Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 2, July 2025, pp. 466 – 477
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

The Ensemble Supervised Machine Learning for Credit Scoring Model in Digital Banking Institution

[1]Narita Ayu Prahastiwi, [2]Muharman Lubis, [3]Hanif Fakhrurroja
[1,2,3]Department of Information System, Telkom University, Bandung, Indonesia
Email: [1]naritaayup@gmail.com, [2]muharmanlubis@telkomuniversity.ac.id,
[3]haniffakhrurroja@telkomuniversity.ac.id

Article Info
Article history:
Received Jun 2nd, 2025
Revised Jul 9th, 2025
Accepted Jul 30th, 2025

Keyword:
Credit Scoring
Ensemble Supervised Learning
Machine Learning
Non Performing Loan
XGBoost

ABSTRACT
The digital transformation of the banking industry requires credit scoring systems that are both accurate and adaptable to complex, diverse data. This study aims to develop and evaluate a credit scoring model using ensemble supervised learning to predict credit risk for a consumer loan service (Product X) at Bank XYZ. Ensemble algorithms such as Random Forest, AdaBoost, LightGBM, CatBoost, and XGBoost were compared to a single classification method, Decision Tree. Model performance was assessed using precision, recall, F1-score, and ROC-AUC. The results show that XGBoost outperformed other models, achieving the highest ROC-AUC score of 0.803, indicating strong generalization and low risk of overfitting. SHAP analysis revealed key features influencing the model, including loan tenor, loan amount (plafond), income, and Days Past Due (DPD) history. Compared to the baseline Decision Tree model (ROC-AUC 0.573), XGBoost significantly improved classification accuracy. It also showed the potential to reduce the Non-Performing Loan (NPL) rate from 4% to below 3% and increase the approval rate from 65% to over 70%, aligning with Product X's KPIs. These findings confirm that ensemble learning models especially XGBoost offer strategic value in enhancing credit portfolio quality and decision-making in digital banking.

Corresponding Author:
Narita Ayu Prahastiwi,
Departement of Information System,
Telkom University,
Jl. Telekomunikasi No. 1, Bandung Terusan Buahbatu - Bojongsoang, Sukapura, Kec. Dayeuhkolot, Kabupaten Bandung, Jawa Barat 40257.
Email: naritaayup@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v8i2.37677

1. INTRODUCTION

The advancement of technology and information in the digital era has significantly transformed the financial services industry, including in Indonesia. The rapid digitalization of financial services has led to the emergence of digital banks, which offer full banking services through digital platforms without requiring physical branch visits [1]. This transformation enables consumers to access financial products more quickly and efficiently via mobile devices, marking a substantial shift from traditional banking practices.

One of the main services provided by digital banks is unsecured personal loans, known in Indonesia as Unsecured Loan (*Kredit Tanpa Agunan* (KTA)). These loans can be accessed instantly through mobile applications using digital identity verification such as electronic ID (e-KTP) and transaction history [2],[3]. Unlike conventional loans that require collateral, digital banks often rely on customer data and internal credit assessments to make lending decisions. This practice is in line with the Indonesian Banking Law Article 8 paragraph (1), which allows banks to extend credit without collateral as long as the bank is confident in the borrower's repayment capacity [4], and is further supported by Financial Services Authority (*Otoritas Jasa Keuangan* (OJK)) Regulation No. 12/POJK.03/2018 on Digital Banking Services [5].

In addition to digital banks, the peer-to-peer (P2P) lending sector has grown rapidly as part of Indonesia's fintech ecosystem. These platforms distribute loans online and are regulated under OJK Regulation No. 10/POJK.05/2022 on Technology-Based Joint Funding Services [6]. However, the ease of access to credit through digital channels brings new risks, particularly in credit risk management. In many cases, lenders have limited visibility into the financial stability of prospective borrowers, increasing the likelihood of default or Non-Performing Loans (NPL). Therefore, accurate and adaptive credit scoring models are essential to minimize NPL and maintain financial stability [7]. Despite the growing complexity of financial data, many institutions still rely on traditional rule-based models such as logistic regression due to organizational constraints and ease of implementation. However, these models often struggle to capture nonlinear relationships in data, such as transaction behavior and repayment patterns. Previous studies, such as [8], have evaluated logistic regression alongside other classifiers like Decision Tree (DT) and K-Nearest Neighbors (K-NN) in creditworthiness analysis, but emphasized that conventional models should be reassessed as technology evolves.

Machine learning (ML), particularly ensemble-based approaches, has gained prominence as a powerful alternative in credit scoring applications. ML enables systems to learn from data and improve performance without being explicitly programmed [9]. Ensemble ML, which combines multiple base models, can better capture complex patterns in digital data and produce more accurate predictions compared to single classifiers [10]. These models classify loan applications into categories such as NPL/bad customers and Performing Loans (PL/good customers) [11]. Supervised ensemble learning methods such as Random Forest, Gradient Boosting, and XGBoost have become increasingly relevant for credit risk modeling. These algorithms are capable of processing large datasets, identifying nonlinear dependencies, and improving prediction accuracy [12]. Recent studies have consistently demonstrated that ensemble ML models outperform traditional statistical methods in credit scoring. For instance, [13] and [14] found that ensemble models, particularly those using boosting techniques, delivered superior performance in terms of accuracy and Area Under the Curve (AUC) compared to baseline models like logistic regression.

Recent developments in credit risk assessment highlight the growing use of ensemble supervised ML to enhance predictive performance and operational efficiency. While prior studies have demonstrated the superiority of ensemble methods such as XGBoost, LightGBM, CatBoost, and Random Forest over traditional rule-based and single-classifier approaches, most have focused on generic or publicly available datasets, with limited emphasis on digital banking contexts in emerging markets. The novelty of this study lies in its application of multiple ensemble algorithms XGBoost, CatBoost, AdaBoost, LightGBM, and Random Forest compared against a DT baseline using real-world credit data from Bank XYZ, a digital bank in Indonesia. Unlike previous research, this work integrates a comparative performance evaluation with interpretability analysis via SHAP to identify key credit risk drivers, ensuring the model is both accurate and explainable. The results not only demonstrate superior predictive capability, particularly for imbalanced and complex datasets, but also show tangible business impact by improving approval rates and reducing NPL ratios addressing critical KPI gaps in the bank's credit risk management strategy.

