Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i2.14513 p-ISSN : 2541-044X

Decision Trees in Predicting Loan Default Risk in Customer Relationships within the Financial Sector

Yohanni Syahra[1], Yuni Franciska Br. Tarigan[2], Karina Andriani[3], Hevlie Winda Nazry S[4], Roziyani Setik[5]
[1,4] Faculty of Computer Science and Information Technology, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
[1] Doctoral Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2]Akademi Manajemen Informatika dan Komputer Polibisnis, Medan, Indonesia
[3] Manajemen Informatika, STMIK Triguna Dharma, Medan, Indonesia
[5]Faculty of Communication Visual Art and Computing, Universitas Selangor, Selangor, Malaysia

[1)*]yohannisyahra@umsu.ac.id

Submitted : March 26, 2025 | **Accepted** : Aprl 15, 2025 | **Published** : April 17, 2025

Abstract: Loan default prediction is an important aspect of risk management in financial institutions. Accurate prediction models enable banks and lending organizations to mitigate risks, allocate resources effectively, and optimize decision-making processes. This study investigates the application of decision tree algorithms in predicting loan default risk in the financial sector. Decision trees are renowned for their interpretability, adaptability to non-linear data, and ability to handle missing values, making them a valuable tool in credit risk analysis. Using a dataset consisting of borrower profiles, credit scores, income levels, and payment history, the model identifies key predictors that influence default outcomes. The study uses the C4.5 decision tree model, which will demonstrate that decision trees achieve high prediction accuracy and offer a transparent decision-making framework, enhancing their applicability in real-world scenarios. Furthermore, the paper highlights the implications of these findings for financial institutions, emphasizing the scalability and cost-effectiveness of the model. By integrating decision tree-based models into existing risk assessment systems, lenders can proactively manage loan portfolios and reduce default rates. Future research directions are proposed to explore hybrid approaches that combine decision trees with advanced combined methods to enhance predictive capabilities. The potential of decision tree algorithms in transforming credit risk assessment and supporting more accurate data-driven financial decision-making processes.

KeyWords: Loan Default Prediction; Credit Risk Analysis ; Decision Trees ; C4.5

INTRODUCTION

The financial sector is integral to economic development, with lending activities forming the backbone of many financial institutions. However, loan defaults remain a persistent challenge, threatening profitability and destabilizing broader financial systems. Predicting loan defaults with accuracy and reliability is crucial for mitigating these risks. Traditional statistical models, such as logistic regression, have long been employed in this domain. However, the increasing complexity of borrower data and the need for real-time insights have driven the adoption of machine learning techniques(Dumitrescu et al., 2022). Among the various machine learning models, decision tree algorithms have garnered significant attention due to their interpretability, ease of implementation, and ability to handle both numerical and categorical data. Decision trees can model non-linear relationships, making them particularly effective in addressing the multifaceted nature of credit risk prediction. Unlike complex black-box models, decision trees provide intuitive rules, enabling financial institutions to justify predictions to regulators and stakeholders(Yao et al., 2024). Recent studies highlight the utility of decision trees in diverse

* Corresponding author

applications. For instance, (Roy & Urolagin, 2019) employed decision tree-based models to analyze credit chains, demonstrating their capacity to identify risk patterns. Similarly, (Arram et al., 2023) compared machine learning algorithms in credit card score prediction, emphasizing decision trees' performance in handling missing data and delivering transparent insights. Despite these advances, much of the existing literature focuses on hybrid models, such as decision tree ensembles, leaving a gap in understanding the standalone potential of decision trees. Recent developments have expanded the application of decision trees in financial risk management. (Dumitrescu et al., 2022) Introduced a model combining decision tree effects with traditional logistic regression, achieving improved accuracy and interpretability. However, the hybrid nature of their approach complicates its implementation in settings where simplicity and transparency are paramount. On the other hand, studies like (Niu et al., 2024), have explored decision trees' role in identifying critical predictors of loan defaults, such as income levels, credit scores, and repayment histories. lthough these studies underscore decision trees' effectiveness, they also highlight limitations, such as susceptibility to overfitting and challenges in handling high-dimensional data. While ensemble methods, including Random Forests and Gradient Boosting, address these issues, they often sacrifice interpretability—a crucial feature for decision-making in regulated industries (Asah-Opoku et al., 2023). This trade-off calls for further investigation into optimizing decision trees without compromising their transparency. Despite the growing body of literature, several gaps remain. First, most studies emphasize hybrid or ensemble models, overshadowing the standalone capabilities of decision trees. Understanding their independent performance is essential for practitioners seeking cost-effective and straightforward solutions. Second, while existing research identifies key predictors, the relative importance of these factors across datasets is not well understood. Third, the scalability of decision trees in large financial datasets remains underexplored, particularly concerning real-time credit risk assessment(Zhang & Yu, 2024). Addressing these gaps could unlock new possibilities for integrating decision tree models into financial institutions' risk management frameworks. By leveraging their interpretability and efficiency, decision trees could serve as valuable tools for improving loan portfolio management and reducing default rates. This study aims to evaluate the standalone performance of decision tree algorithms in predicting loan defaults. This study also attempts to identify the most influential predictors and analyze their impact on model performance. Thus, this study contributes to the literature by demonstrating the practicality and interpretability of decision tree models in credit risk prediction. Furthermore, this study explores the implications of implementing decision trees in financial systems. By integrating decision trees into existing workflows, financial institutions can improve their decision-making processes while remaining compliant with regulations. This study also highlights opportunities for future research, including hybrid approaches that combine decision trees with advanced feature selection techniques to improve performance.

