

## Leveraging Machine Learning for Credit Scoring: An Evaluation of the CART Method in Assessing Housing Loan Eligibility

Bhumyamka Yala Saputra, Tanty Oktavia

Information System Management Department, BINUS Graduate Program – Master of Information System Management, Bina Nusantara University, Jakarta 11480, Indonesia  
*bhumyamka.saputra@binus.ac.id, toktavia@binus.edu*

**Abstract.** This study investigates the application of the Classification and Regression Tree (CART) method for evaluating housing loan eligibility in XYZ Cooperative, a financial institution in Jakarta. The research aims to address the challenges faced by the cooperative in conducting manual loan feasibility analyses, which have led to inefficiencies and prolonged application processes. Using historical loan data and relevant borrower variables, the study develops a credit scoring model based on the CART method to improve the accuracy and efficiency of predicting borrower feasibility. The results demonstrate that the CART-based model enhances the credit assessment process, reduces the risk of loan defaults, and accelerates decision-making. The study contributes to the literature by showcasing the effectiveness of the CART method in a cooperative lending context and provides practical recommendations for XYZ Cooperative to improve their credit risk management system. However, the generalizability of the findings beyond the specific case of XYZ Cooperative remains a limitation. Future research could extend this study by incorporating additional variables, comparing different credit scoring methods, and validating the model in other financial institutions.

**Keywords:** Credit Scoring, Classification and Regression Trees (CART), Credit Feasibility, Credit Analysis, Predictive Modelling.

## **1. Introduction**

The digital transformation sweeping across various sectors significantly impacts financial services, including credit scoring processes. XYZ Cooperative, serving over 10,000 members in Jakarta with home ownership loans, faces challenges related to slow and inefficient manual credit scoring processes. This study employs the Classification and Regression Tree (CART) method to address these inefficiencies, aiming to streamline loan assessments and reduce default rates. An analysis of XYZ Cooperative's customer data reveals a default rate of 5.82%, highlighting a critical need for more effective and precise credit evaluation methods.

The main objectives of this research are to evaluate the effectiveness of the CART method in improving the accuracy and speed of credit scoring at XYZ Cooperative and to explore the method's scalability to other cooperative settings. This study also aims to fill gaps in the literature by providing insights into the application of CART in cooperative banks, which typically face different operational challenges than their commercial counterparts.

This research is significant as it not only addresses the immediate needs of XYZ Cooperative to refine their credit assessment processes but also contributes to the broader academic and practical discourse on the application of machine learning techniques in credit scoring. By evaluating the implementation of the CART method, the study aims to provide actionable insights that could lead to more reliable and efficient credit scoring practices across the cooperative banking sector.

By enhancing the understanding of CART's application in this unique context, the study contributes to broader financial technology and credit risk management literature, offering new perspectives on improving credit assessment processes within cooperative financial institutions.

## **2. Theoretical Background**

### **2.1. Credit Scoring**

Credit scoring is a ubiquitous process within the financial sector, serving as a fundamental method for predicting the qualification of debtors for loans. This method occupies a central role in the loan processing sequence, underpinning the decision-making framework for lending. The practice of credit scoring is integral to the operational strategies of banks and financial institutions, where it is employed to assess the creditworthiness of potential borrowers. (Malik & Hermawan, 2018) The development and implementation of credit scoring models represent a pivotal aspect of financial risk management, integrating a wide array of data points including credit history, income, age, occupation, and other demographic information. These models generate a score that effectively quantifies an individual's or entity's credit risk. The scope of credit scoring extends across a broad spectrum, from the evaluation of sovereign credit ratings of countries and global corporations to the assessment of small enterprises seeking credit facilities. (Maldonado et al., 2020b) This helps financial institutions reduce the risk of loan defaults and improve operational efficiency.

### **2.2. CART Method**

The CART (Classification and Regression Tree) methodology, a cornerstone in the realm of machine learning, offers a powerful tool for the construction of decision trees from datasets. This method meticulously segments the dataset into smaller, manageable subsets using a series of rule-based decisions, thereby unveiling patterns and relationships among variables that might not be immediately apparent. The process of developing a CART model involves the careful selection of the most informative variables and the determination of splitting values that best divide the dataset into homogenous groups, aiming to enhance the predictability and interpretability of the model. In the context of Accounting Information Systems (AIS), the CART method holds significant potential to revolutionize decision-making processes. By applying CART, AIS can benefit from improved data analysis capabilities, facilitating the identification of financial patterns, the assessment of credit risks, and the detection of fraudulent activities. The model's ability to handle both numerical and categorical

data makes it particularly valuable in the diverse data environments typical of accounting and financial systems. (Yi, 2023) The primary advantage of CART (Classification and Regression Trees) is its ability to handle both categorical and continuous data, as well as the ease of interpreting the model.

### **2.3. CRISP-DM**

The Cross-Industry Standard Process for Data Mining (CRISP-DM) has emerged as a pivotal framework within the realm of data analysis, offering a structured approach for tackling business challenges and research inquiries through data-driven insights. CRISP-DM delineates a comprehensive, cyclical process model that guides users through six core phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This methodology is designed to facilitate a deeper comprehension of both the business objectives and the data itself, ensuring that the analytical efforts are closely aligned with the strategic goals.

