Published *online* on the journal's web page : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Development Prediction Model to Optimize Cooperative Loans Based on Machine Learning Algorithms

Hidayatulloh Himawan[1], Tito Pinandita[2], Rizky Ridwan[3], Hilmi Aziz[4]

[1]*Department of Informatics, UPN Veteran Yogyakarta, Indonesia*
[2]*Department of Informatics, Muhammadiyah Purwokerto University, Indonesia*
[3]*Department of Accounting, Cipasung University, Tasikmalaya, Indonesia*
[4]*Institute of Advanced Informatics and Computing, Indonesia*

[1]*if.iwan@upnyk.ac.id,* [2]*titop@ump.ac.id,* [3]*rizkyridwan@uncip.ac.id,* [4]*hilmi@iaico.org*

ARTICLE INFORMATION

Article History:
Received: April 23, 2024
Last Revision: May 10, 2024
Published *Online*: May 13, 2024

KEYWORDS

Decision Tree,
Default Prediction,
Logistic Regression,
Random Forest,
K-NN

CORRESPONDENCE

Phone: +62 812-2732-222

E-mail: if.iwan@upnyk.ac.id

ABSTRACT

Default on loans by borrowers to the cooperative to optimize the cooperative's business performance. In this research, a default prediction model was developed using several quite popular machine learning algorithms, namely decision tree, K-NN, logistic regression, and random forest, then all models with each of these algorithms were compared and evaluated. to find out which algorithm model is the most effective and accurate in predicting loan defaults in cooperatives. Model evaluation is carried out using metrics such as accuracy, precision, recall, and f1-score. The dataset used in this research was obtained from the loan list at one of the Savings and Loans Cooperatives in Tasikmalaya Regency, the contents of which include attributes such as borrower profile, loan amount, number of installments, and others. This dataset is divided into training data and test data to train and evaluate the model. These machine learning algorithms were chosen because they are quite well known among other algorithms for prediction and have been proven in several financial studies. The results of this prediction model can be used by cooperatives to support decisions in providing appropriate loans.

1. INTRODUCTION

In carrying out its main business activities, the Savings and Loans Cooperative always tries to provide loans with the hope that they will be right on target and optimally for members to get profits which will later return to the members [1]. However, savings and loans cooperatives often face the risk of bankruptcy due to the large number of loan defaults that occur [2]. This could threaten the financial stability of the cooperative which could even result in losses for members. From research [3] that in cooperatives there have been no efforts that are deemed effective enough to reduce bad credit in cooperatives because cooperatives do not have reliable credit analysis like banking.

Therefore, it is necessary to carry out risk analysis and develop an accurate and effective loan payment failure prediction model to minimize the risk of loss [4]. The results of this prediction model will later be used as decision support in optimizing lending to improve business performance and minimize the risk of bankruptcy [5].

This prediction model can be created using traditional machine learning algorithms such as Decision Tree, K-Nearest Neighbors, and others. You can also use several Deep Learning algorithms such as Recurrent Neural Network and Convolutional Neural Network. Each of these algorithms has advantages and disadvantages because they can create prediction models with their own characteristics/methods. From research [6] and more complex, which is not suitable for this cooperative dataset, which is quite small. This was also found in credit risk analysis research [7] who found that tree-based models were more stable than models based on multilayer artificial neural networks. The capabilities of traditional machine learning algorithms are also quite good, as according to a

study conducted by [8] using several machine learning algorithms such as decision trees, random forests, and logistic regression can help predict microloan defaults in associations savings and loans in India. The conclusion from their research results shows that the XGBoost algorithm can provide quite good prediction performance compared to other algorithms.

In research Chen et al. [9] said that XGBoost is an effective open-source implementation of the gradient boosting technique, namely a machine learning method that aims to precisely estimate the target variable by combining the results of a series of variables that are weaker and simpler than the model. This is what makes it very effective, more powerful than existing variants, and computationally efficient. A similar thing was also found in research conducted by [10] which shows the results of the Deep Neural Networks model and the XGBoost model have better performance compared to other machine learning approaches in terms of AUC and accuracy. From these studies and considering the relatively small lending dataset at Savings and Loans Cooperatives, the algorithms used to compare the prediction models are decision tree, K-NN, logistic regression, and random forest.

