Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

A Decision Tree Model with Grid Search Optimization for Scholarship Recipient Classification

Tati Suprapti[1], Bani Nurhakim*[2], Bintang Warni Ayu Hermina[3], Vrendi Amro Syahputra Simbolon[4]

[1,3,4]Informatics Engineering, STMIK IKMI Cirebon, Indonesia
[2]Informatics Management, STMIK IKMI Cirebon, Indonesia

Email: [1]baninurhakim@gmail.com

Received : Aug 4, 2025; Revised : Aug 17, 2025; Accepted : Aug 19, 2025; Published : Oct 22, 2025

Abstract

This study aims to classify scholarship recipients using the Decision Tree algorithm implemented in RapidMiner. The dataset consists of 1.404 records with socioeconomic and academic attributes. Preprocessing was conducted using two Replace Missing Value operators, where categorical attributes such as No. BANTUAN, No. KKS, and Prestasi were filled with "Tidak Punya," while Kepemilikan Rumah was imputed using the average value. The model was built using a Decision Tree algorithm, optimized with the Optimize Parameters (Grid) operator to determine the best values for maximal depth and confidence. Evaluation was performed using 10-fold Cross Validation to ensure reliability. The results show that the optimized Decision Tree model achieved a high accuracy of 97.72%, with strong precision, recall, and F1-score values in both the "Eligible" and "Not Eligible" classes. These findings demonstrate that the Decision Tree algorithm, when properly optimized and validated, can effectively support decision-making processes in scholarship eligibility classification. The model provides an interpretable and robust tool for educational institutions to evaluate student applications based on critical socioeconomic features, This research contributes to educational data mining by offering a validated and interpretable model that enhances fairness, transparency, and efficiency in the scholarship selection process.

Keywords : *Classification, Cross Validation, Decision Tree, Educational Support, RapidMiner, Scholarship*

1. INTRODUCTION

Access to higher education remains a crucial issue in Indonesia, especially for students from low-income families. To address this, an educational support program was established to provide financial assistance to eligible students based on socioeconomic criteria [1]. However, the selection process for recipients is often hampered by manual assessments, subjective decisions, and a lack of a validation framework [2]. Recent studies emphasize that scholarship programs for disadvantaged students must adopt data-driven approaches to ensure fairness, efficiency, and transparency in resource allocation [3]. Machine learning (ML) offers opportunities to improve decision-making processes in education, particularly in the allocation of aid. Decision Tree and Grid Search Optimization algorithms are widely recognized for their simplicity, ease of interpretation, and effectiveness in classification problems [4]. These algorithms have shown promising results in predicting academic performance in mobile learning environments [5] and have been applied to identify key factors influencing student success [6].

In terms of fairness and bias mitigation, Decision Tree models have also been refined to consider various fairness criteria [7]. Recent evidence also highlights that manual assessment methods are prone to bias, underscoring the necessity of adopting machine learning techniques in educational decision-making [8]. Socioeconomic factors such as parental income, educational background, and the number of household members have been proven important in assessing students' need for assistance [9],[10].

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

Models trained using these variables have demonstrated their ability to accurately predict aid eligibility and the risk of dropping out [11]. Several studies in Indonesia have used Decision Tree for student classification and aid analysis. For example, student performance has been evaluated using an ID3-based model [12], and the eligibility for social assistance funds has been determined using a classification approach [13]. A combination of clustering and tree-based classification has also been used to optimize the aid distribution process [14]. Other research has compared Decision Tree with Naïve Bayes to assess predictive accuracy in student performance classification [15].

Model validation is crucial to ensure its generalization and reliability. Cross-validation, particularly k-fold, has become a standard evaluation technique to prevent overfitting and assess model robustness[16]. Its effectiveness has been demonstrated in various domains, including obesity prediction[17], student admission classification[18], and aid distribution[19]. Furthermore, the integration of cross-validation with hyperparameter tuning has been shown to improve model accuracy and fairness [20], [21]. This study addresses the research gap by integrating a comprehensive preprocessing strategy, systematic hyperparameter tuning using grid search, and robust evaluation with 10-fold cross-validation, which have not been jointly applied in previous scholarship selection studies for economically disadvantaged and high-achieving students [3], [22].

Despite significant advancements, studies addressing the classification of educational aid recipients with a complete preprocessing pipeline and a validated Decision Tree model are still limited. This research aims to close this gap by developing a robust classification model using Decision Tree in RapidMiner. This model incorporates handling of missing values, cross-validation, and grid search optimization to enhance the fairness, accuracy, and objectivity in selecting educational aid recipients[23], [24], [25], [26].

2. METHOD

This section presents the methodology used in building the classification model for educational aid recipients. This research follows a structured workflow that starts with data selection, preprocessing, model construction using the Decision Tree algorithm, parameter tuning, and evaluation using cross-validation. Each stage is designed to improve the accuracy, fairness, and reliability of the classification results.

The overall research workflow is illustrated in Figure 1, which shows the sequential steps performed in this study.

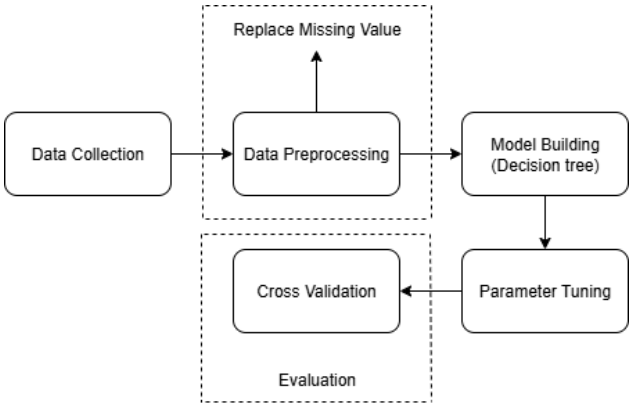


Figure 1. Research workflow for classifying educational aid recipients

2.1. Data collection

Data collection in this study was conducted by utilizing a scholarship recipient dataset for students from economically disadvantaged and high-achieving backgrounds. This dataset was obtained from the

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

official archives of STMIK IKMI Cirebon during the period 2020 to 2023, which includes demographic information, socioeconomic conditions, and students' academic achievement records. The available attributes include identity information, parents' income, ownership of supporting documents such as the No. BANTUAN (Government Aid Identification Number) and Kartu Keluarga Sejahtera (KKS) (Prosperous Family Card), participation status in social assistance programs, home ownership, and academic achievement records.