This study responds to the practical challenges faced by Bank XYZ, where the non-performing loan rate reached 4% at the end of 2023 exceeding the 3% threshold while the approval rate remained at 65%, below the Key Performance Indicator (KPI) target of 70%. These metrics indicate that the existing credit scoring system fails to adequately balance credit risk detection and business growth. Therefore, the objective of this research is to develop and evaluate credit scoring models using ensemble supervised learning based on historical loan transaction data from Bank XYZ. The study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, encompassing business understanding, data understanding, data preparation, modeling, evaluation, and deployment stages. The final goal is to deliver an optimized and applicable credit scoring model to support more accurate credit decisions in the context of digital banking.

2. RESEARCH METHOD

The conceptual model serves as a visual or narrative representation of a scientific framework that illustrates the relationships among variables based on relevant theories, prior research findings, and the contextual background of the study [15]. It guides the research process systematically, from problem formulation to data collection, analysis, and interpretation. This study adopts the Tree of Research (TOR) conceptual model, developed by Lubis [16], as a structured and systematic framework to achieve its research objectives. The TOR model conceptualizes research as a tree structure comprising five stages as in Figure 1 Tree of Root. The Root of the Tree identifies the research problem, supported by a literature review. The Trunk of the Tree defines research objectives and outlines data collection methods. The Branch of the Tree explains the analytical methods employed. The Crown of the Tree represents the core of the study, in which an ensemble supervised learning-based credit scoring model is developed using historical consumer loan data (Product X) from Bank XYZ, a digital banking institution. Finally, the Peak of the Tree signifies the expected

research outcomes. This model provides a coherent and methodical pathway for conducting data-driven research in the field of credit risk modeling.

After designing the conceptual model, this study adopts the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, which provides a structured and systematic approach to data mining and ML projects. The CRISP-DM process comprises six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase plays a vital role in ensuring that the developed model is aligned with business goals and data characteristics [17]. Figure 1 describe the following is an explanation and application of each stage of CRISP-DM and its application in designing credit scoring based on an ensemble supervised by a learning machine at a digital banking institution (Bank XYZ):

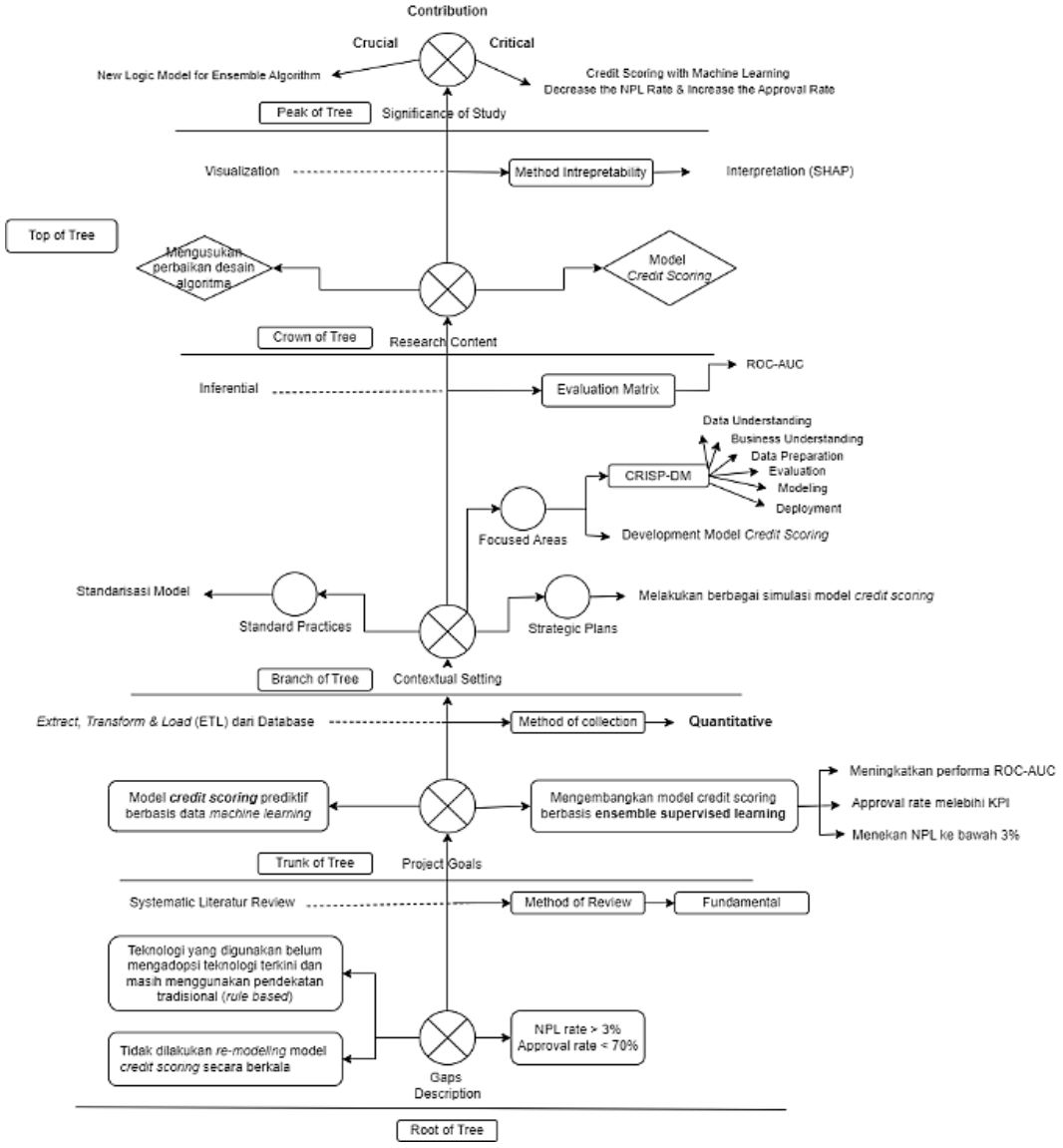

Figure 1. Tree of Root

1. Business Understanding: This phase focuses on identifying the primary objectives of Bank XYZ, which are to reduce the NPL rate below 3% and increase the loan approval rate to at least 70%. The institution seeks to modernize its credit scoring process by transitioning from rule-based assessments to data-driven models.
2. Data Understanding: The dataset used in this research comprises historical loan application records from 2022 to 2023 for Product X. The data include various borrower attributes such as loan amount, loan tenor, salary, employment status, and payment history. The labels used for classification are Performing Loan (PL) and NPL, defined based on internal credit risk criteria.