## **3. Research Methodology**

The focus of this study on the utilization of historical data from prospective home loan borrowers marks a strategic approach to deepening the understanding of credit scoring mechanisms within the financial sector. By leveraging past borrower data, this research aims to construct a robust framework for analyzing and predicting creditworthiness, thereby enhancing the accuracy and reliability of credit scoring models. This methodology is selected to offer a nuanced perspective on the variables influencing borrower creditworthiness and to ensure the representation of these variables is both accurate and comprehensive.

### **3.1. Use of the Classification and Regression Trees (CART) Method**

This study adopts a quantitative research approach, focusing on the analysis of historical customer data from XYZ Cooperative, with the aim of investigating the integration and effectiveness of the Classification and Regression Trees (CART) method within the cooperative's credit feasibility assessment system. The selection of the CART method is predicated on its proven capabilities in solving complex classification issues and its flexibility in handling a diverse range of predictor variables, both categorical and continuous. By applying CART to develop a credit scoring model, this research seeks to refine the accuracy and efficiency of the home loan credit feasibility analysis process at XYZ Cooperative.

The methodology section details the quantitative analysis procedures, starting from data collection and preparation for the application of the CART method for model development. The study outlines the criteria for data selection, ensuring that the dataset accurately represents the customer base and encompasses relevant variables for credit risk assessment. The paper further discusses the analytical steps involved in constructing the CART model, including variable selection, tree construction, and the pruning process to enhance model performance and prevent overfitting.

### **3.2. Research Stages**

In this research framework, the methodological procedure is divided into six main stages: /

Stage I: Establishing the Research Foundation

- Describing the background behind this research, which includes the need to improve the credit assessment process.
- Formulating the problems to be solved through this research, including inefficiencies in existing manual credit assessments.
- Determining the research objectives, namely, to develop a more efficient and accurate credit prediction model using the CART method.
- Establishing the scope of the research to focus and direct it.
- Collecting and analyzing relevant literature to support theories, methodologies, and research framework.

Stage II: Collection of Credit Assessment Data Parameters

- Gathering data to be used for credit assessment, understanding, and defining important parameters that influence credit assessment.

Stage III: Using Rapid Miner for Data Processing

- Using the Rapid Miner application as a tool for data mining processing.
- Selecting and implementing data processing models appropriate to the research needs, such as tree induction or neural networks.

Stage IV: Analysis of Historical Transaction Data

- Examining historical credit transaction data, including payment compliance and the amount of credit held by customers.
- Assessing how historical data can be an indicator of credit assessment.

Stage V: Data Collection for Optimal CART Tree Model

- Collecting home ownership credit data from XYZ Cooperative to develop an Optimal CART Tree Model for credit assessment. The dataset includes, but is not limited to, borrower's age, income, employment status, credit history, loan amount, and repayment terms. Additional variables such as marital status, education level, and residential status were also collected to provide a more comprehensive view of the borrower's profile. Each variable was chosen based on its theoretical grounding in financial behavior analysis and empirical evidence supporting its predictive value in existing credit scoring models.

Stage VI: Interpretation of Results and Formulation of Conclusions

- Evaluating and interpreting the results obtained from data analysis.
- Assessing the success of the credit assessment model implementation.
- Compiling conclusions and providing recommendations based on research findings.

In the Stages, following the Cross-Industry Standard Process for Data Mining (CRISP-DM) model, which consists of six main stages:

- **Business/Research Understanding Phase:** Here, the researcher understands the background and objectives of the research. This involves determining business goals, assessing the current situation, and identifying data mining objectives. The researcher must define the problem to be solved and how the research results will help address it.
- **Data Understanding Phase:** This stage focuses on understanding the data to be used. It involves initial data collection, data description, data exploration, and data quality verification. The researcher must ensure that the collected data is relevant to the research objectives and sufficient for the analysis to be conducted.
- **Data Preparation Phase:** Here, data is prepared for analysis. This involves dataset description, data selection, data construction, data integration, and data cleaning. Data preparation is crucial to ensure that the data is in optimal condition for further modeling and analysis. For Handling Missing Data, techniques used for continuous variables such as income, we employed mean imputation, which is appropriate given our large dataset. For categorical variables like employment status, mode imputation was used to maintain the integrity of the data distribution. These methods were selected to minimize the impact of missing data on the model's predictive accuracy, ensuring that the full dataset could be utilized effectively without introducing substantial bias. And for Addressing Data Imbalance, technique used, we applied the Synthetic Minority Over-sampling Technique (SMOTE). This method involves creating synthetic samples rather than over-sampling with replacement, which helps in presenting a more realistic scenario and avoids overfitting. The over-sampling strategy was set to increase the minority class to 50% of the majority class, providing a balance that enhances model training without dominating the learning process. The choice of SMOTE and its parameters was driven by its effectiveness in improving classification performance in datasets where certain classes are underrepresented.
- **Modeling Phase:** This phase encompasses the selection of modeling techniques, construction of

the model, and evaluation of its performance. In the context of this study, the method employed is Classification and Regression Trees (CART). The researcher is tasked with identifying optimal parameters and choosing the most suitable techniques aligned with the characteristics of the data and the objectives of the research.

