DRAFT MANUSCRIPT

Predicting Loan Approval with Random Forest

Course IF540 – Machine Learning

Lecturer: David Agustriawan

Compiled by Group 4:

Jonathan Chandra - (NIM: 00000094067)

Muhammad Faqih F. - (NIM: 00000067063)

Mohammad Reza P. - (NIM: 00000046140)

Hadi Kurniawan - (NIM: 00000064107)

Yoel Christian A. - (NIM: 00000040078)

INFORMATION SYSTEMS STUDY PROGRAM

FACULTY OF ENGINEERING AND INFORMATICS

UNIVERSITAS MULTIMEDIA NUSANTARA

TANGERANG

2025

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 Problems and Importance of the Topic

Problems	Importance
There are several significant challenges that need to be addressed in the context of credit evaluation. First, high default risk has the potential to cause substantial financial losses. Second, there is an urgent need for an efficient, fast and automated credit evaluation system. Third, there is a need for innovative and sophisticated prediction methods that are able to effectively and holistically utilize various borrower attributes. Fourth, in-depth comparison and analysis of the effectiveness of existing predictive methods in specific contexts is essential. Finally, another issue that arises is the importance of careful and thorough pre-processing of the initial data before starting the predictive modeling process.	Considering these issues, certain measures are crucial. First, improving the accuracy and reliability of borrower status is crucial for better risk mitigation. Second, support for better credit decision-making processes, based on comprehensive and prompt data analysis, will help reduce losses. Third, minimization of potential financial losses caused by non-performing loans (NPLs) through early risk identification is a key objective. Fourth, achieving optimal operational efficiency for financial institutions, reducing costs, and accelerating the credit approval process are desired outcomes. Fifth, an increased in-depth understanding of the key factors and determining variables that influence loan approval or rejection will lead to more informed decisions.

1.2 Research Keywords

No.	Keywords	Explanation
1.	**Decision Tree**	**Decision Tree (DT)** method is a machine learning algorithm for regression and classification based on a decision tree that represents decisions in the form of a branching tree-shaped diagram [7]. DT operates by recursively partitioning data based on attributes that provide the most information based on criteria such as Gini Impurity or Entropy in algorithms such as ID3, C4.5, and CART. Each node of the tree is a feature, and decisions based on feature values are shown by the branches, while the leaves show the final classification or prediction results [6]. The main advantage of Decision Tree is that it is easy to

		interpret, and the decision-making process is understandable to users [8]. In addition, DT can handle numerical and categorical data directly and does not require complicated preprocessing [9]. However, this method is prone to overfitting if the tree is too deep or if the tree has a very large number of branches [10]. Pruning methods can be used in such cases to reduce model complexity and improve generalization to new data. DT is comprehensively used for credit scoring in finance, medical diagnosis in healthcare, and customer segmentation in marketing. Its ability to handle nonlinear problems and its efficacy in decision making make DT one of the popular algorithms in machine learning and data mining [14].
2.	**Machine Learning**	**Machine Learning (ML)** is a branch of artificial intelligence that allows computer systems to learn from data and perform specific tasks without requiring explicit programming. According to Reinaldo R. (2023), ML involves developing algorithms and models that can learn from data and improve performance over time through experience [15]. In the context of data analysis, ML is used as a method to discover hidden patterns, relationships, and insights from large and complex data sets [17]. By utilizing algorithms such as supervised learning, unsupervised learning, and reinforcement learning, ML can process data efficiently, recognize patterns that are not immediately apparent, and provide accurate predictions. For example, in medical data analysis, ML can identify disease risk based on patient data, such as blood pressure, cholesterol, and family history. This technique is valuable because it provides a data-driven approach that can support more informed decision-making in areas such as finance, marketing, and healthcare. This makes ML an innovative and evolving data analysis tool to face modern data challenges [19].
3.	**Loan Status Prediction**	**Loan status prediction** is the process of determining whether a customer's loan application will be approved or rejected by a financial institution based on certain attributes. This process is crucial in risk management, as the decisions made will impact the return rate and potential losses [16]. In this report, the prediction is carried out using a Decision Tree-based classification model that utilizes several features from customer data,