The aim of this research is to find out which model is most effective in predicting loan failure in Savings and Loans Cooperatives. It is hoped that the results of this research can help minimize the risk of bankruptcy and improve the business performance of Savings and Loans Cooperatives. It is also hoped that this research can become a reference for similar research in the future. Apart from that, to obtain better prediction accuracy values, this research also carried out several tests using datasets from the Cooperative. This dataset is created into several more datasets with different attributes. This is done based on the consideration that the estimated importance of these attributes will influence the model in making predictions.

2. RELATED WORK

Research [11] developed a loan default prediction model using several approaches to obtain more optimal performance. The dataset in this study contains more than 115,000 users' original loan data with 102 attributes. The contribution in this research is that Random Forest has the best performance, with 98% accuracy compared to support vector machines and logistic regression. Other research [10] focuses on credit risk models with machine learning that can replace models based on financial domain experts which still dominate. The dataset was taken from a three-year cross-sectional survey, which used 4245 data with 345 variables. As a result, the Deep Neural Network algorithm obtained more promising results than other machine learning algorithms.

Research [8] utilizes machine learning to be used to make credit default predictions to replace less accurate traditional methods. The dataset contains information on 16,1715 loans between January and August 2021. Resulting in the XGBoost algorithm getting the best results with 97% accuracy. The paper [12] discusses the significance of credit risk estimation and portfolio evaluation for financial institutions lending to businesses and individuals. It emphasizes the importance of predicting non-performing loans (NPLs), where customers fail to make scheduled payments. The integration of machine learning algorithms and big data analytics into banking models is highlighted. The study evaluates various machine learning algorithms in addressing NPL prediction, particularly focusing on a dataset from a private bank in Turkey. The research addresses class imbalance using class weights and assesses model performance using metrics like Precision, Recall, F1 Score, Imbalance Accuracy (IAM), and Specificity. Among the tested algorithms, LightGBM emerged as the most effective, outperforming logistic regression, SVM, random forest, bagging classifier, XGBoost, and LSTM for the given dataset.

The research [13] aims to address the challenge of loan approval by leveraging machine learning algorithms on loan data from various sources. By analyzing real bank credit data, the study seeks to develop a model that assists organizations in making informed decisions regarding loan approval. The goal is to create a bank risk automated system that determines the creditworthiness of customers, thus mitigating potential losses associated with loan defaults.

3. METHODOLOGY

The method used in research is as shown in Figure 1. This research was carried out by following sequential steps to achieve the research objectives.

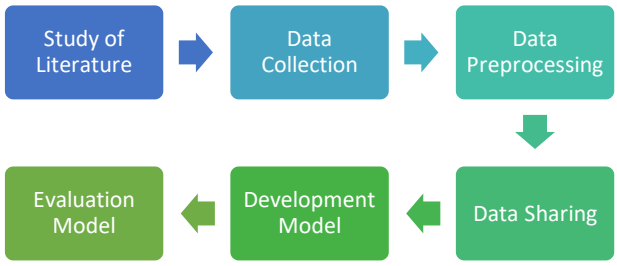

FIGURE 1. RESEARCH METHODOLOGY

3.1 Study of Literature

The first stage of research was carried out by reading and reviewing related research in the form of articles from accredited journals to be used as quality references.

3.2 Data Collection

The dataset used comes from historical loan data from the Pamipiran savings and loan cooperative in the last month, totaling 324 loan data with 9 attributes. This data is taken from the cooperative database which includes some information regarding the borrower's profile, transaction number, credit history, total loan, number of installments, and loan payment status.