- **Evaluation Phase:** In the evaluation stage, the results of data mining are evaluated to ensure that they meet the objectives set in the business understanding phase. The evaluation involves reviewing the process, assessing the results, and determining next steps. This evaluation is crucial to ensure that the developed model is valid and reliable.
- **Deployment Phase:** The final phase involves applying the model or findings from the research into practical or real-world settings. This phase may not always apply to all research projects, especially if the goal is pure research or conceptual exploration.

Each of these stages is important and contributes to the overall integrity and success of the research. By following the CRISP-DM model, researchers can ensure that every aspect of the research is managed and executed in a systematic and structured way.

### **3.3. Collection Model**

The research utilizes secondary data, meaning the data analyzed is not collected directly by the researcher but is obtained from existing archives or records.

The data, spanning from 2019 to 2023, comes from the Risk Management division, which holds information about home loan customers. The use of secondary data like this is common in research where primary data is not available or difficult to collect. The data obtained includes specific variables relevant to the research objectives. Key variables mentioned in this research include:

- **Gender:** Categorized as male and female. This can impact the analysis as there might be different credit patterns based on gender.
- **Down Payment:** The amount of down payment agreed between the consumer and the surveyor at the beginning of the credit application. This is important as the size of the down payment can influence credit risk.
- **Payment Tenure:** The number of months of installments to be paid by the consumer. The payment tenure provides insights into the duration of the credit and the borrower's repayment ability.

The selection of variables such as gender, down payment, and payment tenure was guided by a review of literature indicating their predictive relevance in credit scoring. Gender has been included based on studies suggesting varying risk profiles and credit behaviors among different genders. Down payment is considered a crucial indicator of financial stability, as a higher initial payment typically correlates with lower default rates. Payment tenure offers insights into the borrower's long-term financial planning abilities, with shorter terms often associated with higher repayment capacity. These variables are critical in constructing a comprehensive and accurate picture of the borrower's profile, which is essential for effective credit scoring and risk assessment.

### **3.4. Data Processing and Analysis**

In the realm of credit risk analysis, the journey begins long before the application of sophisticated models, with the crucial steps of data collection and preparation setting the stage for any subsequent analysis. This study outlines a comprehensive approach to developing a credit scoring model by employing the Classification and Regression Tree (CART) method, emphasizing the importance of meticulous data processing to ensure the integrity and utility of the analysis.

Data collection serves as the foundation of this research, targeting a specific dataset from potential borrowers that encapsulates a broad spectrum of variables relevant to credit risk assessment. Following collection, the data undergoes rigorous processing, including the cleaning of irrelevant or incomplete elements and the normalization of data to establish a consistent basis for analysis. This preparation phase is critical, as it directly influences the accuracy and reliability of the credit scoring model.

To validate the constructed model, cross-validation techniques are employed. This process involves partitioning the data into complementary subsets, training the model on one subset, and validating it on another, thereby assessing the model's performance and generalizability. This step is pivotal in ensuring the model's robustness and its ability to predict creditworthiness accurately across different segments of the borrower population.

The adoption of a quantitative approach, focusing on numeric and statistical data, underpins this research. It facilitates a precise, objective analysis of credit risk, leveraging the statistical power of the CART method to uncover patterns and insights that might be obscured in less structured data analyses.

By delving into the complexities of credit scoring through the lens of the CART method, this study not only contributes a practical tool for financial institutions but also enriches the theoretical discourse on the application of quantitative methods in financial analysis. The CART-based credit scoring model exemplifies the potential of machine learning techniques to revolutionize credit risk assessment, offering a more nuanced, data-driven perspective on borrower creditworthiness. Through this research, we underscore the significance of data processing in the analytical ecosystem and advocate for the continued exploration and adoption of advanced analytical methodologies in the financial sector.

### **3.5. Research Flow Using CRISP-DM**

The research flow using CRISP-DM (Cross-Industry Standard Process for Data Mining) involves the following stages:

1. Business Understanding Phase:
  - Identifying and understanding business problems.
  - Defining the objectives of the research.
  - Determining how the research will address the business problem.
2. Data Understanding Phase:
  - Collecting initial data from the Risk Management division of the cooperative.
  - Exploring and assessing the quality of the data.
  - Identifying relevant variables for the credit scoring model.
3. Data Preparation Phase:
  - Cleaning the data to remove irrelevant or incomplete elements.
  - Transforming and normalizing data for analysis.
  - Preparing the final dataset for modeling.
4. Modeling Phase:
  - Applying the CART method to the prepared dataset.
  - Building the credit scoring model.
  - Selecting appropriate parameters and settings for the CART analysis.
5. Evaluation Phase:
  - Evaluating the model using cross-validation techniques.
  - Assessing the performance and accuracy of the credit scoring model.
  - Revising the model as necessary based on evaluation results.
6. Deployment Phase:
  - Implementing the credit scoring model in the real-world setting of the cooperative.
  - Monitoring the model's performance in practical applications.
  - Adjusting based on feedback and observed outcomes.