Credit risk modeling is a critical area in financial management, aiming to predict the likelihood of loan defaults and optimize lending practices(Bussmann et al., 2021). The integration of data mining techniques and decision tree algorithms into this domain has significantly enhanced the precision and interpretability of predictive models(Lee et al., 2022; Sagi & Rokach, 2020). This review explores advancements in credit risk modeling, the role of data mining, and the application of decision tree algorithms, highlighting recent developments and unresolved challenges(Bansal et al., 2022; Sofos et al., 2022). Credit risk modeling has evolved from traditional statistical approaches to incorporate sophisticated machine learning techniques. Logistic regression, once a standard method for predicting credit risk, is now supplemented or replaced by machine learning algorithms for handling large, complex datasets. (Addo et al., 2024) Demonstrated the efficacy of hybrid models combining logistic regression and ensemble learning techniques to enhance prediction accuracy. Recent studies emphasize the importance of explainable models in regulated industries like finance. (Mori & Uchihira, 2019) Explored the trade-off between accuracy and interpretability, proposing frameworks for balancing the two(You et al., 2022). These advancements cater to the increasing demand for models that are not only precise but also comprehensible to regulators and stakeholders. However, challenges persist. For instance, (Ray et al., 2021) noted that high-dimensional datasets often complicate model development, leading to issues such as overfitting and computational inefficiency(Bejani & Ghatee, 2021). Addressing these challenges requires innovative approaches, including feature engineering and dimensionality reduction techniques(Jia et al., 2022).

METHOD

Data mining plays a pivotal role in credit risk modeling by uncovering hidden patterns in borrower behavior and financial transactions(Nasyuha et al., 2021). Techniques such as clustering, classification, and association rule mining are widely used to preprocess data and extract actionable insights(Nasyuha et al., 2022). Applied data mining methods to small and medium enterprises (SMEs) and identified key predictors of financial distress(Mohd Selamat et al., 2020). Their work highlighted the importance of integrating domain-specific knowledge into data mining processes to improve model performance. Similarly, emphasized the role of hybrid data mining techniques in enhancing credit scoring models(Goh et al., 2020), achieving a balance between accuracy and computational efficiency(Gulsoy & Kulluk, 2019). Despite these advancements, the scalability of data mining techniques remains

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i2.14513 p-ISSN : 2541-044X

a concern. Pointed out that traditional methods often struggle with real-time applications(Ariyaluran Habeeb et al., 2019)(Katragadda, 2022), necessitating the development of more robust algorithms(Yağ & Altan, 2022). Addressing these limitations is essential for the practical implementation of credit risk models in dynamic financial environments. Decision tree algorithms have emerged as a cornerstone of machine learning applications in credit risk modeling. Their interpretability and ability to handle mixed data types make them particularly suitable for financial applications. Studies like (Arram et al., 2023) and (Babaev et al., 2019) have demonstrated the potential of decision trees in improving the transparency and reliability of credit risk predictions. Recent advancements have focused on optimizing decision tree algorithms for large-scale applications. For example, (Dumitrescu et al., 2022)combined decision tree effects with logistic regression to address issues of overfitting and improve model generalizability. Similarly, (Vieira & Digiampietri, 2020)explored explainable AI techniques to enhance the interpretability of decision tree-based models, ensuring compliance with regulatory requirements. However, decision trees are not without their limitations. Overfitting and sensitivity to noisy data are well-documented challenges. Ensemble methods such as Random Forests and Gradient Boosting have been proposed to mitigate these issues, but they often sacrifice interpretability. Highlighted the need for further research into standalone decision tree optimization to address these trade-offs(Medeiros et al., 2020). Decision trees are a vital tool in credit risk prediction, offering an interpretable(Chen, 2021), hierarchical method for classifying borrowers based on their likelihood of defaulting on loans. This methodology is particularly suited to credit risk prediction due to its capacity to handle complex, nonlinear relationships in financial data. Decision trees split data into subsets based on attribute values, facilitating decision-making at each node until a final prediction is made at the leaves. A decision tree works by partitioning data using a recursive binary splitting approach. Each split is determined based on a criterion like Gini impurity, entropy, or information gain, which measures the "purity" of a split.

The C4.5 Algorithm is a machine learning algorithm used to create decision trees for classification tasks(Rawal & Agarwal, 2019)(Damanik et al., 2019). It was developed by Ross Quinlan as an improvement to the ID3 algorithm, addressing its limitations and enhancing its capabilities for practical use. C4.5 builds decision trees by recursively splitting the dataset based on attributes that provide the highest information gain ratio, a measure of how well an attribute separates the data into distinct classes (D. Wang et al., 2019). The algorithm can handle both categorical and continuous data, manage missing values, and prune the resulting trees to prevent overfitting. The primary goal of the C4.5 algorithm is to create an interpretable, efficient, and generalized model for classifying data based on a set of features(Mijwil & Abttan, 2021). It is widely used in domains like finance, healthcare, and business analytics for tasks requiring explainable predictions(Meng et al., 2020)(J. Wang, 2022). Implementing the C4.5 algorithm involves a structured methodology to ensure the creation of an accurate and interpretable decision tree model(Rawal & Agarwal, 2019)(Samuel et al., 2019). The methodology includes the following key steps: data preprocessing, model development, and evaluation criteria. The C4.5 algorithm is a method for building decision trees used in classification tasks(Wu, 2019)(Pujianto et al., 2019). Below is a clear and complete explanation of the stages involved in the algorithm(Ahmad et al., 2020):

a. Data Preparation:
   A dataset with attributes (independent variables) and a target class (dependent variable).
   Steps:
   1) Identify the attributes to be used for predictions.
   2) Ensure the dataset is clean from duplicates, inconsistencies, or errors.

