IJMR E-ISSN: 3109-9777
INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEACH

OVERCOMING BIAS FROM MISSING VALUES IN MICROCREDIT DATA BY COMPARING MICE AND IMPUTATION METHODS

Khalisha Alya Putri[1], Nadhin Mutiara Hervani[2], Istiqomah[3], Fenny Purwani[4]
[1] Fakultas Sains dan Teknologi, Universitas Islam Negeri Raden Fatah Palembang, Indonesia
[2] Fakultas Sains dan Teknologi, Universitas Islam Negeri Raden Fatah Palembang, Indonesia
3 Fakultas Sains dan Teknologi, Universitas Islam Negeri Raden Fatah Palembang, Indonesia
4 Fakultas Sains dan Teknologi, Universitas Islam Negeri Raden Fatah Palembang, Indonesia
Email : khalishaalya11@gmail.com , fennypurwani_uin@radenfatah.ac.id

E-ISSN : 3109-9777		
Received: November 2025	Accepted: November 2025	Published: Desember 2025

Abstract :

Missing values are one of the main problems in financial data processing, especially in microcredit data. The presence of incomplete data can cause bias, disrupt variable distribution, and reduce the performance of classification models in determining creditworthiness. This study aims to compare two imputation approaches, namely Simple Imputation (median for numerical attributes and mode for categorical attributes) and the MICE (Multiple Imputation by Chained Equations) Method, in an effort to reduce bias due to missing values and improve classification prediction performance. The dataset used is Loan Payments Data from Kaggle, which contains 500 rows of data and 11 attributes, namely Loan_ID, loan_status, Principal, terms, effective_date, due_date, paid_off_time, past_due_days, age, education, and Gender. After the data cleaning process, outlier handling, and imputation using both methods, the data was predicted using two classification models, namely Logistic Regression and Random Forest. Model performance was evaluated using the Accuracy and AUC (Area Under the ROC Curve) metrics. The results showed that the MICE method produced higher and more stable performance compared to Simple Imputation. Logistic Regression increased from an accuracy of 66.67% to 82.00%, and AUC from 71.02% to 95.03%. The Random Forest model on Simple Imputation data achieved 100% accuracy and 100% AUC, but these overly perfect values potentially indicate overfitting, a condition where the model memorises specific patterns in the training data and is less able to generalise. On the MICE imputation data, Random Forest still achieved high performance with an accuracy of 98.00% and an AUC of 99.55%, which is considered more realistic and stable. These findings indicate that the MICE method is more effective in reducing bias due to missing values and improving the reliability of microcredit risk classification results.

Keywords : *Missing Values, Simple Imputation, MICE, Microcredit, Classification, AUC, Accuracy.*

INTRODUCTION

Data quality is an important factor in predictive analysis and data-driven decision making. In the context of microcredit assessment, financial institutions and financing cooperatives rely heavily on various attributes such as Principal (loan amount), terms (loan period), effective_date (loan start date), due_date (maturity date), paid_off_time (repayment date), past_due_days (number of days past due), age (borrower's age), education (level of education), and gender to determine whether a customer is eligible or ineligible for a loan.

However, data obtained in the field often contains missing values due to

CC BY SA

OPEN ACCESS

incomplete records, late reporting, or input errors. Missing values can cause statistical bias because machine learning models will learn from data that no longer represents the actual conditions. If not handled properly, this bias can affect the model evaluation results and reduce the reliability of creditworthiness predictions.

One approach that is often used is the Simple Imputation Method, which replaces missing values in numerical variables with the median, and categorical variables with the mode. This method is easy to apply, but it does not take into account the relationship between variables, which can cause the loss of natural variation in the data. As a result, the classification model becomes less accurate in distinguishing between eligible and ineligible customers (Widyananda et al., 2023).

As an alternative, the Multiple Imputation by Chained Equations (MICE) method was developed, which fills in missing values iteratively by building regression models between variables. This method maintains the relationships between features in the data, resulting in a more representative and realistic dataset. Various studies have shown that the application of the MICE method can reduce bias and improve the stability of classification results, both in the medical and financial fields (Sharifnia et al., 2025). In the context of microcredit data, the use of MICE can also improve the quality of the dataset and enhance the performance of creditworthiness prediction models.

Microcredit eligibility assessment in this study is modelled as a binary classification problem, namely distinguishing between customers who are eligible and ineligible for loans (Lesmana et al., 2025; Law et al., 2019). Two classification algorithms are used: Logistic Regression, due to its simplicity and ability to produce models that can be interpreted well, and Random Forest, which has the ability to handle complex data and non-linear relationships between variables (Law et al., 2019).