3.3 Data Preprocessing

The data that has been collected then undergoes a preprocessing stage to prepare it so that it is ready for analysis and model creation. Data pre-processing steps include:

a. Data Cleaning: data that is considered less important for the prediction model is removed from the dataset, such as transaction number, borrower name, and borrowing date.

b. Feature/Attribute Selection: the attributes used for predictions in this research are age, working status, credit_value, number of_installments, total_loan, remaining_loan, and default_payment (loan payment status).
c. Missing Values Processing: handles missing values in the data, either by filling in missing values or using imputation techniques.
d. Outliers Processing: Dealing with outlier data that can influence analysis results.
e. Data Transformation: Perform data transformation if necessary, such as normalization or standardization.

3.4 Data Sharing

The preprocessed data is then divided into two subsets, namely 20% training data and the remaining 80% testing data. This division was carried out to test the performance of the model developed on data that had never been seen before.

3.5 Development Model

The machine learning algorithms used in this research are logistic regression, decision tree, random forest, and k-nearest neighbors. The model was created using the Python programming language in the Jupyter Notebook application by utilizing existing libraries such as Pandas, Sklearn and others. Then all these algorithms are then implemented into a model to predict loan default predictions with the following stages:

a. Reading Data: The first stage is reading the dataset from a CSV file that has been processed using the Pandas library. This data is a dataset that contains information about loans and default status.
b. Dataset Sharing: After loading the data, the dataset needs to be divided into training data and testing data as explained in the data division point.
c. Model Initialization: In this process, the four algorithms are initialized using the classes provided by the Sklearn library in Python.
d. Model Training: After the algorithm is initialized for each model, the next stage is that the model is trained using training data using the 'fit' method on the model objects that have been initialized previously. In this process the model can learn patterns and structures from the training data so that the model can make predictions.
e. Prediction: At this stage the model can be used to make predictions using testing data with the 'predict' method so that the results can then be evaluated and compared with each other.

3.6 Evaluation Model

After the model predicts default on the cooperative dataset using all existing attributes, it is then repeated by reducing several attributes in the dataset to find out other results to compare using metrics such as accuracy, precision, recall, and f1-score to evaluate performance. and the model's predictive capabilities. From the results of this evaluation, the model with the best performance can be identified and selected as the most effective and optimal loan default prediction model for predicting loan default in Savings and Loans Cooperatives.

4. RESULT AND DISCUSSION

The loan default prediction model that has been created by implementing several algorithms with all the attributes is then tested by reducing the attributes that are deemed less influential to find out other results from the model created if the attributes change. The attributes used in this modeling include: (U) age, (S) employment_status, (NK) credit_value, (BK) number of_installments, (TP), total_loan, (SP) remaining_loan, (GB) failed_to pay. In this test, the attributes of the ready-made dataset were changed into three datasets with different attributes in them with the following details:

a. Dataset_1: U, SB, NK, BA, TP, SP, GB
b. Dataset_2: SB, NK, BA, TP, SP, GB
c. Dataset_3: NK, BA, TP, SP, GB

The following are the results of model testing with these three datasets:

