JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6437

IMPLEMENTATION MEAN IMPUTATION AND OUTLIER DETECTION FOR LOAN PREDICTION USING THE RANDOM FOREST ALGORITHM

Ni'matul Ma'muriyah[1*]; Richard[2]; Haeruddin[3]

Information Technology Study Program[1, 2, 3]
Universitas Internasional Batam, Batam, Indonesia[1, 2, 3]
https://www.uib.ac.id/ [1, 2, 3]
nimatul@uib.ac,id[1*], 2132011.richard@uib.edu[2], haeruddin@uib.ac.id[3]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Loans and credit are among the most in-demand banking products, making accurate loan prediction systems essential for minimizing bank credit risks and boosting profitability. This study proposed a loan prediction model using the Random Forest algorithm, with mean imputation and 3 outlier detection (Boxplot, Z-score, and Interquartile Range (IQR)) as data pre-processing methods. Using Lending Club loan data from 2014-2021 (466,285 records, split 70/30 for training/testing), model performance was assessed using accuracy, recall, and F1 Score. The proposed approach achieved a 95% prediction accuracy, outperforming previous models at 83%. The best results were obtained using mean imputation with IQR-based outlier detection. However, the determination of the mean imputation mean can be a limitation of this study. This highlights the importance of thorough pre-processing in enhancing prediction accuracy. The study underscores the role of machine learning and financial technology (fintech) in informing credit decisions and support incorporating imputation and outlier handling as standard steps in financial modeling pipeline.*

Keywords: *accuracy, loan prediction, pre-processing, random forest.*

Intisari— *Pinjaman atau kredit salah satu produk yang paling diminati di industri perbankan, sehingga kebutuhan akan sistem prediksi pinjaman dengan akurasi tinggi sangat penting untuk meminimalkan risiko kredit macet dan meningkatkan profitabilitas. Penelitian ini mengusulkan model prediksi pinjaman menggunakan Algoritma Random Forest, dengan mean imputasi dan 3 deteksi data outlier Boxplot, Z-score, dan Interquartile Range (IQR) sebagai metode pre-processing data. Dataset yang digunakan Lending Club dari tahun 2014 hingga 2021, yang mencakup 466.285 data, dibagi menjadi 70% data pelatihan dan 30% data pengujian. Kinerja model dievaluasi menggunakan Confusion Matrix dengan tiga parameter pengukuran: akurasi, recall, dan F1 Score. Pendekatan yang diusulkan menunjukkan peningkatan kinerja, mencapai akurasi prediksi pinjaman sebesar 95%, dibandingkan dengan 83% dalam penelitian sebelumnya. Akurasi tertinggi diperoleh ketika pra-pemrosesan data dilakukan menggunakan metode imputasi rata-rata dan Interquartile Range (IQR) untuk deteksi outlier. Walaupun demikian penetapan penggunaan mean imputation dapat menjadi Batasan dalam penelitian ini. Hasil dari penelitian ini yang perlu di garis bawahi bahwa penggunaan pra proses data dapat meningkatkan keakuratan model. Kontribusi hasil penelitian ini pada pengambilan keputusan keuangan yang terinformasi dan membantu pengguna merencanakan keuangan mereka dengan lebih baik.*

Kata Kunci: *akurasi. prediksi pinjaman, pre-processing, random forest.*

INTRODUCTION

Bank Indonesia (BI) plays a key role in keeping system stable and supporting sustainable economic growth[1]. It sets regulations to maintain stability, even economic downturns that could slow growth. Research indicates that Indonesia's bank depend heavily on interest from credit, making it a vital product. When giving credit, banks consider trust, agreements, time periods, risk, and reward.

Credit is not given easily, it requires several agreements beforehand [2]. Currently Financial Technology (fintech) has variously offer credit loans, this condition as impact of artificial Intelligent (AI) development. To support the platforms Machine Learning (ML) technology take important part to improve the process of appraisal. For the above reason Loan prediction's studies with ML model or algorithm become interesting in last five years.

