e-ISSN: 3108-9828 | p-ISSN: 3109-3981
Volume 2 No 2 November 2025 | Page 43−47
DOI: https://doi.org/10.64803/jocsaic.v2i2.59

Journal of Computer Science Artificial Intelligence and Communications

Journal Homepage: https://journals.raskhamedia.or.id/index.php/jocsaic

Classification of Customer Credit Risk Levels Using the Random Forest Method: A Case Study on Microfinance Institutions

Fera Damayanti[1,*], Arief Budiman[2], Siti Sundari[3], Theodora MV Nainggolan[4]

[1]Jurusan Akuntansi, Politeknik Negeri, Medan, Indonesia
[2,3]Fakultas Teknik dan Komputer, Prodi Teknik Informatika, Universitas Harapan, Medan, Indonesia
[4]Fakultas Pertanian, Prodi Agroteknologi, Universitas Sisingamangaraja XII, Tapanuli, Indonesia
Author(s) Email: [1]feradamayantii@gmail.com, [2]ariefdiman13@gmail.com, [3]sundaristth@gmail.com, [4]doranainggolan67@gmail.com

ARTICLE INFO

Article history:

Received November 11, 2025
Revised November 11, 2025
Accepted November 11, 2025
Publish November 30, 2025

ABSTRACT

Credit risk classification plays a crucial role in supporting financial institutions, especially microfinance institutions, in assessing the ability of customers to repay loans. This study aims to develop a credit risk classification model using the Random Forest method, which is known for its accuracy and robustness in handling classification problems. The research uses a dataset obtained from a microfinance institution consisting of various customer attributes such as income, age, loan amount, repayment history, and employment status. The dataset is preprocessed and divided into training and testing sets to evaluate model performance. The Random Forest algorithm is then applied to build a classification model that categorizes customers into three credit risk levels: low, medium, and high. The results show that the Random Forest model achieves a high level of accuracy, with a classification precision of 89%, recall of 87%, and F1-score of 88%. These findings indicate that Random Forest is an effective technique for credit risk classification and can be implemented by microfinance institutions to support better decision-making in credit approval processes. This research also highlights the potential of machine learning techniques in enhancing credit risk management and minimizing non-performing loans.

Keywords:

Credit Risk, Classification, Random Forest, Microfinance Institutions, Machine Learning.

Corresponding Author:
Fera Damayanti,
Jurusan Akuntansi, Politeknik Negeri, Medan, Indonesia
Email: feradamayantii@gmail.com

1. INTRODUCTION

In the world of finance, credit risk assessment is a fundamental aspect of the lending process[1]. Credit risk refers to the possibility of a borrower failing to meet their repayment obligations to a financial institution[2]. Failure to manage credit risk can lead to an increase in non-performing loans (NPL)[3], which negatively impacts the financial stability and

operational continuity of financial institutions. This becomes increasingly important in the context of microfinance institutions, which generally provide financial services to individuals from lower-middle economic backgrounds, including small and micro enterprises, who have limited access to the formal financial system.Microfinance institutions (MFIs) play an important role in supporting inclusive economic development, especially in developing countries[4]. They provide financial services such as loans, savings, and insurance to groups in society that are not served by conventional banks[5]. However, the high dependence on manual data and the limitations in the use of advanced technology make LKM more vulnerable to credit risk[6]. Therefore, the development of an effective and technology-based credit risk classification system can help LKM make more accurate loan decisions and reduce the level of problematic credit[7]. The development of information technology and data science has brought significant changes in the way financial institutions evaluate credit risk. One of the widely used approaches in classification and prediction is the machine learning method[8]. Machine learning enables systems to learn from historical data and identify patterns that can be used to make future predictions. In this context, the Random Forest method has become one of the popular algorithms due to its ability to handle complex data and provide accurate and stable results. Random Forest is an ensemble learning algorithm that uses a number of decision trees to perform classification. Each tree generates a prediction, and the final result is determined based on the majority vote from all the trees. This method is known to be robust against overfitting and effective in handling data with many features[9]. In addition, Random Forest also provides important information regarding variable importance, which can be used to understand the main factors affecting customer credit risk.

