International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-14, Issue-1, January 2026
DOI: https://doi.org/10.55524/ijircst.2026.14.1.14
Article ID IRP-1714, Pages 122-126
www.ijircst.org

Machine Learning-Based Credit Risk Classification Using German Credit Data

Fitria [1], **Emy Iryanie** [2], **Heldalina** [3], **Muhammad Syahid Pebriadi** [4], **Anhar Khalid** [5]

[1, 2, 3, 4] Department of Accounting Information System, Politeknik Negeri Banjarmasin, Banjarmasin, Indonesia
[5] Department of Mechanical Engineering, Politeknik Negeri Banjarmasin, Banjarmasin, Indonesia

Correspondence should be addressed to Fitria; fitria@poliban.ac.id

Received: 22 December 2025 Revised: 6 January 2026 Accepted: 22 January 2026

ABSTRACT- Credit risk assessment is an important process in financial institutions to minimize potential losses due to non-performing loans. This study aims to apply a machine learning approach in classifying credit risk using the German Credit Dataset. The research method includes exploratory data analysis, data preprocessing, and the application of the Logistic Regression algorithm as a classification method. The dataset consists of numerical and categorical attributes that represent the financial and demographic characteristics of the credit applicant. The test results showed that the built model produced an accuracy rate of 66% on the test data, with a recall value for the high-risk class of 68%. These results suggest that machine learning approaches can be used as a decision support system in credit risk assessment, although further development is still needed to improve classification performance.

KEYWORDS- Credit Risk, Machine Learning, Classification, Logistic Regression, German Credit Dataset

I. INTRODUCTION

Credit risk management has a very important role in maintaining the sustainability of financial institution operations. One of the main challenges faced by banks and finance institutions is identifying credit applicants who have the potential to default. The conventional credit scoring process generally still relies on manual evaluations and certain rules, so it is less able to capture complex patterns in credit data.

The development of information technology encourages the use of machine learning techniques in credit risk analysis [1], [2], [3]. Machine learning is capable of processing large amounts of data and identifying hidden patterns that can improve the quality of decision-making [4], [5]. Various studies have shown that the classification approach can be used to distinguish between low-risk and high-risk credit applicants more systematically.

Based on this background, this study applied the Logistic Regression algorithm to classify credit risk using the German Credit Dataset [6]. Logistic Regression was chosen because of its simple and easy-to-interpret nature, making it suitable for use in the context of financial decision-making [7], [8], [9]. It is hoped that the results of this study can provide an overview of the use of machine learning as a tool in credit risk assessment.

II. RESEARCH METHODOLOGY

This study uses a quantitative approach with a supervised machine learning method to classify credit risk. The research stages are systematically arranged which include data collection, exploratory data analysis (EDA), data preprocessing, classification modeling, and model performance evaluation.

A. Research Dataset

The dataset used in this study is the German Credit Dataset obtained from the Kaggle platform [6]. This dataset consists of 1000 credit applicant data with 20 predictor attributes that reflect the demographic characteristics and financial condition of customers, as well as one target attribute that represents credit risk, namely the good and bad categories.

B. Exploratory Data Analysis (EDA)

The exploratory data analysis stage is carried out to understand the characteristics and quality of the data before the modeling process. This analysis includes examining the data structure, identifying data types, checking for missing values, and analyzing the distribution of the target class. The EDA results showed that the dataset had no lost value, but the distribution of the target class was unbalanced, where the amount of good credit data was greater than the bad credit. This condition shows that there is a class imbalance problem that needs to be considered at the modeling stage. In addition, an analysis of numerical and categorical attributes was carried out to understand the pattern of relationships between features and credit risk. The results of this analysis are used as a basis for determining the appropriate data preprocessing method.

C. Pre-processing Data

In the pre-processing stage of the data, categorical attributes are converted into numerical form using the one-hot encoding technique. The target variable is converted into numerical form for classification modeling purposes. This process aims to ensure that all data can be processed by machine learning algorithms optimally. The dataset was then divided into training data and test data with a comparison of 80% of the training data and 20% of the test data. Data distribution was carried out using stratified sampling techniques to maintain the proportion of target classes in both subsets of data, so that the class distribution remained representative.

D. Classification Modeling

The Logistic Regression algorithm was used as a credit risk classification model in this study. Logistic Regression was chosen because it has a simple and easy-to-interpret model structure, making it suitable for use in the context of decision-making in the financial sector [10], [11]. To overcome the problem of class imbalance, the model is configured with class weight adjustment, so that the model can pay more attention to the bad credit risk class.

E. Model Evaluation

The evaluation of model performance was carried out using several evaluation metrics, namely accuracy, confusion matrix, precision, recall, and F1-score. The recall metric in the bad credit class is the main focus of the evaluation because errors in classifying high-risk customers can have a significant impact on financial losses. The results of the evaluation are used to assess the effectiveness of the model in classifying credit risk as well as a basis for discussion in the results and discussion sections.