According to the literature review, several researchers have implemented machine learning (ML) models such as Decision Tree, Random Forest, XGBoost, and Support Vector Classifier (SVM) to classify or predict loans with better performances. Table 1 shows how ML is widely used in various applications such as detection, classification, and prediction, delivering optimal results in various applications. Some studies have compared two or more ML models to determine the best performance in loan classification or prediction.

Table 1. Resume of Literature review for ML

references	Proposed Models	Better Performance	Accuracy
[3]	3 ML Models	Random Forest Classifier (RFC)	99%
[4]	7 ML Models	XGBoost	79%
[5]	10 ML Models	Adaboost, XGboost	81.71 %, 80.8%
[6]	1 ML Model	LightGBM	73%
[7]	3 ML Models	Support Vector Classifier (SVM)	83%
[8]	2 ML	Random Forest	80%
[9]	1 ML	Random Forest	83%
[10]	1 ML	Random Forest	97%

Source : (Research Results, 2025)

From Table 1, we see that some studies found the Random Forest model delivered the best performance in prediction or classification, with outcomes improving by at least 80%. Therefore, this study uses the Random Forest model. Missing data in dataset occurs frequently which lead to challenges, Primarily to face these issue there are two type of approach includes deletion and imputation, Deletion involves removing observations or features with missing values[11]. Deletion is often the default method due to its simplicity and speed[12]. However, this approach has limitations, including a reduction in dataset size, potential bias, and the loss of important information, especially when a large proportion of the data is missing. Previous studies comparing kNN imputation and Mean Imputation for handling missing data on vulnerability index, they concluded that kNN and mean imputation can handle missing data[13]. Other researchers used combination of deleting technique and Mean, Mode and ANN imputation method to overcomes missing data in heart diseases dataset[14], [15]. Result of these studied shown that implementation Mean imputation method would improve the accuracy.

Another issue within the dataset, aside from missing data, is the presence of outliers. Similar to missing data, outliers can lead to a decrease in the performance of classification or prediction models. Therefore, it is crucial to perform outlier detection on the dataset to improve the performance of classification or prediction tasks. This has been demonstrated in a study on Hybrid Diabetes Diseases prediction, which achieved an accuracy of up to 96%[16]. Other studies stunting classification with Indonesian DHS dataset was implemented Data imputation Method and Outlier Detection and improved the accuracy of classification stunting between 96% - 98%[17].

In previous study titled "Analysis of Prediction of Loan Eligibility with the Random Forest Method "[9], the results showed the accuracy of classification using Random Forest algorithm was 83%. Data preprocessing was performed by eliminating not-matching data. In this study, We propose a new loan prediction model using the Random Forest algorithm, with mean Imputation and outlier detection methods applied as preprocessing step.

Finally, this study aims to improve loan prediction using the Random Forest algorithm by applying the mean imputation method and outlier detection techniques. The dataset used is a public Lending club loan dataset from 2014-2021, available on Kaggle. For data preprocessing, mean imputation will be used to address missing values, while three different outlier detections methods will be employed: Boxplot, Z-Score, and Interquartile Range (IQR). Different combinations will be tested (mean imputation with Boxplot, mean imputation with Z-score, and men imputation with IQR). To evaluate the classification or prediction performance, a confusion matrix will be used three parameters: Accuracy, recall, and F1-score.

MATERIALS AND METHODS

Machine Learning

Machine learning (ML) is a part of computer science that builds a specific algorithm in studying a certain phenomenon based on a dataset collected and then modeled with a specific algorithm. There are 4 learning types of ML; a). Supervised Learning; b). Semi-Supervised Learning; c). Unsupervised Learning; and d). Reinforcement Learning.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6437

In the loan prediction study, the type of machine learning used is supervised learning, where the dataset is well labeled. The algorithm discussed in this study is Random Forest, based on the findings from the literature review.

