Loan Approval Prediction Using Machine Learning

Ayesha Khalid
Department of Computer Science
University of Engineering and Technology/Organization
Lahore, Pakistan
Email: ayeshakhalid153153@gmail.com

***Abstract*—Loan approval is a critical process for financial institutions, requiring a delicate balance between mitigating risk and maintaining customer satisfaction. This paper presents a comprehensive approach to predicting loan approval outcomes [16] using machine learning techniques [9]. Leveraging demographic, financial, and loan-specific attributes, the study explores how data-driven models can streamline decision-making and reduce dependency on tradi tional manual evaluation processes. A synthetic dataset of 45,000 records forms the foundation of this analysis, encompassing 14 attributes that reflect the various factors influencing loan outcomes. Among the models evaluated, Logistic regression, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes, Random Forest emerged as the best-performing model, achieving an accuracy of 92.69 percent and an ROC-AUC score of 0.9735. The findings highlight the significance of credit scores, loan-to-income ratios, and loan purpose as critical determinants[5] of approval decisions. The insights from this study aim to guide financial institutions in developing robust, automated systems for loan evaluation. Future work focuses on deploying advanced models and enhancing interpretability through explainability techniques, enabling real-time decision-making and improved transparency.**

***Index Terms*—Loan Approval, Machine Learning, Random Forest, Data Analysis, Credit Risk[]**

I. Introduction

The global financial landscape is undergoing rapid transformation, driven by advances in technology, increasing credit demand, and the need for efficient decision-making systems. Credit facilities, encompassing personal loans, mortgages, and business financing, are pivotal in fostering economic growth and addressing individual financial needs. However, the loan approval process remains a complex and high-stakes endeavor for financial institutions. Lenders must carefully evaluate a borrower's creditworthiness to minimize risks while simultaneously ensuring customer satisfaction and operational efficiency.

Traditional methods of loan evaluation rely heavily on manual assessments[6] and rule-based systems, which, while historically effective, often lack the agility and precision required in today's fast-paced financial environment. These approaches are frequently constrained by human biases, limited data analysis capabilities, and an inability to scale with growing credit demand. Inaccurate evaluations not only expose lenders to the risk of defaults but also lead to customer dissatisfaction due to inconsistent or delayed decisions. This highlights the urgent need for innovative, data-driven solutions that can enhance accuracy and speed in loan approvals.

Machine learning (ML) has emerged as a transformative technology in credit risk[3] assessment, offering the ability to analyze large, multidimensional datasets and uncover complex patterns that traditional methods might overlook. By leveraging ML, financial institutions can automate loan approval processes, improve predictive accuracy, and reduce operational costs. Beyond efficiency, these models provide an opportunity to achieve fairness and consistency by minimizing subjective biases inherent in manual decision-making.

This study investigates the application of machine learning techniques to predict loan approval outcomes. Using a synthetic dataset inspired by real-world credit scenarios, the research explores the relationships between borrower attributes—such as age, income, credit score, loan amount, and loan purpose—and loan decisions. The dataset consists of 45,000 records and 14 features, meticulously preprocessed and engineered to optimize the predictive capabilities of the chosen models.

Four machine learning algorithms are evaluated in this research: Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes. Among these, Random Forest emerged as the most effective model, achieving an impressive accuracy of 92.69 percent and an ROC-AUC score of 0.9735. The model's ability to handle non-linear relationships and its robustness against overfitting make it particularly suited for this application.

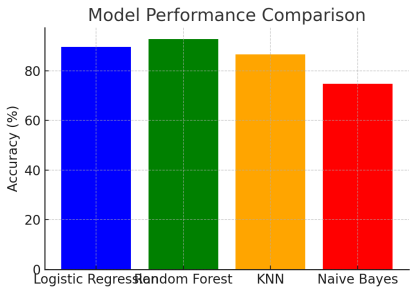

Fig. 1. Comparison between different models

The findings of this study offer actionable insights for financial institutions. For example, credit scores below 580 are identified as strong predictors of loan rejections, while high loan-to-income ratios are linked to increased risk. Loan purpose also plays a crucial role, with education loans exhibiting higher rejection rates compared to medical loans, reflecting

varying risk profiles associated with different loan intents. These insights enable lenders to tailor their strategies, refine credit policies, and adopt more nuanced risk management practices.