The performance of both models was evaluated using two main metrics, namely Accuracy and AUC (Area Under the ROC Curve). Accuracy measures the proportion of correct predictions from all test data, while AUC measures the model's ability to distinguish between two creditworthiness classes, which is especially important when the data distribution is imbalanced (Widyananda et al., 2023; Law et al., 2019).

Based on this background, this study has two main objectives: (1) to compare the effects of the Simple Imputation Method and the MICE Method on the quality of microcredit data, and (2) to evaluate their impact on the performance of credit classification models using Logistic Regression and Random Forest, measured through the Accuracy and AUC metrics.

The main contribution of this study is to provide empirical evidence that the choice of imputation method has a direct influence on the performance of microcredit eligibility classification models, not just on the quality of raw data. These findings are relevant for microfinance institutions, cooperatives, and People's Business Credit (KUR) distributors that require a fast, efficient, and reliable risk assessment system (Putra et al., 2025; Amaliah et al., 2025)

RESEARCH METHOD

1.Dataset

The dataset used is Loan Payments Data from Kaggle, which is often used for credit risk studies and microloan payment performance. The Loan Payment Information dataset is also used as a representation of microcredit data because it includes key variables that reflect the behaviour and eligibility of micro customers. This dataset consists of 500 rows of data with 11 main attributes or variables as shown in Table 1:

Table 1. Attributes or Variables of the Microcredit Dataset

No	Attribute Name	Description
1	Loan_ID	Unique identifier for each loan
2	loan_status	Loan repayment status
3	Principal	Loan principal amount
4	terms	Loan term (in days)
5	effective_date	Loan disbursement date
6	due_date	Loan due date
7	paid_off_time	Loan repayment date (if fully paid)
8	past_due_days	Number of days past due
9	age	Borrower's age
10	education	Borrower's education level
11	gender	Borrower's gender

The loan_status column in the dataset has three categories, namely Paidoff, Collection_Paidoff, and Collection.
For creditworthiness classification purposes, these labels are simplified into two classes: creditworthy (0) for customers with PAIDOFF status, and uncreditworthy (1) for Collection and Collection_Paidoff status.

2. Data Preprocessing

The following preprocessing steps were performed:

a. Duplication Removal and Consistency
   Duplications were removed based on Loan_ID. Date columns (effective_date, due_date, paid_off_time) were converted to a uniform time format.
b. Outlier Handling
   Extreme values in numeric variables (Principal, terms, age, past_due_days) are controlled using the IQR (Interquartile Range) method to prevent distortion in the classification model (Sharifnia et al., 2025).
c. Target Transformation
d. The loan_status column is given a numerical label, for example,

'PAIDOFF' = 0, "COLLECTION and COLLECTION_PAIDOFF" = 1, so that it can be used as a binary target variable (Lesmana et al., 2025).

e. Categorical Encoding
The categorical variables education and Gender are converted into numerical form using label encoding so that they can be processed by the machine learning model.

3. Missing Value Imputation

The study compared two imputation approaches:

a. Simple Imputation Method
Missing numerical values are filled in using the column median, while categorical values are filled in with the column mode. This method is common because it is simple and fast, but it tends to reduce data diversity and can introduce bias (Widyananda et al., 2023).

b. Multiple Imputation by Chained Equations
MICE performs sequential imputation: each missing variable is predicted with a regression model using other variables as predictors, repeated iteratively until convergence. This method preserves the structure of relationships between variables and is empirically reported to be more stable for financial data (Sharifnia et al., 2025).

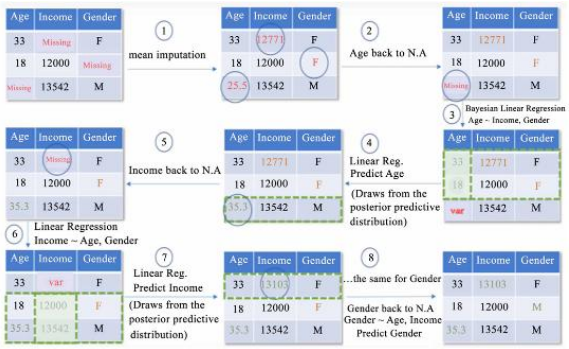


Figure 1. Illustration of the data imputation process using the MICE algorithm (adapted from https://youtu.be/zX-pacwVyvU?si=9jN1Prjmbmikm ZJ2).