Random Forest Algorithm

Random Forest is built from combining the output of several decision trees or combining the prediction results of several decision trees to produce a more stable and stable method. That's why the Random Forest is an ensemble method. A single Decision Tree often overfits the training data, especially when the tree grows deep, but Random Forest reduces overfitting by introducing randomness in two ways:

a. Bootstrapping: Each tree is trained on a different random subset of the data.
b. Random Feature Selection: At each split, only a random subset of features is considered.

Table 2 presents the overall performance comparison between the single Decision Tree (DT) and the Random Forest (RF) algorithms.

Table 2. Comparation Performance DT and RF

Feature	Decision Tree	Random Forest
Generalization	Poor	Strong
Robustness	Low	High
Overfitting Risk	High	Low
Accuracy	Moderate	High
Feature Importance	Limited	Strong

Source : (Research Results, 2025)

From Table 2 we can conclude that the Random Forest algorithms more reliable than decision tree algorithm, so this study, the Random Forest algorithm is used to perform loan classification or prediction. Previous study on loan prediction that compared two algorithms Decision Tree and Random Forest algorithm, the results indicated that the Random Forest algorithm outperformed the Decision Tree algorithm in terms of performance[8],[18], [19]

Dataset

The dataset used in this study is the public Lending Club loan dataset from 2014 to 2021, obtained from Kaggle, consisting of 466,285 records. The data has been split into 70% training data and 30% test data.

The data preparation process involved deleting features with less than 10% completeness or those containing entirely empty values. Only features with the float or integer data types were selected. Ultimately, 32 features were used in the analysis.

Preprocessing Data

In this study, the public dataset Lending Club loan data from 2014 to 2021 contains missing values in certain features, as shown in Figure 2, and outlier data. To address the missing values, mean imputation will be applied. Additionally, three methods for outlier detection: Boxplot, Z-score, and Interquartile Range (IQR) will be used to identify outliers in the dataset.

1. Imputation Data

Data imputation is the process of replacing missing values, which can also be beneficial in maintaining the completeness of a dataset. There are three types of missing data: 1. MCAR (Missing Completely At Random), 2. MAR (Missing At Random), and 3. MNAR (Missing Not At Random). Based on the definition, the missing data in the Lending Club loan dataset was categorized as MAR. The Data Imputation methods has potential limitation such as:

a. Distortion of data distribution
b. Model too complex than dataset
c. Misinterpretation

The Data Imputation Method used in this study was mean imputation, Mean Imputation are a suitable imputation method for filing missing data at a relatively small and single value is required for replacement[12], [17]. Mean imputation Method uses mean values to replace the null data with the average calculation of the data with the formula (1) below.

$$Mean(X) = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (1)$$

The rationale using Mean Imputation is the simplest statistical method for numeric features, if the missing data less or equal 1% of total data Mean imputation can replaced the missing data to become complete dataset, but if the missing data in domain more than 10% can lead produce biased estimates for each data prediction. In this study, 60% of the missing data in features has amounted less than 1%, therefore the use of the Mean imputation method is a good choice even though there is still a risk of data bias[12], [14].

2. Outlier Detection Methods

Outlier Data can indicate a set of points that significantly differ from the majority of the other data, This often referred to as an outlying observation or contaminant in the data, Outliers can result from sources like an error in the data compilation, editing or coding[20].

The outlier detection methods used in this study are Boxplot, Z-Score, and Interquartile Range (IRQ). these three methods have been implemented in previous study for stunting classification[17].

Boxplot method a visual tool to describes outlier in dataset. Boxplot provides an overview of data distribution based on essential statistics including median, quartiles, and potential outliers.