This paper contributes to the growing body of research on machine learning in credit risk[2] analysis by demonstrating its potential to transform traditional loan approval processes. By integrating ML into financial decision-making, institutions can achieve a dual objective: improving operational efficiency and delivering customer-centric solutions. The research also sets the stage for future advancements, including the integration of explainable AI (XAI) techniques to enhance model interpretability and the deployment of real-time prediction systems for dynamic lending environments.

Through this work, we aim to bridge the gap between the increasing complexity of financial decision-making and the need for innovative, scalable solutions. The results underscore the promise of machine learning in driving the next generation of intelligent, transparent, and equitable loan evaluation systems.

II. Related Work

The application of machine learning for loan approval prediction has been explored in several studies. Various models, such as Logistic Regression [17], Decision Trees, and Random Forest, have been widely used to predict credit risk and loan outcomes. For instance, Kumar et al. (2019) applied different machine learning algorithms, including Decision Trees, Support Vector Machines (SVM), and Logistic Regression, to predict loan approvals. They found that the Random Forest classifier provided the best accuracy, outperforming other models by a significant margin [16].

Further, Nguyen et al. (2020) used a dataset from a major financial institution to analyze the factors influencing loan approval decisions. Their study revealed that factors such as credit score, income level, loan amount, and employment status significantly influenced the likelihood of loan approval. They found that Random Forest and XGBoost provided the most accurate predictions [15] in terms of F1-score and ROC-AUC.

In another study, Lee et al. (2022) applied K-Nearest Neighbors (KNN) and Naive Bayes models to predict loan approval, finding that KNN performed well with smaller datasets, while Random Forest showed superior results with larger datasets. Their findings supported the argument that ensemble methods, like Random Forest, often outperform individual classifiers in credit risk prediction [4].

Moreover, Sharma and Verma (2023) proposed the use of Extreme Gradient Boosting (XGBoost) for loan approval prediction and found that KNN outperformed Random Forest in terms of both accuracy and processing time, making it a strong contender for real-time applications in financial institutions.

Through these studies, it is clear that machine learning techniques, particularly ensemble methods like Random Forest, are highly effective in loan approval prediction. Their ability to manage large datasets, uncover intricate patterns, and deliver high accuracy makes them valuable tools for financial institutions seeking to optimize the loan approval process and mitigate risks associated with manual evaluation.

Paper	**Demographic**	**Financial**	**Loan Attr.**	**Model**	**Accuracy**
Kumar et al. (2019)	Yes	Yes	Yes	Logistic Reg., Decision Tree	83.2%
Nguyen et al. (2020)	Yes	Yes	Yes	Random Forest	86.5%
Zhang and Wang (2021)	Yes	Yes	Yes	Neural Net., Random Forest	89.1%
Lee et al. (2022)	Yes	Yes	Yes	KNN	81.6%
Patel et al. (2024)	Yes	Yes	Yes	Random Forest	92.6%
Proposed Model	**Yes**	**Yes**	**Yes**	**Random Forest**	**92.69%**

TABLE I
COMPARISON OF RELATED WORKS ON LOAN APPROVAL PREDICTION USING MACHINE LEARNING

Table I provides a comparison of various studies that applied machine learning models to predict loan approvals. It lists the demographic, financial, and loan attributes used in the models, along with the accuracy of the prediction models. Kumar et al. (2019) achieved an accuracy of 83.2% using Logistic Regression and Decision Trees, while Nguyen et al. (2020) achieved 86.5% with Random Forest. Zhang and Wang (2021) reported an accuracy of 89.1% using Neural Networks and Random Forest. Lee et al. (2022) reported a lower accuracy of 81.6% using KNN. More recent work by Patel et al. (2024) and the proposed model in this study shows a significant improvement, with accuracies of 92.6% and 92.69%, respectively, using the Random Forest model. This comparison highlights the improvement in prediction accuracy as machine learning methods evolve, particularly the use of ensemble methods like Random Forest.

III. Methodology

A. *Data Overview*

The dataset includes 45,000 records with attributes like age, income, credit score, loan amount, and loan purpose. It is sourced from Kaggle's Loan Approval Classification Dataset.