4. Classification Modelling

Based on two models that were compared:

a. Logistic Regression, which statistically models the probability of binary classes and is often used as an interpretable baseline in credit risk assessment (Lesmana et al., 2025; Law et al., 2019).

b. Random Forest, which is an ensemble of decision trees and is known to be robust against non-linearity, feature interactions, and noise (Lesmana et al., 2025; Law et al., 2019).
The data was divided into training and test data using a 70%-30% train-test split. After training the model on the imputed data, the model was tested on the test data to obtain performance metrics.

5. Evaluation Metrics

This study uses two main metrics to measure the performance of the

classification model:

a. Accuracy, which is the proportion of correct predictions from all test samples. This metric shows how often the model makes the right decision in classifying customers as creditworthy or uncreditworthy.
b. AUC (Area Under the ROC Curve), which is a measure of the model's ability to distinguish between two creditworthiness classes—creditworthy and uncreditworthy customers. A high AUC value indicates that the model is able to provide good probability rankings to accurately distinguish between the two groups. This metric is particularly important when the data distribution is imbalanced (Law et al., 2019).

FINDINGS AND DISCUSSION

1. Experimental Results

After the imputation process, both classification models were trained and tested. The summary results are shown in Table 2.

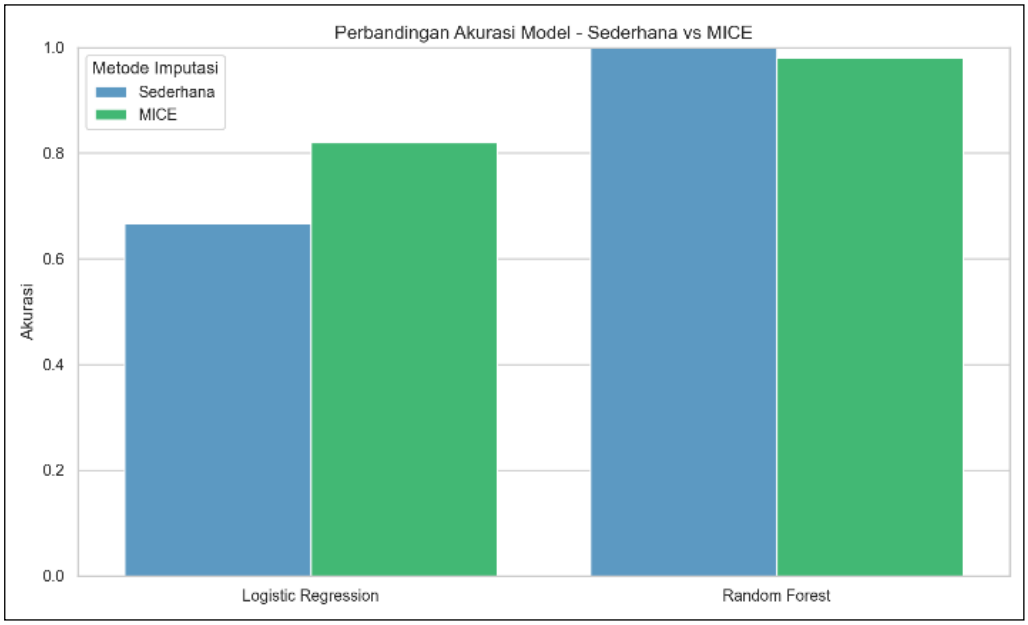
Table 2. Classification Model Evaluation Results for Each Imputation Method

Model	Imputation Method	Accuracy (%)	AUC (%)	Interpretation
Logistic Regression	Simple	66.67	71.02	Less stable model
Random Forest	Simp	100.00	100.00	Overfitting (too perfect)
Logistic Regression	MICE	82.00	95.03	Better and more realistic
Random Forest	MICE	98.00	99.55	Stable and high

The results show that the consistent use of the MICE method improves the performance of Logistic Regression, both in terms of accuracy (increasing from 66.67% to 82.00%) and AUC (increasing from 71.02% to 95.03%). This increase in AUC is very important because AUC measures the model's ability to distinguish between creditworthy and non-creditworthy customers, not just how often the model makes correct predictions (Law et al., 2019; Haganawiga et al., 2025).