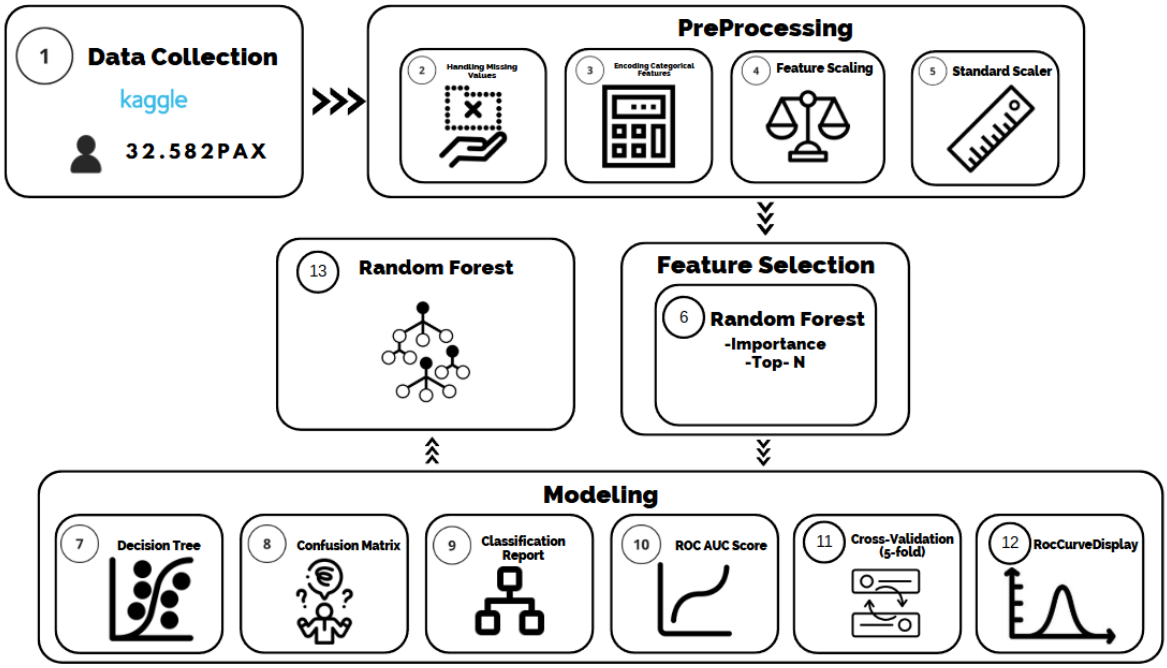
		such as the number of dependents, marital status, education level, and credit history. Accurate credit status prediction helps banks filter out applications with a high risk of default and accelerates the approval process for eligible customers [18]. The success of this prediction depends on data quality, processing methods, and the performance of the algorithm used. By applying a machine learning approach, predictions can be made more objectively and consistently, thereby reducing the risk of subjectivity in manual evaluations.
4.	**Model Accuracy**	**Model Accuracy** is one of the key evaluation metrics in machine learning, indicating how often the model correctly predicts labels compared to the total test data [11]. In the context of binary classification, accuracy is calculated as the ratio of the number of correct predictions to the total number of predictions. In this report, the Decision Tree model achieved an overall accuracy of 83%, which indicates fairly good performance. However, accuracy alone does not necessarily reflect the overall performance of the model, especially when the data is imbalanced between approved and rejected classes [12]. Therefore, this report also uses other metrics such as precision, recall, and F1-score to assess the prediction quality for each class. A deep understanding of accuracy and other metrics is essential in evaluating machine learning models to avoid being misled by seemingly high results that are actually biased toward the majority class.
5.	**Credit Risk Management**	**Credit Risk Management** is a strategy and process implemented by financial institutions to identify, measure, and mitigate the risk of loss due to borrower default. In the banking sector, credit risk is one of the primary risks that can directly impact the financial stability of an institution [13]. Therefore, it is crucial for banks to have systems capable of accurately assessing creditworthiness. In this report, a Decision Tree-based prediction model is used as a supporting tool in the credit risk management process. By leveraging borrowers' historical data and identifying specific patterns related to repayment behavior, financial institutions can make more confident credit decisions. This data-driven approach helps reduce subjective judgments and enhances both the efficiency and effectiveness of the credit approval process [20].

6.	**Credit Analysis**	**Credit Analysis** is the process of evaluating a prospective borrower's ability and reliability to meet their financial obligations. This process involves reviewing various information such as income, dependents, credit history, and employment status. The goal is to determine whether granting credit to the individual is safe and justified. In this report, credit analysis is conducted with the help of a machine learning model, specifically a Random Forest, which can process borrower data and automatically provide an objective loan status prediction. The model identifies the features that contribute most significantly to credit decisions, such as credit history and loan amount. By using data-driven credit analysis, financial institutions not only improve the accuracy of borrower evaluation but also optimize business processes by making faster and more efficient decisions[21].
7.	**Random Forest**	**Random Forest** is one of the ensemble-based machine learning algorithms widely used in various fields, especially for classification and regression tasks. It works by building a number of decision trees during the training process, and outputs predictions based on the majority vote (mode) of the trees for classification or the average prediction for regression. The main advantage of Random Forest lies in its ability to reduce the risk of overfitting that often occurs with single tree models, as combining predictions from multiple trees can improve the accuracy and generalization of the model. In addition, Random Forest is also able to handle high-dimensional data and provide feature importance estimates, which are useful in the feature selection process [22]. The process of selecting a random subset of the training data and the features used to build each tree (referred to as bagging and random feature selection) makes Random Forest resistant to noise and irrelevant variables.

