IT Journal Research and Development (ITJRD)
Vol.10, No.1, August 2025, E-ISSN : 2528-4053 | P-ISSN : 2528-4061
DOI : 10.25299/itjrd.2025.18209

Prediction of People's Business Credit Loan Acceptance Using Machine Learning Model: Application of Random Forest Algorithm

Algies Rifkha Fadillah[1], Mohamad Nurkamal Fauzan[2]
Departement of Electrical and Computer Engineering, Universitas Logistics Business International
algies.rifkha@gmail.com[1], m.nurkamal.f@ulbi.ac.id[2]

Article Info

Article history:

Received Jul 19, 2024
Revised Nov 18, 2024
Accepted Jul 18, 2025

Keyword:

Machine Learning
Random Forest
Credit Loan Acceptance
Credit Risk

ABSTRACT

This research aims to develop a Machine Learning model using the Random Forest algorithm to predict the receipt of People's Business Credit loans. These loans play a crucial role in supporting economic growth in Indonesia, especially for small and medium enterprises (SMEs). Credit risk is a major challenge in providing such loans, making accurate prediction of loan acceptance essential. This study uses a synthetic dataset consisting of 1,000 records and 13 raw attributes, which were reduced to nine predictive features through correlation analysis. The developed Random Forest model achieved an accuracy rate of 95.0 percent, with confusion matrix analysis and a classification report providing a detailed overview of the model's performance, including precision, recall, and F1-score. In addition to demonstrating strong predictive capabilities, this study addresses several practical implementation challenges. These include reliance on synthetic data calibrated from real distributions, potential feature drift in live systems, and the computational overhead associated with ensemble methods used for real-time scoring. The results of this study are expected to support better decision-making in evaluating loan applications, improving credit risk management, and optimizing the allocation of financial resources. Ultimately, the implementation of this model in operational environments is expected to contribute to economic growth and reduce credit risk effectively.

Corresponding Author:

Muhammad Nurkamal Fauzan
Departement of Electrical and Computer Engineering
Universitas Logistics Business International
Sariasih Street No. 54, Sarijadi, Sukasari, Bandung City, West Java 40151, Indonesia
Email: m.nurkamal.f@ulbi.ac.id

1. INTRODUCTION

Credit lending is considered the most important aspect of most financial institutions [1]. People's Business Credit loans are a financial product that plays a significant role in supporting economic growth in Indonesia [2]. The government program aims to provide access to capital to small and medium-sized enterprises (SMEs) that often have difficulty obtaining loans from conventional banks [3].

The government program contributes positively to improving the business continuity of SMEs [4]. The success or failure of a financial institution depends on the amount of the loan, i.e. the ability of the customer to repay the loan [5]. It cannot be denied that providing loans also carries risks for financial institutions [6]. One of the main risks faced is credit risk, which is the possibility that loan recipients will not be able to repay their loans according to a predetermined schedule [2]. Credit scoring, which helps assess a customer's ability to repay a loan, is a key issue for lending institutions [7]. Therefore, it is important to have an effective tool to predict loan acceptance.

Precise predictions are essential for maximizing profits [8]. In this case, machine learning technology becomes a powerful tool for analysis and prediction [9][10]. Building prediction models for high-risk decisions such as credit loan acceptance requires high model prediction accuracy [11]. Machine learning models can be used in the context of historical data on loan recipients to predict the chances of loan success or failure [12]. With this approach, it can improve credit risk management, identify potential borrowers more accurately, and optimize the loan portfolio.

Therefore, this research will focus on developing a machine learning model to predict loan acceptance. In this context, the Random Forest algorithm is selected due to its proven effectiveness in handling high-dimensional data, resistance to overfitting, and capability to capture complex nonlinear patterns through ensemble learning. This algorithm works by constructing multiple decision trees and aggregating their outputs, thus improving prediction accuracy and model stability. The model is trained using a synthetic dataset consisting of 1,000 records. However, the use of synthetic data [13], although necessary to address privacy and accessibility constraints, presents limitations in terms of generalizability to real-world applicants. Additionally, practical implementation of this model may face challenges such as feature drift over time, computational overhead in real-time systems, and the need for interpretability in regulated lending environments. By combining machine learning technology with SME-focused lending systems, this study aims to contribute not only to more effective credit risk management but also to broader economic empowerment. The discussion of these challenges and limitations is intended to ensure that the model is not only accurate, but also practical and adaptable to real-world financial decision-making contexts.