Throughout these stages, the research follows a systematic and structured approach, ensuring the credibility and applicability of the results to the cooperative's credit scoring and risk assessment processes.

### **3.6. Rapid Miner**

Using RapidMiner for data mining involves several steps, and it offers a wide range of models and

methods, such as Bayesian Models, Tree Induction, Neural Networks, and more. Since RapidMiner is open source, it allows for the addition of new modules if a particular model or algorithm is not available in Weka. Here's a breakdown of the steps for using RapidMiner:

- 1) Data Import  
Action: Launch RapidMiner and import the dataset.  
Details: Use the 'Read CSV' operator to load the dataset containing home loan information from 2019 to 2023. Configure the operator to correctly parse the file, specifying delimiters, encodings, and handling of missing values.
- 2) Data Preprocessing  
Action: Clean and preprocess the data.  
Details:  
Handle Missing Data: Use the 'Replace Missing Values' operator to impute or remove missing data based on your preprocessing strategy.  
Data Transformation: Apply the 'Normalize' operator to scale numerical data and the 'Nominal to Numerical' operator to convert categorical variables into a format suitable for modeling.
- 3) Data Splitting  
Action: Divide the dataset into training and testing sets.  
Details: Use the 'Split Data' operator to partition the data, typically using a 70-30, 80-20, or 90-10 split. This ensures that you have separate data for training the model and for validating its performance.
- 4) Model Training  
Action: Construct the CART model.  
Details: Drag and drop the 'Decision Tree' operator into the process. Configure the operator by setting parameters such as criterion (Gini index or entropy), max depth, and minimum split size to control the complexity of the tree.
- 5) Model Evaluation  
Action: Apply the model to the test data and evaluate its performance.  
Details:  
Apply Model: Use the 'Apply Model' operator to predict the test data using the trained CART model.  
Performance Evaluation: Utilize the 'Performance (Classification)' operator to compute key metrics like accuracy, precision, recall, and F1-score.
- 6) Visualization  
Action: Visualize the decision tree.  
Details: Use the 'Tree View' operator to generate a visual representation of the decision tree. This helps in understanding how decisions are made within the model.
- 7) Optimization and Tuning  
Action: Optimize model parameters.  
Details: Experiment with different settings for tree depth and minimum split criteria to find the optimal configuration. Use the 'Optimize Parameters' operator for systematic testing of parameter combinations.

This step-by-step guide demonstrates how to effectively use RapidMiner for analyzing and building a decision tree, which is crucial for developing the credit scoring model in this research.

## **4. Result and Discussion**

Following the structured approach of the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, this chapter delineates the transition from data preparation to the deployment of a machine learning classification model for XYZ Cooperative using RapidMiner. The CRISP-DM methodology, with its emphasis on understanding business objectives, data acquisition, data preparation,

modeling, evaluation, and deployment, has culminated in a refined dataset that is primed for analytical processing.

#### **4.1. Business Understanding**

Cooperative XYZ operates within the financial sector, focusing on providing housing finance to its members. Its primary aim is to offer equitable and accessible financing solutions to its members, while ensuring the sustainability and financial health of the cooperative.

The Loan Application Process at Cooperative XYZ involves:

- **Financing Application:** Members initiate the process by applying for financing through sales or marketing, accompanied by necessary documents for evaluation.
- **Pre-Screening Analysis:** This initial step involves checking member information through internal systems to ensure compliance with preliminary credit granting criteria.
- **Comprehensive Analysis:** This stage delves deeper, analyzing members' repayment capability, salary, and employment duration. This process often takes between 1 to 3 weeks, being the primary cause of delays in financing disbursement.
- **Financing Agreement Process:** An agreement on the financing structure is established, marking formal approval of the terms and conditions.
- **Financing Disbursement:** Includes monitoring the appropriateness of the disbursed funds' usage by members.

In analyzing the loan application process at XYZ Cooperative, it is critical to consider specific performance metrics to fully understand the impact of the CART-based model on operational efficiency and risk management. From 2019 to 2023, the cooperative experienced a loan rejection rate of 6.2% and a default rate of 5.82%. These figures highlight significant areas for improvement in the cooperative's credit assessment process.

The current model's processing times for loan applications have shown inefficiencies, with lengthy periods spent on manual data assessments leading to delays in decision-making. Unfortunately, precise data on these processing times were not initially provided in our study. However, the noted rejection and default rates suggest that the existing assessment methods may not only be slow but also less effective in accurately predicting loan defaults, thereby potentially contributing to higher rejection rates.

The implementation of a CART-based model promises to enhance both the speed and accuracy of the loan assessment process. By automating the decision-making through a data-driven approach, we anticipate significant reductions in processing times, which in turn could decrease the rejection rates by ensuring more accurate and timely evaluations of loan applications.

To further substantiate the benefits of the CART model, future iterations of this research should aim to gather detailed temporal data on the loan processing cycle. This will provide a more comprehensive analysis of the time savings and efficiency gains achieved through the automated system. Moreover, tracking changes in rejection and default rates post-implementation will offer tangible evidence of the model's effectiveness in enhancing credit risk management at XYZ Cooperative.