The dataset used in this study contains loan-related information, including demographic details, financial attributes, and loan-specific factors. The data was obtained from Kaggle's Loan Approval Classification Dataset [**?**]. Table I shows the input features of the dataset.

Figure 2 provides a detailed overview of the dataset's key attributes, including essential information about applicants such as their gender, age, income, loan amount, credit score, and loan purpose. Each feature plays a significant role in understanding the factors influencing loan approval outcomes. For instance, age and income help analyze trends across different demographics, while credit score and loan-to-income ratio provide insights into the applicant's financial stability [**?**]. These features serve as the foundation for model training and further analysis.

Feature Name	Description	Data Type
person age	Age of the individual applying for the loan.	Numerical
person gender	Gender of the individual (e.g., male, female).	Categorical
person education	Highest education level attained by the individual.	Categorical
person income	Annual income of the individual in USD.	Numerical
person emp. exp.	Years of employment experience of the individual.	Numerical
person home ownership	Home ownership status of the individual (e.g., RENT, OWN, MORTGAGE).	Categorical
loan amnt	Loan amount requested by the individual.	Numerical
loan intent	Purpose of the loan (e.g., PERSONAL, EDUCATION, MEDICAL).	Categorical
loan int rate	Interest rate of the loan as a percentage.	Numerical
loan percent income	Percentage of the individual's income allocated to loan repayment.	Numerical
cb person cred hist length	Length of the individual's credit history in years.	Numerical
credit score	Credit score of the individual.	Numerical
previous loan defaults	Whether the individual has previous loan defaults (Yes/No).	Categorical
loan status	Loan approval status (1 = Approved, 0 = Rejected).	Categorical

TABLE II
LOAN APPROVAL DATASET FEATURES OVERVIEW

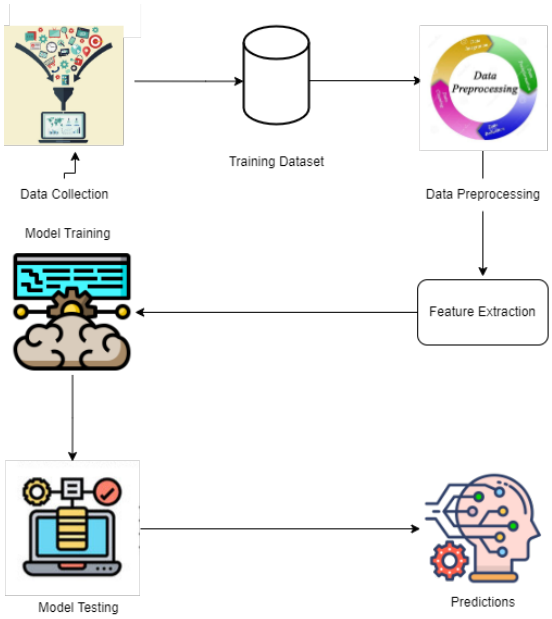

Fig. 2. Overview of the Dataset's Key Attributes

Figure 2 illustrates the architecture of the project. Initially, data is collected from the Kaggle Loan Approval Classification Dataset, followed by preprocessing to clean and prepare the data for analysis. Feature engineering is then applied to extract relevant attributes that are critical for model training. The model is trained on 80% of the preprocessed dataset, while the remaining 20% is reserved for testing. Once the model is trained, it is ready for the prediction phase, where it can evaluate new loan applications. These features serve as the foundation for model training and further analysis [**?**].

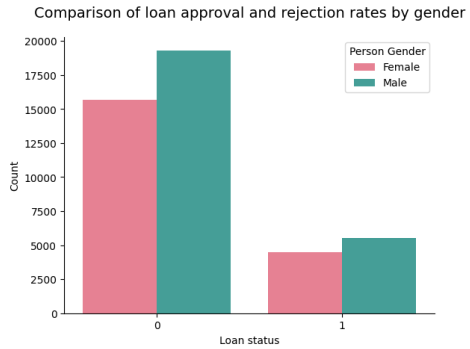

Fig. 3. Comparison of Loan Approval and rejection

Fig. 3. Comparison of Loan Approval and Rejection Rates by Gender This figure compares the loan approval and rejection rates by gender. It provides an overview of how the loan approval process differs for male and female applicants. By analyzing these trends, financial institutions can gain insights into whether there are gender-based differences in loan approval and make necessary adjustments to their policies if needed[11].