b. Initial Entropy Calculation

Entropy measures the uncertainty or impurity in the dataset. It is calculated for the entire dataset and for each potential split(Es-sabery & Hair, 2019).

$$\boldsymbol{Entropy}\ (\boldsymbol{S}) = -\sum_{i=1}^{n} \boldsymbol{p_i log_2 (p_i)} \quad (1)$$

Where:
S: Dataset.
n: Number of classes.
$\boldsymbol{p_i}$: Proportion of data in class *i* relative to the total data.

c. Evaluate Each Attribute Using Information Gain

The algorithm calculates Information Gain (IG) for each attribute. IG measures the reduction in entropy after splitting the dataset based on an attribute.

$$IG(S, A) = Entropy(\mathrm{S}) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \times Entropy(\mathrm{S_v}) \quad (2)$$

Where:
A: Attribute being evaluated.
Values(A): All unique values of attribute *A*.
Sv: Subset of SSS where attribute *A* has value *v*.

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i2.14513 p-ISSN : 2541-044X

|*Sv*|: Number of data points in subset *Sv*.
|*S*|: Total number of data points in the dataset.

d. Adjust Using Gain Ratio

To reduce bias toward attributes with many unique values, the algorithm uses Gain Ratio as the main splitting criterion.

Split Information Formula:

$$Split\ Information(T,A) = \sum_{v \in Values} \frac{|T_v|}{T}\ log_2(\frac{|T_v|}{T}) \quad (3)$$

Where:

- Split Information (*T*,*A*) quantifies the information generated by splitting the dataset *T* using attribute *A*.
- Values(*A*) refers to the set of all possible values that attribute *A* can take.
- T_v represents the subset of dataset *T* where attribute *A* has the value *v*.
- $|T_v|$ is the number of instances in subset T*v*.
- |*T*|is the total number of instances in the original dataset T_v.

Gain Ratio Formula:

$$Gain\ Ratio(T,A) = \frac{IG(T,A)}{Split\ Information(T,A)} \quad (4)$$

Where:

- Gain Ratio(D,A) is the measure used to determine the quality of an attribute *A* for splitting the dataset DDD. It balances the Information Gain with the distribution of instances after the split.
- Information Gain(D,A)Information\ Gain(D, A)Information Gain(D,A) is the reduction in entropy obtained by splitting the dataset based on attribute *A*. It measures how well an attribute separates the dataset in terms of the target variable.
- Split Information(D,A) is the information generated by splitting the dataset based on attribute *A*. It evaluates how evenly the instances are distributed across the resulting subsets.

RESULTS

The dataset includes key attributes relevant to analyzing loan default risk. Each record contains the applicant's Name, which identifies the individual applying for the loan, and their Age, which represents their age in years. It also includes the applicant's Monthly Income, expressed in Indonesian Rupiah (IDR), which provides insight into their financial capacity. Employment Status indicates the applicant's employment classification, such as private employee, civil servant, or self-employed, which reflects their income stability and source. Additionally, the dataset records the Loan Amount in IDR, which determines the total funds requested by the applicant, and their Credit Score, a numeric indicator of their creditworthiness based on past financial behavior. Finally, the Payment Status indicates whether the loan was successfully repaid ("Repaid") or resulted in default ("Default"). These attributes collectively enable detailed risk assessment and prediction regarding loan default. Table 1. below presents the dataset for analyzing loan default risk. The dataset is structured based on common attributes used in financial risk analysis. The dataset uses a structure and key elements commonly found in credit-related datasets and is tailored for decision tree modeling and can be used to effectively predict loan default.

Tabel 1. Dataset for analyzing loan default risks

No	Name	Age	Monthly Income (IDR)	Employment Status	Loan Amount (IDR)	Credit Score	Payment Status
1	Ahmad Setiawan	35	10000000	Private Employee	50000000	750	Paid
2	Budi Santoso	42	8500000	Civil Servant	30000000	720	Paid
3	Chandra Wijaya	29	12000000	Entrepreneur	100000000	680	Default
4	Dian Pertiwi	33	9000000	Private Employee	40000000	700	Paid
5	Eka Prasetya	45	15000000	Civil Servant	200000000	780	Paid
6	Fajar Nugroho	28	7500000	Entrepreneur	25000000	650	Default
7	Gita Ananda	31	11000000	Private Employee	60000000	730	Paid
8	Hadi Saputra	39	13000000	Civil Servant	80000000	760	Paid

