tree and Logistic Regression were not far behind, each achieving AUC scores of 0.97, which still indicates excellent classification ability.

A pie chart was used to present the proportion of approved versus not-approved loan applications. As illustrated in Fig. 8., the distribution appears relatively balanced, with a modest skew favoring approved applications. This balanced class representation likely contributed to the robust performance observed across all three predictive models, as the classifiers were exposed to a sufficiently diverse set of outcomes during training. A balanced class distribution also helps reduce bias and improves the model's ability to perform well on unseen dataset.

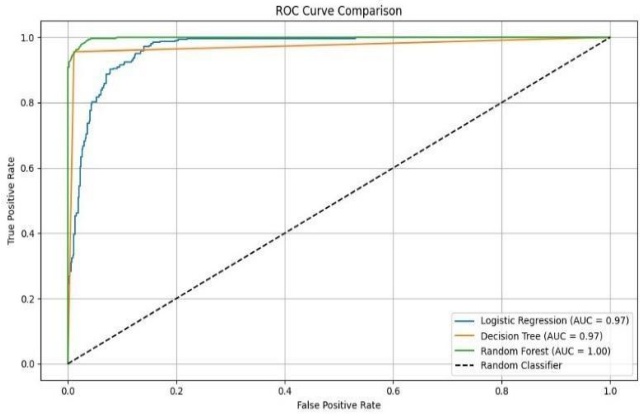


Fig. 7. *Representation of ROC Curve Comparison*

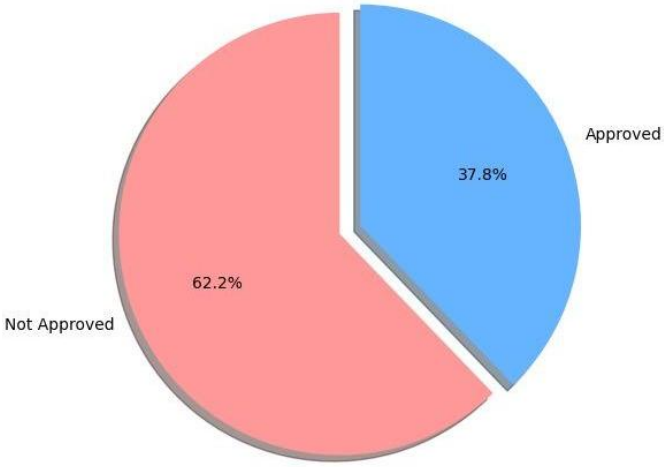


Fig. 8. *The initial data visualization, highlighting the imbalance in the dataset.*

Discussion

The success of the Random Forest model can be associated to its ensemble approach. By combining 'n' number of decision trees and averaging their outputs, Random Forest counters the possibility of overfitting and provides more stable, accurate predictions.

Logistic Regression, despite being a simpler linear model, performed well due to standardized features and a well-structured dataset. It is also highly interpretable, which can be a major advantage in real-world financial applications.

The Decision Tree classifier proved to be a strong, interpretable baseline. While slightly more prone to overfitting, it still delivered solid results and may be especially useful in applications where decision rules need to be easily explained

Conclusion

This study showcased a comparative analysis of three machine learning models Random Forest, Logistic Regression, and Decision Tree for predicting loan approval outcomes. The dataset was carefully pre-processed, and the models were evaluated using standard performance metrics like recall, precision, F1-score, and ROC-AUC.

The results depict that all models performed well; however, Random Forest consistently met the highest accuracy and reliability in general. Its ensemble approach helps it handle complex data patterns while minimizing the danger of overfitting, making it best fit for real-world applications.

Decision Tree and Logistic Regression also demonstrated strong performance and may be preferred in scenarios where interpretability and simplicity are important.

In conclusion, machine learning techniques offer a powerful approach to improving loan approval systems by making them faster, more consistent, and data-driven. Future research can focus on incorporating more advanced models and ensuring fairness in automated decision-making processes.

References

1. Sirishma, A. G., et al. (2024). *Machine Learning for Risk Assessment: A Comparative Study of Models Predicting Loan Approval*. SSRN.
2. Khan, A., et al. (2021). *Loan Approval Prediction Model: A Comparative Analysis*. Advances and Applications in Mathematical Sciences, 20(3).
3. Uddin, N., et al. (2023). *An Ensemble Machine Learning-Based Bank Loan Approval Prediction System*. Int. J. Cognitive Computing in Engineering, 4, 327–339.
4. Dansana, D., et al. (2023). *Impact of Loan Features on Bank Loan Prediction Using Random Forest*. Engineering Reports.
5. Yasaswini, P., et al. (2023). *Forecasting of Bank Loan Approval Data Using ML Algorithms.*
6. Singh, et al. (2023). *The Use of Responsible AI Techniques in Loan Approval*. Int. J. Human–Computer Interaction, 39(7), 1543–1562
7. Sinap, V. (2024). *A Comparative Study of Loan Approval Prediction Using Machine Learning Methods*. Gazi University Journal of Science.
8. Chen, G., et al. (2024). *Predicting Loan Eligibility Approval Using Machine Learning Techniques*. Proceedings of SCITEPRESS.
9. Haque, F. M., et al. (2024). *Bank Loan Prediction Using Machine Learning Algorithms*. arXiv.
10. Singh, D. P., & Kaushik, B. (2025). *Predictive Modeling for Bank Loan Approval: From Data to Decisions*. Procedia Computer Science.
11. Pandey, N., et al. (2021). *Loan Approval Prediction Using Machine Learning Algorithms Approach*. International Journal of Innovative Research in Technology.
12. Sengaliappan, M., & Sangaiah, P. (2024). *Loan Approval Prediction Using Machine Learning*. IJSREM.
13. Velvigneswar, J., & Kumari, P. S. (2025). *Loan Approval Prediction Using Machine Learning*. International Journal of Advanced Research in Computer and Communication Engineering.
14. Shahzad, A. (2025). *Enhancing Loan Approval Accuracy Through Machine Learning Techniques*. Preprints.org.
15. Juyal, A., et al. (2023). *Comparative Study of Machine Learning Models in Loan Approval Prediction*. Taylor & Francis.
16. Shinde, A., et al. (2022). *Loan Prediction System Using Machine Learning*. ITM Conferences.