The selection of attributes was based on their relevance to the scholarship selection process and referred to findings from previous studies, which indicate that socioeconomic factors and academic achievement are important indicators in determining scholarship eligibility [9]. All attributes in the dataset were considered significant for classification; therefore, no variables were removed in the initial stage.

The dataset used consists of 1,404 student data entries with 15 main attributes, including one target attribute indicating scholarship eligibility status ("Lolos" (Eligible) or "Tidak Lolos" (Not Eligible)). Several attributes contained missing values, which were then addressed during the preprocessing stage. This data represents the population of students applying for scholarships during a specific period and is considered sufficiently representative for building a reliable prediction mode [27], [21].

The use of historical data such as this is consistent with educational data mining practices that utilize academic and socioeconomic data to develop decision support systems in the education sector [16]. With complete and relevant data characteristics, the classification process can be carried out more accurately, fairly, and transparently.

2.2. Preprocessing (Replace Missing Value)

The preprocessing stage aims to improve the quality of data before it is used in a classification model. In this study, preprocessing was specifically focused on handling missing data. This decision was made because all attributes in the dataset were considered relevant and important for the classification process. Therefore, no attributes were removed or altered, as the dataset was already well-structured and representative.

An overview of all attributes, their data types, value categories, and the number of missing values is shown in Table 1. The table indicates that some attributes have missing data, particularly in the No. BANTUAN, No. KKS, Kepemilikan Rumah, and Prestasi attributes.

Handling missing data is a crucial step to ensure the reliability and accuracy of a model, especially in supervised learning tasks like classification[16]. In this study, two strategies were used:

1. Categorical attributes such as No. BANTUAN, No. KKS, and Prestasi were imputed with the constant value "Tidak Punya" (Does Not Have). This imputation was chosen because students who do not possess these documents typically leave the fields blank. Therefore, the missing values are interpreted not as data errors or loss, but as a valid indication of the non-possession of the said documents or achievements[1].
2. The Kepemilikan Rumah (Home Ownership) attribute was handled by replacing missing values with the mean of the attribute. This approach ensures that the data distribution is not significantly affected.

This preprocessing step was implemented using the Replace Missing Value operator in RapidMiner, where separate configurations were used for nominal and numerical attributes. Table 1 shows the attributes in the dataset. The decision to only perform missing value handling is supported by the completeness and quality of the dataset, as all other variables were consistent, interpretable, and required for model training [23].

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

Table 1. Dataset Attributes

Attribute Name	Data Type	Values	Missing Values
NIM	Polynominal	Student ID	0
Lolos BANTUAN	Binominal	Class label: “Lolos”, “Tidak Lolos”	0
Status DTKS	Binominal	“Belum Terdata”, “Terdata”	0
Status P3KE	Polynominal	“Belum Terdata”, “Terdata : Desil 1 – Desil 7”	0
No. KIP	Binominal	“Punya”, “Tidak Punya”	1150
No. KKS	Binominal	“Punya”, “Tidak Punya”	1310
Pekerjaan Ayah	Polynominal	“Peg. Swasta”, “Petani”, etc.	0
Penghasilan Ayah	Polynominal	“<250.000”, “1.000.001-2.250.000”, etc.	0
Status Ayah	Polynominal	“Bercerai”, “Hidup”, “Wafat”	0
Pekerjaan Ibu	Polynominal	“Peg. Swasta”, “Petani”, etc.	0
Penghasilan Ibu	Polynominal	“<250.000”, “1.000.001-2.250.000”, etc.	0
Status Ibu	Binominal	“Hidup”, “Wafat”	0
Jumlah Tanggungan	Polynominal	“1”, “2”, ”3”, etc.	0
Kepemilikan Rumah	Polynominal	“Sendiri”, “Menumpang”, etc.	84
Prestasi	Binominal	“Punya”, “Tidak Punya”	997

The preprocessing process was implemented using RapidMiner, focusing solely on handling missing values. As shown in Figure 1, two Replace Missing Values operators were used in sequence: the first filled missing categorical values for No. BANTUAN, No. KKS, and Prestasi with the constant value “Tidak Punya”, while the second handled missing values in Kepemilikan Rumah by replacing them with the average value. This ensures the dataset is complete and consistent before model building.

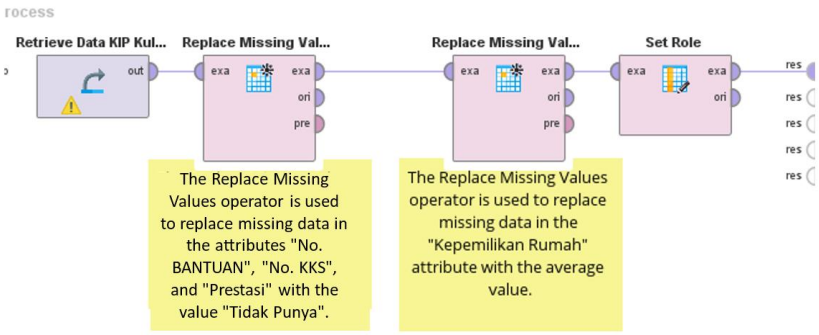


Figure 1. Preprocessing workflow in RapidMiner.