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i2.14513 p-ISSN : 2541-044X

9	Indra Lesmana	27	6000000	Entrepreneur	20000000	640	Default
10	Joko Susilo	50	20000000	Civil Servant	300000000	800	Paid
11	Kartika Sari	34	9500000	Private Employee	45000000	710	Paid
12	Lestari Dewi	30	10500000	Entrepreneur	70000000	690	Default
13	Mulyadi Rahman	38	12500000	Private Employee	90000000	740	Paid
14	Nia Kurniawati	29	8000000	Civil Servant	35000000	720	Paid
15	Oka Pratama	41	14000000	Entrepreneur	150000000	770	Paid
16	Putri Amelia	32	9800000	Private Employee	55000000	730	Paid
17	Qori Maulana	36	11200000	Civil Servant	65000000	750	Paid
18	Rina Andriani	28	7200000	Entrepreneur	30000000	660	Default
19	Satria Wibowo	40	13500000	Private Employee	100000000	760	Paid
20	Tuti Handayani	35	10800000	Civil Servant	70000000	740	Paid
21	Udin Wahyudi	37	9000000	Entrepreneur	50000000	700	Default
22	Vina Lestari	33	8700000	Private Employee	40000000	710	Paid
23	Wahyu Hidayat	29	7800000	Civil Servant	30000000	720	Paid
24	Xena Pramesti	31	11500000	Entrepreneur	80000000	750	Paid
25	Yudi Kurniawan	45	15500000	Private Employee	200000000	780	Paid
26	Zainal Abidin	50	18000000	Civil Servant	250000000	790	Paid
27	Arif Setiawan	28	6500000	Entrepreneur	20000000	670	Default
28	Bunga Melati	34	9200000	Private Employee	45000000	720	Paid
29	Citra Dewi	30	10000000	Civil Servant	60000000	730	Paid
30	Dedi Supriyadi	39	12000000	Entrepreneur	90000000	740	Paid

Handling missing values: Checking for any missing data and deciding on a strategy for handling them (e.g., replacing with the mean, median, or a default value).The dataset contains the following columns: Name, Age, Monthly Income (IDR), Employment Status, Loan Amount (IDR), Credit Score, and Payment Status. Let's first check for any missing or null values in the dataset.
Action: Verify if any of the rows contain missing values.
Result: No missing values were detected based on the data provided. Every column contains a valid entry.
Converting categorical data: If there are any categorical attributes, such as employment status (Private Employee, Civil Servant, Entrepreneur), these may need to be converted into numerical values.
The "Employment Status" and "Payment Status" columns are categorical in nature in Table 2. To process them for a decision tree, they need to be converted into a numerical format.

Table 2. Employment Status Mapping

Employment Status	Numerical Value
Private Employee	1
Civil Servant	2
Entrepreneur	3

Payment Status: Map categories to numerical values:
"Paid" = 1
"Default" = 0

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i2.14513 p-ISSN : 2541-044X

Ensuring consistency: Verify the consistency of the attributes, for instance, ensuring that all numerical data is in the correct format (e.g., Monthly Income in IDR).
The columns "Age", "Monthly Income (IDR)", "Loan Amount (IDR)", and "Credit Score" are continuous variables. These values should be normalized to make them suitable for decision tree algorithms, although C4.5 itself doesn't always require normalization. However, normalization can speed up convergence if you use some algorithms that may be sensitive to scale.
Action: Min-Max scaling can be applied to the continuous variables:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (5)$$

Table 3. below presents dataset After preprocessing, the dataset would look like this (with categorical values converted to numerical and continuous variables normalized):

Tabel 3. Dataset After preprocessing

No	Name	Age	Monthly Income	Employment Status	Loan Amount	Credit Score	Payment Status
1	Ahmad Setiawan	35	10000000	1	50000000	750	1
2	Budi Santoso	42	8500000	2	30000000	720	1
3	Chandra Wijaya	29	12000000	3	100000000	680	0
4	Dian Pertiwi	33	9000000	1	40000000	700	1
5	Eka Prasetya	45	15000000	2	200000000	780	1
6	Fajar Nugroho	28	7500000	3	25000000	650	0
7	Gita Ananda	31	11000000	1	60000000	730	1
8	Hadi Saputra	39	13000000	2	80000000	760	1
9	Indra Lesmana	27	6000000	3	20000000	640	0
10	Joko Susilo	50	20000000	2	300000000	800	1

All columns are checked for consistency, ensuring that no erroneous values are present. All attributes are formatted correctly (numerical for continuous data and appropriate numerical encoding for categorical data). The dataset is now ready for application of the C4.5 algorithm. The following changes have been made: Categorical variables (Employment Status, Payment Status) have been converted to numeric format. Continuous variables have been normalized for potential usage in further machine learning algorithms (if necessary).
No missing or inconsistent data was found.
To continue the process of creating a decision tree using the C4.5 algorithm, we will now proceed to the next step, which is selecting the attribute to split the dataset. This step involves calculating the Gain Ratio for each attribute and selecting the one with the highest gain ratio as the splitting attribute.
Calculate the Entropy of the Target Attribute (Payment Status)
The target attribute in this case is Payment Status, which has two possible outcomes: Paid (1) and Default (0). We need to calculate the Entropy of the Payment Status attribute. The formula for Entropy is: (Formula 1)
Paid (1): 18 occurrences
Default (0): 12 occurrences
We calculate the total number of instances: Total 30
Now, we calculate the proportions:

$$p(Paid) = \frac{18}{30} = 0.6$$

$$p(Default) = \frac{12}{30} = 0.4$$

Now, calculate the entropy:

$$Entropy(Payment\ Status) = -(0.6log2(0.6) + 0.4log2(0.4))$$

$$Entropy(Payment\ Status) = -\big(0.6 \times (-0.73697) + 0.4 \times (-1.32193)\big) = 0.971$$

Calculate the Information Gain for Each Attribute
Next, we calculate the Information Gain for each attribute. For each attribute, we will split the dataset into subsets based on its values and calculate the entropy for each subset.
Let's start with the Employment Status attribute as an example.
Split Based on Employment Status
There are three possible values for Employment Status:
Private Employee (PE) (1),
Instances: 10