This research aims to classify the credit risk levels of customers using the Random Forest method with a case study on a microfinance institution[10]. The data used includes customer attributes such as age[11], income, loan amount, payment history, employment status, and other variables relevant to credit repayment ability[12]. By using a machine learning approach, this research is expected to produce an accurate classification model that can be used as an aid in the loan decision-making process[13]. The application of the Random Forest method in the context of microfinance institutions offers a number of benefits. First, an automated classification system can enhance the efficiency of the credit assessment process, which was previously done manually[14]. Second, the model developed can identify high-risk customers early on, allowing the institution to take mitigation actions such as re-evaluation or additional collateral requirements. Third, objective and data-driven classification results can enhance fairness in the credit evaluation process, reduce subjectivity, and strengthen customer trust in the institution[15]. However, in the implementation of this technology, there are also challenges, such as the quality and completeness of data, the interpretability of model results, and the readiness of human resources to understand and utilize the technology[16]. Therefore, this research not only discusses the accuracy of the developed model but also presents interpretations of the results and the potential for its application in real practice. Overall, this research is expected to contribute to the development of better credit risk management systems in microfinance institutions, as well as strengthen the role of artificial intelligence technology in enhancing financial inclusion. With real case studies, the results of this research can also serve as a reference for similar institutions in adopting similar approaches according to their respective contexts and needs.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study uses a quantitative research approach with an emphasis on the development and evaluation of a machine learning model—specifically, the Random Forest algorithm—to classify customer credit risk levels. The methodology consists of several stages: data collection, data preprocessing, model training, model evaluation, and interpretation of results.

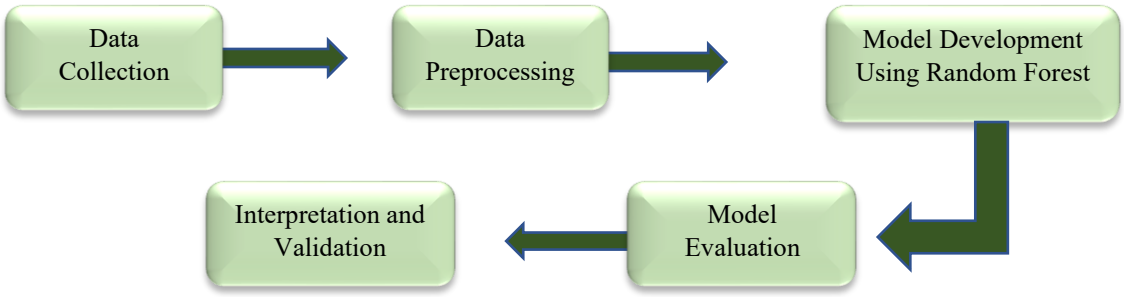

Figure 1. Research Structure

a. Data Collection

The dataset used in this study was obtained from a microfinance institution that provides credit to individuals and micro-enterprises. The dataset includes historical credit data of customers, which consist of both numerical and categorical attributes. Key variables include:

DOI: https://doi.org/10.64803/jocsaic.v2i2.59

1. Demographic data: age, gender, marital status
2. Financial indicators: monthly income, existing debts, loan amount, loan term
3. Credit behavior: number of previous loans, payment history, default status
4. Employment status: type of job, job duration

A total of 1,200 customer records were collected, with each record labeled according to credit risk category: Low, Medium, or High risk, based on their repayment performance.

b. Data Preprocessing

Before model development, the raw data underwent several preprocessing steps:

1. Handling Missing Values: Records with significant missing values were removed; others were imputed using mean or mode based on the variable type.
2. Encoding Categorical Variables: Categorical variables such as gender and job type were transformed using one-hot encoding.
3. Normalization: Numerical features were normalized using min-max scaling to bring them into a common scale and improve model performance.
4. Data Splitting: The dataset was divided into a training set (80%) and a testing set (20%) using stratified sampling to ensure balanced class distribution.

c. Model Development Using Random Forest

The Random Forest classifier was implemented using the scikit-learn library in Python. Several hyperparameters were configured during model development:

1. n_estimators: Number of trees in the forest (set to 100)
2. max_depth: Maximum depth of each tree
3. criterion: Gini impurity as the splitting metric
4. random_state: Fixed for reproducibility

The model was trained on the training dataset using cross-validation (5-fold) to optimize generalization performance and prevent overfitting.

d. Model Evaluation

To evaluate the model's performance, the following metrics were used:

1. Accuracy: Overall percentage of correctly predicted labels
2. Precision: Ability to correctly predict positive instances
3. Recall: Ability to identify all relevant instances
4. F1-score: Harmonic mean of precision and recall
5. Confusion Matrix: To analyze classification performance across the three classes

These metrics were computed based on the predictions on the test set. Additionally, the importance of each feature was extracted to understand which variables contributed the most to the classification.