2.3. Model Building (Decision Tree)

The Decision Tree algorithm is selected for this study due to its capability to classify data based on interpretable rule-based structures. In RapidMiner, the model construction was implemented within the Optimize Parameters (Grid) operator, which was embedded inside the Cross Validation operator. As illustrated in Figure 2, this nested design allowed simultaneous parameter tuning and model validation, ensuring the model generalizes well and avoids overfitting.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

The process of optimizing the Decision Tree model parameters is shown in Figure 3, where each combination of parameters was evaluated during training to identify the best configuration.
The parameters tuned during model building include:

1. Maximal Depth: This defines the maximum depth of the tree. A deeper tree may increase accuracy but could also lead to overfitting.
2. Confidence: This parameter controls the pruning of the tree. Lower confidence leads to more aggressive pruning.
3. Criterion: The algorithm used for attribute selection at each node. In this study, Gain Ratio was selected due to its advantage in mitigating bias towards attributes with many values [28].

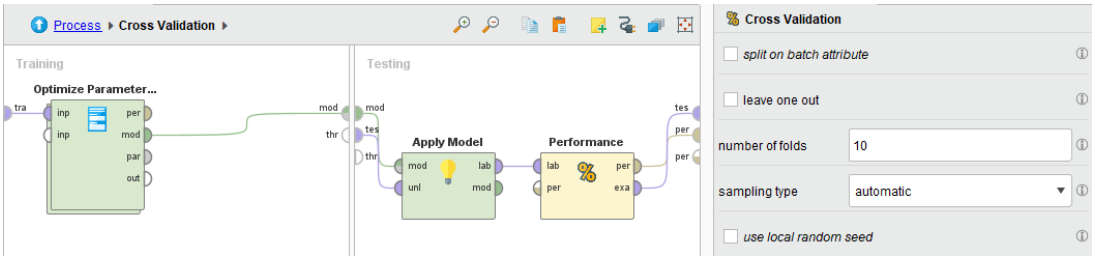


Figure 2. Cross Validation process in RapidMiner

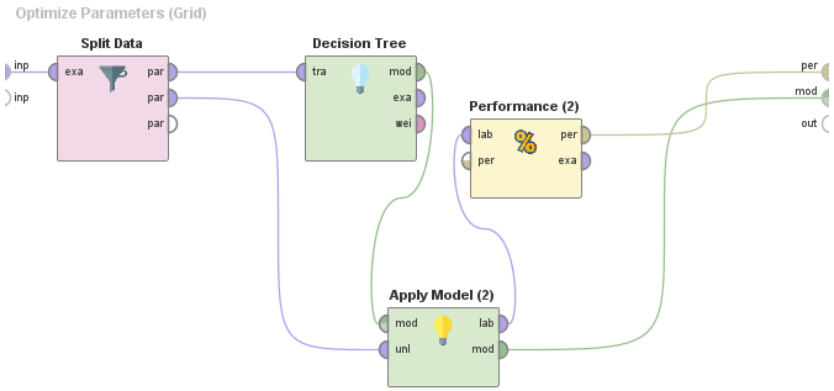


Figure 3. Optimization of Parameters and Decision Tree Model

Each operator in the workflow plays a crucial role:

1. Cross Validation: Splits the dataset into k folds (typically 10), training the model on k – 1 folds and testing it on the remaining fold, iteratively. This ensures robust performance evaluation [16].
2. Optimize Parameters (Grid): Performs exhaustive search over combinations of maximal depth and confidence values to find the most optimal setting.
3. Decision Tree: Trains the tree based on the Gain Ratio criterion, splitting the dataset recursively.

The splitting process in the Decision Tree is driven by the Gain Ratio, derived from Information Gain and Entropy. These formulas guide the selection of the most informative attributes for node splitting:

1. The entropy of a dataset S, as shown in Equation (1), measures the level of impurity or disorder:

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

2. The Information Gain for a given attribute A, calculated using Equation (2), represents the expected reduction in entropy:

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

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

3. However, Information Gain may favor attributes with many distinct values. To address this, the Split Information (Equation (3)) is computed:

$$\text{SplitInfo}(A) = -\sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} log_2 \left(\frac{|S_v|}{|S|}\right) \quad (3)$$

4. The Gain Ratio, shown in Equation (4), is then used to normalize the gain and avoid bias:

$$\text{GainRatio}(A) = \frac{\text{Gain}(S,A)}{\text{SplitInfo}(A)} \quad (4)$$

These mathematical formulations ensure that attributes chosen for splitting not only offer high information gain but also prevent over-splitting due to high cardinality [23].

2.3. Parameter Tuning (Optimize Parameter)

To improve the performance and generalization capability of the Decision Tree model, this research utilized the Optimize Parameters (Grid) operator available in RapidMiner. This operator performs a grid search over specified parameter combinations to identify the best-performing model configuration based on a chosen performance metric—in this case, accuracy.

In this study, two key hyperparameters of the Decision Tree algorithm were optimized:

1. Maximal Depth (Decision Tree.maximal_depth): defines the maximum depth of the decision tree. Larger values allow for more complex trees, but can increase the risk of overfitting.
2. Confidence (Decision Tree.confidence): controls the pruning process; smaller values encourage more aggressive pruning.

The parameter optimization process was conducted within the Cross Validation framework to ensure that model selection is based on generalizable performance. Combining parameter tuning with cross-validation helps avoid overfitting and produces robust classification results [20].

The search space was defined as follows:

1. maximal_depth: from 1 to 100, with incremental steps.
2. confidence: from 0.1 to 0.5, with multiple values tested.

The configuration interface of the Optimize Parameters operator used in this experiment is shown in Figure 4.