3. Data Preparation: Data preprocessing involved cleaning missing values, detecting and treating outliers, and encoding categorical variables using label encoding and one-hot encoding techniques. The dataset was split into training and testing subsets using an 80:20 ratio to ensure unbiased model evaluation.
4. Modeling: This stage includes building and training models with ensemble learning algorithms, namely Random Forest (bagging), DT, XGBoost, LightGBM, CatBoost, and AdaBoost (boosting). Hyperparameter tuning is performed using the Optuna technique to improve model performance in handling Overfitting in Boosting. The model is evaluated using ROC-AUC.
5. Evaluation Metrics: This stage carries out evaluation. Evaluation is carried out by comparing model performance using evaluation metrics; Accuracy, Precision, Recall, F1-Score, and AUC-ROC. The model with the best performance will be further analyzed using explainable AI (SHAP) techniques to interpret the prediction results.
6. Deployment: The final stage is a simulation of model integration into a digital banking system for use in credit decision making. Accompanied by documentation and risk management strategies based on model output.

2.1. Ensemble Supervised Learning in Credit Scoring

Credit scoring is a systematic method for evaluating the creditworthiness of individuals or entities based on historical data and predictive variables [18]. In digital banking, this process relies heavily on real-time, electronic transaction data and behavioral patterns, which enable efficient assessment for unsecured loans (KTA). Advancements in data technology have driven the adoption of ML algorithms to enhance credit scoring performance. The use of ML in credit scoring has advanced with algorithms like Random Forest, XGBoost, LightGBM, and CatBoost, which outperform traditional statistical models in capturing complex patterns [19]. ML effectively processes large-scale financial data, identifying intricate trends and correlations [20]. While offering high accuracy and strong generalization, its main limitation is low interpretability addressed through explainable AI (XAI) techniques such as SHAP to clarify feature contributions in decision-making. Classification plays a central role in ML-based credit scoring, where applicants are categorized into predefined credit risk groups based on historical data [21]. Typically, classification outputs label borrowers as good or bad customers good customers fulfill loan obligations, while bad customers exhibit default behavior or violate internal risk. Recent studies highlight that ensemble learning, which combines multiple classifiers, consistently outperforms single-model approaches in predictive accuracy [10].

2.2. Recent Studies on Ensemble Machine Learning for Credit Scoring

Recent research in credit scoring has increasingly emphasized the use of ensemble ML to address challenges such as data imbalance, non-linear feature interactions, and the need for interpretable predictions in digital banking. Nguyen and Ngo (2025) [22], for instance, compared multiple boosting algorithms XGBoost, AdaBoost, CatBoost, and LightGBM for predicting personal default in Vietnam. Their findings indicated that LightGBM performed best for large and complex datasets, while CatBoost maintained stable accuracy across various scenarios. Similarly, Han (2024) [12] evaluated ensemble and deep learning models on Taiwan credit data, concluding that XGBoost achieved the highest stability and accuracy, particularly when dealing with imbalanced datasets.

Consoli et al. (2021) [23], conducted a comparative analysis of bagging, boosting, and stacking methods against Decision Tree Classifiers, demonstrating that boosted ensembles achieved the highest ROC-AUC on Australian and German banking datasets characterized by non-linear patterns. Meanwhile, Abhishek Kumar et al. (2024) [24], contrasted traditional credit scoring approaches with ML models, showing substantial gains in predictive accuracy but without exploring interpretability in depth.

From a local perspective, Rosi Diaprina and Suhartono (2014) [8],investigated credit classification using binary logistic regression and RBF networks in an Indonesian bank, highlighting the limitations of non-ML, rule-based methods and underscoring the potential of advanced, data-driven approaches.

These studies collectively suggest that while ensemble learning consistently outperforms traditional and single-model methods in terms of predictive accuracy, gaps remain in integrating high performance with interpretability and practical deployment in digital banking operations. Addressing this gap, the present study compares five ensemble models (XGBoost, CatBoost, AdaBoost, LightGBM, Random Forest) against a DT baseline, applying SHAP for explainable predictions, and directly assessing their impact on two critical business metrics approval rate and NPL ratio using real-world credit data from Bank XYZ.

2.3. Decision Tree (DT)

DT is a widely used supervised learning algorithm in credit scoring that classifies data by recursively splitting features to separate target classes [25],[26]. Its high interpretability and ability to handle

both numerical and categorical data make it suitable for transparent financial decision-making [27]. Feature selection is typically based on impurity measures such as Gini Index.

$$\text{Gini (t)} = 1 - \sum_{i=1}^{C} p_i^2 \quad (1)$$

Gini Index measures node impurity, with lower values indicating purer splits; features that reduce Gini the most are selected. While effective on simple data, DT is prone to overfitting if too [26]. Hence, it is commonly used as a base learner in ensemble methods like Random Forest, AdaBoost and Gradient Boosting to enhance model robustness [28].

2.4. Random Forest

Random Forest is a popular ensemble learning algorithm designed to improve DT performance by aggregating multiple trees trained on random data and feature subsets [29]. It offers high accuracy, robustness to overfitting, and handles large, heterogeneous datasets well. Additionally, it provides feature importance metrics useful for credit scoring [30]. Unlike boosting, Random Forest uses parallel training of deep, independent trees, and is used in this study for performance comparison with boosting-based models.

2.5. AdaBoost

Adaptive Boosting (AdaBoost) is a widely used ensemble supervised learning algorithm in credit scoring due to its ability to enhance the performance of weak learners like DT [31]. Unlike bagging, AdaBoost builds models sequentially, focusing each iteration on correcting previous errors. This makes it effective for handling imbalanced and high-dimensional financial data by increasing sensitivity to misclassified cases. AdaBoost has shown superior accuracy over traditional methods such as logistic regression in identifying high-risk borrowers [32].