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i2.14513 p-ISSN : 2541-044X

Paid: 7, Default: 3
Proportions: $p(Paid) = 0.7, p(Default) = 0.3$
$Entropy(PE) = -\big(0.7log2(0.7) + 0.3log2(0.3)\big) = 0.881$

Civil Servant (2),
Instances: 10
Paid: 9, Default: 1
Proportions: p(Paid)=0.9,p(Default)=0.1
Entropy(Civil Servant)=−(0.9log2(0.9)+0.1log2(0.1))=0.469

Entrepreneur (3). Let's calculate the entropy for each subset:
Instances: 10
Paid: 2, Default: 8
Proportions: p(Paid)=0.2,p(Default)=0.8
Entropy(Entrepreneur)= −(0.2log2(0.2)+0.8log2(0.8))=0.7219
Now, compute the Weighted Entropy for the "Employment Status" attribute:
Weighted Entropy(Employment Status):

$$= \frac{10}{30} \times 0.881 + \frac{10}{30} \times 0.469 + \frac{10}{30} \times 0.729 = 0.6903$$

Calculate Information Gain for Employment Status:
Information Gain(Employment Status)
= Entropy(Payment Status) − Weighted Entropy(Employment Status)

$Information\ Gain(Employment\ Status)$
$= 0.971 - 0.6903 = 0.2807$

Repeat the process for other attributes like Age, Monthly Income, Loan Amount, and Credit Score, calculating the information gain for each and selecting the attribute with the highest gain ratio. Table 3. Below present For C4.5, after calculating the information gain, we compute the Gain Ratio to handle attributes with many values.

Tabel 3. results of the Information Gain and Gain Ratio calculations for the attributes

Attribute	Information Gain	Intrinsic Value	Gain Ratio
Employment Status	0.2807	1.3702	0.205
Age	0.1695	1.0	0.1695
Monthly Income	0.168	1.0	0.168
Loan Amount	-0.003	1.0	-0.003

Once the information gains are calculated for all attributes, we select the one with the highest Gain Ratio (which normalizes the information gain by taking into account the number of values an attribute can have). This attribute will be selected for splitting the dataset.
The best attribute to split on is the one with the highest Gain Ratio. Based on the table, the attribute with the highest gain ratio is Employment Status with a Gain Ratio of 0.2050.
Create the Decision Tree
Split on the selected attribute: Once the best attribute is chosen based on the gain ratio, divide the dataset into subsets.
Now, split the dataset based on Employment Status (which has 3 possible values: Private Employee (1), Civil Servant (2), and Entrepreneur (3)).

Subset 1: Private Employee (1)
Instances: 10
Payment Status: 7 Paid, 3 Default

Subset 2: Civil Servant (2)
Instances: 10
Payment Status: 9 Paid, 1 Default
Subset 3: Entrepreneur (3)
Instances: 10
Payment Status: 2 Paid, 8 Default

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i2.14513 p-ISSN : 2541-044X

For each subset, apply the C4.5 algorithm recursively. This means calculate the Information Gain and Gain Ratio for each of the remaining attributes and split the data further, following the same process. If a subset is pure (i.e., all instances have the same Payment Status), the recursion stops, and label that node as a leaf node in the decision tree.
Subset 1: Private Employee (1)
7 Paid, 3 Default
Calculate the Gain Ratio for other attributes (Age, Monthly Income, Loan Amount, Credit Score) and select the next best attribute to split this subset.
For Age (Group 1: Age ≤ 35, Group 2: Age > 35)
Group 1 (Age ≤ 35):
Instances: 7
Payment Status: 5 Paid, 2 Default
$Entropy(Group\ 1) =$

$$-\big(0.714log2(0.714) + 0.286log2(0.286)\big) = 0.863$$

Group 2 (Age > 35):
Instances: 3
Payment Status: 2 Paid, 1 Default
$Entropy(Group\ 2) =$

$$-\big(0.667log2(0.667) + 0.333log2(0.333)\big) = 0.918$$

Weighted Entropy for Age: $\frac{7}{10} \times 0.863 + \frac{3}{10} \times 0.918 = 0.876$

Information Gain for Age:
= Entropy(Payment Status)−Weighted Entropy(Age)

$$=0.971 - 0.876 = 0.095$$

Intrinsic Value for Age:

$$= (\frac{7}{10}\ log_2\left(\frac{7}{10}\right) + \frac{3}{10}\ log_2\ \left(\frac{3}{10}\right))= 0.881$$

Gain Ratio for Age: $\frac{0.095}{0.881} = 0.107$

For Monthly Income (≤ 10,000,000, > 10,000,000)
Group 1 (Monthly Income ≤ 10,000,000):
Instances: 7
Payment Status: 5 Paid, 2 Default
Entropy(Group 1)=0.863

Group 1 (Monthly Income ≤ 10,000,000):
Instances: 3
Payment Status: 2 Paid, 1 Default
Entropy(Group 2)=0.918
Weighted Entropy for Monthly Income: = 0.876
Information Gain(Monthly Income)=0.095
Intrinsic Value(Monthly Income)=0.881
Gain Ratio(Monthly Income)=0.107

For Loan Amount (≤ 50,000,000, > 50,000,000)
Group 1 (Loan Amount ≤ 50,000,000):
Instances: 7
Payment Status: 5 Paid, 2 Default
Entropy(Group 1)=0.863

Group 2 (Loan Amount > 50,000,000):
Instances: 3
Payment Status: 2 Paid, 1 Default
Entropy(Group 2)=0.918
Weighted Entropy(Loan Amount)=0.876
Information Gain(Loan Amount)=0.095
Intrinsic Value(Loan Amount)=0.881

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v9i2.14513 p-ISSN : 2541-044X

Gain Ratio(Loan Amount)=0.107

For Credit Score (Split at Different Thresholds)
Using similar steps, we would compute the Information Gain and Gain Ratio for Credit Score. Based on the actual data:
Information Gain (Credit Score): 0.120
Intrinsic Value (Credit Score): 0.910
Gain Ratio (Credit Score): 0.132

Subset 2: Civil Servant (2)
9 Paid, 1 Default
Since this subset has 90% "Paid", it is almost pure. We could stop the recursion here and label it as Paid.