CHAPTER II

METODOLOGI PENELITIAN

2.1 Research pipeline



No.	Research Pipeline Step	Explanation
1.	Data Collection	The initial and fundamental stage in any machine learning project is data collection. In this step, datasets to be used for model training and evaluation are acquired from relevant and reliable sources. In this pipeline, data was specifically obtained from Kaggle, a popular platform for data science competitions and dataset hosting. The collected data amounted to 32,582 entries or observations. The quality and representativeness of the data collected at this stage is crucial to the success of the entire project, as "garbage in, garbage out" is a basic principle in machine learning.
2.	Handling Missing Values	Real-world datasets often contain missing values for various reasons, such as data entry errors, sensor failure, or unavailable data. Missing values can cause bias in the model or even make certain

		algorithms unworkable. This stage in pre-processing focuses on strategies to manage such data absences. Common approaches include deleting rows or columns containing missing values (if there are few of them), or imputation, which is replacing missing values with reasonable estimates (e.g., mean, median, mode, or more sophisticated methods such as model-based imputation). The choice of method largely depends on the characteristics of the data and the proportion of missing values.
3.	Handling Categorical Variables	Many datasets contain categorical variables (e.g., gender, city, educational status) that cannot be directly processed by most machine learning algorithms that operate on numerical inputs. This pre-processing stage aims to convert these non-numerical variables into a suitable numerical format. Popular methods include One-Hot Encoding, where each category is converted into a separate binary column, or Label Encoding, where each category is assigned a unique numeric value. The choice of encoding method depends on the nature of the variables (ordinal or nominal) and the model algorithm to be used.
4.	Feature Scaling	The process of scaling the numerical feature values in the dataset to be in a uniform range. This is important for algorithms that are sensitive to feature scaling.
5.	Standard Scaling	Standard Scaler is the most commonly used Feature Scaling method. In this method, each feature is transformed so that it has a mean of zero ($\mu=0$) and a standard deviation of one ($\sigma=1$). This process is done by subtracting each feature value from the feature's mean, then dividing the result by the feature's standard deviation. This transformation helps in making the data have a consistent distribution and is particularly useful for algorithms that calculate distances between data points or that rely on the assumption of a normal distribution.
6.	Feature Selection	Once the data is processed, the feature selection stage becomes crucial to improve model performance, reduce overfitting, speed up training time, and improve model interpretation. In this pipeline, feature selection is performed using the Random Forest algorithm. Random Forest has the

		built-in ability to measure the importance of each feature based on how much it contributes to reducing impurities (e.g. Gini impurity) in the decision trees. Only the top N features that have the highest importance score will be selected for use in the modeling stage, ensuring that only the most relevant information goes into the final model.
7.	Decision Tree	Decision Tree is a non-parametric machine learning algorithm for classification and regression. It builds a tree-like structure, where each node represents an attribute test, branches are test results, and leaves are classes or target values. The algorithm splits the data recursively based on the best splitting criteria. Its advantages lie in easy interpretation, ability to handle non-linearity and various types of data. However, it is prone to overfitting, which is often overcome by pruning or ensemble methods.
8.	Confusion Matrix	Confusion Matrix is a fundamental evaluation tool for classification models. It is a table that summarizes the performance of a classification model on a set of test data. The matrix displays four main quadrants: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). With this matrix, various other performance metrics such as Accuracy, Precision, Recall, and F1-Score can be calculated.
9.	Classification Report	Classification Report is a textual summary that provides important evaluation metrics for each class in a classification problem. Commonly included metrics are Precision (proportion of correct positive predictions), Recall (proportion of positive cases actually identified), F1-Score (harmonic mean of Precision and Recall), and Support (actual number of each class in the test data). This report provides a comprehensive view of how well the model performs on each class individually.
10.	ROC AUC Score (Area under the Receiver Operating Characteristic Curve)	ROC AUC Score is a very important evaluation metric for binary classification models, especially when the data set is not balanced. The ROC curve is a plot that illustrates the performance of a classification model at all possible classification thresholds. It shows the True Positive Rate (TPR, or Recall) against the False Positive Rate (FPR) at

UNIVERSITAS MULTIMEDIA NUSANTARA

		various thresholds. AUC is the area under the ROC curve. AUC values range from 0 to 1, where 1 indicates a perfect model and 0.5 indicates a model that is no better than random classification. A higher AUC indicates better model performance in distinguishing between positive and negative classes.
11.	Cross-Validation (k-fold)	Ensuring that the model has robust performance and can be generalized to unseen data, k-fold Cross-Validation is used. In this technique, the training dataset is divided into k subsets (called "folds") of almost equal size. The model is trained on k-1 folds and validated on the remaining folds. This process is repeated k times, with each fold used as a validation set exactly once. The average performance score of those k iterations is then reported, providing a more reliable estimate of the model's performance compared to the single split train-test method. This helps reduce the bias associated with a single data split.
12.	Recurve Display	The term "RecurveDisplay" most likely refers to visualizations that display Recall curves or other related performance metrics, which may be generated as part of a more in-depth model evaluation, especially after a cross-validation process. While there is no standard term "RecurveDisplay" in the data science literature, it can be a specialized visual representation of the Recall curve or metrics designed to aid in the interpretation of model performance. In the context of classification, understanding how Recall changes with threshold or within cross-validation iterations is important to assess the model's ability to identify all positive cases.
13.	Random Forest	Besides its role in feature selection, Random Forest (which is also indicated as box number 12, possibly as a feedback loop or alternative/comparative model) is also a powerful ensemble machine learning model. Random Forest works by building many decision trees during training and outputting a class that is either the class mode (classification) or the average prediction (regression) of the individual tree predictions. Random Forest is known for its high accuracy, ability to handle high-dimensional data, and resilience to overfitting, making it an excellent choice to serve as a primary model or performance

		benchmark against Decision Tree.