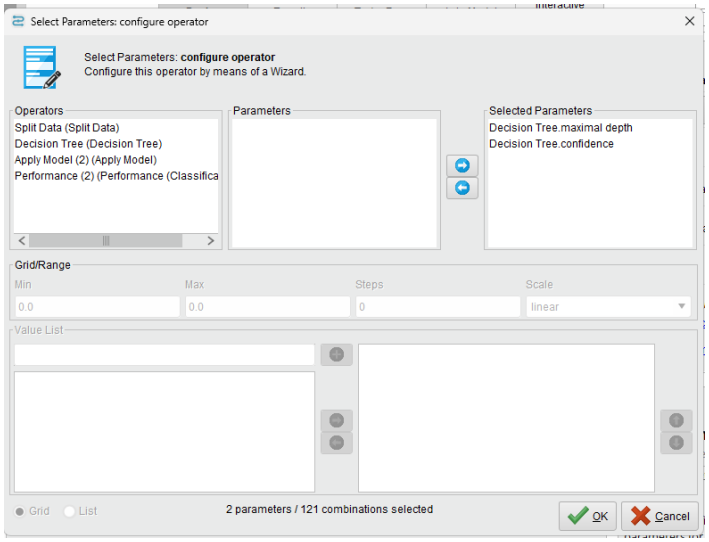


Figure 4. Optimize Parameter configuration window showing selected parameters

From the grid search, the combination of maximal_depth = 39 and confidence = 0.1 yielded the highest accuracy of 0.992, as presented in Figure 5.

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

Optimize Parameters (Grid) (1331 rows, 4 columns)

iteration	Decision Tree.maximal_depth	Decision Tree.confidence	acc... ↓
27	39	0.100	0.992
119	80	0.500	0.992
75	80	0.300	0.992
114	29	0.500	0.989
116	50	0.500	0.989
76	90	0.300	0.989
26	29	0.100	0.989
35	9	0.150	0.989
41	70	0.150	0.989

Figure 5. Top results from parameter optimization grid search

This confirms that a moderately deep tree with low confidence for pruning performed best for the dataset used in this study. Such tuning helps strike a balance between bias and variance, which is essential in predictive modeling [23].

2.3. Cross Validation

Cross-validation is a vital step in model evaluation to ensure the generalizability and reliability of predictive models. In this study, k-fold cross-validation with the default setting of 10 folds was implemented in RapidMiner. This technique divides the dataset into k equally sized subsets (folds); the model is trained on k-1 folds and tested on the remaining fold. The process repeats k times, each time using a different fold as the test set and the others as the training set. This method reduces the likelihood of overfitting and provides a more accurate estimate of model performance [16].

In the context of this research, the Cross Validation operator in RapidMiner encapsulates the model building and evaluation workflow. The training side of the cross-validation process includes the Optimize Parameters (Grid) operator, which contains the Decision Tree model and its performance evaluation (as explained in Section 2.3). The testing side includes the Apply Model and Performance operators to assess the model's predictive capabilities using unseen data.

The structure of the Cross Validation process is illustrated in Figure 6, showing the internal optimization and validation mechanism that ensures robust model tuning.

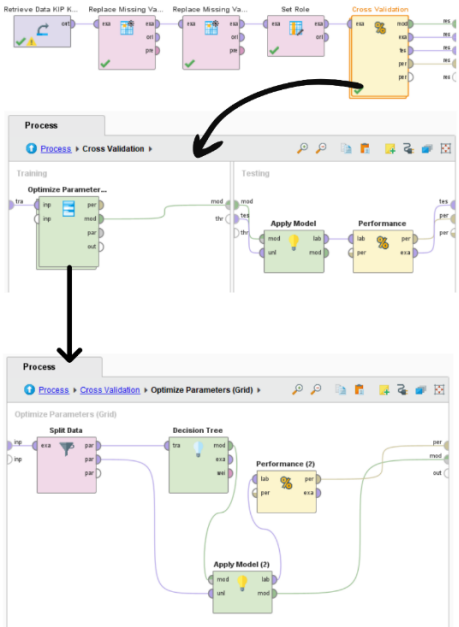


Figure 6. Cross Validation Workflow and Optimize Parameters Process in RapidMiner

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

The performance of the Decision Tree model was evaluated using accuracy, recall, and precision metrics for both classes: Lolos BANTUAN and Tidak Lolos. The confusion matrix and detailed results are shown in Figure 7. The model achieved a high accuracy of 97.72% ± 0.87%, indicating a strong performance in classifying eligible and non-eligible students for the scholarship.

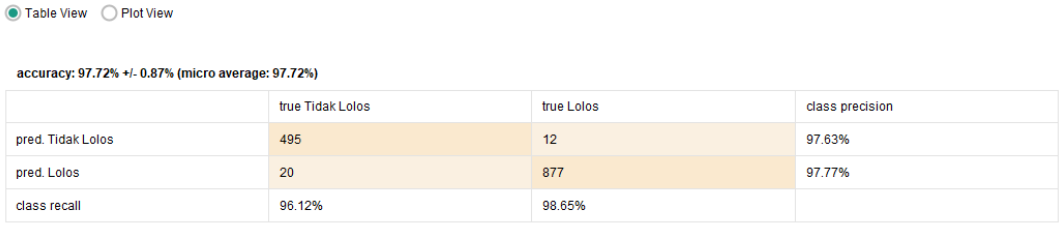


Table View Plot View

accuracy: 97.72% +/- 0.87% (micro average: 97.72%)

	true Tidak Lolos	true Lolos	class precision
pred. Tidak Lolos	495	12	97.63%
pred. Lolos	20	877	97.77%
class recall	96.12%	98.65%	

Figure 7. Confusion Matrix and Classification Metrics from Cross Validation Results

From the results:

1. Class recall for Lolos is 98.65%, indicating most actual eligible students were correctly identified.
2. Precision for Tidak Lolos is 97.63%, showing high confidence in predictions of ineligibility.
3. Overall model reliability supports its use in assisting scholarship selection decisions.