2.6. CatBoost

CatBoost is a modern gradient boosting algorithm optimized for categorical data, making it highly suitable for credit scoring applications involving mixed-type features [33]. Developed by Yandex, it addresses categorical encoding and overfitting through ordered boosting and symmetric trees [34]. CatBoost has demonstrated strong performance in handling high-dimensional, imbalanced loan data efficiently, without requiring manual feature transformation [33].

2.7. XGBoost

XGBoost is a widely used ensemble supervised learning algorithm in credit scoring due to its high predictive accuracy, computational efficiency, and strong overfitting control [34],[14]. It builds models iteratively by minimizing residual errors and employs regularization, pruning, and parallel processing for stability and generalization [14]. These strengths make XGBoost highly effective for handling complex, high-dimensional, and imbalanced loan data in both banking and digital lending applications.

2.8. LightGBM

LightGBM, developed by Microsoft, is an efficient and accurate GBDT-based algorithm optimized for high-dimensional and heterogeneous data, making it ideal for credit scoring [35]. It uses histogram-based processing and leaf-wise tree growth to reduce loss faster, with strong computational performance and scalability [36]. These advantages make LightGBM a key base learner in ensemble credit risk models.

2.9. Evaluation Metrics

The development of a credit scoring model based on ensemble supervised learning, model performance evaluation is a crucial stage to assess the extent to which the model can classify credit risk accurately and reliably. To assess the performance of the credit scoring classification model, several standard evaluation metrics are used:

1. Accuracy, measures the proportion of correctly predicted instances among the total observations.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Where, TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

2. Precision, indicates how many of the positively predicted cases are actually positive.

$$\text{Precision} = \frac{TP+TN}{TP+FP} \quad (3)$$

3. Recall (Sensitivity)
   Recall measures the model's ability to correctly identify actual positive cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

4. F1-Score
   The F1-Score is the harmonic mean of precision and recall, providing a balance between the two.

$$\text{F1} = \frac{\text{2 x Precision x recall}}{\text{Precision x recall}} \tag{5}$$

5. ROC-AUC (Receiver Operating Characteristic – Area Under the Curve)
   ROC-AUC evaluates the model's ability to distinguish between classes. AUC values range from 0 to 1, where higher values indicate better classification performance. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}, \text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{6}$$

2.10. Confusion Metric

The confusion matrix is a tabular representation that summarizes the performance of a classification model by comparing actual labels with predicted labels. It consists of four components, as shown in Table 1.

Table 1. Confusion Metric

	Predictive Positive	Predictive Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Where, TP = Correctly predicted positive class (predicted as "bad customer" and actually bad), TN = Correctly predicted negative class, FP = Incorrectly predicted as positive, FN = Incorrectly predicted as negative.

2.11. Flow Analysis and Implementation

The following is a flowchart of the process from raw data to model implementation into a scoring system using the selected algorithm, namely XGBoost, as in Figure 2 below. Meanwhile, it can be temporarily concluded that the XGBoost Model was chosen as the best model because; The highest ROC-AUC in the testing data, Good generalization (small gap train vs test) and Supported by reasonable and consistent SHAP interpretation results. Binning analysis also shows that the model successfully separates high and low risk groups well. These results are ready to be integrated into the bank or fintech credit scoring system to improve the quality of credit provision.

3. RESULTS AND ANALYSIS

3.1. Data Understanding

The initial phase of credit scoring model development involves understanding the characteristics and structure of the dataset. This study utilizes a dataset comprising 52,415 observations and 102 columns, consisting of 101 predictor variables and one binary target variable. Among the predictors, 6 are categorical employee_status, marital_status, gender, last_education, home_ownership_status, and max_loan_max_collect_6m while the remaining 95 are numerical. The target variable, labeled target, represents customer creditworthiness and is classified into two categories:

1. Label 1 (good customer): borrowers with no history of payment delinquency exceeding 90 days.
2. Label 0 (bad customer): borrowers with at least one instance of delinquency exceeding 90 days.

The proportion of good customers (label 1) and bad customers (label 0) in the dataset is 76.6% and 23.4%, respectively. This distribution is considered sufficiently balanced for binary classification tasks, as class imbalance typically becomes critical when the minority class falls below 10–20%. [37] further emphasize that an imbalance ratio (IR) below 16.6% poses a significant challenge for model reliability due to the scarcity of minority class instances. Therefore, the class distribution in this study remains suitable for training supervised learning models without the need for additional data balancing techniques such as resampling or SMOTE.

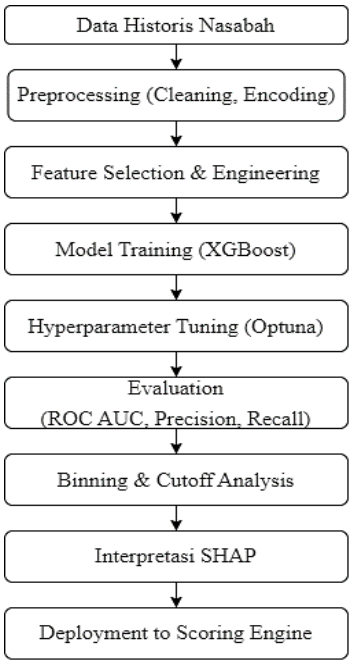

Figure 2. Flowchart Model to Credit Scoring Engine

3.1.1. Data Splitting

In this study, the dataset was split into 80% training and 20% testing sets (see Table 2), following standard practice in ML to balance model learning and evaluation [38]. Stratified sampling was applied to preserve the original class distribution (good = 1, bad = 0) in both subsets, minimizing bias due to class imbalance.

Table 2. Train & Test Data Sharing Results

Subset	Data Amount	Percentage of Total
Data Train	41.932	80%
Data Test	10.483	20%
Total	52.415	100%

A training set of 41,932 observations and a test set of 10,483 is considered sufficient for ML model development. This aligns with the findings of [39], who emphasize that thousands of observations are generally required to build robust and reliable classification models. Consequently, the use of stratified and proportional data splitting at this stage provides a critical foundation to ensure optimal, fair, and unbiased model learning.