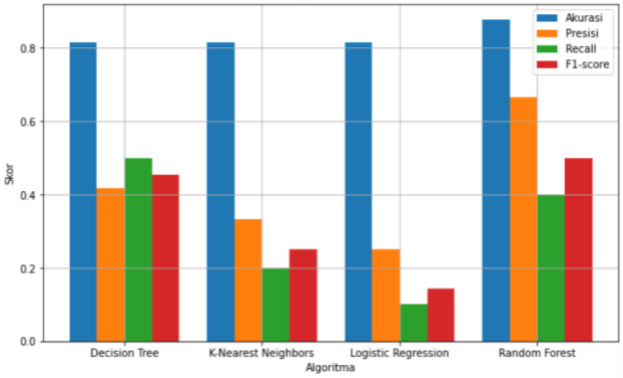

FIGURE 2. COMPARISON OF EVALUATION METRICS DATASET 1

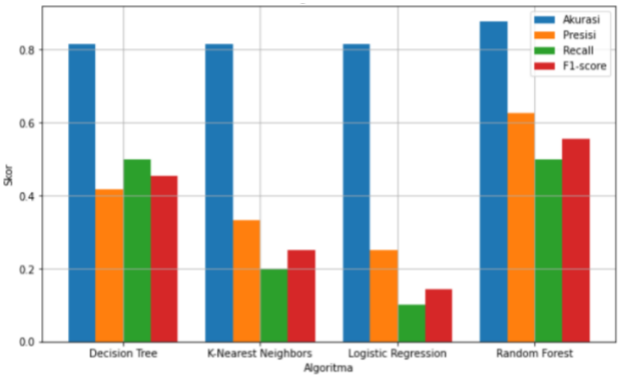

FIGURE 3. COMPARISON OF EVALUATION METRICS DATASET 2

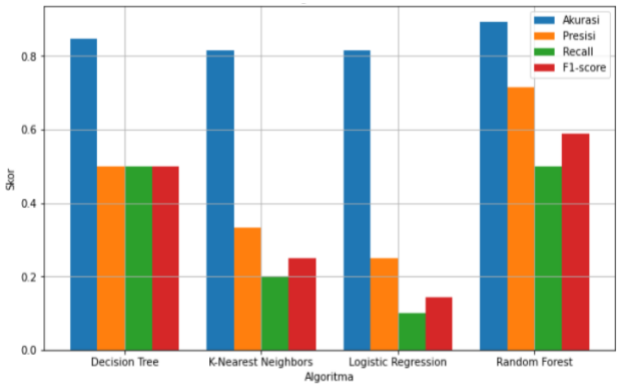

FIGURE 4. COMPARISON OF EVALUATION METRICS DATASET 3

From the test results above in Figures 2, 3, 4 the results for each model can be described as follows:

A. Model Performance Decision Tree Algorithm

The first model implements the Decision Tree algorithm to build a loan default prediction model. After training the model using training data, the model performance is evaluated using testing data. The following table 1 is an evaluation with effectiveness measurement metrics. The results above show that the model with the Decision Tree algorithm succeeded in achieving the best accuracy of 84% in predicting loan default for a dataset without the 'age' and 'working_status' attributes

TABLE 1. MODEL EVALUATION WITH THE DECISION TREE ALGORITHM

No.	Attributes on the dataset	Accuracy	Precision	Recall	F1-Score
1.	U, SB, NK, BA, TP, SP, GB	0.8153	0.4166	0.5	0.4545
2.	SB, NK, BA, TP, SP, GB	0.8153	0.4166	0.5	0.4545
3.	NK, BA, TP, SP, GB	0.8461	0.5	0.5	0.5

TABLE 2. MODEL EVALUATION WITH THE K-NEAREST NEIGHBORS ALGORITHM

No.	Attributes on the dataset	Accuracy	Precision	Recall	F1-Score
1.	U, SB, NK, BA, TP, SP, GB	0.8153	0.3333	0.2	0.25
2.	SB, NK, BA, TP, SP, GB	0.8153	0.3333	0.2	0.25
3.	NK, BA, TP, SP, GB	0.8153	0.3333	0.2	0.25

TABLE 3. MODEL EVALUATION WITH THE LOGISTIC REGRESSION ALGORITHM

No.	Attributes on the dataset	Accuracy	Precision	Recall	F1-Score
1.	U, SB, NK, BA, TP, SP, GB	0.8153	0.25	0.1	0.1428
2.	SB, NK, BA, TP, SP, GB	0.8153	0.25	0.1	0.1428
3.	NK, BA, TP, SP, GB	0.8153	0.25	0.1	0.1428

TABLE 4. MODEL EVALUATION WITH THE RANDOM FOREST ALGORITHM

No.	Attributes on the dataset	Accuracy	Precision	Recall	F1-Score
1.	U, SB, NK, BA, TP, SP, GB	0.8769	0.6666	0.4	0.5
2.	SB, NK, BA, TP, SP, GB	0.8769	0.625	0.5	0.5555
3.	NK, BA, TP, SP, GB	0.8923	0.714	0.5	0.5882

B. Model Performance K-NN Algorithm

Following this, the K-Nearest Neighbors algorithm was employed for the model. Table 2 presents an evaluation of the outcomes derived from the conducted tests. As illustrated in the test results above, all three trials utilizing the K-Nearest Neighbors algorithm yielded identical outcomes, demonstrating consistency even amidst attribute reduction within the dataset. Notably, the accuracy metric consistently registers at 81%, specifically 0.8153, underscoring the reliability and robustness of the model's predictive capabilities across varying conditions.