2. PREVIOUS RESEARCH

In LiYu's research, "Credit Risk Prediction Based on Machine Learning Methods," found that the XGBoost algorithm outperforms logistic regression in credit risk prediction [14]. In M. Sheikh's study, "Prediction of Loan Approval using Machine Learning Algorithm," logistic regression indicates that applicants with low credit scores face difficulties in loan approval, while those with high income and lower loan requests have a higher likelihood of approval [15]. P. Sanika Narayanpethkar's research on "Loan Prediction Using Machine Learning Algorithm" suggests the use of machine learning for automated loan approval, resulting in a quicker lending process and immediate approval for eligible borrowers [16]. Gupta Ashwin Arunkumar's study on "Predictive Analysis in Banking using Machine Learning" highlights the effectiveness of support vector machines (91% accuracy) and Random Forest (92% accuracy) in credit card fraud detection and loan approval prediction [17]. In Miraz Al Mamun's research, "Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis," the Gradient Boosting model achieves the highest accuracy in predicting bank loan eligibility [18].

Research results that focus on predicting credit loan acceptance have been grouped based on researchers who present the accuracy value of the various models used. Information related to the results of this research can be found in Table 1.

Table 1. Accuracy Results of Previous Research

Research	Model	Accuracy
LiYu [14]	XGBoost	**88,06%**
	Logistic Regression	83,5%
M. Sheikh [15]	Logistic Regression	**81,1%**
	K-Nearest Neighbors	82.46%

Research	Model	Accuracy
P. Sanika Narayanpethkar [16]	SVM	82.46%
	Decision Tree	75,97%
	Logistic Regression	83,76%
	Naïve Bayes	83,11%
	Random Forest	83,76%
Gupta Ashwin A. [17]	SVM	91%
	Naïve Bayes	83%
	Random Forest	92%
Miraz Al M. [18]	XgBoost	91,80%
	AdaBoost	91,87%
	LightGBM	91,89%
	Random forest	91,88%
	Decision tree	84,97%
	K-Nearest Neighbors	91,67%

Although each model shows varying accuracy levels, these results provide crucial insights for selecting the most suitable algorithm for credit risk prediction tasks. Models such as Random Forest, XGBoost, AdaBoost, and LightGBM have consistently demonstrated high performance across different datasets. However, many previous studies rely on real-world datasets that are often proprietary and lack transparency in preprocessing steps, making reproducibility and practical application in different institutional contexts challenging. In contrast, this research contributes by developing a Random Forest model based on a synthetic dataset calibrated to reflect the characteristics of actual loan applicants. This approach enables open experimentation while addressing privacy concerns. Furthermore, our study fills an important gap by evaluating not only the predictive performance but also discussing implementation constraints such as data representativeness, feature drift, and computational feasibility, which are often overlooked in prior work. Thus, this research not only confirms the effectiveness of Random Forest in credit scoring but also extends its applicability to contexts with limited access to sensitive financial data.

3. RESEARCH METHOD

This research adopts the Crisp-DM method as its system development model. Crisp-DM, which stands for Cross-Industry Standard Process for Data Mining, is a widely utilized and standardized methodology in the field of data mining [19]. It comprises six iterative phases that delineate the lifecycle of a data mining project [20]. The flow of the Crisp-DM model is illustrated in Figure 1:

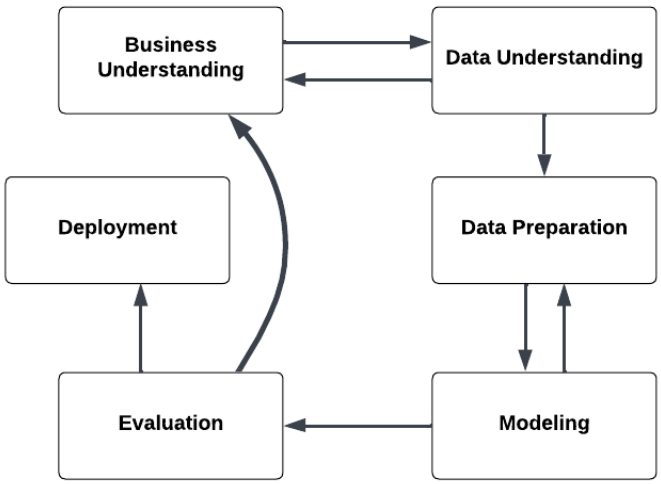

Figure 1. CRISP-DM

3.1. Business Understanding

The business understanding stage will focus on the problem to be solved, namely the prediction of loan receipts. The business objective is to improve the efficiency of loan management and reduce credit risk.