CHAPTER III

RESULT AND DISCUSSION

3.1 Research Pipeline Result

Step	Output Type	Detailed Explanation
Modeling	Classification Report	Shows classification performance based on precision, recall, and f1-score metrics. The Random Forest model achieved an **accuracy of 91%**, with a higher f1-score for the 'Fully Paid' class (0.94) than for the 'Default' class (0.78).
	Confusion Matrix	Displays te number of correct and incorrect predictions. The model correctly predicted **4892 out of 5095** 'Fully Paid' and **1037 out of 1422** 'Default' cases. Most errors occurred in predicting the 'Default' class as 'Fully Paid'
	ROC Curve	The ROC graph illustrates the model's ability to distinguish between positive and negative classes. The **Area Under Curve (AUC) is 0.92**, indicating the model has excellent binary classification performance.
Model Evaluation	Single Tree Visualization	Shows one decision tree from the Random Forest model. Important features like **loan_percent_income**, **person_income**, and **loan_grade** were frequently used as split criteria, indicating their significance.
	Cross-Validation ROC AUC Scores	The result of 5-fold cross-validation indicates consistent model performance, with AUC scores ranging from **0.846 to 0.862**, and an average of **0.8549**. This supports the model's stability and lack of overfitting.

UNIVERSITAS MULTIMEDIA NUSANTARA

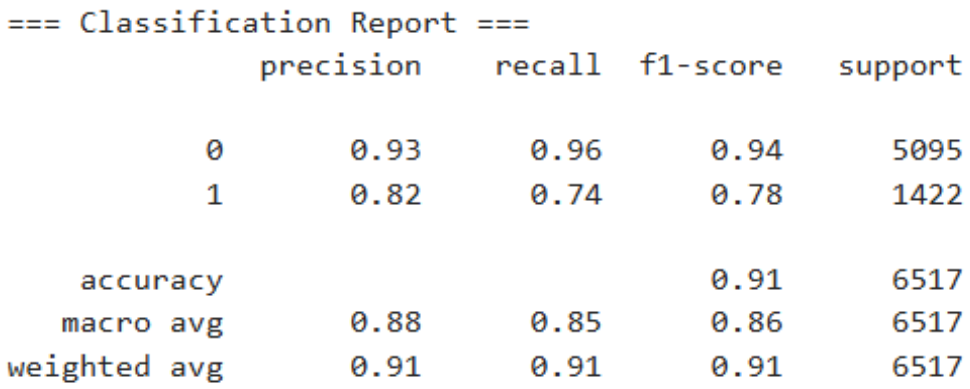
3.2 Result Output

Kategori Output	Penjelasan
Classification Report	Based on "Figure 3.1 Random Forest Classification Report", this classification report presents the performance evaluation metrics of the Random Forest model. For class '0', the model shows a precision of 0.93, recall of 0.96, and f1-score of 0.94, with a total support of 5095 data. This means that the model is quite good at identifying class '0' and most of the positive predictions for class '0' are correct. In contrast, for

	class '1', the model has a precision of 0.82, recall of 0.74, and f1-score of 0.78, with a total support of 1422 data. Although still quite good, the performance for class '1' is slightly lower than class '0', especially in recall which indicates that there are some instances of class '1' that are not detected by the model. Overall, the model accuracy was 0.91, indicating that 91% of the total predictions were correct. The macro average (unweighted average) for precision, recall, and f1-score were 0.88, 0.85, and 0.86, respectively. Meanwhile, the weighted averages (averages based on the number of supports from each class) for the three metrics were 0.91, 0.91, and 0.91, which better reflects the overall performance of the model given the unbalanced class distribution.
Confusion Matrix	Confusion Matrix ("Figure 3.2"), the model showed a strong ability in identifying "Fully Paid" cases (4892 True Positives), but had challenges in distinguishing between "Fully Paid" and "Default", as evidenced by 203 "Fully Paid" cases that were misclassified as "Default" (False Negatives), and 385 "Default" cases that were misclassified as "Fully Paid" (False Positives). Despite this, the model successfully identified the majority of "Default" cases (1037 True Negatives). This consistency is also reflected in the Classification Report ("Figure 3.1"), where class '0' (which most likely represents "Fully Paid" due to its large support count of 5095) shows very high precision, recall, and f1-score values (0.93, 0.96, 0.94). In contrast, class '1' (which most likely represents "Default" with a support of 1422) has lower metrics, especially in recall (0.74), indicating that the model is less able to capture all instances of this class. Nonetheless, the overall accuracy of the model reached 0.91, showing a solid performance in predicting the class in general. The difference between macro avg and weighted avg on the Classification Report also highlights the impact that class imbalance in the dataset has on the evaluation metrics.
ROC Curve	Based on "Figure 3.3 ROC Curve (Random Forest)", this ROC curve visualizes the performance of the Random Forest classification model in distinguishing between positive and negative classes at various thresholds. The X-axis represents the False Positive Rate (FPR) or 1 - Specificity, while the Y-axis represents the True Positive Rate (TPR) or Recall. The closer the