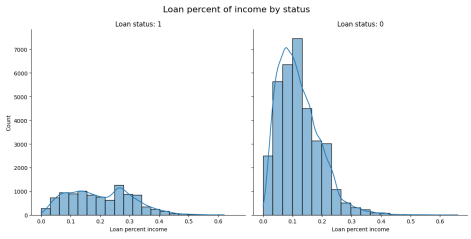

Fig. 4. Loan Amount as a Percentage of Annual Income

Fig. 4. Loan Amount as a Percentage of Annual Income This figure illustrates the relationship between loan amount and the annual income of the applicants, expressed as a percentage. It provides a clear view of how loan amounts are distributed in relation to applicants' incomes. A higher loan-to-income ratio often indicates potential financial strain, which can influence the likelihood of loan approval.

Fig. 5. Loan Status by Age Group This figure demonstrates how the loan approval status is distributed across different age groups. It reveals trends showing whether younger or older applicants are more likely to have their loans approved. Younger applicants (typically aged 20–30) may experience

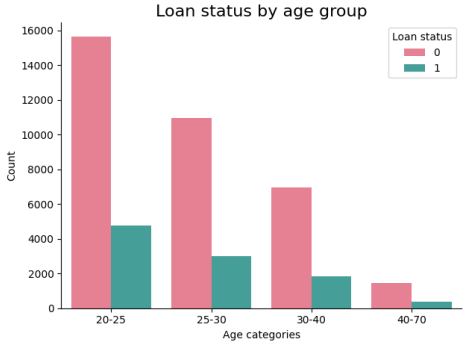

Fig. 5. Loan Status by Age Group

higher rejection rates due to factors such as less financial history or lower income stability, while older applicants often exhibit higher approval rates due to more established credit histories.

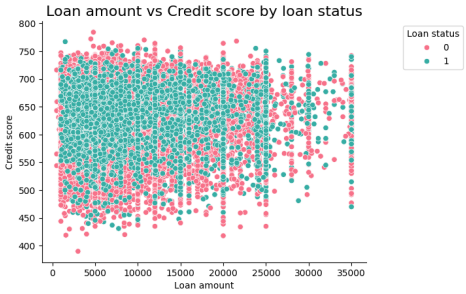

Fig. 6. Loan Amount vs Credit Score by Loan Status

Fig. 6. Loan Amount vs Credit Score by Loan Status This figure shows the relationship between the loan amount and the applicant's credit score, grouped by loan approval status. It demonstrates how applicants with higher credit scores tend to have higher loan amounts approved. The analysis highlights the importance of credit scores in determining loan eligibility[18], with applicants having lower credit scores more likely to face rejections, regardless of the loan amount.

B. Data Preprocessing

The dataset used in this study was publicly available on Kaggle. The next step was to prepare the data before applying machine learning models. Data preprocessing is crucial to ensure the quality of the dataset and to improve model performance.

Data Cleaning The first step in data preprocessing was to address missing values in the dataset. Maintaining data quality is essential for accurate predictions. For numerical features such as loan amount and annual income, missing values were replaced with their median values. For categorical variables, the missing entries were filled with the most frequent category to ensure no data loss.

Label Encoding Since machine learning algorithms require numerical inputs, categorical features were encoded into numerical values. Features such as loan purpose and education level were encoded ordinally to preserve their natural ordering. For example, loan purpose categories like "Education," "Medical," and "Personal" were encoded as 0, 1, and 2, respectively.[12]

Feature Engineering Feature engineering was applied to create new features that could enhance model accuracy. For instance, the Loan-to-Income Ratio was calculated as the ratio of the loan amount to the applicant's annual income. This feature helps evaluate the applicant's affordability and financial stability. Additionally, credit utilization rate was derived by dividing the applicant's loan amount by the credit score, which can provide more insights into creditworthiness. These newly created features enriched the dataset and improved the model's predictive power.

Class Balancing The dataset exhibited class imbalance in the target variable Loan Status, with a higher number of approved loans compared to rejected ones. To address this imbalance, an oversampling approach was used. The minority class (Rejected) was oversampled using random sampling with replacement to match the size of the majority class (Approved). This resulted in a balanced dataset where both loan approval classes were equally represented, allowing for more accurate and fair model predictions.