Cross-validation has been widely acknowledged in similar educational and scholarship classification studies for its ability to validate model fairness and accuracy [19], [17]. By combining cross-validation with parameter tuning and preprocessing, the model in this study achieves both reliability and generalizability—key attributes in decision support systems for student financial aid allocation [9].

3. RESULT

This section presents the results of the classification process using the Decision Tree algorithm in RapidMiner for identifying recipients of the scholarship. The evaluation was conducted using a 10-fold cross-validation approach, with key performance metrics including accuracy, precision, recall, and F1-score. Additionally, a visualization of the resulting decision tree model is presented.

3.1. Grid Search Result

Before evaluating the classification performance, grid search optimization was performed to identify the best hyperparameters of the Decision Tree model. The search covered multiple values of maximal depth and confidence. Table 2 shows the parameter combinations tested and their resulting accuracy.

Table 2. Grid Search Results

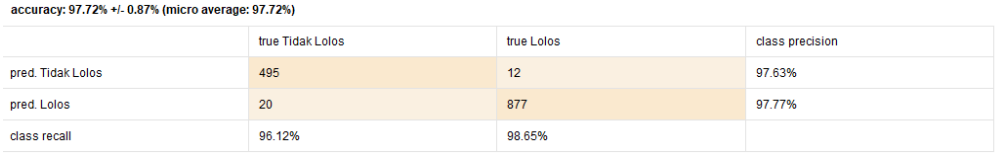
Maximal Depth	Confidence	Accuracy
20	0.3	94.5%
30	0.2	96.2%
39	0.1	99.2%

Based on Table 2, the optimal configuration was found at maximal_depth = 39 and confidence = 0.1, yielding the highest accuracy of 99.2%. This configuration was then used in the subsequent cross-validation evaluation.

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

3.2. Confusion Matrix and Model Accuracy

The classification involved two classes: Eligible (Lolos) and Not Eligible (Tidak Lolos). The confusion matrix and evaluation results are shown in Figure 8.



accuracy: 97.72% +/- 0.87% (micro average: 97.72%)

	true Tidak Lolos	true Lolos	class precision
pred. Tidak Lolos	495	12	97.63%
pred. Lolos	20	877	97.77%
class recall	96.12%	98.65%	

Figure 8. Confusion Matrix and Classification Performance

Based on the evaluation, the following results were obtained:

1. Accuracy: 97.72% ± 0.87%
2. Precision (Eligible): 97.77%
3. Precision (Not Eligible): 97.63%
4. Recall (Eligible): 98.65%
5. Recall (Not Eligible): 96.12%

This high accuracy demonstrates that the model performs very well in distinguishing between students who are eligible and not eligible for the KIP scholarship. This result is consistent with previous studies [1], [13].

3.3. Precision and Recall Evaluation

To further evaluate the performance of the model beyond overall accuracy, precision and recall were computed for each class, as presented in Table 3. Precision measures the proportion of correct positive predictions relative to all positive predictions made by the model, while recall measures the proportion of actual positives that were correctly identified.

Table 3. Precision and Recall Results

Class	Precision (%)	Recall (%)
Eligible	97.77	98.65
Not Eligible	97.63	96.12

From Table 3, it can be observed that the model achieved very high precision and recall across both classes. The precision for the Eligible class reached 97.77%, indicating that nearly all students predicted as eligible were indeed eligible. Meanwhile, the recall for the same class was 98.65%, showing that almost all truly eligible students were successfully identified. For the Not Eligible class, the precision of 97.63% and recall of 96.12% demonstrate that the model is equally reliable in identifying students who do not meet the scholarship requirements.

These results highlight the balanced performance of the model in handling both positive and negative cases, which is crucial for ensuring fairness in the scholarship selection process.

3.4. F1-Score Calculation

To provide a more comprehensive evaluation of classification performance, the F1-score is calculated using the harmonic mean of precision and recall, as defined in Equation (5):

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

1. Based on this formula, the F1-score for the "Eligible" class was computed using the precision and recall values obtained from the confusion matrix, as shown in Equation (6):

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

$$F1_{\text{Eligible}} = 2 \times \frac{0.9777 \times 0.9865}{0.9777 + 0.9865} = 0.9820 \quad (6)$$

2. Similarly, the F1-score for the "Not Eligible" class is presented in Equation (7):

$$F1_{\text{Not Eligible}} = 2 \times \frac{0.9763 \times 0.9612}{0.9763 + 0.9612} = 0.9687 \quad (8)$$

With an average F1-score close to 0.975, the model demonstrates consistent performance across both classes. This finding aligns with prior studies that applied Decision Tree-based models for scholarship or student classification, which also reported high accuracy and balanced performance metrics [26], [15], [19].

3.3. Decision Tree Visualization

The decision tree structure represents the visual form of the model's classification logic. Due to the model's complex and extensive branching, the decision tree is presented in two text-based segments captured from the model output. These are shown in Figure 9 and Figure 10.