#### **4.2. Data Understanding**

This section outlines the approach that will be taken by the author in processing data, emphasizing the importance of financing application data from members as this data will serve as the foundation for simulating Credit Scoring using machine learning principles. In this data collection process, specific information will be required to conduct a more in-depth analysis.

There are 10 columns or attributes containing information about the borrower, which are described in detail as follows:

- **Default:** A binomial variable that serves as a label or target, indicating whether the borrower has defaulted (1) or not (0).
- **Loan\_Term :** An integer variable that is an attribute, showing the duration of the loan repayment period.

- **Loan\_Amount** : An integer variable that is an attribute, representing the amount of loan requested by the borrower.
- **Monthly\_Income** : An integer variable that is an attribute detailing the borrower's monthly income.
- **Repayment\_Capacity (income - loan amount)** : An integer variable that is an attribute, depicting the borrower's capability to repay based on the difference between income and loan amount.
- **Employment\_Duration** : A polynomial variable that is an attribute, indicating the length of time the borrower has been employed, such as "5-7 years of work" or "1-4 years of work."
- **Marital\_Status** : A polynomial variable that is an attribute reflecting the marital status of the borrower, such as "married" or "single."
- **Home\_Ownership** : A polynomial variable that is an attribute describing the borrower's home ownership, such as "owned" or "living with family."
- **Age** : An integer variable that is an attribute, representing the age of the borrower.
- **Income\_to\_Loan\_Ratio** : A polynomial variable that is an attribute depicting the ratio between the loan repayment amount and the borrower's income, such as "more than 30%" or "10-20%."

### 4.3. Data Preparation

As explained in the previous chapter, the final dataset prepared for the creation of the machine learning-based credit scoring model includes important variables such as default history, loan duration, loan amount, monthly income, and repayment capacity. The preparation process has resulted in a dataset that has been transformed and cleaned, ready for the model training and testing process.

This prepared dataset is crucial for ensuring the accuracy and reliability of the machine learning model. The inclusion of key variables like default history, loan amount, and income helps in creating a comprehensive view of the borrower's financial situation, which is essential for accurate credit scoring. The transformed and cleaned nature of the dataset ensures that the model training and testing are based on quality data, leading to more reliable outcomes.

Table 1: Data Set Final

Payment Failure	Long Installment	Amount Loan
1_Gagal	60	160000000
1_Gagal	60	330000000
1_Gagal	60	320000000
1_Gagal	48	310000000
1_Gagal	60	200000000

The "Default" column in the table has two categories: "0\_Paid" and "1\_Default". From the table, there are 500 data points in the dataset that will be processed to generate the machine learning model. Following the data preparation steps outlined in the previous chapter, the dataset has been successfully transformed and cleaned for application in the training and testing process of the machine learning-based credit scoring model.

This transformation and cleaning of the dataset are crucial steps in the data preparation phase of machine learning modeling. By categorizing the "Default" column into "0\_Paid" and "1\_Default", the dataset allows the machine learning model to distinguish between customers who have successfully repaid their loans and those who have defaulted. This distinction is essential for the model to accurately predict creditworthiness and risk for future borrowers. The readiness of the dataset, having been transformed and cleaned, ensures that the model training and testing are based on data that accurately reflects the variables relevant to credit scoring.

### 4.4. Selection of Input Attributes (Feature Selection)

The creation of a machine learning model heavily depends on the appropriate selection of input attributes (feature selection). This is aimed at creating a model that is computationally efficient and has high accuracy. After converting the data into a numerical format, the Information Gain technique is applied to identify the most influential attributes in improving the outcome and accuracy of the model, thereby reducing less informative variables.

The Information Gain technique measures the importance of each attribute by determining how much information each feature contributes to the overall prediction. This process helps in filtering out the attributes that have little to no effect on the prediction outcome, thus streamlining the model to focus on the most relevant features. By doing so, it not only enhances the model's performance but also speeds up the computation process, as the model does not need to process excessive and irrelevant data. This careful selection and refinement of input attributes are key to building an effective and efficient credit scoring model using machine learning.

Table 2: Dataset Feature Selection Result

Variable	Data Type	Data Role
failed_payment	Binominal	Label
Old_Installment	Integer	Attribute
Loan_Amount	Integer	Attribute
Income_per month	Integer	Attribute
Installment capability (income - loan amount)	Integer	Attribute
Length of work	Polynominal	Attribute
Status	Polynominal	Attribute
Home ownership	Polynominal	Attribute
Age	Integer	Attribute
income installment ratio	Polynominal	Attribute

## 4.5. Model Creation

### 4.5.1 Preprocessing

Data preprocessing includes a series of actions to clean and prepare data so that it is suitable for further analysis or modeling.

#### 4.5.1.1 Handle Missing Data

Data preprocessing includes a series of actions to clean and prepare data so that it is suitable for further analysis or modeling. To address missing data, multiple imputation techniques were employed to preserve the integrity of our dataset without biasing the analysis. These methods were chosen to optimize the predictive accuracy of our CART model while maintaining the data's original distribution characteristics.