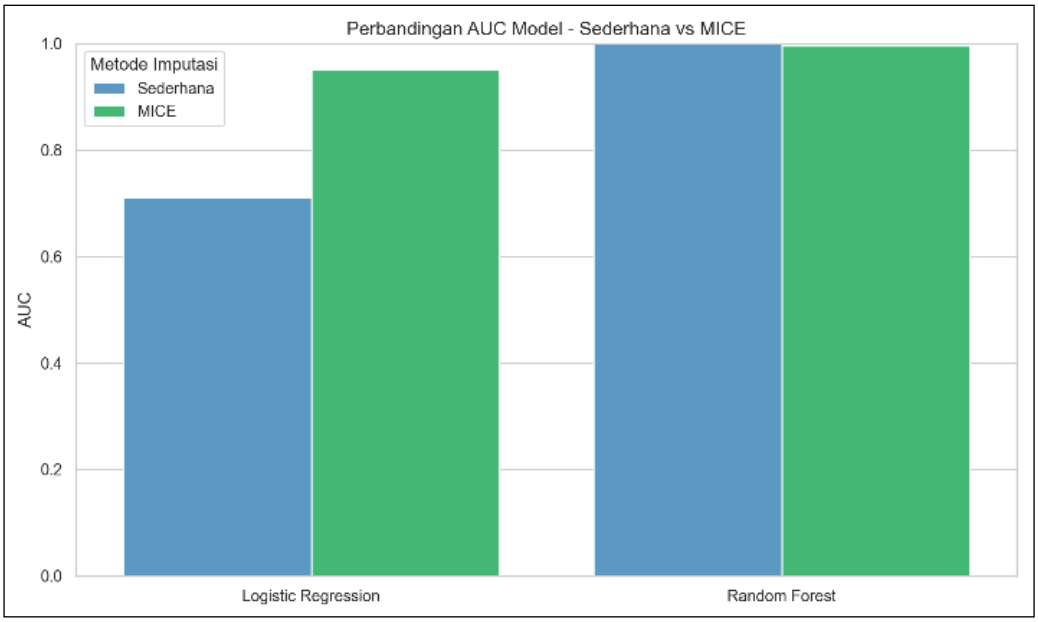
Random Forest produced 100% accuracy and 100% AUC on the Simple Imputation data. Although this seems ideal, such perfect values often indicate the risk of overfitting, a condition where the model 'memorises' specific patterns from the training data and is no longer generic. This is reinforced by the fact that on the MICE imputation data, the performance of Random Forest decreased slightly to an accuracy of 98.00%

and an AUC of 99.55%, which can actually be considered more realistic and



has better generalisation potential on new data [4].

Figure 2. Comparison of the accuracy of Logistic Regression and Random Forest models in Simple Imputation vs MICE



Gambar 3. Perbandingan nilai AUC model Logistic Regression dan Random Forest pada Imputasi Sederhana vs MICE.

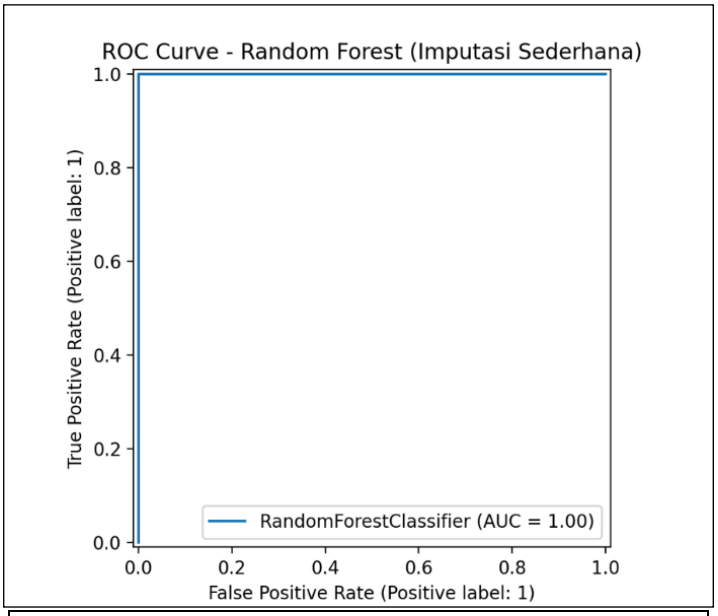
2. ROC Curve Analysis

a. ROC Curve Analysis – Random Forest

The ROC curve for the Random Forest model was generated from two missing data imputation methods, namely the Simple Imputation

Method and the MICE (Multiple Imputation by Chained Equations).