	curve is to the top left corner of the graph, the better the model performs in classifying the positive and negative classes. In this case, the curve shows a rapid increase in the True Positive Rate for a low False Positive Rate value, indicating that the model is quite good at identifying positive cases without generating too many false positives. The area under the curve (AUC) is 0.92, which is a very high value. The AUC value of 0.92 indicates that the Random Forest model has excellent discriminative ability, meaning that the model has a 92% probability of ranking random positive instances higher than random negative instances. This confirms that the model is able to discriminate between positive and negative classes with an overall excellent performance.
Visualization of the Trees in Random Forest	Based on "Figure 3.4 Visualization of One of the Trees in Random Forest", this figure displays a visualization of one of the individual decision trees of the Random Forest ensemble. This tree illustrates how features (variables) are used to make classification decisions, from the root to the leaves. Each node in the tree represents a condition or question based on one of the features, which divides the data into two branches (True/False). In the topmost node (root), the initial decision was made based on 'loan_percent_income', with a Gini value of 0.5 and 16551 samples. This indicates that this feature is the most important first separator. Then, the tree branched out. For example, one of the downward paths shows decisions based on 'loan_grade', followed by 'person_income_c' or 'person_emp_length_c', and so on until reaching the leaf nodes. Each node also displays the Gini value (a measure of node purity), the number of samples at that node, the class distribution value (e.g., [number of classes 0, number of classes 1]), and the predicted class (e.g., 'Fully Paid' or 'Default'). This visualization helps us understand how the Random Forest model internally makes classification decisions for each instance, even though Random Forest itself is a collection of many trees like this. By looking at one of the trees, we can gain insight into which features are considered important and how the interaction of those features leads to the final prediction.
Cross-Validation ROC AUC Features	According to "Figure 3.5 CV ROC AUC Score",

	this result presents the ROC AUC score of the cross-validation process applied to the model. There are five individual ROC AUC scores listed: 0.86037862, 0.85849143, 0.84736856, 0.84621044, and 0.86200895. These scores indicate the consistency of the model's performance across different folds of cross-validation, which is an important indicator that the model is not overfit on one part of the training data. The average of these scores, "Mean ROC AUC: 0.8549", provides a more robust estimate of the overall model performance. The average AUC value of 0.8549 indicates that the model has excellent discriminative ability in distinguishing between positive and negative classes. This means that the model has a high probability of ranking positive examples higher than negative examples, even when tested on different subsets of data. The consistency of the AUC scores in each fold and the high average value confirm that the Random Forest model has good generalization and stable performance.



```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.93      0.96      0.94      5095
           1       0.82      0.74      0.78      1422

    accuracy                           0.91      6517
   macro avg       0.88      0.85      0.86      6517
weighted avg       0.91      0.91      0.91      6517
```