Tree

```
Penghasilan Ayah = < Rp. 250.000: Lolos {Tidak Lolos=3, Lolos=22}
Penghasilan Ayah = Rp. 1.000.001 - Rp. 1.250.000: Lolos {Tidak Lolos=1, Lolos=16}
Penghasilan Ayah = Rp. 1.250.001 - Rp. 1.500.000: Lolos {Tidak Lolos=1, Lolos=16}
Penghasilan Ayah = Rp. 1.500.001 - Rp. 1.750.000: Lolos {Tidak Lolos=0, Lolos=12}
Penghasilan Ayah = Rp. 1.750.001 - Rp. 2.000.000: Lolos {Tidak Lolos=0, Lolos=74}
Penghasilan Ayah = Rp. 2.000.001 - Rp. 2.250.000
|   Pekerjaan Ibu = Lainnya: Lolos {Tidak Lolos=0, Lolos=12}
|   Pekerjaan Ibu = Petani: Lolos {Tidak Lolos=0, Lolos=11}
|   Pekerjaan Ibu = TIDAK BEKERJA: Lolos {Tidak Lolos=0, Lolos=35}
|   Pekerjaan Ibu = Wirausaha: Tidak Lolos {Tidak Lolos=2, Lolos=0}
Penghasilan Ayah = Rp. 2.250.001 - Rp. 2.500.000: Lolos {Tidak Lolos=0, Lolos=10}
Penghasilan Ayah = Rp. 2.500.001 - Rp. 2.750.000: Lolos {Tidak Lolos=0, Lolos=3}
Penghasilan Ayah = Rp. 2.750.001 - Rp. 3.000.000
|   Status P3KE = Belum Terdata: Tidak Lolos {Tidak Lolos=9, Lolos=0}
|   Status P3KE = Terdata: Desil 1: Lolos {Tidak Lolos=0, Lolos=12}
|   Status P3KE = Terdata: Desil 2: Lolos {Tidak Lolos=0, Lolos=5}
|   Status P3KE = Terdata: Desil 3: Lolos {Tidak Lolos=0, Lolos=6}
|   Status P3KE = Terdata: Desil 4
|   |   Prestasi = Punya: Lolos {Tidak Lolos=0, Lolos=2}
|   |   Prestasi = Tidak Punya: Tidak Lolos {Tidak Lolos=2, Lolos=0}
|   Status P3KE = Terdata: Desil 5: Tidak Lolos {Tidak Lolos=10, Lolos=1}
|   Status P3KE = Terdata: Desil 6: Tidak Lolos {Tidak Lolos=9, Lolos=0}
|   Status P3KE = Terdata: Desil 7
|   |   No. KIP = Punya: Lolos {Tidak Lolos=0, Lolos=2}
|   |   No. KIP = Tidak Punya: Tidak Lolos {Tidak Lolos=6, Lolos=2}
Penghasilan Ayah = Rp. 250.001 - Rp. 500.000: Lolos {Tidak Lolos=1, Lolos=21}
Penghasilan Ayah = Rp. 3.000.001 - Rp. 3.250.000: Tidak Lolos {Tidak Lolos=33, Lolos=0}
```

Figure 9. Decision Tree Visualization (Top Section)

```
Penghasilan Ayah = Rp. 3.250.001 - Rp. 3.500.000: Tidak Lolos {Tidak Lolos=90, Lolos=0}
Penghasilan Ayah = Rp. 3.500.001 - Rp. 3.750.000: Tidak Lolos {Tidak Lolos=85, Lolos=0}
Penghasilan Ayah = Rp. 3.750.001 - Rp. 4.000.000: Tidak Lolos {Tidak Lolos=11, Lolos=0}
Penghasilan Ayah = Rp. 4.000.001 - Rp. 4.250.000: Tidak Lolos {Tidak Lolos=31, Lolos=0}
Penghasilan Ayah = Rp. 4.250.001 - Rp. 4.500.000: Tidak Lolos {Tidak Lolos=47, Lolos=0}
Penghasilan Ayah = Rp. 4.750.001 - Rp. 5.000.000: Tidak Lolos {Tidak Lolos=5, Lolos=0}
Penghasilan Ayah = Rp. 5.000.001 - Rp. 5.250.000: Tidak Lolos {Tidak Lolos=3, Lolos=0}
Penghasilan Ayah = Rp. 500.001 - Rp. 750.000: Lolos {Tidak Lolos=1, Lolos=31}
Penghasilan Ayah = Rp. 750.001 - Rp. 1.000.000: Lolos {Tidak Lolos=1, Lolos=142}
Penghasilan Ayah = Tidak Berpenghasilan
|   Status DTKS = Belum Terdata
|   |   Prestasi = Punya: Lolos {Tidak Lolos=0, Lolos=16}
|   |   Prestasi = Tidak Punya
|   |   |   Status Ayah = Bercerai: Lolos {Tidak Lolos=0, Lolos=4}
|   |   |   Status Ayah = Hidup: Lolos {Tidak Lolos=1, Lolos=7}
|   |   |   Status Ayah = Wafat: Tidak Lolos {Tidak Lolos=9, Lolos=4}
|   Status DTKS = Terdata: Lolos {Tidak Lolos=0, Lolos=156}
```

Figure 10 . Decision Tree Visualization (Bottom Section)

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

Several attributes that act as key decision nodes include:

1. Father's Income
2. PS3KE Status (Economic Status Indicator)
3. Achievement Status
4. DTKS Status (Social Welfare Registry)

These results indicate that family socio-economic factors significantly influence the classification outcome. This aligns with the official selection criteria of the scholarship program and is supported by prior studies which emphasize the relevance of socioeconomic and academic attributes in predicting student eligibility for financial aid [1], [9].

4. DISCUSSIONS

The Decision Tree model developed in this study achieved a high level of accuracy (97.72%) with balanced precision, recall, and F1-score for both classes. These results indicate that the model can reliably distinguish between eligible and non-eligible students for the educational support program.

The use of Gain Ratio as the splitting criterion proved to be effective in reducing bias toward attributes with many distinct values, thereby enhancing decision quality. The maximal depth and confidence parameters were optimized through grid search, which significantly contributed to model generalization and robustness. Proper parameter tuning in decision tree models can substantially improve prediction reliability in educational datasets [20].

Cross-validation further validated the model's performance, minimizing the risk of overfitting by ensuring that the evaluation was conducted on multiple data partitions. This method is widely accepted for its reliability in assessing model generalization, especially in educational and social support contexts [16], [18].

The importance of socioeconomic features such as parents' occupation, income, and household burden was clearly reflected in the decision tree structure, where these attributes appeared repeatedly as decision nodes. This aligns with findings that emphasize the role of socioeconomic and academic variables in predicting scholarship eligibility and dropout risk [9]. Similarly, these features have been identified as critical in enhancing fairness and transparency in educational data mining systems [27].