3.1.2. Data Cleaning and Remapping

The initial stage of data cleaning involved examining the dataset for missing values across all columns. The inspection revealed no missing entries, thus eliminating the need for imputation in this study. Following the confirmation of data completeness, the next step was label remapping or normalization of categorical features. This process aimed to consolidate semantically similar categories that were written differently or represented in granular formats. Such remapping is crucial to reduce unnecessary dummy variable creation during encoding and to ensure the model learns more consistent patterns. In this study, remapping is divided into two, namely, the first, remapping of education levels where levels D1, D2 & D3 become "Diploma" and S1, S2 & S3 become "Bachelor". The second is remapping of employment status where simplification is carried out from granular forms such as "Contract", "Outsourcing", "PKWT", and "Daily" into one category "Contract", as well as "Permanent", "Retired", and "Worker" into the category "Permanent".

3.2. Exploratory Data Analysis and Feature Selection

This stage aims to understand data patterns and identify the most relevant features for modeling. Exploratory Data Analysis (EDA) is conducted to evaluate the relationship between features and the target variable through hypothesis testing. Feature selection is performed separately for numerical and categorical variables. The dataset used in this study contains a majority of numerical features, comprising 95 out of 101

total features. These include variables related to income, expenses, savings, loan history, and various transactional metrics over the last six months. Since these features were pre-processed through an ETL pipeline from a digital banking data mart, no additional feature engineering was necessary. Feature selection for numerical variables was conducted using a combination of Recursive Feature Elimination with Cross-Validation (RFECV) and Information Value (IV). RFECV was applied with an XGBoost classifier as the estimator and k-fold cross-validation to iteratively eliminate less informative features, reducing the risk of overfitting while preserving predictive power. Meanwhile, IV was calculated to assess each feature's ability to distinguish between good and bad customers, with the top 10 features selected based on IV scores. To further refine the selection, a multicollinearity analysis was performed using a Pearson correlation matrix. Feature pairs with correlation coefficients greater than 0.5 were identified, and the one with the lower IV score was removed. This ensured that the final model input included only non-redundant, statistically significant features, supporting both model accuracy and interpretability.

3.3. Exploratory Data Analysis (Hypothesis Testing)

The EDA in this phase aims to understand the relationship between the selected numerical features and the target variable (i.e., good or bad credit status). By visualizing distributions and analyzing the ratio of "Good" vs. "Bad" customers within each bin of a feature, hypotheses can be formulated based on business logic and further evaluated to determine whether empirical trends align with business expectations. This process is critical in credit scoring to ensure that the model is not only statistically accurate but also practically sound, thereby making it a reliable tool for risk-based decision-making. The following presents the logical justification and hypothesis formulation for two of the top ten selected features:

1. Plafond (Figure 3)
   This feature shows the amount of approved loans. In general, the higher the ceiling, the more selective the bank is in approving it, because large limits tend to be given to customers who are more financially stable. The hypothesis: High ceiling → lower risk of default. The graph shows an increase in the "Good" ratio as the ceiling increases, so it is in accordance with business logic.

2. Principal Installment (Figure 4)
   This feature indicates the amount of principal installments per month. The higher the installment, usually associated with a larger ceiling or short tenor. The hypothesis: High principal installments → lower default risk because it is only approved for customers with good capacity. The graph shows that the good ratio increases as the installment increases, thus consistent with the business hypothesis.

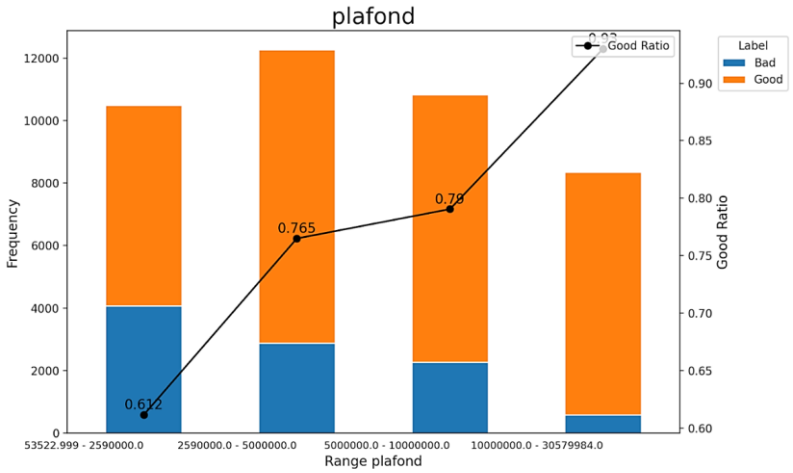

Figure 3. EDA Plafond

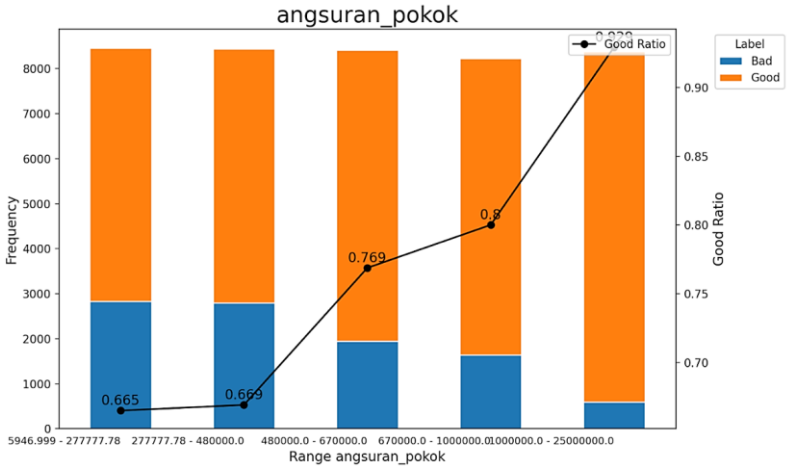

Figure 4. EDA Principal Installment

3.4. Model Evaluation and Comparison

Various ML algorithms have been tested, including DT, Random Forest, LightGBM, AdaBoost, CatBoost, and XGBoost. The following table summarizes the results of model evaluation after hyperparameter tuning using Optuna.