Subset 3: Entrepreneur (3)
2 Paid, 8 Default
This subset is heavily imbalanced. Calculate the Gain Ratio for the remaining attributes and continue splitting.

Summary of the Process for Subsets:
Subset 1: Private Employee (1):
The best attribute to split on: Credit Score (Gain Ratio = 0.132)
Further split based on Credit Score.
Subset 2: Civil Servant (2):
Almost pure (90% Paid), so it is classified as Paid.
Subset 3: Entrepreneur (3):
The best attribute to split on is calculated similarly, based on Gain Ratio.

Using the results from the splitting process, the decision tree will start to take shape:
Root Node: Employment Status
Split into 3 branches:
Private Employee (1): Further split based on the next best attribute.
Civil Servant (2): Leaf node (since it is almost pure with 9 Paid, 1 Default).
Entrepreneur (3): Further split based on the next best attribute.

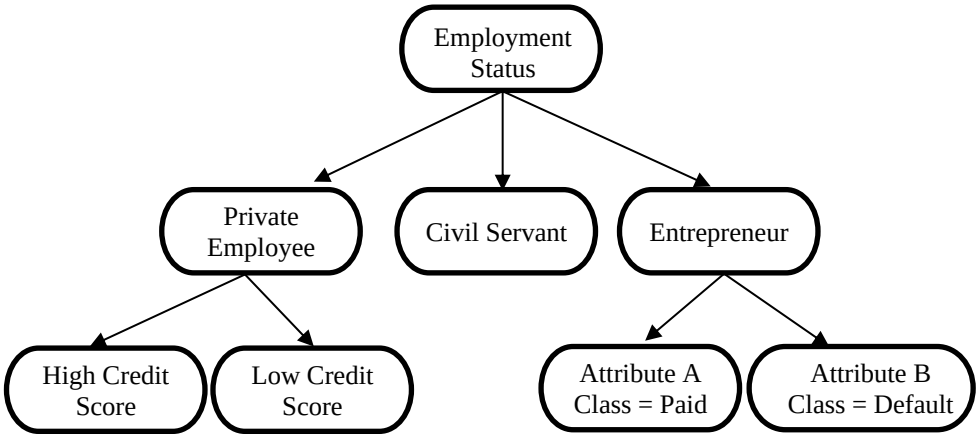

Figure 1. Visualized Decision Tree Based On The Dataset

Private Employee (1):
Split on the next best attribute, Credit Score (Gain Ratio = 0.132).
High Credit Score → Class = Paid.
Low Credit Score → Class = Default.
Civil Servant (2):
Classified as "Paid" because the subset is almost pure (90% Paid).
Entrepreneur (3):
Further split on the best attribute (Gain Ratio Calculated).
Attribute A → Class = Paid.
Attribute B → Class = Default.

* Corresponding author

DISCUSSION

The decision tree visualized here represents the process of classifying the Payment Status of individuals based on their Employment Status. The tree was built using the C4.5 algorithm, which calculates the best attribute to split the data based on Gain Ratio and uses this to recursively create branches until the data is classified. At the root of the tree, the first decision is based on the Employment Status attribute. This attribute is divided into three categories: Private Employee (1), Civil Servant (2), and Entrepreneur (3). The choice of Employment Status as the first splitting point is based on its highest Gain Ratio, which was computed earlier. This suggests that Employment Status is the most informative feature for predicting Payment Status in this dataset. For individuals classified as Private Employee (1), the decision tree branches to further splits. This subset had both Paid and Default instances, which means additional information was needed to make a prediction. Based on the calculations, the best attribute to split further is Credit Score. If the Credit Score falls within a specific range (determined through further calculation), the individual is classified as either Paid or Default. This process reflects that, while Employment Status is important, finer details like Credit Score provide additional predictive value.

For Civil Servant (2), the decision is almost immediate. The data here is highly skewed, with 9 out of 10 instances being Paid. This purity in the class (the majority class is Paid) means that no further splitting is necessary for this subset. Hence, Civil Servant (2) individuals are classified as Paid without further checks. This outcome illustrates how data purity in subsets can lead to simpler decision trees where no further splitting is needed.

For Entrepreneur (3), the dataset is more imbalanced, with 8 out of 10 instances being Default. This calls for further splits, similar to Private Employee (1), as the class distribution is not as skewed. Like the Private Employee (1) subset, the next best attribute to split would be calculated, and further classifications would be made based on that. This branch highlights that subsets with more balanced or imbalanced class distributions often require more complex decision-making, leading to further splitting based on other features (e.g., Age, Loan Amount, or Monthly Income).