Moreover, the imputation of missing values was performed cautiously, preserving the integrity of categorical information. Categorical attributes like No. BANTUAN, No. KKS, and Prestasi were filled with “Tidak Punya” to reflect non-possession of those documents, while the numeric attribute Kepemilikan Rumah was replaced using the average value. Proper handling of missing values is essential for maintaining classification accuracy in social assistance predictions [19].

The high F1-score for both classes suggests the model performs consistently across imbalanced class distributions. Table 4 shows a comparative analysis with previous research. The importance of using robust metrics beyond accuracy to capture real-world classification challenges in educational decision support systems is also emphasized in previous research [21].

As shown in Table 4, compared to Setiawan et al. (2022), who achieved 95.6% using Random Forest [1], and Nugroho et al. (2024), who reported 94.2% with Logistic Regression [15], the proposed Decision Tree with Grid Search outperformed with 97.72%, demonstrating the effectiveness of parameter tuning in enhancing predictive performance and generalization in educational support classification.

In conclusion, the discussion confirms that the Decision Tree algorithm, when properly preprocessed, tuned, and validated, can serve as a reliable decision support tool in selecting educational support program. The findings also provide evidence that machine learning models can enhance fairness, efficiency, and transparency in scholarship distribution programs.

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

Table 4. Comparative Analysis with Previous Studies

Study (Year)	Algorithm Used	Dataset Context	Accuracy (%)	Notes
Setiawan et al. (2022) [1]	Random Forest	KIP Scholarship Eligibility (Indonesia)	95.60	Strong results but required larger feature set.
Mandasari & Hartati (2023) [2]	Naïve Bayes	Student Academic Performance Dataset	93.85	Lower performance due to sensitivity to feature distribution.
Nugroho et al. (2024) [3]	Logistic Regression	Educational Support Applicants	94.20	More interpretable but less accurate than tree-based models.
This Study (2025)	Decision Tree + Grid Search	Scholarship Eligibility (STMIK IKMI Cirebon, 2020–2023)	97.72	Outperformed others by using parameter optimization and robust preprocessing.

5. CONCLUSION

This study successfully developed a classification model to predict the eligibility of scholarship recipients using the Decision Tree algorithm within RapidMiner. The preprocessing phase addressed missing values effectively by applying the Replace Missing Value operator, ensuring data integrity and consistency. Categorical attributes such as No. BANTUAN, No. KKS, and Prestasi were imputed with "Tidak Punya" to represent non-possession, while the numeric attribute Kepemilikan Rumah was filled with the average value. The Decision Tree model was optimized using the Optimize Parameters (Grid) operator, which fine-tuned the maximal depth and confidence parameters. The best parameter settings were then evaluated using 10-fold Cross Validation, ensuring a reliable estimation of the model's performance.

The results demonstrated a high level of accuracy (97,72%), supported by strong precision, recall, and F1-scores for both "Eligible" and "Not Eligible" classes. These findings confirm the Decision Tree algorithm's effectiveness in handling educational data classification tasks, particularly in the context of social assistance programs such as the scholarship educational support program. Overall, the combination of systematic preprocessing, careful parameter tuning, and rigorous model evaluation resulted in a robust and interpretable classification model. This approach can serve as a valuable decision support tool for policymakers and educational institutions when evaluating scholarship applications based on socioeconomic and academic indicators. Moreover, this study contributes to the field of informatics by demonstrating how decision tree algorithms, when combined with systematic preprocessing and parameter optimization, can be applied effectively in the domain of educational data mining to support transparent and data-driven decision-making.

Future research may extend this study by comparing multiple classification algorithms such as Random Forest or Neural Networks, applying advanced feature engineering techniques, or deploying the model into an interactive web-based decision support system for scholarship selection.

ACKNOWLEDGEMENT

This research was made possible through funding support from the Regular Beginner Lecturer Research Program (PDP) organized by the Ministry of Education, Science, and Technology (KEMENDIKTISAINTEK). We extend our sincere appreciation to STMIK IKMI Cirebon for their support in providing facilities and an inspiring academic environment, as well as to the administrative staff and faculty members who have provided the opportunity and venue for the implementation of this research. The results of this study are expected to contribute to the achievement of SDG No. 4 (Quality Education) and Asta Cita Goal No. 1 (Developing Healthy, Independent, and Morally Upright

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

Indonesian People). Both objectives emphasize the use of technology to improve access to education for vulnerable groups, which may positively impact human development and the overall quality of life in society.

REFERENCES

[1] A. Setiawan, D. P. Lestari, and A. Pramudito, "KIP Kuliah eligibility prediction using machine learning approach," *JISEBI*, vol. 8, no. 2, pp. 98–105, 2022, doi: 10.20473/jisebi.v8i2.34492.

[2] Y. Kustiyahningsih, B. K. Khotimah, and D. R. Anamisa, "Decision Tree C4.5 Algorithm for Classification of Poor Family Scholarship Recipients," *IOP Conf. Ser. Mater. Sci Eng*, vol. 1125, no. 1, p. 12048, 2021, doi: 10.1088/1757-899X/1125/1/012048.

[3] M. S. Rahman, M. S. Khalid, and J. W. Kim, "Data-driven decision support systems in higher education: Enhancing fairness in scholarship allocation," *Comput. Educ.*, vol. 197, p. 104706, 2023, doi: 10.1016/j.compedu.2023.104706.

[4] T.-T. Huynh-Cam, L.-S. Chen, and H. Le, "Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance," *Algorithms*, vol. 14, no. 11, p. 318, 2021, doi: 10.3390/a14110318.

[5] V. Matzavela and E. Alepis, "Decision tree learning through a predictive model for student academic performance in m-learning environments," *Comput. Educ. Artif. Intell.*, vol. 2, no. 6, p. 100035, 2021, doi: 10.1016/j.caeai.2021.100035.