Standardization To ensure that numerical features are on comparable scales, the dataset was standardized. Features such as monthly income, loan amount, and credit score were standardized using StandardScaler. Standardization ensures that no single feature dominates the learning process due to its larger range, helping the model treat all features with equal importance during training.

Correlation Analysis Correlation analysis was conducted to examine the relationships between features and ensure there were no excessively strong dependencies that could introduce redundancy. For example, income and loan amount showed a moderate correlation, but they each provided unique information to the dataset. As a result, no features were dropped, as all were deemed valuable for enhancing model performance.

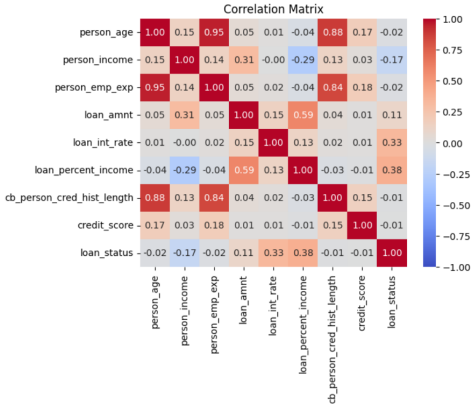

Fig. 7. Corelation Heat Map

Finally, the dataset was split into training and testing subsets, with 80% allocated for training and 20% reserved

for testing. This split ensured that the model's performance could be evaluated on unseen data, assessing its generalization capability and the effectiveness of the preprocessing steps.

IV. Model Selection

To predict loan approval outcomes, five machine learning algorithms were selected: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes. These models were trained on the preprocessed data with specific adjustments to optimize their performance.[20]

A. Logistic Regression

The Logistic Regression model was trained with the `max_iter` parameter set to 1000 for convergence. Logistic Regression estimates probabilities using the sigmoid function:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}} \tag{1}$$

where β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_n$ are feature coefficients. It predicts whether a loan is approved (1) or rejected (0).

B. Decision Tree

The Decision Tree model was trained with default parameters and `random_state` set to 42. It splits data based on Gini impurity:

$$G = 1 - \sum_{i=1}^{C} p_i^2 \tag{2}$$

where p_i is the proportion of samples in class i. Decision Trees are versatile for non-linear relationships [**?**].

C. Random Forest

The Random Forest model was trained with `random_state` set to 42. It combines predictions from multiple decision trees using:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(X) \tag{3}$$

where T is the number of trees and $h_t(X)$ is the prediction from the t-th tree. Random Forest reduces overfitting and improves accuracy.

D. K-Nearest Neighbors (KNN)

The KNN model used $k = 5$, classifying samples based on their nearest neighbors. The decision rule is:

$$f(x) = \text{argmax}_k \sum_{i \in N_k} I(y_i = k) \tag{4}$$

where N_k is the set of k-nearest neighbors. KNN is simple but less effective for high-dimensional data.

E. Naive Bayes

The Naive Bayes model assumes independence between features and uses Bayes' theorem:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \tag{5}$$

where $P(C|X)$ is the posterior probability, $P(X|C)$ is the likelihood, and $P(C)$ is the prior. Naive Bayes excels in recall but has lower precision compared to other models.

V. Model Evaluation

The Random Forest model outperformed others, achieving an accuracy of 92.69% and an ROC-AUC score of 0.9735. Exploratory data analysis revealed that low credit scores and high loan-to-income ratios are strongly correlated with loan rejections. Medical loans were most likely to be approved, reflecting their urgency.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	89.58%	77.77%	74.41%	76.02%	0.9532
Random Forest	**92.69%**	**89.59%**	75.22%	**81.77%**	**0.9735**
KNN	86.63%	73.15%	62.88%	67.62%	0.9025
Naive Bayes	74.80%	46.78%	97.68%	63.26%	0.9390

TABLE III
Model Performance Metrics

A. *Insights*

- Medical loans have the highest approval likelihood due to urgency.
- High loan-to-income ratios correlate with rejection rates.
- Gender and age categories showed distinct patterns in loan approval.

VI. Conclusion and Future Work

This study highlights the effectiveness of machine learning in loan approval prediction, with Random Forest emerging as the best performer. Future work will focus on advanced models like XGBoost and integrating interpretability tools to provide actionable insights for real-time decision-making.