Z-Score method uses two estimators mean and standard deviation to identify the outliers, The Z-Score can be calculated using the following Equation (2) as follows:

$$Z_i = \frac{x_i - \bar{x}}{sd} \quad (2)$$

Where:
x_i = i-th data value
$\bar{x}$ = Mean value of the dataset
Sd = Standard deviation from the dataset

Interquartile Range (IQR) outlier detect data points that fall significantly outside the range of most other value in the dataset, The data range of IQR is the range between the first (Q1) and the third quartile (Q3) with the following Equation (3) for IQR.

$$IQR = Q_3 - Q_1 \quad (3)$$

The formula below is how IQR detects an outlier data :

$x_i < Q_1 - 2{,}5\ x\ IQR$ Or $x_i > Q_3 + 2{,}5\ x\ IQR$

Another method of detecting outlier is boxplot, Which is a visually represents the IQR, The difference is data that are outside data points which extend from the box edges are considered outliers, The value of whiskers is determined by the following Equation (4) as follows

$$Q_1 - 1{,}5\ x\ IQR \text{ And } Q_3 + 1{,}5\ x\ IQR \quad (4)$$

In this study, the loan prediction experiment was conducted using three outlier detection methods separately. The determination of the outlier detection method to be used in the proposed model was based on the experiment results, selecting the one with the best performance accuracy..

Model Performance Assessment

Models that have been built using training data need to be tested by running the model with test data that has never been run with the model. If the model can make predictions well, then the model built is good. Although the model was running well, but still necessary to assess the model performance with sufficient Metric and tools. Metric and Tools that mostly used to assess the model performance consist of:

a. Confusion Metrix
b. Accuracy
c. Cost-sensitive accuracy,
d. Precision/Recall, and
e. Area under the ROC curve

In this study, model performance assessment for loan prediction was conducted using a Confusion Matrix with three parameters: Accuracy, Recall, and F1-Score. These three parameters were chosen because the F1-score is calculated as the harmonic mean of precision and recall, providing a balanced measure of a model's performance. The formula is shown in Equation (9).

The result of the Lending Club classification or prediction will be evaluated using a confusion matrix. A confusion matrix is a tool that clearly and effectively displays a classifier's performance [21].

To determine accuracy, we will use a confusion matrix, as shown in Table 3. The confusion matrix contains four different combinations: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) which are very useful for measuring Accuracy, Specificity, Recall, Precision and F1-score.

F1-score is a metric that combine precision and recall into a single model performance, Particularly useful to detect uneven class distribution, A high F1-score indicate the model is effective in detecting a significant proportion of true lending cases (high recall) while also maintaining a low rate of false positives (high precision).

Table 3. Confusion Matrix

Confusion matrix		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Source : (Research Results, 2025)

Base on the data from Table 3 the parameters below can be calculated using the equations (5) – (9):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

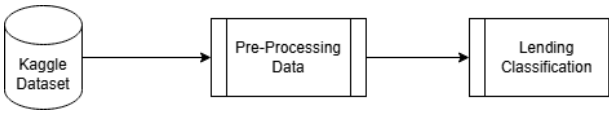
$$\text{F1 Score} = 2x\frac{Precision\ x\ Recall}{Precision+Recall} \quad (9).$$

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6437

RESULTS AND DISCUSSION

Before discussing the results of the prediction, the first step is to examine the design of the experiment, as described in Figure 1. Figure 1 illustrates the flow of the experiment, where the Loan Dataset undergoes preprocessing before the classification process