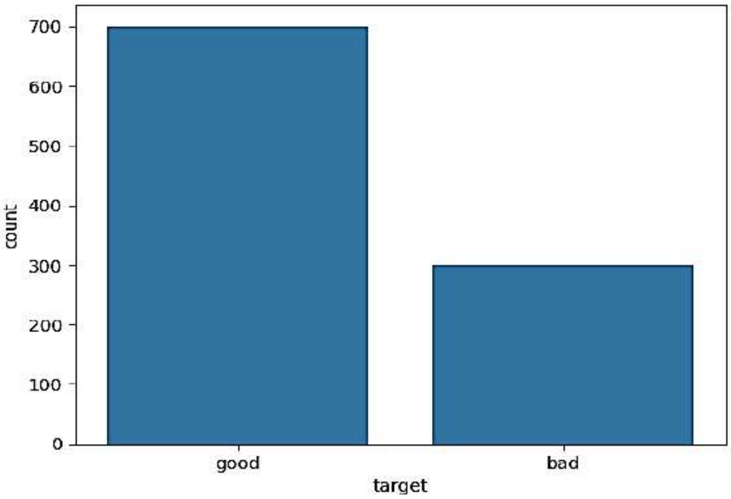

Figure 1: Credit Risk Distribution in German Credit Dataset

III. RESULTS AND DISCUSSION

A. Credit Risk Distribution Analysis

The initial stage of analysis was carried out to determine the distribution of credit risk classes in the German Credit Dataset. This distribution is important for understanding the characteristics of the data as well as identifying potential class imbalances that may affect the performance of the classification model.

Figure 1 shows the distribution of credit risk which consists of two classes, namely good credit (good) and bad credit (bad). It can be seen that the amount of good credit data is more dominant than bad credit. This condition indicates a class imbalance in the dataset, where the proportion of good credit is greater than bad credit.

This class imbalance needs to be considered in the modeling process because it can cause the model to tend to predict the majority class. Therefore, at the modeling stage, a class imbalance handling strategy was applied to improve the model's ability to detect high-risk credit.

B. Analysis of Categorical

Features on Credit RiskIn addition to analyzing the distribution of the target class, this study also analyzed several categorical features to see their relationship with credit risk. This analysis aims to identify early patterns that can provide an understanding of the factors that affect credit risk.

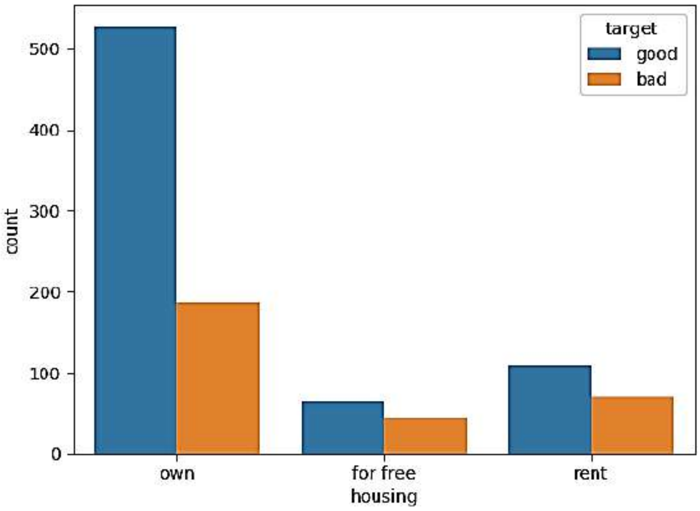

Figure 2: Housing Ownership vs Credit Risk

Home Ownership- Figure 2 shows the relationship between homeownership status and credit risk. Based on this graph, credit applicants who own their own tend to have a higher proportion of good credit compared to applicants who rent a house. Meanwhile, applicants who live for free show a relatively smaller amount of data. These results show that home ownership can be one of the indicators of the financial stability of credit applicants. Applicants with own home ownership generally have a lower credit risk compared to applicants who do not own a home.