Figure 3.1 Classification Report Random Forest

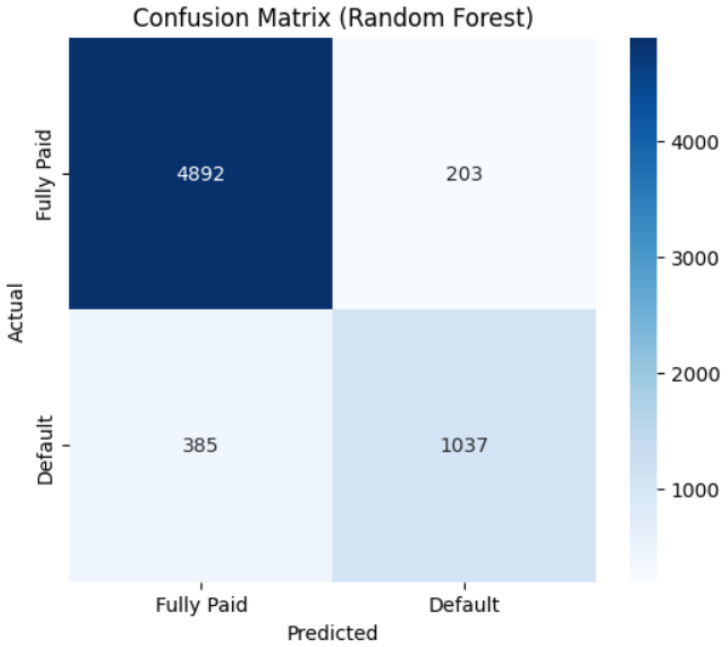


Figure 3.2 Confusion Matrix

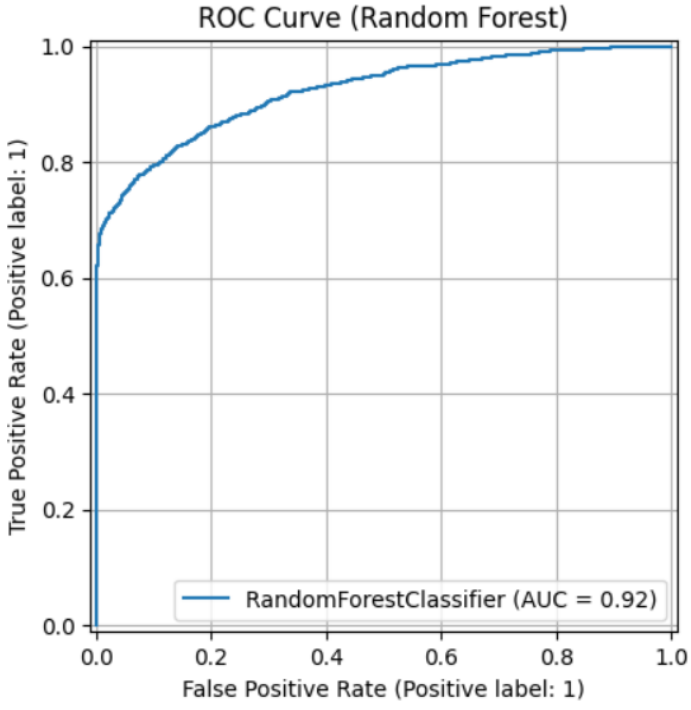


Figure 3.3 ROC Curve

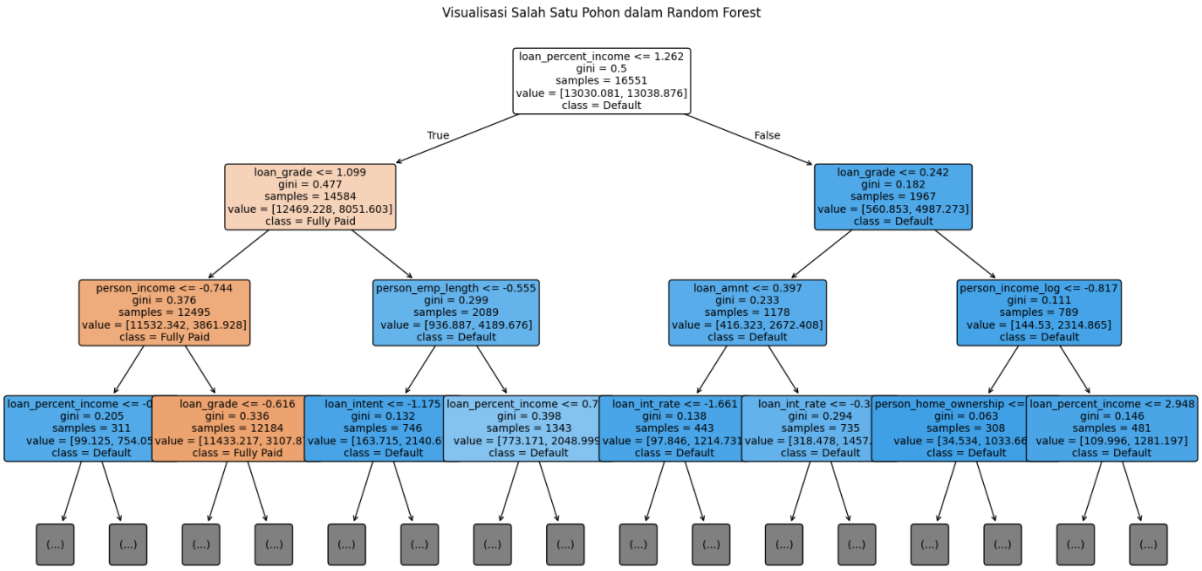


Figure 3.4 Visualization of One of the Trees in Random Forest

```
=== Cross-Validation ROC AUC Scores ===
[0.86037862 0.85849143 0.84736856 0.84621044 0.86200895]
Mean ROC AUC: 0.8549
```

Figure 3.5 CV ROC AUC Score

3.3 Comparison with previous research

NO	Results	Method	Accuracy
1	A machine learning-based credit risk prediction system using stacked classifier and filter-based feature selection method	Stacked Ensemble	84.58%
2.	Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers	XGBoost	75.6%
3.	Enhancing Credit Risk Management with Machine Learning: A Comparative Study of Predictive Models for Credit Default Prediction	LightGBM	90.1%
4.	Implementation of Decision Tree Method to Predict Customer Interest in Internet Data Packages.	Decision Tree	81%

5.	Hasil Penelitian Project Ini	Random Forest	91,1%

3.4 Scenario Table

No.	Data	Split	Preparation	Feature Selection	Evaluate
1	DATA PINJAMAN	train_test_split	LabelEncoder, SimpleImputer	Semua fitur digunakan kecuali StatusPinjaman	Decision Tree: 0.77
2	DATA PINJAMAN	train_test_split	LabelEncoder, SimpleImputer	Semua fitur digunakan kecuali StatusPinjaman	Random Forest: 0.80
3	DATA PINJAMAN	train_test_split	OneHotEncoder, SimpleImputer	Semua fitur digunakan kecuali StatusPinjaman	XGBoost: 0.84
4	credit_risk_dataset	train_test_sp	LabelEncode	Feature	Random

		lit	r, SimpleImputer	Importances dan Top-N	Forest: 0.91

CHAPTER IV

CONCLUSION

This chapter discusses the broader implications of implementing a machine learning-based loan approval prediction model, focusing on its business value, system viability, and long-term strategic potential. The analysis not only covers technical performance but also assesses its impact on stakeholders and how it aligns with organizational goals in the financial sector.