3.2. Data Understanding

Data Understanding In this stage, historical data on loan recipients is collected and understood. This data includes the predictive variables that will be used to build the prediction model. The data utilized in this analysis originates from synthetic datasets generated from real data. Python, along with libraries like Pandas, Faker, and the datetime module, was employed to create this synthetic dataset. The dataset encompasses details about customers, loans, and other relevant parameters, comprising a total of 1000 data entries and 13 attributes. This approach allows for experimentation without compromising data privacy. However, the use of synthetic data presents limitations, particularly in capturing the full variability and complexity of real-world borrower behavior. To mitigate noise and redundancy, a correlation analysis was conducted to select nine key predictive features. Table 2 shows the description of the dataset attributes used.

Table 2. Dataset Attributes

Attributes	Description
Credit_Number	Unique identification number for each credit, used to distinguish one credit from another.
CIF_Number	Unique identification number for each customer or client.
Customer_Name	Full name of the customer associated with the credit.
Marital_Status	Marital status of the customer, such as "Married," "Single," or other marital status types.
Number_of_Dependents	Number of dependents or family members that the customer is financially responsible for.
Customer_Income	Monthly income of the customer, a crucial factor in assessing their ability to repay the loan.
Credit_History	Customer's credit history, indicating whether it is Good and Bad Credit History
Loan_Amount	The amount of the loan requested by the customer or the total loan amount granted.
Application_Date	Date when the credit was applied for by the customer
Credit_Date	Date when the credit was approved or the loan amount disbursed.
Due_Date	Due date or the deadline for loan repayment.
Tenure	The loan tenure or the period over which the loan is to be repaid, for example, 12 months, 24 months, etc.
Loan_Status	Final status of the loan, such as "Approved" or "Rejected."

Data filling involves a random process guided by specific rules. Marital status is determined probabilistically, the number of dependents is associated with marital status, and customer income is set within the original data's maximum and minimum values. Other attributes like credit history, loan amount, and loan status are determined based on conditions from the original data. The resulting dataset is structured in a DataFrame format using Pandas, facilitating analysis and further data processing.

Figure 2. Comparison of Customer Loan Receipt Status

In Figure 2, shows the results of data exploration, as much as 494 data has 'N' status which means the loan application is not accepted and 506 data has 'Y' status which means the loan application can be accepted.

3.3. Data Preparation

In the data preparation stage, the focus is on ensuring that the historical data to be used for analysis is properly prepared to be suitable for model training. The data preparation process involves several key steps, including:

a. **Data Cleaning:** Identification and handling of missing values, outliers, or anomalies in the data. This may involve filling in missing values or removing invalid or irrelevant data.
b. **Data Transformation:** Processing of variables that may require transformation, such as normalization to rescale variables or encoding to manage categorical variables.
c. **Variable Selection:** Selection of variables to be used in the analysis. This process may involve statistical analysis or other variable selection methods to select the most relevant and influential variables for prediction. The following Figure 3 shows the correlation matrix generated from the analysis of the dataset.