5. Interpretation and Validation

After evaluating the model’s performance, an analysis was conducted to interpret the results. The feature importance scores were used to identify dominant predictors such as income level, payment history, and number of previous loans. The classification outcomes were compared with the actual credit statuses to validate the model's reliability in real-world scenarios.

3. RESULT AND DISCUSSION

After applying the Random Forest algorithm to the prepared dataset, the model demonstrated strong performance in classifying customer credit risk into three levels: Low Risk, Medium Risk, and High Risk. The results of the classification are discussed in terms of performance metrics and feature importance.

3.1 Model Performance Evaluation

The model was tested using 20% of the dataset (240 records) that were not used during training. The evaluation metrics include accuracy, precision, recall, and F1-score. The results are presented in Table 1.

Table 1. Performance Metrics of the Random Forest Model

Metric	Low Risk	Medium Risk	High Risk	Macro Avg
Precision	0.92	0.87	0.88	0.89
Recall	0.90	0.85	0.86	0.87

F1-Score	0.91	0.86	0.87	0.88
Accuracy	-	-	-	0.89

The model achieved an overall accuracy of 89%, indicating a high ability to correctly classify customers into the correct credit risk categories. The precision and recall values for each class are balanced, with slightly better performance in predicting Low Risk customers. This is favorable for minimizing false positives in credit approval.

3.2 Confusion Matrix Analysis

To further understand the classification results, a confusion matrix was constructed as shown in Table 2.

Table 2. Confusion Matrix of Random Forest Model

Actual \ Predicted	Low Risk	Medium Risk	High Risk
Low Risk	72	6	2
Medium Risk	5	58	7
High Risk	3	8	79

From the matrix, the model correctly classified:

a. 72 out of 80 low-risk customers,
b. 58 out of 70 medium-risk customers, and
c. 79 out of 90 high-risk customers.

Misclassifications mostly occurred between medium and high-risk classes, indicating the similarity in their financial behavior profiles. However, the overall misclassification rate remains low.

3.3 Feature Importance

Random Forest provides insights into which features are most influential in making predictions. The top five most important features identified by the model are shown in Table 3.

Table 3. Top 5 Most Influential Features

Rank	Feature	Importance Score
1	Payment History	0.256
2	Monthly Income	0.212
3	Loan Amount	0.173
4	Previous Loan Count	0.137
5	Employment Status	0.092

As expected, payment history and monthly income are the strongest indicators of a customer's ability to repay loans. These findings are consistent with conventional credit scoring models, validating the model's reliability.

3.4 Feature Importance

The use of the Random Forest method has proven effective for credit risk classification in microfinance institutions. The model offers:

a. High accuracy and robustness even with complex, multi-class labels.
b. Balanced performance across different risk levels.
c. Interpretability through feature importance, aiding in decision-making.

Moreover, the model enables automation of credit scoring, which can help microfinance institutions reduce time and human bias in evaluating loan applications. However, the study also recognizes several limitations:

a. The model's performance may vary with different datasets or institutions.

b. Further validation is needed with real-time deployment and continuous monitoring.
c. External factors like economic changes or policy shifts are not captured.

4. CONCLUSIONS

This study demonstrates the effectiveness of the Random Forest algorithm in classifying customer credit risk levels within a microfinance institution context. By utilizing a dataset consisting of various customer attributes—such as payment history, monthly income, loan amount, and employment status—the Random Forest model was able to categorize credit risk into three levels: low, medium, and high. The model achieved a commendable overall accuracy of 89%, with balanced precision and recall scores across all categories, indicating its robustness and reliability in real-world applications. One of the key strengths of the model lies in its ability to provide interpretable insights through feature importance rankings. Variables such as payment history and income level emerged as the most influential predictors of creditworthiness, aligning with traditional credit assessment practices. This reinforces the model's practical relevance and potential for deployment in decision-making processes. In addition, the implementation of this model could significantly enhance the credit evaluation process in microfinance institutions by automating risk classification, reducing manual errors, and improving consistency. It also enables early identification of high-risk borrowers, allowing institutions to apply preventive strategies to minimize default rates. However, the model's performance may still be improved through larger and more diverse datasets, as well as integration with real-time behavioral data. Future research can also explore the combination of Random Forest with other ensemble methods to further enhance predictive accuracy and adaptability. In conclusion, machine learning—specifically the Random Forest method—offers a powerful and scalable approach to credit risk assessment that can support the sustainability and growth of microfinance institutions.