Source : (Research Results, 2025)
Figure 1. Lending Classification Process

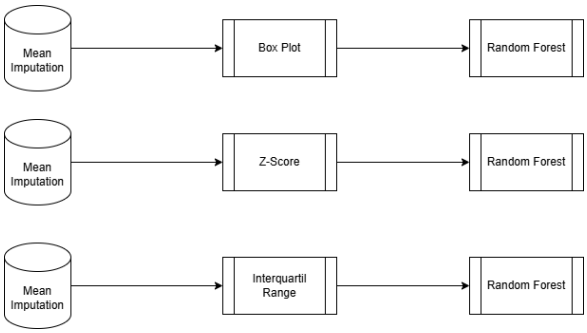
Preprocessing data of The Lending club 2014 – 2021 dataset, started with implementing Mean imputation method to the dataset so the missing data will be replaced with mean imputation data until all the features has 466,285 numbers of data. The dataset that is utilized spans from 2014-2021 with 32 features.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 33 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   loan_amnt                   466285 non-null  int64
 1   funded_amnt                 466285 non-null  int64
 2   funded_amnt_inv             466285 non-null  float64
 3   int_rate                    466285 non-null  float64
 4   installment                 466285 non-null  float64
 5   annual_inc                  466281 non-null  float64
 6   loan_status                 466285 non-null  int64
 7   dti                         466285 non-null  float64
 8   delinq_2yrs                 466256 non-null  float64
 9   inq_last_6mths              466256 non-null  float64
 10  mths_since_last_delinq      215934 non-null  float64
 11  mths_since_last_record      62638 non-null   float64
 12  open_acc                    466256 non-null  float64
 13  pub_rec                     466256 non-null  float64
 14  revol_bal                   466285 non-null  int64
 15  revol_util                  465945 non-null  float64
 16  total_acc                   466256 non-null  float64
 17  out_prncp                   466285 non-null  float64
 18  out_prncp_inv               466285 non-null  float64
 19  total_pymnt                 466285 non-null  float64
 20  total_pymnt_inv             466285 non-null  float64
 21  total_rec_prncp             466285 non-null  float64
 22  total_rec_int               466285 non-null  float64
 23  total_rec_late_fee          466285 non-null  float64
 24  recoveries                  466285 non-null  float64
 25  collection_recovery_fee     466285 non-null  float64
 26  last_pymnt_amnt             466285 non-null  float64
 27  collections_12_mths_ex_med  466140 non-null  float64
 28  mths_since_last_major_derog 98974 non-null   float64
 29  acc_now_delinq              466256 non-null  float64
 30  tot_coll_amt                396009 non-null  float64
 31  tot_cur_bal                 396009 non-null  float64
 32  total_rev_hi_lim            396009 non-null  float64
dtypes: float64(29), int64(4)
memory usage: 117.4 MB
```

Source : (Research Results, 2025)
Figure 2. Dataset Before Imputation

Figure 2 shows that 15 features have missing data, with 9 of them having less than 1% missing values. This was the reason for implementing the Mean Imputation method. The missing data posed a challenge to the integrity and reliability of the dataset.

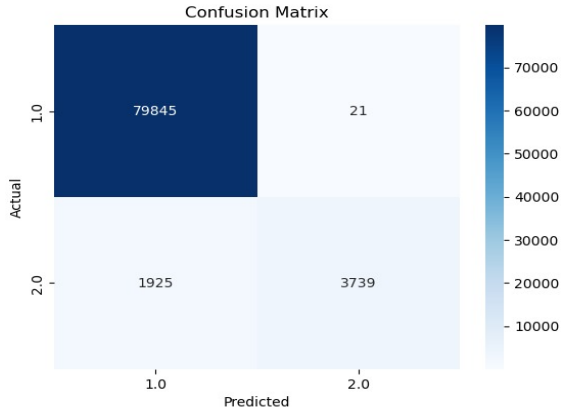
The next step is to evaluate the dataset using three different methods of data outlier detection: Boxplot, Z-Score, and Interquartile Range. After that, the dataset will be classified using the Random Forest algorithm. The model's performance will be evaluated using a Confusion Matrix by measuring three parameters; Accuracy, Recall and F1 Score. This study defines two loan classification categories: approved loan classification and rejected loan classification.



Source : (Research Results, 2025)
Figure 3. Lending Classification with 3 Outlier Detection Method

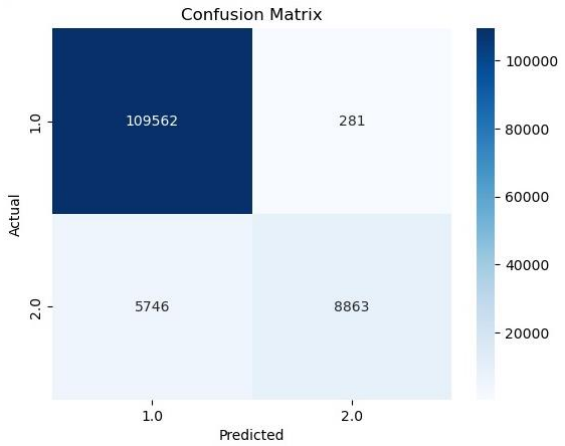