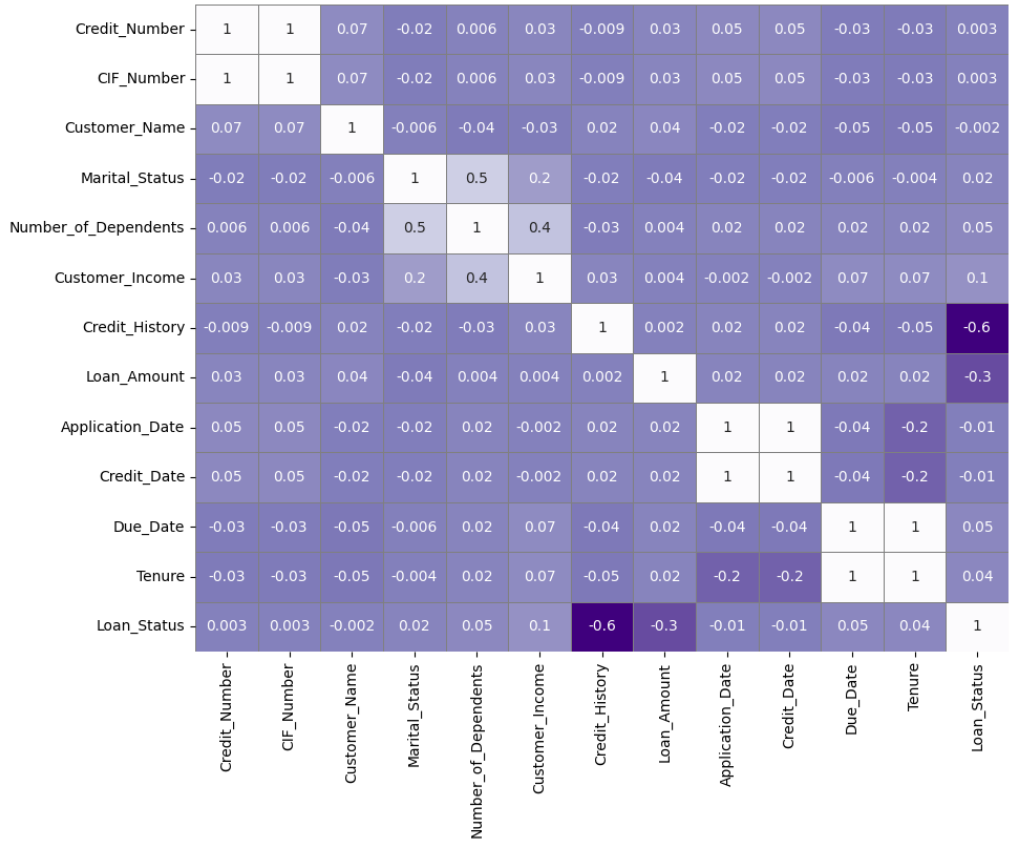

Figure 3. Correlation Matrix

From Figure 3, the features Credit_Number, CIF_Number, Customer_Name are selected for deletion. Marital_Status, Number_of_Dependents, Customer_Income, Credit_History, Loan_Amount, Application_Date, Credit_Date, Due_Date, and Tenure are the nine selected attributes that will be used in prediction.

d. Data Split: The division of data into a training set and a testing set to test the performance of the model on data not used in training. The separation is divided into 80% for training data and 20% for testing data, respectively.

3.4. Modeling

In the modeling stage, the primary objective is to create a machine learning model for predicting loan acceptance. This involves selecting and fine-tuning predictive variables to maximize accuracy. Random Forest is a suitable machine learning algorithm for this purpose. Random Forest is a versatile machine learning algorithm employed for data classification, regression, and clustering tasks [21]. This approach constructs numerous decision trees during training and consolidates their predicted results to generate the final output [22]. The algorithm operates by comprehending the decision tree, a fundamental component of the Random Forest algorithm that forms a tree-shaped structure [23][24]. In predicting loan acceptance, Random Forest proves beneficial by discerning intricate patterns in historical data, thereby enhancing the model's accuracy in making predictions. While this study does not compare Random Forest directly with other algorithms, its selection is grounded in prior literature showing consistent performance in similar contexts. Further research could explore algorithmic benchmarking using real-world datasets to complement this synthetic-data approach.

3.5. Evaluation

In the evaluation stage, the emphasis is on a comprehensive assessment of the developed model's performance, specifically in predicting loan acceptance. The evaluation process aims to gauge the model's effectiveness in accurately predicting loan acceptance.

3.6. Deployment

In the implementation phase, the verified and assessed machine learning model will be implemented in an operational environment. The model will be an important tool in optimizing the company's loan portfolio and assisting in identifying more discerning potential borrowers.

4. RESULTS AND ANALYSIS

Based on the results of the machine learning model research to predict the acceptance of People's Business Credit loans, the implementation of the Random Forest model has an accuracy rate of 95%. Random Forest provides the potential to increase the efficiency and accuracy of assessing the customer's ability to repay loans.

4.1. Classification Report

The classification report offers a detailed assessment of the model's performance for each identified class, including statistical evaluations such as precision, recall, and F1-score. These metrics provide a more nuanced understanding of the model's effectiveness in handling different classes within the dataset.