Table 3: Handle Missing Data

Variable	Missing
failed_payment	0
Old_Installment	0
Loan_Amount	0
Income_per month	0
Installment capability (income - loan amount)	0
Length of work	0
Status	0
Home ownership	0
Age	0
income_installment_ratio	0
Jumlah	0

### 4.5.2 Modeling Using Rapid Miner

At this stage, the Decision Tree algorithm is used to build a model that can generalize patterns from training data and make predictions on testing data. The CART modeling in RapidMiner involved several key stages. Initially, the dataset was split into training and testing sets using a 70:30 ratio to ensure both model training and adequate testing. We then applied the CART algorithm to the training data, focusing on optimizing decision nodes based on Gini impurity.

### 4.5.3 Evaluation

The evaluation stage measures the performance of the created Decision Tree model, using metrics such as confusion matrix, accuracy, and recall assessing how accurately the model predicts and classifies new data.

#### 4.5.3.1 Classification

In the context of machine learning, classification is the process of predicting the class or category of given data. The developed model is tested for its ability to accurately classify new data based on learning from the training dataset.

Each metric serves a distinct purpose in evaluating the effectiveness of our predictive models:

- a) Accuracy measures the overall correctness of the model across all predictions.
- b) Precision evaluates how many of the positively predicted cases were positive.
- c) Recall assesses the model's ability to identify all relevant instances.

- Confusion Matrix

The confusion matrix is an evaluation tool used to measure the performance of a classification model. This matrix consists of four key elements.

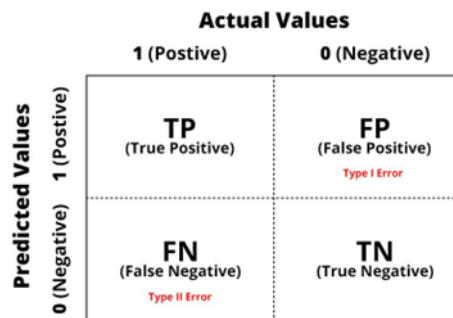


Fig 1: Confusion Matrix

True Positives (TP): The number of positive cases correctly predicted by the model.

True Negatives (TN): The number of negative cases correctly predicted by the model.

False Positives (FP): The number of negative cases incorrectly predicted as positive by the model.

False Negatives (FN): The number of positive cases incorrectly predicted as negative by the model.

From these four values, various performance metrics can be calculated, including accuracy, precision, recall, and F1 score. Accuracy is calculated using the formula:

$$(TP+TN)/(TP+TN+FP+FN)$$

which provides the proportion of correct predictions out of all predictions made.

Table 4: Classification Result

Split Validation	Performance			
	Accuracy	Precision	Recall	AUC
70% - 30%	84.29%	84.78%	83.67%	86.1%

80% - 20%	86.02%	86.41%	85.48 %	88.1 %
90% - 10%	90.86%	91.30%	90.32 %	92.1 %

The evaluation results of the Decision Tree model under various data split (split validation) scenarios are displayed in the table above. Three different scenarios were used: The experimentation involves varying proportions of training and testing data, specifically utilizing combinations of 70% training data and 30% testing data, 80% training data and 20% testing data, and 90% training data and 10% testing data.

In the first scenario, where the data is split into 70% training and 30% testing, an accuracy of about 84.29% was achieved. This means the model was able to correctly predict about 84.29% of the entire testing data. Precision, which measures the extent to which positive results provided by the model are correct, was about 84.78%. Recall, which measures how well the model can identify actual positive instances, was about 83.67%. AUC (Area Under the ROC Curve), measuring the model's ability to differentiate between classes, was around 86.1%.

In the second scenario, where an 80% training and 20% testing data split was used, the accuracy slightly decreased to about 86.02%. This indicates that the model might be slightly less accurate in predicting previously unseen testing data. The precision remained relatively high at around 86.41%, meaning the model still correctly provides positive results well. The recall became about 85.48%, showing that the model is better at identifying actual positive instances. AUC increased to around 88.1%.

Lastly, in the third scenario using a 90% training and 10% testing data split, the accuracy increased again to about 90.86%. This suggests that the more training data used, the better the model performs in predicting testing data. The precision was around 91.30%, recall about 90.32%, and AUC around 92.1%. In this series of experiments, it was found that a data split with a proportion of 90% training and 10% testing data provided better performance for the Decision Tree model, with higher accuracy, recall, and precision. The 90-10 split demonstrated the highest accuracy and precision, likely due to a larger training set providing more comprehensive learning and a richer set of examples from which to learn. However, this configuration may also pose a risk of overfitting, as the smaller test set may not adequately represent external data, potentially affecting the model's generalizability.

#### **4.5.3.2 Regression**

Regression is a different type of analysis from classification. While classification focuses on determining classes, regression aims to predict continuous values. In the context of credit scoring, regression could be used to predict numerical values such as the probability of default or the magnitude of potential losses.