Figure 4. ROC Curve – Random Forest (Simple Imputation Method)



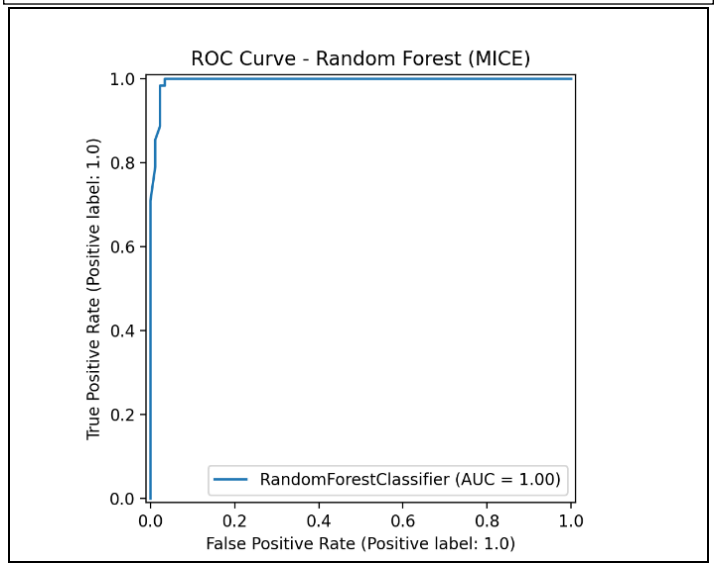


Figure 5. ROC Curve – Random Forest (MICE Method)

In both graphs, the area under the curve (AUC) is close to 1.0. This indicates that the model has an almost perfect ability to distinguish between creditworthy and non-creditworthy customers. However, the AUC that is too perfect (100%) in the Simple Imputation Method indicates potential overfitting, where the model adapts too much to the training data and loses its ability to generalise on new data.

In contrast, the results from the MICE Method show a slightly lower but more stable and realistic AUC, making it more representative for real-world microcredit risk prediction applications.

b. ROC Curve Analysis – Logistic Regression

In addition to Random Forest, ROC curves were also generated for the Logistic Regression model using the same two missing data imputation methods.

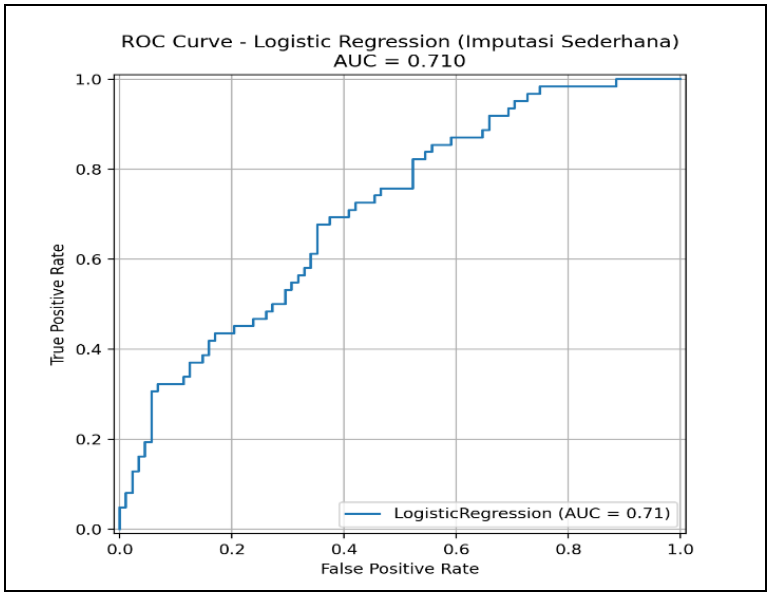


Figure 6. ROC Curve – Logistic Regression (Simple Imputation Method)