Figure 3 shows the next step, which is the recognition of outlier data. We combined mean imputation with 3 different methods: Boxplot, Z-Score, and Interquartile Range. Each combination produces different loan classification accuracy. The results of the loan classification are evaluated using three parameters: accuracy, recall, and F1-score. Accuracy indicates how correct the classification is, regardless of True Positives or False Negatives.

Figure 4 through 6 show the results of model performance assessment using confusion matrix for three outlier detection methods: Boxplot, Z-Score, and Interquartile Range (IQR). Using formulas (6) to (10), Accuracy, Recall, and F1 Score can be calculated.



Source : (Research Results, 2025)
Figure 4. Prediction Result Using Boxplot

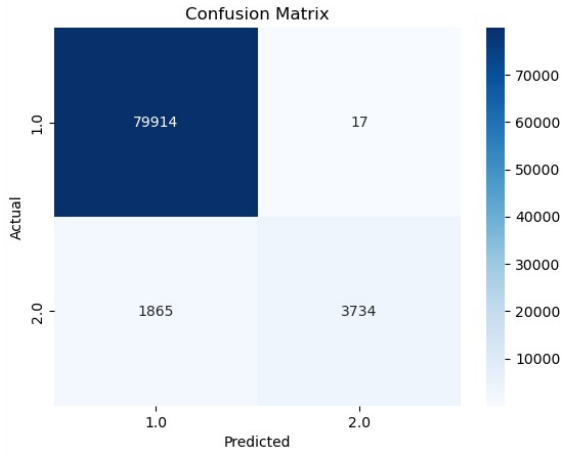
Figure 4 shows the combination of mean imputation and Boxplot. The confusion matrix results are as follows: True Positive (TP) 79,845 predictions, False Positive (FP) 1,925 predictions, False Negative (FN) 21 predictions, and True Negative (TN) 3,739 predictions.



Source : (Research Results, 2025)

Figure 5. Prediction Result Using Z-Score

Figure 5 shows the combination of mean imputation and Z-score. The confusion matrix results are as follows: True Positive (TP) 109,562 predictions, False Positive (FP) 5,746 predictions, False Negative (FN) 281 predictions, and True Negative (TN) 8,863 predictions.



Source : (Research Results, 2025)

Figure 6. Prediction Result Using IQR

Figure 6 shows the combination of mean imputation and Interquartile Range (IQR). The confusion matrix results are as follows: True Positive (TP) 79,914 predictions, False Positive (FP) 1,865 predictions, False Negative (FN) 17 predictions, and True Negative (TN) 3,734 predictions.

Table 4 presents the results of Loan prediction based on values of TP, FP, FN, and TN from Figures 4 to 6. Using Equations (5) to (9), the results are summarized according to Accuracy, Recall, and F1 Score as follows:

Table 4. Result of Loan Prediction

Outlier Detection Method	Accuracy	Recall	F1 Score
Boxplot	97.72%	97.64%	98.79%
Interquartile Range	97.80%	97.72%	98.84%
Z-Score	95.16%	95.01%	97.31%

Source : (Research Results, 2025)

In terms of accuracy, the Z-Score method achieved the lowest accuracy at 95,16%, while the Highest accuracy achieved Interquartile Range method at 97,80%. In term of recall, the Z-Score method achieved the lowest value at 95,01%, while the Interquartile Range method achieved the highest at 97,72%. In term of F1-Score, the Z-Score method achieved the lowest score at 97,31%, and the Interquartile Range method achieved the higher at 98,84%. Both the Boxplot method and Interquartile Range (IQR) method achieved exceed 97% accuracy, but overall considering accuracy, recall, and F1-score the Interquartile method performed better than the Boxplot method. This is because outlier detection using the Interquartile method determines the upper and lower bounds using formula (6). A data point is classified as an outlier if its value falls below the lower bound or above the upper bound. Since outlier detection is applied to each individual data point, the Interquartile Range method achieves better accuracy compared to the Boxplot method.