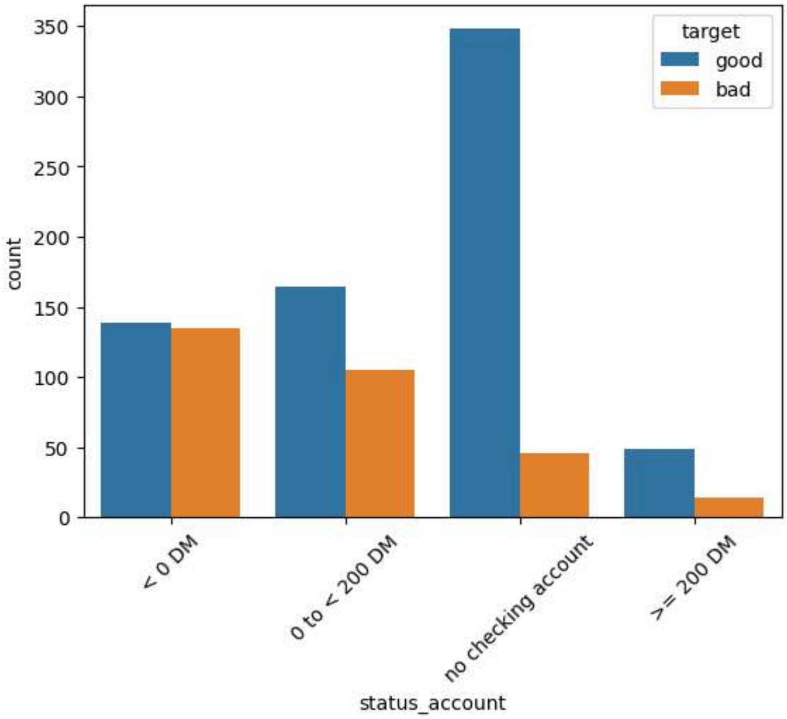

Figure 3: Account Status vs Credit Risk

Account Status- The next analysis is carried out on the account status of the credit applicant. Figure 3 shows the relationship between account status and credit risk. It can be seen that applicants with better account status, such as sufficient account balances, have a higher proportion of good credit. On the other hand, applicants with low or no checking account status have a larger proportion of bad credit. This shows that account status is one of the important factors in assessing credit risk, as it reflects the applicant's financial condition and financial behavior.

C. Model Classification and Evaluation Results

After initial data analysis and pre-processing of data, the next stage is the development of a credit risk classification model using the Logistic Regression algorithm. Model evaluation is conducted using test data to measure the model's ability to classify data that has never been seen before.

Table 1: Performance Evaluation of Logistic Regression Model

Credit Risk Class	Precision	Recall	F1-Score
Bad Credit	0.46	0.68	0.55
Good Credit	0.83	0.65	0.73
Accuracy			**0.66**

In the above table 1, the results of the evaluation showed that the model produced an accuracy rate of 66%. Based on the confusion matrix, the model is able to classify most credit data correctly, both in good credit and bad credit grades. The recall value for bad credit grades reached 68%, which suggests that the model is quite effective in detecting high-risk credit applicants.

In the context of credit risk management, the ability to detect non-performing loans is more important than just the level of accuracy. Therefore, although the accuracy value of the model is still moderate, the recall results on bad credit grades show that the model has the potential to be used as an initial aid in the credit risk assessment process.

However, the precision value of the bad credit class is still relatively low, indicating that there are low-risk credit applicants who are incorrectly classified as high-risk. This indicates that the model still has limitations and needs to be further developed.

D. Discussion

Overall, the results show that the application of machine learning with the Logistic Regression algorithm can be used to classify credit risk based on the financial and demographic characteristics of the applicant. Categorical feature analysis shows the relationship between home ownership, account status, and credit risk, which is in line with the concept of credit scoring in banking practice. The obtained classification performance is consistent with previous studies reporting that classical machine learning models remain effective for credit risk assessment [4], [12]. However, the resulting model still has limitations in terms of the accuracy of classification, especially in bad credit grades. Therefore, this model is more appropriate to be used as a decision support system that assists credit analysts in conducting initial screening of credit applicants, rather than as the sole basis for decision-makingn.

IV. CONCLUSIONS AND SUGGESTIONS

A. Conclusion

Based on the results of the research that has been conducted, it can be concluded that the application of the Logistic Regression algorithm in the German Credit Dataset can be used to classify credit risk. The model was built to produce an accuracy rate of 66% on the test data, with a recall value for the high-risk credit class (bad) of 68%. These results show that the machine learning approach has the potential to assist the credit risk assessment process, particularly in identifying potentially problematic credit applicants.

However, the classification results obtained still have limitations, especially in the accuracy of the classification of high-risk credit classes. Therefore, the proposed model is more appropriately used as a decision support system, not as the only basis for credit decision-making.