Precision (P), Recall (R), F1-score (F1) and Accuracy are calculated as follows:

Precision: $\frac{TP}{TP + FP}$ Recall: $\frac{TP}{TP + FN}$ F1−score: $\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$ Accuracy: $\frac{TP + TN}{TP + TN + FP + FN}$

Where:

- TP (True Positive) is the number of correct positive predictions.
- TN (True Negative) is the number of correct negative predictions.
- FP (False Positive) is the number of incorrect positive predictions.
- FN (False Negative) is the number of incorrect negative predictions.

In the following, there is a Classification Report that illustrates the results of using the Random Forest model, as shown in Figure 4.

Random Forest Classification Report:

	precision	recall	f1-score	support
0.0	0.98	0.92	0.95	103
1.0	0.92	0.98	0.95	97
accuracy			0.95	200
macro avg	0.95	0.95	0.95	200
weighted avg	0.95	0.95	0.95	200

Figure 4. Classification Report

a. **Precision**: For class 0.0, the precision is 98%, indicating how precise the model is in classifying data that actually belongs to that class. For class 1.0, the precision is 92%.
b. **Recall**: Recall measures the extent to which the model can identify all actual cases. The model has a recall of 92% for class 0.0 and 98% for class 1.0.
c. **F1-Score**: F1-Score, as the harmonic mean of precision and recall, reached 95% for both classes. This indicates a good balance between precision and recall.

In addition to precision, recall, and F1-score, it is important to highlight the model's balanced performance across the different classes. The high precision for class 0.0 indicates that the model is very reliable in identifying rejected loans with minimal false positives. However, the recall for class 0.0 is slightly lower, meaning there are some rejected loans that the model failed to detect, leading to false negatives. Conversely, for class 1.0, the model shows a stronger recall, which signifies that most accepted loans are correctly identified, minimizing the risk of missing out on valid loan applications. This balance between precision and recall across both classes highlights the Random Forest model's robustness and its potential to minimize risk in a real-world loan approval system.

These results highlight that the model performs slightly better in predicting accepted loans, which may be attributed to the higher recall for class 1.0. This demonstrates that while the model is effective, there is still room for improvement in reducing false positives, especially for class 0.0

4.2. Confusion Matrix

The confusion matrix provides a detailed overview of the model's predictions for each class. The confusion matrix includes information about the number of true positives, true negatives, false positives, and false negatives for each class, which is very useful in evaluating the model's performance in more depth. In the following, the Confusion matrix illustrates the results about the predictions of the Random Forest model for each class, as shown in Figure 5.

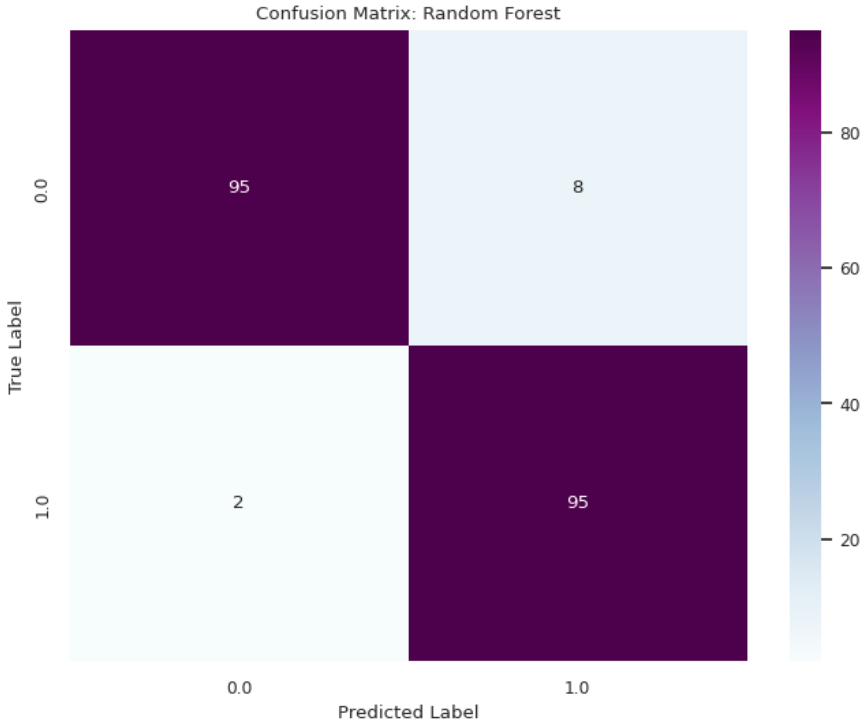

Figure 5. Confusion Matrix

a. Out of 103 cases in class 0.0, the model correctly predicted 95 cases and incorrectly predicted 8 cases in class 1.0.
b. Out of 97 cases in class 1.0, the model correctly predicted 95 cases and incorrectly predicted 2 cases in class 0.0.