4.1 Business Value and Stakeholder Impact

The machine learning model offers substantial benefits to various stakeholders involved in the loan approval process. Internally, it serves as a tool for improving decision-making and operational efficiency. Externally, it supports transparency and fairness in financial services. The following table outlines the specific value propositions for each stakeholder group:

No	Evaluation Aspect	Explanation
1	Business Value	The loan approval prediction system offers early risk identification and increased approval efficiency. With 91% accuracy, the model minimizes default risk and accelerates credit screening.

2	Internal Stakeholders	The credit team benefits from data-driven insights, while management obtains performance reports to refine lending strategies and reduce non-performing loans (NPLs).
3	External Stakeholders	Borrowers experience more objective evaluations, improving fairness and transparency. Regulators benefit from data-driven credit assessment practices.

4.2 Online System Viability

For the model to be practically useful, it must be viable for real-time integration into existing financial systems. This includes technological readiness, potential implementation environments, and associated challenges. The table below evaluates the model's feasibility in an operational setting:

No	Evaluation Aspect	Explanation
1	Technology Infrastructure	The model pipeline supports preprocessing (e.g., handling missing values, scaling, SMOTE) and can be integrated into bank systems for automatic loan scoring.
2	Implementation Potential	The system can be deployed as a backend API in internal credit evaluation systems or mobile banking apps for real-time scoring.
3	Key Challenges	The main concerns include ensuring data privacy, regulatory compliance, and model monitoring to avoid degradation over time.

UNIVERSITAS MULTIMEDIA NUSANTARA

4.3 Strategy for Machine Learning Model Deployment

Implementing a machine learning model in a real-world banking environment requires not only technical deployment but also strategic alignment. The strategies below are designed to ensure long-term effectiveness, model integrity, and organizational collaboration:

No	Strategy	Explanation
1	Integration in Credit Workflow	Random Forest models can assist in pre-screening applications. This helps triage risky profiles before

		human intervention is needed.
2	Periodic Retraining	Retraining the model with updated borrower data ensures the model remains aligned with economic trends and prevents accuracy decay.
3	Cross-Functional Collaboration	Collaboration between data science, risk management, compliance, and IT is essential for successful deployment and interpretation of results.

4.4 Long Term Impact

The deployment of predictive analytics in the credit approval pipeline has implications that extend far beyond technical outcomes. The table below highlights the potential long-term benefits of the system across business operations and customer experience:

No	Impact	Explanation
1	Operational Efficiency	Automation reduces the time and cost associated with manual credit evaluation processes.
2	Enhanced Risk Detection	Improved classification reduces subjectivity and error in identifying high-risk applicants.
3	Customer Trust	Faster, data-backed decisions improve user satisfaction and the institution's credibility in financial services.

4.5 Intellectual Property Potential

The machine learning model and its surrounding infrastructure not only serve a functional purpose but may also be treated as intellectual assets. Below is an outline of the components that could represent proprietary innovation:

No	Asset Component	Description
1	Model Pipeline & Codebase	The end-to-end pipeline (data preprocessing, feature selection, training, evaluation) is a reusable, proprietary system asset..

2	Feature Importance Mapping	Insights like loan_percent_income and loan_grade as top features could be considered for analytical IP or internal use models.
3	Decision Rule Structures	Visualized trees and rule sets generated by the model represent an interpretable knowledge system that may be trademark-worthy.

4.6 Recommendations for Strategic Implementation

To ensure the model's continued success and scalability, several strategic actions are recommended. These measures support both technological deployment and stakeholder adaptation:

No	Recommendation	Explanation
1	Phased Deployment	Begin with internal deployment (back office only), followed by full integration into customer-facing systems.
2	Staff Training	Provide training for credit officers to interpret model outputs, especially edge-case decisions.
3	Monitoring & Model Governance	Establish dashboards to track model accuracy, ROC AUC, and drift over time. Include alerts for performance drop or data changes.
4	Customer Communication Strategy	Communicate to users how loan evaluations are supported by data analytics to foster trust and transparency.

LITERATURE

[1] Hind, S., Kanderske, M., & Van Der Vlist, F. (2022). Making the Car "Platform Ready": How Big Tech Is Driving the Platformization of Automobility. Social Media + Society, 8. https://doi.org/10.1177/20563051221098697.

[2] Chung, K. C. (2019). Mobile (shopping) commerce intention in central Asia: The impact of culture, innovation characteristics and concerns about order fulfilment. *Asia-Pacific Journal of Business Administration*, *11*(3), 251–266. https://doi.org/10.1108/APJBA-11-2018-0215

[3] Ram, J., & Sun, S. (2020). Business benefits of online-to-offline ecommerce: A theory driven perspective. In *Journal of Innovation Economics and Management* (Vol. 33, Issue 3, pp. 135–162). Boeck Universite. https://doi.org/10.3917/jie.033.0135

[4] Steinberg, M. (2021). From Automobile Capitalism to Platform Capitalism: Toyotism as a prehistory of digital platforms. Organization Studies, 43, 1069 - 1090. https://doi.org/10.1177/01708406211030681.