C. Model Performance Logistic Regression Algorithm

Moving forward, the model employing the Logistic Regression algorithm, as delineated in Table 3, underwent evaluation utilizing metrics such as accuracy, precision, recall, and f1-score, both with the entirety of attributes and after attribute reduction. The outcomes depicted above illustrate a remarkable consistency across all metrics, mirroring the pattern observed in the K-Nearest Neighbors model. Notably, akin to its predecessor, the Logistic Regression model exhibited a commendable accuracy rate of 81%, underscoring its efficacy and reliability in predictive performance under varying conditions. This uniformity in results across differing attribute configurations reinforces the model's robustness and underscores its potential for practical application in optimizing cooperative loans.

D. Model Performance Random Forest Algorithm

Then the final algorithm used for the model is Random Forest and the following is table 4 evaluating the results obtained from the tests carried out. The results above show that the model with the Random Forest algorithm succeeded in achieving the best accuracy of 89% for a dataset whose attributes ignored the attributes 'age' and 'employment_status'. The test results are different for each dataset with different attributes. There was a slight increase in accuracy, precision, recall and f1-score values that occurred when attributes that were felt to have less influence were reduced.

Based on the performance evaluation results of the four models used, it can be concluded that the model that implements the Random Forest algorithm produces the best value in predicting loan default on all metrics. The best accuracy of this model is 89%, higher than models using logistic regression algorithms (accuracy 81%), decision trees (accuracy 84%), and k-nearest neighbors (accuracy 81%). This highest value was obtained after making predictions using dataset 3 with the attributes NK, BA, TP, SP, GB. Choosing the right attributes in the dataset can increase the accuracy of the model predictions.

Achieving high accuracy in predicting loan defaults is paramount for cooperatives to effectively pinpoint high credit risks and implement suitable measures. However, it's crucial to recognize that the selection of the appropriate algorithm hinges on several factors, including the nuances of the dataset, its scale, the complexity of the problem, and the attainable prediction objectives.

Therefore, it is recommended to conduct further experiments and studies to validate these results and consider other factors such as computational time, model interpretation, and related factors. In future research, it is recommended to examine and compare several other machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, and/or even with Deep Learning algorithms such as Neural Networks with more complex architectures. Further research will provide a deeper understanding of algorithms for creating the most appropriate loan default prediction models and more detailed guidance for other financial institutions in their lending decisions.

5. CONCLUSIONS

Based on the comprehensive analysis of the test results for each model, it becomes evident that the Random Forest algorithm, particularly when applied to dataset 3, emerges as the optimal choice for predicting loan defaults within savings and loans cooperatives, especially when dealing with relatively small datasets. This algorithm's superiority underscores its robustness and efficacy in handling the intricacies of cooperative loan prediction tasks. Moreover, the importance of attribute selection within the dataset cannot be overstated. The attributes chosen significantly to impact the predictive performance of the model. Through meticulous selection and curation of attributes, predictive accuracy can be substantially enhanced. This study underscores the pivotal role played by attribute selection in refining prediction models and highlights the need for careful consideration in this aspect of model development. Furthermore, the research corroborates the effectiveness of attribute reduction strategies in bolstering prediction accuracy. By identifying and omitting attributes deemed less influential or redundant, the model can focus on the most pertinent features, thereby refining its predictive capabilities. This finding underscores the importance of feature engineering and underscores its potential to optimize predictive model performance. In essence, this study provides valuable insights into the intricacies of predictive modeling for loan defaults in savings and loans cooperatives. It not only identifies the Random Forest algorithm as the preferred choice for such tasks but also emphasizes the critical role of attribute selection and reduction in enhancing predictive accuracy. These findings hold significant implications for practitioners in the financial sector, offering actionable strategies to improve loan default prediction models and mitigate associated risks.