These results suggest that the model is particularly effective in predicting accepted loans, likely due to the higher recall observed in class 1.0. While this is beneficial in minimizing the risk of overlooking eligible applicants, the slightly lower recall for class 0.0 means that some risky applications may be incorrectly classified as acceptable. This poses a potential challenge for loan risk mitigation, as false negatives (i.e., risky applicants misclassified as eligible) can contribute to an increase in non-performing loans.

A critical analysis of these outcomes reveals that although the model performs well overall, it could be further improved by addressing its vulnerability to class imbalance and borderline decision cases. The reliance on synthetic data, although beneficial for experimentation, may not fully capture the nuances of real-world credit behavior, particularly in borderline approval scenarios where contextual factors play a larger role.

As a next step, future enhancements could include fine-tuning the classification threshold to reduce false negatives in class 0.0, applying cost-sensitive learning to assign higher penalties for misclassifying risky applications, and expanding the dataset with real-world examples to improve generalization. Additionally, implementing explainable AI techniques (e.g., SHAP or LIME) could help improve interpretability and trustworthiness of the model in real deployment settings. These steps would further strengthen the model's applicability in operational credit evaluation systems and ensure more reliable outcomes in practice.

5. CONCLUSION

This research focuses on developing a machine learning model using the Random Forest algorithm to predict loan acceptance. The model has an accuracy rate of 95.0%, demonstrating its ability to correctly classify the test data. Confusion matrix analysis provides a more detailed picture of the model's performance for each class, with good precision, recall, and F1-score for classes 0.0 and 1.0. The model is expected to assist in optimizing loan portfolios, improving credit risk management, and providing more accurate predictions of loan receipts, which in turn can support economic growth and reduce credit risk.

REFERENCES

[1] A. K. Sarthak Choudhary, Chandra Shekhar Tyagi, Sakshi Gaur Choudhary, "Loan Payment Date Prediction Model Using Machine Learning Regression Algorithms," *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, vol. 01, pp. 1–6, 2022, doi: https://www.doi.org/10.1109/ICDSAAI55433.2022.10028838.

[2] Moh. R. A. Velin Diani, "Implementasi produk pembiayaan kredit usaha rakyat (kur) mikro dalam mengembangkan usaha mikro nasabah perbankan syariah," *JOURNAL OF MANAGEMENT Small and Medium Enterprises (SME's)*, vol. 16, no. 2, pp. 299–311, 2023, doi: https://www.doi.org/10.35508/jom.v16i2.8515.

[3] R. M. Suryani I, "Pengaruh Program KUR dan BLT terhadap Kinerja UMKM dengan Strategi Diferensiasi sebagai Variabel Mediasi," *Jurnal Manajemen dan Keuangan*, vol. 12, no. 1, pp. 195–213, 2023, doi: https://dx.doi.org/10.33059/jmk.v12i1.5310.

[4] Suhaila Zulkifi, Naomi Christine, Steven Lim, and Wilfredo Leeson, "the Provision of Unsecured People'S Business Credit (Kur) Based on Law No. 10 of 1998 (Study At Pt. Bank Bri Pulo Brayan Branch Unit I)," *Awang Long Law Review*, vol. 5, no. 1, pp. 134–138, 2022, doi: 10.56301/awl.v5i1.543.

[5] V. Singh, A. Yadav, R. Awasthi, and G. N. Partheeban, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach," *2021 International Conference on Intelligent Technologies, CONIT 2021*, pp. 21–24, 2021, doi: 10.1109/CONIT51480.2021.9498475.

[6] C. Badarau and I. Lapteacru, "Bank Risk, Competition and Bank Connectedness with," *Res Int Bus Finance*, 2019, doi: 10.1016/j.ribaf.2019.03.004.