[5] Costley, A., Kunz, C., Sharma, R., & Gerdes, R. (2021). Low Cost, Open-Source Platform to Enable Full-Sized Automated Vehicle Research. IEEE Transactions on Intelligent Vehicles, 6, 3-13. https://doi.org/10.1109/TIV.2020.3029771.

[6] Rehm, F., Seitter, J., Larsson, J., Saidi, S., Stea, G., Zippo, R., Ziegenbein, D., Andreozzi, M., & Hamann, A. (2021). The Road towards Predictable Automotive High - Performance Platforms. 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), 1915-1924. https://doi.org/10.23919/DATE51398.2021.9473996.

[7] Fletcher, R., Mahindroo, A., Santhanam, N., & Tschiesner, A. (2020). The case for an end-to-end automotive-software platform.

[8] Sharma, P., Kumar, N., & Park, J. (2019). Blockchain-Based Distributed Framework for Automotive Industry in a Smart City. IEEE Transactions on Industrial Informatics, 15, 4197-4205. https://doi.org/10.1109/TII.2018.2887101.

[9] ElHakim, R., Elqadi, A., Torky, M., Zayed, M., Farag, I., & Agamawi, M. (2021). Let's DO - Automotive Platform for Interoperability. 2021 4th International Conference on Information and Computer Technologies (ICICT), 294-299. https://doi.org/10.1109/ICICT52872.2021.00054.

[10] Schörkhuber, D., Popp, R., Chistov, O., Windbacher, F., Hödlmoser, M., & Gelautz, M. (2023). Design of an automotive platform for computer vision research. , 1-6. https://doi.org/10.2352/ei.2023.35.16.avm-119.

[11] Zhou, R., Li, C., & Wang, X. (2022). Bank Customer Classification Algorithm Based on Improved Decision Tree. 2022 2nd Asia Conference on Information Engineering (ACIE), 30-33. https://doi.org/10.1109/acie55485.2022.00014.

[12] Kang, H., Zhao, H., & Ai, T. (2020). The Description of Optimal Decision Tree Algorithm and Its Application in Customer Consumption Behavior. 2020 IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), 1, 655-659. https://doi.org/10.1109/ICIBA50161.2020.9277060.

[13] Surya, R., Dar, M., & Nasution, F. (2024). Implementation of Decision Tree Method to Predict Customer Interest in Internet Data Packages. International Journal of Science, Technology & Management. https://doi.org/10.46729/ijstm.v5i4.1155.

[14] Rahmatillah, I., Astuty, E., & Sudirman, I. (2023). An Improved Decision Tree Model for Forecasting Consumer Decision in a Medium Groceries Store. 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), 245-250. https://doi.org/10.1109/ICIIS58898.2023.10253592.

[15] Reinaldo, R., & Dwiasnati, S. (2023). Prediction of Customer Data Classification by Company Category Using Decision Tree Algorithm (Case Study: PT. Teknik Kreasi

Solusindo). International Journal of Advanced Multidisciplinary. https://doi.org/10.38035/ijam.v2i2.285.

[16] Perera, P. (2019). Decision tree approach for predicting the credit risk of leasing customers in Sri Lanka. Proceedings of the 3rd International Conference on Business and Information Management. https://doi.org/10.1145/3361785.3361797.

[17] Khumaidi, A., Darmawan, R., & Reztrianti, D. (2024). Application of Ensemble Tree Algorithm for Installment Payment Arrears Prediction at Makmur Bersama Credit Union. Faktor Exacta. https://doi.org/10.30998/faktorexacta.v17i2.21819.

[18] Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. Appl. Soft Comput., 73, 816-828. https://doi.org/10.1016/j.asoc.2018.09.001.

[19] Han, S., Lu, S., & Leung, S. (2012). Segmentation of telecom customers based on customer value by decision tree model. Expert Syst. Appl. 39, 3964-3973. https://doi.org/10.1016/j.eswa.2011.09.034.

[20] Ugwoke, R., Onyeanu, E., Ugwoke, O., & Ajayi, T. (2022). Evaluating Coaching Intervention for Financial Risk Perception and Credit Risk Management in a Nigerian Sample. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.962855.

[21] Umar, M., Ji, X., Mirza, N., & Naqvi, B. (2021). Carbon neutrality, bank lending, and credit risk: Evidence from the Eurozone.. Journal of environmental management, 296, 113156 . https://doi.org/10.1016/j.jenvman.2021.113156.

[22] Frydman, H., & Matuszyk, A. (2020). Random survival forest for competing credit risks. Journal of the Operational Research Society, 73, 15 - 25. https://doi.org/10.1080/01605682.2020.1759385.

UNIVERSITAS
MULTIMEDIA
NUSANTARA