From Table 3, we conclude that the impact of preprocessing techniques has improved loan classification and prediction. The accuracy of loan prediction exceeded 95%, Recall exceeded 95% and F1-Score exceeded 97%. These results are much better than a previous study, which reported an accuracy 83% [9]. The best prediction was achieved by implementing the Mean Imputation method combined with Interquartile Range (IQR) for outlier detection, resulting in an accuracy of 97.71%, a recall of 99.97%, and an F1 Score of 98.82%. Implication of finding, for banking industry or financial institutions as follows:

1. Provide better assess the creditworthiness of the customers, and avoid bad credit risk.
2. Enhance costumers trust.
3. Speed-up the process and evaluates loan request from customers.

We encourage researchers to conduct further studies, as there is still room for improvement by

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6437

incorporating different data pre-processing techniques, algorithms, and datasets to enhance the performance of the loan prediction model. For future research, we plan to develop an IoT-based loan approval application. This application will require customers to provide some metadata or financial information. Once the customer submits the metadata, the ML model will process it to make a loan prediction. Customers will be able predict whether their loan application is likely to be disapproved based on their current financial condition, allowing them to better plan their finances.

CONCLUSION

Based on the experimental results and analysis of loan prediction using the Lending Club dataset from 2014 to 2021, it can be concluded that the optimal design for a loan prediction model using the Random Forest algorithm is achieved when data preprocessing includes the Mean Imputation method for handling missing values and the Interquartile Range (IQR) method for outlier detection. The proposed approach demonstrated excellent performance, achieving accuracy rates above 95%, recall exceeding 95%, and an F1-score surpassing 97%, highlighting its effectiveness in enhancing loan prediction outcomes.

The decision to use mean imputation to handle missing data could be a limitation in the loan prediction results. For future projects, the pre-processing process could be revisited by implementing various data imputation methods such as kNN imputation, Max Imputation, and others. Additionally, a study on feature selection for loan prediction is important, as It is crucial to identify which features have a significant impact on the prediction. This will help establish the essential metadata that customers must provide in the development of an IoT-based loan approval prototype, simplifying data processing.

REFERENCE

[1] E. Yudisthira and M. Barthos, "Key Factors and Legal Obstacles in Banking Loan Approval," European Alliance for Innovation n.o., Feb. 2022. doi: 10.4108/eai.30-10-2021.2315743.

[2] S. Alvionita, "Sistem Informasi Pengajuan Pinjaman Kredit Usaha Rakyat (KUR) Pada Bank Rakyat Indonesia (BRI) Unit Sukarame," *Ilmudata.org*, vol. 2, no. 2, pp. 1–13, 2022.

[3] H. Haeruddin, E. Erick, and H. W. Aripradono, "Perbandingan Support Vector Machine, Random Forest Classifier, dan K-Nearest Neighbour dalam Pendeteksian Anomali pada Jaringan DDos," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 7, no. 1, pp. 23–33, Jan. 2025, doi: 10.35746/jtim.v7i1.628.

[4] B. Yi, "P2P Investment Data Analytics: A Case Study of Lending Club," 2023.

[5] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, Feb. 2022, doi: 10.1007/s00607-021-00908-y.

[6] X. Dong, "Loan Default Prediction based on Machine Learning (LightGBM Model)," 2022.