[7] G. Arutjothi and C. Senthamarai, "Credit Risk Analysis Using Fuzzy Logic with Machine Learning Models," *International Journal For Multidisciplinary Research*, vol. 5, no. 3, pp. 1–7, 2023, doi: https://www.doi.org/10.36948/ijfmr.2023.v05i03.3298.

[8] Z. Dong, "A study on personal credit risk assessment model based on integrated learning," *East China University of Science and Technology*, vol. 32, 2022, doi: https://www.doi.org/10.54691/bcpbm.v32i.2994.

[9] M. Rahmaty, "Machine learning with big data to solve real-world problems," *Journal of Data Analytics*, vol. 2, no. 1, pp. 9–16, 2023, doi: 10.59615/jda.2.1.9.

[10] M. Grebovic, M. Vukotic, L. Filipovic, T. Popovic, and I. Katnic, "Machine Learning Models for Statistical Analysis," *The International Arab Journal of Information Technology*, vol. 20, no. 3, pp. 505–514, 2023, doi: https://www.doi.org/10.34028/iajit/20/3a/8.

[11] Md. G. Kibria and M. Sevkli, "Application of Deep Learning for Credit Card Approval: A Comparison with Two Machine Learning Techniques," *Int J Mach Learn Comput*, vol. 11, no. 4, pp. 286–290, 2021, doi: 10.18178/ijmlc.2021.11.4.1049.

[12] N. H. G. Dr. SHANKARAGOWDA B B, "Loan Default Prediction Using Machine Learning Techniques 1," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 07, no. 07, pp. 2021–2024, 2023, doi: 10.55041/IJSREM24519.

[13] S. M. Cherian and K. Rajitha, "Random forest and support vector machine classifiers for coastal wetland characterization using the combination of features derived from optical data and synthetic aperture radar dataset," *Journal of Water and Climate Change*, vol. 15, no. 1, 2024, doi: 10.2166/wcc.2023.238.

[14] Y. Li, "Credit risk prediction based on machine learning methods," *14th International Conference on Computer Science and Education, ICCSE 2019*, no. Iccse, pp. 1011–1013, 2019, doi: 10.1109/ICCSE.2019.8845444.

[15] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, no. Icesc, pp. 490–494, 2020, doi: 10.1109/ICESC48915.2020.9155614.

[16] U. S. Pandey N, Gupta R, "Loan Approval Prediction using Machine Learning Algorithms Approach," *International Journal of Innovative Research in Technology*, vol. 8, no. 1, pp. 898–902, 2021.

[17] Gupta Ashwin Arunkumar, Chaurasiya Ravi Panchuram, Khambati Mohammed Aadil Afzal, Niraj Sureshchand Yadav, and Uma Goradiya, "Predictive Analysis in Banking using Machine Learning," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3307, pp. 434–439, 2023, doi: 10.32628/cseit2390247.

[18] M. Al Mamun, "Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis," *IEOM Society International*, pp. 1423–1432, 2023, doi: 10.46254/na07.20220328.

[19] W. Chang, "An Integrated Development Model of Cultural and Creative Industry and Rural Tourism Industry Based on Data Mining," *Wirel Commun Mob Comput*, vol. 2022, 2022, doi: 10.1155/2022/2351906.

[20] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[21] J. Aditya Fajar Mulyana, Wiyanda Puspita, "Classification to predict student academic performance using a Random Forest," *Journal of Student Research Exploration*, vol. 1, no. 2, pp. 94–103, 2023.

[22] E. Darnila, M. Maryana, K. Mawardi, M. Sinambela, and I. Pahendra, "Supervised models to predict the Stunting in East Aceh," *International Journal of Engineering, Science and Information Technology*, vol. 2, no. 3, pp. 33–39, 2022, doi: 10.52088/ijesty.v2i3.280.

[23] M. Ibrahim, "Evolution of Random Forest from Decision Tree and Bagging: A Bias-Variance Perspective," *Dhaka University Journal of Applied Science and Engineering*, vol. 7, no. 1, pp. 66–71, 2023, doi: 10.3329/dujase.v7i1.62888.

[24] W. W. Hsieh, *Decision Trees, Random Forests and Boosting*. Cambridge University, 2023. doi: https://www.doi.org/10.1017/9781107588493.015.