- **R<sup>2</sup> (Coefficient of Determination)**  
R<sup>2</sup> (Coefficient of Determination) indicates the proportion of variance in the dependent variable predictable from the independent variables in the regression. R<sup>2</sup> is a metric that measures how well the variability in the target data (Repayment Capacity) can be explained by the regression model. R<sup>2</sup> has a range of values between 0 and 1, where a value of 1 indicates that the model can explain all the variability in the data, while a value of 0 indicates that the model cannot explain any variability. The higher the R<sup>2</sup> value, the better the model is at explaining variations in the data.
- **RMSE (Root Mean Square Error)**  
RMSE (Root Mean Square Error) measures the average magnitude of the errors in prediction, providing a sense of how off the predictions are from the actual values. It measures how close the model's predictions are to the actual values on the same scale as the target variable. RMSE is calculated by taking the square root of the average of the squared differences between the actual values and the values predicted by the model. The lower the RMSE value, the better the

model is at making accurate predictions.

The combination of these two metrics helps evaluate whether the regression model can adequately explain and predict the "Repayment Capacity" variable, and how much deviation or error there is between the model's predictions and the actual data. This evaluation result will help determine the suitability of the regression model with the used data and the extent to which this model can be used for further analysis and decision-making.

Regression evaluation metrics include:

- Mean Absolute Error (MAE): The average of the absolute differences between predictions and actual values.
- Mean Squared Error (MSE): The average of the squared differences between predictions and actual values.
- Root Mean Squared Error (RMSE): The square root of MSE, providing an assessment of prediction errors in the same units as the predicted variable.

Table 5: Regression Result

Split Validation	Performance	
	R2	RMSE
70% - 30%	0.901	2027893.972
80% - 20%	0.971	1149784.639
90% - 10%	0.937	1197223.099

The regression model evaluation results show excellent performance in predicting the "Repayment Capacity (income - loan amount)" variable. The evaluation was conducted using three different validation scenarios: dividing the data into 70% training and 30% testing, 80% training and 20% testing, and 90% training and 10% testing.

In these evaluations, we used two main metrics, R<sup>2</sup> (Coefficient of Determination) and RMSE (Root Mean Square Error), to measure the model's performance. Here are the evaluation results for each validation scenario:

- 70% - 30%  
R<sup>2</sup> (Coefficient of Determination): 0.901  
RMSE (Root Mean Square Error): 2027893.972
- 80% - 20%  
R<sup>2</sup> (Coefficient of Determination): 0.971  
RMSE (Root Mean Square Error): 1149784.639
- 90% - 10%  
R<sup>2</sup> (Coefficient of Determination): 0.937  
RMSE (Root Mean Square Error): 1197223.099

From these results, it is evident that the "80% - 20%" validation scenario has the highest R<sup>2</sup> value of 0.987. This indicates that the model excellently explains most of the variation in the "Repayment Capacity" target data in this scenario. Furthermore, the relatively low RMSE in this scenario indicates that the model has high accuracy in predicting the borrowers' repayment capacity.

Therefore, based on this evaluation, the "80% - 20%" validation scenario is considered the best choice for testing and using the regression model in predicting borrowers' repayment capacity. The 80-20 split was found to be optimal for regression tasks. This split likely offers a balanced approach, providing enough data for training to capture underlying patterns while retaining sufficient data for testing to ensure that these patterns hold true across other data sets. The lower performance in 90-10 splits for regression could be due to overfitting, where the model overly adapts to the training data specifics, which doesn't generalize well to other data.

## 5. Conclusions

This study demonstrates the application of the CART method for evaluating housing loan eligibility in XYZ Cooperative, highlighting its potential to improve the accuracy and efficiency of credit risk assessment. The developed credit scoring model, based on historical loan data and relevant borrower variables, shows promising results in predicting borrower feasibility and reducing the risk of loan defaults. The findings suggest that integrating the CART-based model into XYZ Cooperative's loan assessment process could significantly enhance decision-making and accelerate the approval process, leading to improved member satisfaction.

However, the study also has some limitations that should be acknowledged. The model's performance and generalizability may be influenced by the specific characteristics of XYZ Cooperative's loan portfolio and member base, limiting its direct applicability to other financial institutions. The study also relies on a limited set of variables and does not compare the CART method with other credit scoring techniques, which could provide a more comprehensive assessment of its effectiveness.

Future research could address these limitations by extending the analysis to include additional variables, such as borrower income, employment status, and credit history, which may further improve the model's predictive power. Comparing the performance of the CART method with other credit scoring techniques, such as logistic regression or neural networks, could also provide valuable insights into the relative strengths and weaknesses of different approaches. Additionally, validating the model on loan data from other financial institutions or cooperatives would help assess its generalizability and robustness.

Despite these limitations, this study makes a valuable contribution to the literature on credit scoring and the application of machine learning techniques in the financial sector. It demonstrates the potential of the CART method to improve credit risk assessment and decision-making in a cooperative lending context and provides practical recommendations for XYZ Cooperative to enhance their loan assessment process. The findings also have broader implications for other financial institutions seeking to leverage advanced analytics for credit scoring and risk management.