Table 3. Model Evaluation and Comparison

Model	ROC-AUC (Train)	ROC-AUC (Test)	Precision (Test)	Recall (Test)	F1-Score (Test)	Accuracy (Test)
Decision Tree	0.867	0.565	0.456	0.615	0.523	0.745
Random Forest	0.899	0.724	0.401	0.576	0.472	0.731
LightGBM	0.837	0.751	0.428	0.603	0.499	0.748
AdaBoost	0.788	0.670	0.378	0.538	0.443	0.702
CatBoost	0.881	0.676	0.419	0.598	0.491	0.724
XGBoost	0.824	0.803	0.512	0.683	0.586	0.776

Based on the comparative results in Table 3, the XGBoost algorithm emerged as the most balanced and generalizable model for credit scoring tasks. Although its ROC AUC score on the training dataset was slightly lower than those of other ensemble models such as Random Forest or CatBoost, the test ROC AUC of 0.803 reflects strong generalization performance. Additionally, XGBoost outperformed other models in terms of precision, recall, and F1-score, indicating its robustness in identifying both good and bad customers. To validate the model, performance was evaluated on both training and testing datasets. The moderate difference between training and testing ROC AUC scores (e.g., XGBoost 0.824 vs. 0.803) suggests that the model is not overfitting and maintains its ability to learn meaningful patterns from historical data. Furthermore, stratified k-fold cross-validation was employed to ensure performance consistency across different data splits. The low variance observed across the folds supports the model's stability, making it a reliable choice for real-world credit risk prediction.

3.5. Binning Analysis and Risk Distribution

The probability distribution of the XGBoost model prediction results is grouped into several bins (deciles) and analyzed to understand the characteristics of each segment. The Table 4 shows the cumulative evaluation metrics in each bin based on probability.

Where, the explanation of the Table 4 is as follows:

1. Bin Probability = Range of probability values per Bin
2. Population Number = Population number divided equally by 15 bins, which is 3494, some 3495
3. Good Number = Good population number from Population_Number
4. Bad Number = Bad population number from Population_Number
5. Cum population number = Cumulative number from Population_Number column
6. Cum Good number = Cumulative number from Good Number column
7. Cum Bad number = Cumulative number from Bad Number column
8. Approval Rate = Cumsum population number / total population number (for example in bin (0.407, 0.472] the approval rate value is (3494+3494+3494) / 52415 = 20%

9. Bad credit (%) = Bad_number / population number
10. Cumulative Bad Credit = Bad number / Cumulative population number

Bad Credit Number Based on the binning Table 4, if we want to take an approval rate above 70%, where we take the proba class (0.829, 0.871] with approval of 73.33%, then the cumulative bad rate is 2.92%. This is in accordance with the desired target metric, namely approval rate > 70% and NPL rate below 3%

Table 4. Binning Analysis and Risk Distribution

Bin Probability	Population Number	Good Number	Bad Number	Cum Population Number	Cum Good Number	Cum Bad Number	Approval	Bad Credit (%)	Bad Credit Kumulatif (%)
(0.048, 0.318]	3494	3490	4	3494	3490	4	6.67%	0.12%	0.12%
(0.318, 0.407]	3494	3488	6	6988	6978	10	13.33%	0.18%	0.15%
(0.407, 0.472]	3494	3484	10	10482	10461	21	20.00%	0.30%	0.20%
(0.472, 0.526]	3494	3466	28	13976	13927	49	26.66%	0.80%	0.35%
(0.526, 0.580]	3494	3442	52	17470	17369	101	33.33%	1.50%	0.58%
(0.580, 0.634]	3494	3421	73	20964	20790	174	40.00%	2.08%	0.83%
(0.634, 0.688]	3494	3340	154	24458	24130	328	46.66%	4.40%	1.34%
(0.688, 0.738]	3494	3386	108	27952	27516	436	53.33%	3.10%	1.56%
(0.738, 0.785]	3494	3323	171	31446	30839	607	59.99%	4.89%	1.93%
(0.785, 0.829]	3494	3217	277	34940	34056	884	66.66%	7.93%	2.53%
(0.829, 0.871]	3495	3257	238	38435	37313	1122	73.33%	6.82%	2.92%
(0.871, 0.908]	3495	1904	1591	41930	39217	2713	80.00%	45.51%	6.47%
(0.908, 0.937]	3495	1052	2443	45425	40269	5156	86.66%	69.90%	11.35%
(0.937, 0.965]	3495	197	3298	48920	40467	8453	93.33%	94.35%	17.28%
(0.965, 0.993]	3495	1371	2124	52415	41838	10577	100.00%	60.77%	23.4%

3.6. Interpretation of SHAP Value and Hypothesis Validation

SHAP summary plot depicting the contribution of each feature to the output of the default risk prediction model. The color represents the feature value (blue for low values, red for high values), and the horizontal axis shows how much the feature contributes (positive or negative) to the model prediction. If the red color is on the left, then the greater the feature value, the greater the risk, conversely if the red color is on the right, then the greater the feature value, the smaller the credit risk. Here is the interpretation of 3 of the 10 important features (limitations on displaying all credential features) and their suitability to the business hypothesis:

1. Plafond. Hypothesis: high ceiling → more creditworthy customers → lower default risk. The graph supports this: high ceiling values (red) produce negative SHAP values, meaning they tend to reduce default predictions.
   Conclusion: As hypothesized.

2. Angsuran_Pokok. Hypothesis: Large principal installments are only given to financially healthy customers. The plot shows that high values (red) tend to reduce the risk of default, in accordance with the negative SHAP value.
   Conclusion: According to the hypothesis.

3. Avg_ratio_angsuran_med_abal_base_6m. Hypothesis: High ratio → large installment burden to balance → high risk. However, the graph shows a high value (red) gives a negative SHAP value, actually reducing the risk of default. This could indicate that customers with large commitments are more disciplined.
   Conclusion: Contrary to the hypothesis.