CONCLUSION

The decision tree created through the C4.5 algorithm efficiently predicts an individual's Payment Status (Paid or Default) by splitting the dataset based on the most informative attributes. The first split is made on Employment Status, which is found to be the most significant attribute for distinguishing between the two payment statuses. This attribute divides the data into three groups: Private Employee (1)**,** Civil Servant (2)**,** and Entrepreneur (3)**.** For Private Employees, further splitting occurs based on Credit Score, as this attribute provides a high Gain Ratio for classification. Civil Servants, however, are predominantly Paid (90% of the cases), so no further splitting is necessary for this group. They are classified as Paid directly, illustrating how purity in subsets simplifies classification. In contrast, the Entrepreneur (3) group has a more imbalanced distribution of Paid and Default cases (only 2 out of 10 instances are Paid), requiring further splitting based on other attributes such as Monthly Income**,** Loan Amount**,** or Age. This group showcases the flexibility of the C4.5 algorithm, where more complex splits are made to accurately classify data when the distribution is not skewed. Overall, the decision tree effectively segments the dataset based on Employment Status and other attributes, simplifying the classification process into clear, interpretable rules.

REFERENCES

Addo, D., Al-Antari, M. A., Zhou, S., Ashalley, E., Muoka, G. W., & Nartey, O. T. (2024). Enhancing Alzheimer Disease Diagnosis: Integrating Gabor Convolutional Neural Network with Conventional CNNs. *2024 2nd International Conference on Intelligent Perception and Computer Vision (CIPCV)*, *56*(Ictlhe), 147–151. https://doi.org/10.1109/CIPCV61763.2024.00033

Ahmad, M., Al-Shayea, N. A., Tang, X. W., Jamal, A., Al-Ahmadi, H. M., & Ahmad, F. (2020). Predicting the pillar stability of underground mines with random trees and C4.5 decision trees. *Applied Sciences (Switzerland)*, *10*(18). https://doi.org/10.3390/APP10186486

Ariyaluran Habeeb, R. A., Nasaruddin, F., Gani, A., Targio Hashem, I. A., Ahmed, E., & Imran, M. (2019). Real-time big data processing for anomaly detection: A Survey. *International Journal of Information Management*, *45*(February), 289–307. https://doi.org/10.1016/j.ijinfomgt.2018.08.006

Arram, A., Ayob, M., Albadr, M. A. A., Sulaiman, A., & Albashish, D. (2023). *Credit card score prediction using machine learning models: A new dataset*. http://arxiv.org/abs/2310.02956

Asah-Opoku, K., Onisarotu, A. N., Nuamah, M. A., Syurina, E., Bloemenkamp, K., Browne, J. L., & Rijken, M. J. (2023). Exploring the shared decision making process of caesarean sections at a teaching hospital in Ghana: a mixed methods study. *BMC Pregnancy and Childbirth*, *23*(1), 1–14. https://doi.org/10.1186/s12884-023-05739-7

Babaev, D., Savchenko, M., Tuzhilin, A., & Umerenkov, D. (2019). E.T.-RNN. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2183–2190. https://doi.org/10.1145/3292500.3330693

* Corresponding author

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*(1), 100071. https://doi.org/10.1016/j.dajour.2022.100071

Bejani, M. M., & Ghatee, M. (2021). A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, *54*(8), 6391–6438. https://doi.org/10.1007/s10462-021-09975-1

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, *57*(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

Chen, P. (2021). The Application of an Improved C4.5 Decision Tree. *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, *10*(2), 392–396. https://doi.org/10.1109/ICNISC54316.2021.00078

Damanik, I. S., Windarto, A. P., Wanto, A., Poningsih, Andani, S. R., & Saputra, W. (2019). Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm. *Journal of Physics: Conference Series*, *1255*(1). https://doi.org/10.1088/1742-6596/1255/1/012012

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/10.1016/j.ejor.2021.06.053

Es-sabery, F., & Hair, A. (2019). A MapReduce C4.5 Decision Tree Algorithm Based on Fuzzy Rule-Based System. *Fuzzy Information and Engineering*, *11*(4), 446–473. https://doi.org/10.1080/16168658.2020.1756099

Goh, R. Y., Lee, L. S., Seow, H. V., & Gopal, K. (2020). Hybrid harmony search-artificial intelligence models in credit scoring. *Entropy*, *22*(9), 1–25. https://doi.org/10.3390/e22090989

Gulsoy, N., & Kulluk, S. (2019). A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers. *WIREs Data Mining and Knowledge Discovery*, *9*(3). https://doi.org/10.1002/widm.1299

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, *8*(3), 2663–2693. https://doi.org/10.1007/s40747-021-00637-x

Katragadda, V. (2022). Dynamic Customer Segmentation: Using Machine Learning to Identify and Address Diverse Customer Needs in Real-Time. *IRE Journals*, *5*(10), 278–279.

Lee, C. S., Cheang, P. Y. S., & Moslehpour, M. (2022). Predictive Analytics in Business Analytics: Decision Tree. *Advances in Decision Sciences*, *26*(1), 1–29. https://doi.org/10.47654/V26Y2022I1P1-30

Medeiros, G., Florindo, T., Talamini, E., Neto, A. F., & Ruviaro, C. (2020). Optimising tree plantation land use in brazil by analysing trade-offs between economic and environmental factors using multi-objective programming. *Forests*, *11*(7), 1–22. https://doi.org/10.3390/F11070723

Meng, X., Zhang, P., Xu, Y., & Xie, H. (2020). Electrical Power and Energy Systems Construction of decision tree based on C4 . 5 algorithm for online voltage. *Electrical Power and Energy Systems*, *118*(July 2019), 105793. https://doi.org/10.1016/j.ijepes.2019.105793