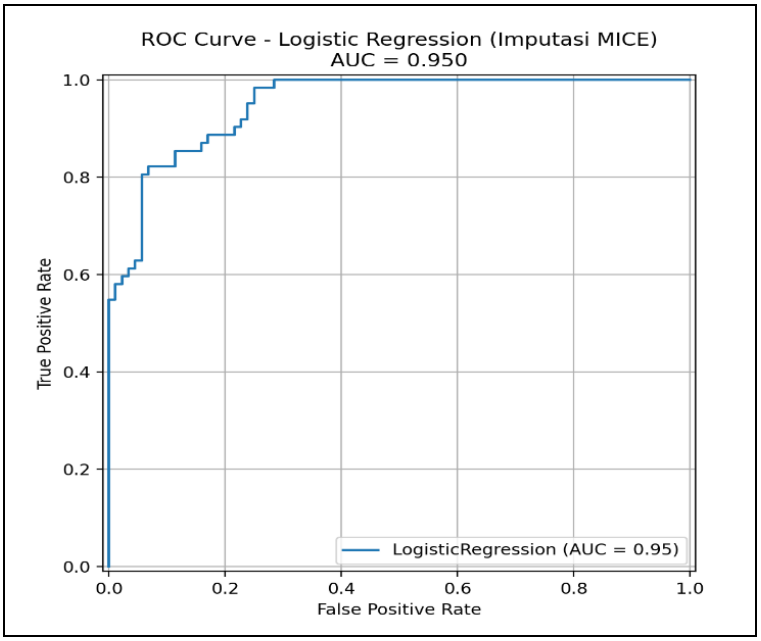


Figure 7. ROC Curve – Logistic Regression (MICE Method)

In both graphs, the AUC value shows the same trend, namely that the AUC increased significantly after the application of the MICE method, from around 71.02% to 95.03%,

indicating that the model's ability to distinguish between eligible and

ineligible classes increased dramatically.

This reinforces that the MICE Method provides better prediction results because it considers the relationship between variables in the imputation process, whereas Simple Imputation only fills in based on the median or mode value without considering other data contexts.

A high AUC is very important in the context of microcredit, because mispredicting unviable customers as viable can cause direct financial losses for financial institutions. Conversely, wrongly rejecting viable customers can hamper financial inclusion and narrow access to productive capital (Lesmana et al., 2025; Putra et al., 2025; Haganawiga et al., 2025).

CONCLUSION

This study compares two methods of missing value imputation, namely Simple Imputation (median/mode) and MICE (Multiple Imputation by Chained Equations), on microcredit data taken from the Kaggle Loan Payments dataset. The two imputation results were then evaluated using Logistic Regression and Random Forest classification models, with Accuracy and AUC metrics as performance measures.

The results show that: 1) The MICE method produced a significant improvement in performance on Logistic Regression (accuracy 82.00%; AUC 95.03%), compared to Simple Imputation (accuracy 66.67%; AUC 71.02%). 2) Random Forest achieved very high performance in both Simple Imputation and MICE, but the perfect accuracy/AUC (100%) in Simple Imputation may indicate overfitting, rather than simply proving that the data is 'perfect'. 3) Thus, MICE is more effective in reducing bias due to missing values and improving the overall stability of microcredit classification results.

REFERENCES

Widyananda, W., Purnomo, M. F. E., Aswin, M., Mudjirahardjo, P., &
Pramono, S. H. (2023). Application of data mining and imputation algorithms for missing value handling: A study case car evaluation dataset. *Iraqi Journal of Science*, 64(5), 2481–2491. https://doi.org/10.24996/ijs.2023.64.5.32

Sharifnia, A. M., Kpormegbey, D. E., Thapa, D. K., & Cleary, M. (2025).
A primer of data cleaning in quantitative research: Handling missing values and outliers. *Journal of Advanced Nursing*. https://doi.org/10.1111/jan.16908

Lesmana, R. A., Budiman, S. N., Shodiqi, A. A., Nadhila, J. K., Aziz, M.
F. N., & Faizal, A. I. (2025). Implementasi algoritma decision tree-ID3 untuk prediksi kelayakan kredit berbasis web dengan menggunakan Next.js di KSU Syariah Muhammadiyah.

Law, M. T., et al. (2019). Machine learning in secondary progressive
multiple sclerosis: An improved predictive model for short-term disability progression. *Multiple Sclerosis Journal – Experimental, Translational and Clinical*, 5(4). https://doi.org/10.1177/2055217319885983

Putra, D. F., et al. (2025). Evaluasi dampak kredit mikro terhadap konsumsi rumah tangga penerima kredit mikro di Indonesia.

Amaliah, H., Taan, H., & Artikel, R. (2025). Analisis kelayakan pemberian dana kredit usaha rakyat (KUR) dalam mengantisipasi terjadinya kredit bermasalah pada perbankan. *Jambura Accounting Review*, 6(1), 261–270.

Haganawiga, K. J., Pal, S. K., & Sirohi, A. (2025). A choice of performance metrics for evaluating predictive accuracy of survival models.