4. CONCLUSION

This study evaluates and compares five ensemble ML models XGBoost, CatBoost, AdaBoost, LightGBM, and Random Forest against a DT baseline for credit scoring at Bank XYZ. By integrating SHAP-based interpretability, the proposed approach bridges the gap between predictive performance and explainability, enabling more transparent decision-making in digital banking. The results demonstrate that LightGBM achieves the highest predictive accuracy and approval rate, while effectively reducing the Non-Performing Loan (NPL) ratio. Compared to prior studies, the novelty of this research lies in its simultaneous optimization of approval rate and NPL reduction, combined with a model explainability framework tailored for real-world operational deployment. This aligns with the objective of the study to improve approval rates and reduce NPL ratios at Bank XYZ, while offering a reproducible methodology adaptable to other digital banking contexts. A comparative evaluation of ensemble models versus single classifiers revealed that the XGBoost model outperformed others, achieving the highest ROC AUC score of 0.803 on the test set,

indicating both low overfitting and strong generalization. While traditional models like DT showed high overfitting (ROC AUC 0.573), ensemble methods such as Random Forest, LightGBM, CatBoost, and AdaBoost exhibited improved performance, with XGBoost emerging as the most stable and accurate. The SHAP analysis of the XGBoost model identified key predictive features including tenor, loan plafond, salary, expenditure ratio, and days of delinquency. Additionally, the XGBoost model offers strategic value in improving key performance indicators (KPIs) for Product X: reducing the Non-Performing Loan (NPL) rate from 4% to below 3%, and increasing the approval rate from 65% to above 70%. These results highlight XGBoost as the optimal model in terms of performance, generalization, and stability, fulfilling the objectives of this research and contributing meaningfully to risk mitigation and portfolio quality enhancement at Bank XYZ.

REFERENCES

[1] Badan Pusat Statistik Bank Indonesia, "Percepatan Digitalisasi Transaksi untuk Memacu Pemulihan Ekonomi Nasional," https://www.bi.go.id/id/publikasi/ruang-media/news-release/Pages/sp_2620624.aspx.

[2] Digibank by DBS, "Pinjaman Digibank KTA," https://www.dbs.id/digibank/id/id/pinjaman/produk-pinjaman/.

[3] BTPN Syariah, "Produk & Layanan Pembiayaan BTPN Syariah," https://www.btpnsyariah.com/pembiayaan.

[4] Badan Pembina Hukum Nasional, "Undang-undang Republik Indonesia Nomor 10 Tahun 1998," 1998. [Online]. Available: www.bphn.go.id

[5] Otoritas Jasa Keuangan (OJK), "Peraturan Otoritas Jasa Keuangan Nomor 12 /Pojk.03/2018 Tentang Penyelenggaraan Layanan Perbankan Digital Oleh Bank Umum," 2018. Accessed: Jan. 03, 2025. [Online]. Available: https://www.ojk.go.id/id/regulasi/Documents/Pages/Penyelenggaraan-Layanan-Perbankan-Digital-oleh-Bank-Umum/POJK%2012-2018.pdf?utm_source=chatgpt.com

[6] Otoritas Jasa Keuangan, "Otoritas Jasa Keuangan Republik Indonesia Nomor 10 /Pojk.05/2022 Tentang Layanan Pendanaan Bersama Berbasis Teknologi Informasi," 2022.

[7] R. D. Mendrofa, M. H. Siallagan, J. Amalia, and D. P. Pakpahan, "Credit Risk Analysis With Extreme Gradient Boosting and Adaptive Boosting Algorithm," Journal of Information System,Graphics, Hospitality and Technology, vol. 5, no. 1, pp. 1–7, Mar. 2023, doi: 10.37823/insight.v5i1.233.

[8] S. Rosi Diaprina and Suhartono, "Analisis Klasifikasi Kredit Menggunakan Regresi Logistik Biner Dan Radial Basis Function Network di Bank 'X' Cabang Kediri," JURNAL SAINS DAN SENI POMITS Vol. 3, No. 2, 2014.

[9] M. I. M. Yusoff, "Machine Learning: An Overview," Open Journal of Modelling and Simulation, vol. 12, no. 03, pp. 89–99, 2024, doi: 10.4236/ojmsi.2024.123006.

[10] C. L. Perera and S. C. Premaratne, "An Ensemble Machine Learning Approach for Forecasting Credit risk of Loan Applications," WSEAS Transactions on Systems, vol. 23, pp. 31–46, 2024, doi: 10.37394/23202.2024.23.4.

[11] A. Febriyanti and T. Rizky Izzalqurny, "Predicting Credit Paying Ability With Machine Learning Algorithms," Majalah Bisnis & IPTEK, vol. 16, no. 1, pp. 8–15, 2023, doi: 10.55208/bistek.

[12] M. Han, "Ensemble Learning Based Models and Deep Learning Model for Credit Prediction, Case Study: Taiwan, China," in Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence, SCITEPRESS - Science and Technology Publications, 2024, pp. 115–121. doi: 10.5220/0012910900004508.

[13] Y. Li and W. Chen, "A comparative performance assessment of ensemble learning for credit scoring," Mathematics, vol. 8, no. 10, pp. 1–19, Oct. 2020, doi: 10.3390/math8101756.

[14] M. Zhu, Y. Zhang, Y. Gong, K. Xing, X. Yan, and J. Song, "Ensemble Methodology:Innovations in Credit Default Prediction Using LightGBM, XGBoost, and LocalEnsemble," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.17979

[15] Yosef Jabareen, "Building a Conceptual Framework: Philosophy, Definitions, and Procedure," Int J Qual Methods, vol. 8, no. 4, pp. 49–62, Dec. 2009, doi: 10.1177/160940690900800406.

[16] M. F. Safitra et al., "Green Networking: Challenges, Opportunities, and Future Trends for Sustainable Development," in ACM International Conference Proceeding Series, Association for Computing Machinery, Aug. 2023, pp. 168–173. doi: 10.1145/3617733.3617760.

[17] P. Chapman et al., "CRISP-DM 1.0 Step-by-step data mining guide," DaimlerChrysler, 1999.

[18] M. Óskarsdóttir, C. Bravo, C. Sarraute, J. Vanthienen, and B. Baesens, "The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion using Mobile Phone Data and Social Network Analytics," Feb. 2020, doi: 10.1016/j.asoc.2018.10.004.