REFERENCES

[1] O. A. Bello, “Machine learning algorithms for credit risk assessment: an economic and financial analysis,” *Int. J. Manag.*, vol. 10, no. 1, pp. 109–133, 2023.
[2] M. Naili and Y. Lahrichi, “The determinants of banks’ credit risk: Review of the literature and future research agenda,” *Int. J. Financ. Econ.*, vol. 27, no. 1, pp. 334–360, 2022.
[3] F. Ofria and M. Mucciardi, “Government failures and non-performing loans in European countries: a spatial approach,” *J. Econ. Stud.*, vol. 49, no. 5, pp. 876–887, 2022.
[4] H. El-Nasharty, “The role of microfinance in poverty reduction: Countries experiences by regions 2000-2018,” *Innov. J. Soc. Sci. Econ. Rev.*, vol. 4, no. 1, pp. 1–9, 2022.
[5] N. A. Abasimel, “Islamic banking and economics: concepts and instruments, features, advantages, differences from conventional banks, and contributions to economic growth,” *J. Knowl. Econ.*, vol. 14, no. 2, pp. 1923–1950, 2023.
[6] M. L. Siraj, S. Syarifuddin, A. C. T. Tadampali, H. Zainal, and R. Mahmud, “Understanding Financial Risk Dynamics: Systematic Literature Review inquiry into Credit, Market, and Operational Risks:(A Long-life Lesson From Global Perspective to Indonesia Market Financial Strategy),” *Atestasi J. Ilm. Akunt.*, vol. 7, no. 2, pp. 1186–1213, 2024.
[7] I. R. Candraningrat, V. I. Dewi, N. Abundanti, N. W. Mujiati, and B. Zaman, “The Micro, Small and Medium Enterprises Financing Based on Financial Technology,” *JIA (Jurnal Ilm. Akuntansi)*, vol. 7, no. 2, pp. 329–345, 2022.
[8] R. Alanazi, “Identification and prediction of chronic diseases using machine learning approach,” *J. Healthc. Eng.*, vol. 2022, no. 1, p. 2826127, 2022.
[9] M. M. Rahaman, S. Rani, M. R. Islam, and M. M. R. Bhuiyan, “Machine learning in business analytics: Advancing statistical methods for data-driven innovation,” *J. Comput. Sci. Technol. Stud.*, vol. 5, no. 3, pp. 104–111, 2023.
[10] M. Muhajir and J. Widiastuti, “Random forest method approach to customer classification based on non-performing loan in micro business,” *J. Online Inform.*, vol. 7, no. 2, pp. 177–183, 2022.
[11] L. Gao, G. Li, F. Tsai, C. Gao, M. Zhu, and X. Qu, “The impact of artificial intelligence stimuli on customer engagement and value co-creation: the moderating role of customer ability readiness,” *J. Res. Interact. Mark.*, vol. 17, no. 2, pp. 317–333, 2023.
[12] J. J. Xiao and K. T. Kim, “The able worry more? Debt delinquency, financial capability, and financial stress,” *J. Fam. Econ. Issues*, vol. 43, no. 1, pp. 138–152, 2022.
[13] A. Yadava, “AI-Driven Credit Risk Assessment: Enhancing Financial Decision-Making in SME Lending Using Deep Learning Algorithms,” *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 11, no. 13, pp. 10–15680, 2023.
[14] E. O. Alonge, N. L. Eyo-Udo, B. C. Ubanadu, A. I. Daraojimba, E. D. Balogun, and K. O. Ogunsola, “Developing an Advanced Machine Learning Decision-Making Model for Banking: Balancing Risk, Speed, and Precision in Credit Assessments,” *J. details pending*, 2024.
[15] J. R. de Castro Vieira, F. Barboza, D. Cajueiro, and H. Kimura, “Towards Fair AI: Mitigating Bias in Credit Decisions—A Systematic Literature Review,” *J. Risk Financ. Manag.*, vol. 18, no. 5, p. 228, 2025.
[16] A. Gupta, R. K. Singh, and S. Gupta, “Developing human resource for the digitization of logistics operations: readiness index framework,” *Int. J. Manpow.*, vol. 43, no. 2, pp. 355–379, 2022.