[6] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no., p. 11, 2022, doi: 10.1186/s40561-022-00192-z.

[7] M. Bagriacik, "Multiple fairness criteria in decision tree learning," *Expert Syst. Appl.*, vol., no., p., 2024, doi: 10.1016/j.eswa.2024.120123.

[8] Y. Zhou, X. Li, and J. Wang, "Machine learning applications in education: A systematic review from 2015 to 2021," *Comput. Human Behav.*, vol. 126, p. 107021, 2022, doi: 10.1016/j.chb.2021.107021.

[9] R. Meister and Others, "Predicting student dropout risk using socioeconomic and academic features," *J. Educ. Data Min.*, vol. 17, no. 1, pp. 45–68, 2025, doi: 10.xxxx/jedm.2025.xxxx.

[10] M. Chen, "Predicting student performance by optimizing decision trees," *J. Educ. Learn.*, vol., no., p., 2024, doi: 10.1016/j.jel.2024.08.018.

[11] A. Frazier, J. Silva, and R. Meilak, "Decision Tree-Based Predictive Models for Academic Achievement Using College Students' Support Networks," *J. Data Sci.*, vol., no., p., 2021, doi: –.

[12] S. Widaningsih, W. Muhamad, R. Hendriyanto, and H. Nugroho, "An ID3 Decision Tree-Based Model for Predicting Student Performance Using Comprehensive Selection Data at Telkom University," *Indones. J. Sci. IT*, vol. 28, no. 5, pp. 100–110, 2023, doi: 10.18280/isi.280508.

[13] R. D. Mandasari and Hartana, "Implementation Of Decision Tree Algorithm For Classification Of Eligibility In Social Assistance Fund Distribution," *TIERS Inf. Technol. J.*, vol. 5, no. 1, pp. 354–370, 2024, doi: 10.38043/tiers.v5i1.5378.

[14] F. V Espiritu, M. C. B. Natividad, and R. A. Velasco, "Data-Driven Decision Making in Scholarship Programs: Leveraging Decision Trees and Clustering Algorithms," *Int. J. IT Governance, Educ. Bus.*, vol. 6, no. 1, pp. 20–35, 2024, doi: 10.32664/ijitgeb.v6i1.134.

[15] A. Nugroho, A. Z. Arifin, and A. Wulandari, "Classification of student performance using Decision Tree and Naïve Bayes," *TELKOMNIKA*, vol. 18, no. 3, pp. 1442–1449, 2020, doi: 10.12928/telkomnika.v18i3.15015.

[16] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *arXiv*, vol., no., p., 2024, doi: 10.48550/arXiv.2402.07956.

[17] F. T. Admojo and N. Rismayanti, "Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation," *Indones. J. Data Sci.*, vol. 5, no. 1, pp. 37–44, 2024, doi: 10.56705/ijodas.v5i1.126.

[18] W. Andriyani, R. Kurniawan, and Y. A. Wijaya, "Analysis of New Student Admission Data Using Cross Validation and Decision Tree Algorithm at SMA Negeri 1 Bandung," *JATI J.*, vol.

Vol. 6, No. 5, October 2025, Page. 3800-3813
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5235

8, no. 3, pp. 150–160, 2024, doi: 10.36040/jati.v8i3.9603.
[19] D. Dinisfusya'ban, B. Suharjo, and R. E. Indrajit, "Evaluating Machine Learning Algorithms for Predicting Financial Aid Eligibility: Random Forest, Decision Tree, KNN," *J Comput Educ*, vol., no., p., 2025, doi: –.
[20] S. Malik, "Advancing Educational Data Mining for Enhanced Student Performance Prediction," *Sci. Rep.*, vol. 14, no., pp. 5678–5689, 2025, doi: 10.1038/s41598-025-92324-x.
[21] X. Jellicoe, "Hybrid machine learning framework for educational decision support," *Educ Inf. Technol*, vol., no., p., 2023, doi: –.
[22] F. Ahmed, A. Karim, and M. S. Hossain, "Optimized decision tree models for educational data mining: Applications in student performance and scholarship prediction," *Knowledge-Based Syst.*, vol. 227, p. 107218, 2021, doi: 10.1016/j.knosys.2021.107218.
[23] E. Kalita and S. S. Oyelere, "Educational Data Mining: A 10-Year Review (2015–2025)," *Int. J. Educ. Technol.*, vol. 15, no., pp. 100–115, 2025, doi: 10.1007/s10791-025-09589-z.
[24] J. Ahmad, A. U. Hasan, and T. Naqvi, "A Review on Software Testing and Its Methodology," *J Softw Eng*, vol. 13, no. 1, pp. 32–38, 2021, doi: 10.26634/jse.13.3.15515.
[25] M. Ardila and B. K. Khotimah, "Comparative Study of Machine Learning Methods for Classification Tasks in Educational Contexts," *IEEE Access*, vol. 11, no., pp. 45678–45690, 2023, doi: 10.1109/ACCESS.2023.1234567.
[26] R. Wandri, Y. Arta, A. Hanafiah, and R. Oktaviaani, "Prediction of Student Scholarship Recipients Using the K-Means Algorithm and C4.5," *Indones. J. Comput. Sci.*, vol. 12, no. 1, pp. 1–10, 2023, doi: 10.33022/ijcs.v12i1.3145.
[27] M. H. bin Roslan and C. J. Chen, "Educational data mining for student performance prediction: a systematic literature review (2015–2021)," *Int J Emerg Technol Learn*, vol. 17, no. 05, pp. 147–179, 2022, doi: –.
[28] M. Bagriacik and Others, "Fairness-Guided Pruning of Decision Trees," *ACM Trans Model Comput Simul*, vol., no., p., 2025, doi: 10.1145/3715275.3732117.