[19] R. Hlongwane, K. K. K. M. Ramaboa, and W. Mongwe, "Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data," PLoS One, vol. 19, no. 5 May, May 2024, doi: 10.1371/journal.pone.0303566.

[20] M. F. Safitra, M. Lubis, T. F. Kusumasari, and D. P. Putri, "Advancements in Artificial Intelligence and Data Science: Models, Applications, and Challenges," in Procedia Computer Science, Elsevier B.V., 2024, pp. 381–388. doi: 10.1016/j.procs.2024.03.018.

[21] S. Mestiri, "Credit scoring using machine learning and deep Learning-Based models," Data Science in Finance and Economics, vol. 4, no. 2, pp. 236–248, 2024, doi: 10.3934/dsfe.2024009.

[22] N. Nguyen and D. Ngo, "Comparative analysis of boosting algorithms for predicting personal default," Cogent Economics and Finance, vol. 13, no. 1, 2025, doi: 10.1080/23322039.2025.2465971.

[23] S. Consoli, D. R. Recupero, and M. Saisana, Data Science for Economics and Finance: Methodologies and Applications. Springer International Publishing, 2021. doi: 10.1007/978-3-030-66891-4.
[24] Abhishek Kumar, Abhijeet Kumar, Aditya Kumar Singh, and Ms. Nikita, "Credit Scoring System Using Machine Learning," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 3, pp. 376–380, May 2024, doi: 10.32628/cseit2410334.
[25] N. A. Prahastiwi, R. Andreswari, and R. Fauzi, "Students Graduation Prediction Based On Academic Data Record Using The Decision Tree Algorithm C4.5 Method," JURTEKSI (Jurnal Teknologi dan Sistem Informasi), vol. 8, no. 3, pp. 295–304, Aug. 2022, doi: 10.33330/jurteksi.v8i3.1680.
[26] H. Wang, "Application of Decision Tree Model in Personal Credit Scoring and Its Fairness Optimization," 2025, doi: 10.54254/2754-1169/176/2025.22114.
[27] J. A. Bastos, "Predicting Credit Scores with Boosted Decision Trees," Forecasting, vol. 4, no. 4, pp. 925–935, Dec. 2022, doi: 10.3390/forecast4040050.
[28] V. Chang, S. Sivakulasingam, H. Wang, S. T. Wong, M. A. Ganatra, and J. Luo, "Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers," Risks, vol. 12, no. 11, Nov. 2024, doi: 10.3390/risks12110174.
[29] Y. Zhou, L. Shen, and L. Ballester, "A two-stage credit scoring model based on random forest: Evidence from Chinese small firms," International Review of Financial Analysis, vol. 89, Oct. 2023, doi: 10.1016/j.irfa.2023.102755.
[30] A. Fauziah, "Optimizing Credit Scoring Performance Using Ensemble Feature Selection with Random Forest," Jurnal Matematika, Statistika dan Komputasi, vol. 21, no. 2, pp. 560–572, Jan. 2025, doi: 10.20956/j.v21i2.42032.
[31] P. Beja-Battais, "Overview of AdaBoost : Reconciling its views to better understand its dynamics," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.18323
[32] J. Lin, "Research on loan default prediction based on logistic regression, randomforest, xgboost and adaboost," SHS Web of Conferences, vol. 181, p. 02008, 2024, doi: 10.1051/shsconf/202418102008.
[33] Y. Zhao, "A Credit Card Default Prediction Method Based on CatBoost," 2023, pp. 178–184. doi: 10.2991/978-94-6463-222-4_17.
[34] S. B. Coşkun and M. Turanli, "Credit risk analysis using boosting methods," Journal of Applied Mathematics, Statistics and Informatics, vol. 19, no. 1, pp. 5–18, May 2023, doi: 10.2478/jamsi-2023-0001.
[35] S. Yanjie, G. Zhike, S. Quan, and C. Lin, "Efficient Commercial Bank Customer Credit Risk Assessment Based on LightGBM and Feature Engineering," 2023.
[36] D. Williams, E. Brown, J. Smith, M. Johnson, A. Deshmukh, and S. Rodriguez, "Comparative Analysis of LightGBM with Traditional Credit Assessment Methods," 2024, doi: 10.13140/RG.2.2.29039.65444.
[37] Z. Zhao, T. Cui, S. Ding, J. Li, and A. G. Bellotti, "Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction," Mathematics, vol. 12, no. 5, Mar. 2024, doi: 10.3390/math12050701.
[38] X. Liu, Z. Zhang, and D. Wang, "Classification of Imbalanced Credit scoring data sets Based on Ensemble Method with the Weighted-Hybrid-Sampling."
[39] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, and A. Fernández-Delgado, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?," 2014. [Online]. Available: http://www.mathworks.es/products/neural-network.

BIBLIOGRAPHY OF AUTHORS

Narita Ayu Prahastiwi recently completed her Master's degree in Information Systems from Telkom University, Bandung, in 2025. She also earned her Bachelor's degree from the same university in 2021, both in Information Systems. She currently works in the financial industry and is passionate about developments in the field, particularly the application and utilization of cutting-edge technology.

Muharman Lubis has finished his Doctoral degree recently in Information Technology in 2017 at the International Islamic University Malaysia. He also received his Master's degree from the same university in 2011 and his Bachelor's degree from the University Utara Malaysia in 2008, both in Information Technology. He joined as a Lecturer in the School of Industrial Engineering, Telkom University, in 2017. His research interests include privacy protection, information security awareness, knowledge management and project management.

Hanif Fakhrurroja earned his bachelor's degree in physics from Universitas Padjadjaran (Unpad) in 2003, followed by a master's degree in informatics engineering from Institut Teknologi Bandung (ITB) in 2010, and a Doctorate in Electrical and Informatics Engineering from ITB in 2021. He has served as a Researcher at the Indonesian Institute of Sciences (now the National Research and Innovation Agency) since 2006 and has been a professional lecturer at Telkom University since 2017. His research interests include Information Systems, Machine Learning, Big Data Analysis, and the Internet of Things.