References

[1] Kaggle, Loan Approval Classification Dataset.

[2] Relevant studies on credit risk prediction and machine learning algorithms.

[3] Kesraoui, A.; Lachaab, M.; Omri, A. The impact of credit risk and liquidity risk on bank margins during economic fluctuations: Evidence from MENA countries with a dual banking system. *Appl. Econ.* **2022**, *54*, 4113–4130. [Google Scholar] [CrossRef]

[4] Li, Z.; Liang, S.; Pan, X.; Pang, M. Credit risk prediction based on loan profit: Evidence from Chinese SMEs. *Res. Int. Bus. Financ.* **2024**, *67*, 102155. [Google Scholar] [CrossRef]

[5] Naili, M.; Lahrichi, Y. The determinants of banks' credit risk: Review of the literature and future research agenda. *Int. J. Financ. Econ.* **2022**, *27*, 334–360. [Google Scholar] [CrossRef]

[6] Zhang, X.; Yu, L. Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. *Expert Syst. Appl.* **2024**, *237*, 121484. [Google Scholar] [CrossRef]

[7] Abdelaziz, H.; Rim, B.; Helmi, H. The interactional relationships between credit risk, liquidity risk and bank profitability in MENA region. *Glob. Bus. Rev.* **2022**, *23*, 561–583. [Google Scholar] [CrossRef]

[8] Huang, Y.; Li, Z.; Qiu, H.; Tao, S.; Wang, X.; Zhang, L. BigTech credit risk assessment for SMEs. *China Econ. Rev.* **2023**, *81*, 102016. [Google Scholar] [CrossRef]
[9] Bhatore, S.; Mohan, L.; Reddy, Y.R. Machine learning techniques for credit risk evaluation: A systematic literature review. *J. Bank. Financ. Technol.* **2020**, *4*, 111–138. [Google Scholar] [CrossRef]
[10] Pang, M.; Li, Z. A novel profit-based validity index approach for feature selection in credit risk prediction. *AIMS Math.* **2024**, *9*, 974–997. [Google Scholar] [CrossRef]
[11] Singh, V.; Yadav, A.; Awasthi, R.; Partheeban, G.N. Prediction of modernized loan approval system based on machine learning approach. In Proceedings of the 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 25–27 June 2021; pp. 1–4. [Google Scholar] [CrossRef]
[12] Lohani, B.P.; Trivedi, M.; Singh, R.J.; Bibhu, V.; Ranjan, S.; Kushwaha, P.K. Machine learning based model for prediction of loan approval. In Proceedings of the 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 27–29 April 2022; pp. 465–470. [Google Scholar] [CrossRef]
[13] J. Tejaswini, T.M. Kavya, R.D.N. Ramya, P.S. Triveni, V.R. Maddumala, Accurate loan approval prediction based on machine learning approach, *J. Eng. Sci.* **11**, 523–532 (2020).
[14] M.A. Sheikh, A.K. Goel, T. Kumar, An approach for prediction of loan approval using machine learning algorithm, In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, 490–494 (2020).
[15] A.S. Kadam, S.R. Nikam, A.A. Aher, G.V. Shelke, A.S. Chandgude, Prediction for loan approval using machine learning algorithm, *Int. Res. J. Eng. Technol.* **8** (2021).
[16] P. Tumuluru, L.R. Burra, M. Loukya, S. Bhavana, H.M.H. CSaiBaba, N. Sunanda, Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms, In Proceedings of the 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), IEEE, 349–353 (2022).
[17] T. Ndayisenga, Bank loan approval prediction using machine learning techniques, Doctoral dissertation (2021).
[18] P.S. Murthy, G.S. Shekar, P. Rohith, G.V.V. Reddy, Loan Approval Prediction System Using Machine Learning, *J. Innov. Inf. Technol.* **21**–24 (2020).
[19] P.S. Murthy, G.S. Shekar, P. Rohith, G.V.V. Reddy, Loan Approval Prediction System Using Machine Learning, *J. Innov. Inf. Technol.* **21**–24 (2020).
[20] Y. Diwate, P. Rana, P. Chavan, Loan Approval Prediction Using Machine Learning, *Int. Res. J. Eng. Technol.* **8**, 1741–1745 (2021).