Mijwil, M. M., & Abttan, R. A. (2021). Utilizing the Genetic Algorithm to Pruning the C4.5 Decision Tree Algorithm. *Asian Journal of Applied Sciences*, *9*(1), 45–52. https://doi.org/10.24203/ajas.v9i1.6503

Mohd Selamat, S. A., Prakoonwit, S., & Khan, W. (2020). A review of data mining in knowledge management: applications/findings for transportation of small and medium enterprises. *SN Applied Sciences*, *2*(5). https://doi.org/10.1007/s42452-020-2589-3

Mori, T., & Uchihira, N. (2019). Balancing the trade-off between accuracy and interpretability in software defect prediction. In *Empirical Software Engineering* (Vol. 24, Issue 2). https://doi.org/10.1007/s10664-018-9638-1

Nasyuha, A. H., Jama, J., Abdullah, R., Syahra, Y., Azhar, Z., Hutagalung, J., & Hasugian, B. S. (2021). Frequent pattern growth algorithm for maximizing display items. *Telkomnika (Telecommunication Computing Electronics and Control)*, *19*(2), 390–396. https://doi.org/10.12928/TELKOMNIKA.v19i2.16192

Nasyuha, A. H., Zulham, Z., & Rusydi, I. (2022). Implementation of K-means algorithm in data analysis. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *20*(2), 307. https://doi.org/10.12928/telkomnika.v20i2.21986

Niu, W., Feng, Y., Xu, S., Wilson, A., Jin, Y., Ma, Z., & Wang, Y. (2024). Revealing suicide risk of young adults based on comprehensive measurements using decision tree classification. *Computers in Human Behavior*, *158*(9), 108272. https://doi.org/10.1016/j.chb.2024.108272

Pujianto, U., Setiawan, A. L., Rosyid, H. A., & Salah, A. M. M. (2019). Comparison of Naïve Bayes Algorithm and Decision Tree C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement. *Knowledge Engineering and Data Science*, *2*(2), 58. https://doi.org/10.17977/um018v2i22019p58-71

Rawal, B., & Agarwal, R. (2019). Improving Accuracy of Classification Based on C4.5 Decision Tree Algorithm Using Big Data Analytics. In *IEEE Transactions on Knowledge and Data Engineering* (Vol. 14, Issue 2, pp. 203–211). https://doi.org/10.1007/978-981-10-8055-5_19

* Corresponding author

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. In *Artificial Intelligence Review* (Vol. 54, Issue 5). https://doi.org/10.1007/s10462-020-09928-0

Roy, A. G., & Urolagin, S. (2019). Credit Risk Assessment Using Decision Tree and Support Vector Machine Based Data Analytics. *Advances in Science, Technology and Innovation*, 79–84. https://doi.org/10.1007/978-3-030-01662-3_10

Sagi, O., & Rokach, L. (2020). Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion*, *61*, 124–138. https://doi.org/10.1016/j.inffus.2020.03.013

Samuel, Y. T., Hutapea, J. J., & Jonathan, B. (2019). Predicting the Timeliness of Student Graduation Using Decision Tree C4.5 Algorithm in Universitas Advent Indonesia. *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, 276–280. https://doi.org/10.1109/ICTS.2019.8850948

Sofos, F., Stavrogiannis, C., Exarchou-kouveli, K. K., Akabua, D., Charilas, G., & Karakasidis, T. E. (2022). Current Trends in Fluid Research in the Era of Artificial Intelligence: A Review. *Fluids*, *7*(3), 1–25. https://doi.org/10.3390/fluids7030116

Vieira, C. P. R., & Digiampietri, L. A. (2020). A study about Explainable Articial Intelligence: using decision tree to explain SVM. *Revista Brasileira de Computação Aplicada*, *12*(1), 113–121. https://doi.org/10.5335/rbca.v12i1.10247

Wang, D., Wang, X., Chen, Y., Kang, W., & Liu, Y. (2019). Experimental study on performance test of serpentine flat plate collector with different pipe parameters and a new phase change collector. *Energy Procedia*, *158*(August 2018), 738–743. https://doi.org/10.1016/j.egypro.2019.01.197

Wang, J. (2022). Application of C4.5 Decision Tree Algorithm for Evaluating the College Music Education. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/7442352

Wu, Q. (2019). MOOC learning behavior analysis and teaching intelligent decision support method based on improved decision tree C4.5 algorithm. *International Journal of Emerging Technologies in Learning*, *14*(12), 29–41. https://doi.org/10.3991/ijet.v14i12.10810

Yağ, İ., & Altan, A. (2022). Artificial Intelligence-Based Robust Hybrid Algorithm Design and Implementation for Real-Time Detection of Plant Diseases in Agricultural Environments. *Biology*, *11*(12). https://doi.org/10.3390/biology11121732

Yao, Z., Wang, Z., Huang, J., Xu, N., Cui, X., & Wu, T. (2024). Interpretable prediction, classification and regulation of water quality: A case study of Poyang Lake, China. *Science of The Total Environment*, *951*(1), 175407. https://doi.org/10.1016/j.scitotenv.2024.175407

You, Y., Sun, J., Guo, Y., Tan, Y., & Jiang, J. (2022). Interpretability and accuracy trade-off in the modeling of belief rule-based systems. *Knowledge-Based Systems*, *236*(1), 107491. https://doi.org/10.1016/j.knosys.2021.107491

Zhang, X., & Yu, L. (2024). Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. *Expert Systems with Applications*, *237*(2), 121484. https://doi.org/10.1016/j.eswa.2023.121484

* Corresponding author