[7] R Nancy Deborah, S Alwyn Rajiv, A Vinora, C Manjula Devi, S Mohammed Arif, and G S Mohammed Arif, "An Efficient Loan Approval Status Prediction Using Machine Learning," in *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, Mumbai, India: IEEE Xplor, Oct. 2023.

[8] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," *IOP Conf Ser Mater Sci Eng*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012042.

[9] B. Prasojo and E. Haryatmi, "Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman dengan Metode Random Forest," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 2, pp. 79–89, 2021, doi: 10.25077/teknosi.v7i2.2021.79-89.

[10] K. B. Simarmata, K. D. Hartomo, and K. D. Hartomo, "Analisa Rekomendasi Fitur Persetujuan Pinjaman Perusahaan Financial Technology Menggunakan Metode Random Forest," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 9, no. 3, pp. 2055–2070, 2022, doi: 10.35957/jatisi.v9i3.2258.

[11] A. Mirzaei, S. R. Carter, A. E. Patanwala, and C. R. Schneider, "Missing data in surveys: Key concepts, approaches, and applications," *Research in Social and Administrative Pharmacy*, vol. 18, no. 2, pp. 2308–2316, 2022, doi: 10.1016/j.sapharm.2021.03.009.

[12] A. Desiani, N. R. Dewi, A. N. Fauza, N. Rachmatullah, M. Arhami, and M. Nawawi, "Handling Missing Data Using Combination of Deletion Technique, Mean, Mode and

Artificial Neural Network Imputation for Heart Disease Dataset," *Science and Technology Indonesia*, vol. 6, no. 4, pp. 303–312, 2021, doi: 10.26554/sti.2021.6.4.303-312.

[13] H. Nugroho, N. Priya utama, and K. Surendro, "kNN Imputation Versus Mean Imputation for Handling Missing Data on Vulnerability Index in Dealing with Covid-19 in Indonesia," in *The 2023 12th International Conference on Software and Computer Applications*, Kuantan, Malaysia, Feb. 2023, pp. 20–25.

[14] L. O. Joel, W. Doorsamy, and B. S. Paul, "On the Performance of Imputation Techniques for Missing Values on Healthcare Datasets," pp. 1–20, 2024, [Online]. Available: http://arxiv.org/abs/2403.14687

[15] K. Seu, M.-S. Kang, and H. Lee, "An Intelligent Missing Data Imputation Techniques: A Review," *INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION*, vol. 6, no. May, pp. 278–283, May 2022, [Online]. Available: www.joiv.org/index.php/joiv

[16] A. K. Srivastava, Y. Kumar, and P. K. Singh, "Hybrid diabetes disease prediction framework based on data imputation and outlier detection techniques," *Expert Syst*, vol. 39, no. 3, Mar. 2022, doi: 10.1111/exsy.12785.

[17] Ni'matul Ma'muriyah, P. Purwanto, E. Noersasongko, S. Winarno, and M. I. Ashiddiq, "XG Boost Based Data Imputation and Outlier Detection Methods for Classification of Stunting," in *7th International Seminar on Research of Information Technology And Intelligent Systems (ISRITI)*, yogjakarta, Dec. 2024, p. 109.

[18] N. Sri Sai Venkata Subba Rao, S. John Justin Thangaraj, and Saveetha, "Flight Ticket Prediction using Random Forest Regressor Compared with Decision Tree Regressor," in *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India : IEEE, Apr. 2023.

[19] Q. Zhang, "Financial Data Anomaly Detection Method Based on Decision Tree and Random Forest Algorithm," *Journal of Mathematics*, vol. 2022, 2022, doi: 10.1155/2022/9135117.

[20] S. Saleem, M. Aslam, and M. Rukh Shaukat, "A REVIEW AND EMPIRICAL COMPARISON OF UNIVARIATE OUTLIER DETECTION METHODS," 2021.

[21] H. Yun, "Prediction model of algal blooms using logistic regression and confusion matrix," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2407–2413, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.