B. Suggestions

Based on the results and limitations of the research, some suggestions for further research can be submitted as follows. First, research can be developed by comparing several machine learning algorithms to obtain more optimal classification performance. Second, optimization of model parameters as well as the application of further data imbalance handling techniques can be considered to improve credit risk detection capabilities. In addition, the use of larger and varied datasets can also be the focus of future research so that the resulting model has better generalization capabilities.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGMENT

The author(s) would like to thank colleagues and reviewers for their valuable suggestions and support during the completion of this research. The German Credit Dataset used in this study was obtained from a publicly available data repository.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Á. Paz, B. Crawford, E. Monfroy, J. Barrera-García, Á. Peña Fritz, R. Soto, F. Cisternas-Caneo, and A. Yáñez, "Machine learning and metaheuristics approach for individual credit risk assessment: A systematic literature review," *Biomimetics*, vol. 10, no. 5, 2025, Available from: https://doi.org/10.3390/biomimetics10050326

[2] J. Wang, W. Rong, Z. Zhang, and D. Mei, "Credit debt default risk assessment based on the XGBoost algorithm: An empirical study from China," *Mathematical Problems in Engineering*, vol. 2022**,** 2022. https://doi.org/10.1155/2022/8005493

[3] N. Chai, M. Z. Abedin, X. Wang, and B. Shi, "Growth potential of machine learning in credit risk predicting of farmers in the Industry 4.0 era," *International Journal of Finance & Economics*, vol. 29**,** no. **1**, 2024. https://doi.org/10.1002/ijfe.3010

[4] S. Shi, R. Tse, W. Luo, S. d'Addona, and G. Pau, "Machine learning-driven credit risk: A systemic review," *Neural Computing and Applications*, vol. 34, no. 22, pp. 19787–19805, 2022. https://doi.org/10.1007/s00521-022-07472-2

[5] S. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1178–1193, 2022. https://doi.org/10.1016/j.ejor.2021.06.053

[6] "German Credit Dataset," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/jumpingdino/german-credit-dataset

[7] S.-H. Kyeong and J.-H. Shin, "Two-stage credit scoring using Bayesian approach," *Journal of Big Data*, vol. 9, no. 1, 2022. https://doi.org/10.1186/s40537-022-00665-5

[8] X. Ran, "Credit anomaly detection method based on Bayesian networks," 2024. https://doi.org/10.61173/pxsg8h72

[9] M. Alaraj, M. Abbod, and M. Majdalawieh, "Modelling customers' credit card behaviour using bidirectional LSTM neural networks," *Journal of Big Data*, vol. 8, no. 1, 2021. https://doi.org/10.1186/s40537-021-00461-7

[10] M. Zhu, B. Shia, M. Su, and J. Liu, "Consumer default risk portrait: An intelligent management framework of online consumer credit default risk," *Mathematics*, vol. 12, no. 10, 2024. https://doi.org/10.3390/math12101582

[11] P. K. Roy and K. Shaw, "A multicriteria credit scoring model for SMEs using hybrid BWM and TOPSIS," *Financial Innovation*, vol. 7, no. 1, 2021. https://doi.org/10.1186/s40854-021-00295-5

[12] L. T. Trinh, "A comparative analysis of consumer credit risk models in peer-to-peer lending," *Journal of Economics, Finance and Administrative Science*, vol. 29, no. 57, pp. 1–20, 2024. https://doi.org/10.1108/JEFAS-04-2021-0026

ABOUT THE AUTHORS

Fitria received her bachelor's degree in informatics engineering and master's degree in informatics engineering. She is currently a lecturer at Politeknik Negeri Banjarmasin. Her research interests include machine learning, data mining, Internet of Things (IoT), information systems, and accounting information systems. She has published several research articles in national and international journals and actively participates in academic research and community service activities.

Emy Iryanie is a dedicated lecturer at Politeknik Negeri Banjarmasin, where she contributes to the development of accounting information system. She earned her bachelor's degree in accounting from STIEI Banjarmasin in 2007 and went on to pursue a master's degree in magister accounting at Universitas Diponegoro, graduating in 2009. Her academic and research interests focus on accounting, management and Financial Technology

Heldalina is a dedicated lecturer at Politeknik Negeri Banjarmasin, where she contributes to the development of accounting information system. She earned his bachelor's degree in sharia accounting from Purwakarta Islamic University in 2005 and went on to pursue a master's degree in magister management at Lambung Mangkurat University, graduating in 2013. Her academic and research interests focus on accounting, management and Financial

Muhammad Syahid Pebriadi is a dedicated lecturer at Politeknik Negeri Banjarmasin, where he contributes to the development of computer science education. He earned his bachelor's degree in computer science from Lambung Mangkurat University in 2014 and went on to pursue a master's degree in the same field at Bogor Agricultural University, graduating in 2017. His academic and research interests focus on Artificial Intelligence, Computer Vision, and Financial Technology.

Anhar Khalid is currently a permanent lecturer in the Mechanical Engineering Program at Politeknik Negeri Banjarmasin. His academic interests include Energy Conversion, Mechanical Engineering, and Materials. He earned his master's degree in mechanical engineering from Pancasila University, Jakarta. He has published several international journal articles and is actively involved in research and community service projects related to mechanical engineering and industrial applications.