REFERENCES

[1] A. S. Ningsih, D. D. Suprapti, and N. Fibrianti, "The Importance of Applying the Membership Value Toward Savings and Loans Cooperatives in Indonesia," *Sriwijaya Law Review*, vol. 3, no. 2, p. 225, Jul. 2019, doi: 10.28946/slrev.Vol3.Iss2.235.pp225-234.

[2] C.-M. Kang, M.-C. Wang, and L. Lin, "Financial Distress Prediction of Cooperative Financial Institutions—Evidence for Taiwan Credit Unions," *International Journal of Financial Studies*, vol. 10, no. 2, p. 30, Apr. 2022, doi: 10.3390/ijfs10020030.

[3] D. Adi Setya Rahardjo, "The Role of Indonesian Credit Cooperatives Towards Strengthening Financial Literacy and Improving Financial Behavior."

[4] A. Nursyahriana, M. Hadjat, and I. Tricahyadinata, "Analisis Faktor Penyebab Terjadinya Kredit Macet," *FORUM EKONOMI*, vol. 19, no. 1, 2017.

[5] D. Máté, H. Raza, and I. Ahmad, "Comparative Analysis of Machine Learning Models for Bankruptcy Prediction in the Context of Pakistani Companies," *Risks*, vol. 11, no. 10, p. 176, Oct. 2023, doi: 10.3390/risks11100176.

[6] S. Syafrudin, R. A. Nugraha, K. Handayani, S. Linawati, and W. Gata, "Prediksi Status Pinjaman Bank dengan Deep Learning Neural Network," *Jurnal Teknik Komputer*, vol. 7, no. 2, pp. 130–135, Jul. 2021, doi: 10.31294/jtk.v7i2.10474.

[7] P. Addo, D. Guegan, and B. Hassani, "Credit Risk Analysis Using Machine and Deep Learning Models," *Risks*, vol. 6, no. 2, p. 38, Apr. 2018, doi: 10.3390/risks6020038.

[8] Kadek Dwi Pradnyana and Raden Aswin Rahadi, "Loan Default Prediction in Microfinance Group Lending with Machine Learning," *International Journal of Business and Technology Management*, Jan. 2023, doi: 10.55057/ijbtm.2022.4.4.8.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2018, pp. 785–794. doi: 10.1145/2939672.2939785.

[10] L. Munkhdalai, T. Munkhdalai, O.-E. Namsrai, J. Lee, and K. Ryu, "An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments," *Sustainability*, vol. 11, no. 3, p. 699, Jan. 2019, doi: 10.3390/su11030699.

[11] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Comput Sci*, vol. 162, pp. 503–513, 2019, doi: 10.1016/j.procs.2019.12.017.

[12] S. I. Serengil, S. Imece, U. G. Tosun, E. B. Buyukbas, and B. Koroglu, "A Comparative Study of Machine Learning Approaches for Non Performing Loan Prediction," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, IEEE, Sep. 2021, pp. 326–331. doi: 10.1109/UBMK52708.2021.9558894.

[13] A. S. Aphale and S. R. Shinde, "Predict Loan Approval In Banking Systemmachine Learning Approach for Cooperative Banks Loan Approval." [Online]. Available: www.ijert.org

AUTHORS

Hidayatulloh Himawan
He is a senior lecturer in the informatics engineering department of UPN Yogyakarta and is pursuing PhD at UTeM Malaysia. His research field focuses on information systems research, data communications, and information technology.

Tito Pinandita
He is a senior lecturer in the Informatics engineering department and deputy dean of engineering at Muhammadiyah University Purwokerto and is pursuing a PhD at UTeM Malaysia. His research field focuses on Computer vision, Augmented Reality, and Artificial Intelligence.

Rizky Ridwan
He is a lecturer in the accounting department study program at Cipasung Tasikmalaya University. Active as editor in chief managing the Invest journal and others. His research field focuses on public accounting research.

Hilmi Aziz
Currently he is a researcher at the Institute of Advanced Informatics and Computing (IAICO), Indonesia. His research focuses on the fields of information systems, information technology and AI.