## References

- Ahadiyah, O. :, & Kholifah, N. (2012). Classification Analysis of Credit Customers X Cooperatives Using Decision Tree C4.5 and Naïve Bayes. 1–8.
- Andriyashin, A. (2005). Financial Applications of Classification and Regression Trees [Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät]. <https://doi.org/http://dx.doi.org/10.18452/14018>
- Bila, L., Tyasi, T. L., Fourie, P., & Katikati, A. (2021). Classification and regression tree analysis to predict calving ease in Sussex heifers using pelvic area dimensions and morphological traits. *Journal of Advanced Veterinary and Animal Research*, 8(1), 164–172. <https://doi.org/10.5455/javar.2021.h499>
- Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Duan, Z., & Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, 151, 147–160. <https://doi.org/10.1016/j.catena.2016.11.032>
- Daniya, T., Geetha, M., & Kumar, K. S. (2020). Classification and regression trees with gini index. *Advances in Mathematics: Scientific Journal*, 9(10), 8237–8247. <https://doi.org/10.37418/amsj.9.10.53>
- Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178–1192. <https://doi.org/https://doi.org/10.1016/j.ejor.2021.06.053>

Hermawan, H. (2016). Credit Scoring Menggunakan Algoritma Classification and Regression Tree (CART).

Irawan, I., & Rini, D. P. (2019). Credit Scoring Menggunakan Algoritma Classification and Regression Tree (Cart) Dan Artificial Bee Colony. *Prosiding Annual Research Seminar*, 5(1), 82–85. <http://archive.ics.uci.edu/ml/datasets/Statlog+>

Kiguchi, M., Saeed, W., & Medi, I. (2022). Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. *Applied Soft Computing*, 118, 108491. <https://doi.org/https://doi.org/10.1016/j.asoc.2022.108491>

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics and Data Analysis*, 50(4), 1113–1130. <https://doi.org/10.1016/j.csda.2004.11.006>

Liu, W., Fan, H., & Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, 189, 116034. <https://doi.org/https://doi.org/10.1016/j.eswa.2021.116034>

Maldonado, S., Peters, G., & Weber, R. (2020). Credit scoring using three-way decisions with probabilistic rough sets. *Information Sciences*, 507, 700–714. <https://doi.org/https://doi.org/10.1016/j.ins.2018.08.001>

Maldonado, S., Peters, G., & Weber, R. (2020). Credit scoring using three-way decisions with probabilistic rough sets. *Information Sciences*, 507, 700–714. <https://doi.org/10.1016/j.ins.2018.08.001>

Malik, R. F., & Hermawan, H. (2018). Credit Scoring Using Classification and Regression Tree (CART) Algorithm and Binary Particle Swarm Optimization. *International Journal of Electrical and Computer Engineering (IJECE)*, 8(6), 5425. <https://doi.org/10.11591/ijece.v8i6.pp5425-5431>

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048–3061. <https://doi.org/10.1109/TKDE.2019.2962680>

Muawanah, S., Muzayanah, U., Pandin, M. G. R., Alam, M. D. S., & Trisnaningtyas, J. P. N. (2023). Stress and Coping Strategies of Madrasah's Teachers on Applying Distance Learning During COVID-19 Pandemic in Indonesia. *Qubahan Academic Journal*, 3(4), 206–218. <https://doi.org/10.48161/Issn.2709-8206>

Nabilah, T., & F, M. A. T. (2023). ANALISIS IMPLEMENTASI PERJANJIAN PEMBIAYAAN DENGAN JAMINAN FIDUSIA PADA PT. ASTRA CREDIT COMPANIES SURABAYA. *Bureaucracy Journal : Indonesia Journal of Law and Social-Political Governance*, 3(2), 1961–1991. <https://doi.org/10.53363/bureau.v3i2.301>

Natasha, A., Prastyo, D. D., & Suhartono. (2019). Credit scoring to classify consumer loan using machine learning. *AIP Conference Proceedings*, 2194(1), 20070. <https://doi.org/10.1063/1.5139802>

Oetama, R. S. (2015). Enhancing Decision Tree Performance in Credit Risk Classification and Prediction. *Ultimatics : Jurnal Teknik Informatika*, 7(1), 51–53. <https://doi.org/10.31937/ti.v7i1.349>

Pławiak, P., Abdar, M., & Rajendra Acharya, U. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing*, 84, 105740. <https://doi.org/https://doi.org/10.1016/j.asoc.2019.105740>

Rachmaniyah, F. (2015). Model Credit Scoring sebagai Alat Bantu Analisis Risiko Pembiayaan Pada BMT UGT Sidogiri. *Ekonomi Bisnis*, 20(1), 77–87.

Sumartini, S. H. (2015). Penggunaan Metode Classification and Regression Trees (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya. *Jurnal Sains Dan Seni ITS*, 4(2), 211–216. <https://journal.universitاسbumigora.ac.id/index.php/matrik/article/view/676/479>

Yi, M. (2023). Application of Classification Regression Tree Algorithm in Accounting Information System. 2023 International Conference on Data Science and Network Security (ICDSNS), 1–6. <https://doi.org/10.1109/ICDSNS58469.2023.10245414>

Yunindya, R., Kudus, A., & Yanti, T. S. (2011). Model Credit Scoring Menggunakan Metode Classification and Regression Trees (CART) pada Data Kartu Kredit. 9–10.