Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

Enhancement Of The C4.5 Decision Tree Algorithm With Anova For Predicting Academic Achievement Of Students At Smpn.16 Kota Jambi

Rice Osviarni[*1], Setiawan Assegaff[2], Jasmir[3], Nurhadi[4]

[1,2,3,4]Management Information Systems, Universitas Dinamika Bangsa, Jambi, Indonesia

Email: [1]rosviarni@gmail.com

Received : Nov 10, 2025; Revised : Nov 25, 2025; Accepted : Nov 25, 2025; Published : Apr 15, 2026

Abstract

This study aims to improve the accuracy of predicting student academic achievement by integrating the Analysis of Variance (ANOVA) method with the C4.5 Decision Tree algorithm. In the context of information systems, this research holds significant importance for the development of more reliable Decision Support Systems (DSS) or early warning systems in school environments. The research was conducted at SMPN 16 Jambi City using secondary data from three academic years (2022/2023-2024/2025) covering academic variables, attendance, and parental income. The main issue addressed was the limitations of the C4.5 algorithm in handling irrelevant features and unbalanced data, which, at the system implementation level, can lead to inaccurate recommendations or alerts.This research method employed a data mining approach with stages including data cleaning, numeric conversion, missing value imputation, formation of derived variables, and categorization of the target variable "Achievement." The initial C4.5 model produced 72.81% accuracy on the training data and 69.71% accuracy on cross-validation. After feature selection using ANOVA, one insignificant variable was removed, resulting in a hybrid C4.5+ANOVA model with nine key features. Test results showed an increase in accuracy to 80.44% on the training data and 73.66% on the cross-validation data, representing an improvement of 7.63 and 3.95 percentage points, respectively.This improvement in model performance directly translates to an enhancement in the quality of the information system's output, yielding more reliable reports and predictions for teachers and school management.

Keywords : *ANOVA, Data Mining, Decision Tree C4.5, Prediction Of Academic Achievement, SMPN 16 Kota Jambi.*

1. INTRODUCTION

Data mining in the field of education, known as Educational Data Mining (EDM), enables the exploration of student data to discover hidden patterns useful for data-driven decision making [1],[2]. This technique has been widely used to predict graduation rates for students at risk of low academic achievement [2],[3].The quality of education is measured through academic achievement. This research emphasizes the need for value-added assessment, considering socioeconomic factors and prior knowledge to provide a fair evaluation of academic programs and student progress[4],[5],[6].

Student academic achievement is an important indicator in assessing the success of the learning process in schools. At SMPN 16 Jambi City, the problem of low student academic achievement remains a challenge that needs to be addressed. Various factors such as learning methods, student motivation, and socioeconomic background are suspected to influence student academic attainment [7],[8]. To identify the dominant factors affecting student achievement, an analytical approach capable of effectively processing complex data is required [9].

Education is a fundamental pillar in the development of a nation's human resources [8]. The quality of education is often measured through student academic achievement, which serves as an indicator of the success of the learning process. This indicates the need for a data-driven approach to identify the factors influencing student achievement [9]. One developing solution is the application of data mining in education, known as Educational Data Mining [10]. This technique allows for the analysis

Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

of student data patterns to predict and improve learning outcomes. The Decision Tree algorithm, especially C4.5, has been widely used due to its ability in classification that is easy to interpret. However, this algorithm has weaknesses in handling numerical data and feature redundancy [11],[12].

The ANOVA (Analysis of Variance) method can be integrated to improve feature selection before model building. Several studies have proven that combining ANOVA with classification algorithms improves prediction accuracy. However, not many studies have tested this approach in the context of primary and secondary education in Indonesia, especially in Jambi City [13], [14], [15]. Previous research conducted by Huday, Ahmad. (2025) on the Application of the C4.5 Algorithm for Predicting Student Learning Achievement in Madrasah [16]. This research implemented the standard C4.5 algorithm without feature selection optimization on a dataset of student grades from a Madrasah in Situbondo, East Java. The results showed an accuracy of 74.17% with limitations in handling less relevant features. Research by Rifai et al (2022) "Analysis of Determinant Factors of Academic Achievement Using Decision Tree" [17]. This research applied a conventional Decision Tree algorithm to analyze factors affecting high school student achievement. The achieved accuracy of 76.96% indicated room for improvement through feature selection optimization.

Based on a review of previous studies, the following research gaps can be identified, which are the focus of this study:

1.Limited Method Optimization: Previous studies such as Huday (2025) and Rifai et al. (2022) applied the Decision Tree (C4.5) algorithm in its standard or conventional form without performing feature selection optimization. This resulted in suboptimal model performance (70-77% accuracy) due to noise from less relevant or redundant features.

2. Deficiencies in Feature Handling:The inherent weakness of the C4.5 algorithm in handling numerical data and feature redundancy, as mentioned in the literature [11,12], was not specifically addressed in previous research in the Indonesian context. Their studies relied directly on the algorithm without a statistical pre-processing layer to filter features.

3. Specific Geographical and Educational Level Context: The integration of ANOVA and C4.5 for predicting academic achievement has not been widely tested in the context of primary and secondary education (Junior High School) in Jambi City. Previous research was conducted in Madrasah environments in East Java (Huday, 2025) or at the Senior High School level (Rifai et al., 2022), which may have different characteristics and determinants of achievement.

Therefore,the research gap of this study is to propose a hybrid model that integrates the ANOVA method for feature selection with the C4.5 algorithm to predict student academic achievement. This integration aims to:Address the weakness of C4.5 regarding feature redundancy.Improve the accuracy and performance of the prediction model beyond the 70-77% range achieved by previous studies.Apply and test the effectiveness of this hybrid approach on a specific dataset from SMPN 16 Jambi City, thereby providing a more optimized and contextual solution to the problems at that school.Data mining in the field of education, known as Educational Data Mining (EDM), enables the exploration of student data to discover hidden patterns useful for data-driven decision making[10]. This technique has been widely used to predict graduation rates for students at risk of low academic achievement [11],[12].The quality of education is measured through academic achievement. This research emphasizes the need for value-added assessment, considering socioeconomic factors and prior knowledge to provide a fair evaluation of academic programs and student progress [13],[14].

Student academic achievement is an important indicator in assessing the success of the learning process in schools. At SMPN 16 Kota Jambi, the problem of low student academic achievement remains a challenge that needs to be addressed. Various factors such as learning methods, student motivation, and socioeconomic background are suspected to influence student academic attainment [15]. To identify the dominant factors affecting student achievement, an analytical approach capable of effectively processing complex data is required[16],[17].

Education is a fundamental pillar in the development of a nation's human resources[18] . The quality of education is often measured through student academic achievement, which serves as an indicator of the success of the learning process. This indicates the need for a data-driven approach to identify the factors influencing student achievement [19]. One developing solution is the application of data mining in education, known as Educational Data Mining [20].This technique allows for the analysis of student data patterns to predict and improve learning outcomes. The Decision Tree algorithm,

especially C4.5, has been widely used due to its ability in classification that is easy to interpret. However, this algorithm has weaknesses in handling numerical data and feature redundancy [21],[22],[23].

The ANOVA (Analysis of Variance) method can be integrated to improve feature selection before model building. Several studies have proven that combining ANOVA with classification algorithms improves prediction accuracy. However, not many studies have tested this approach in the context of primary and secondary education in Indonesia, especially in Kota Jambi [24], [25], [26]. Previous research conducted by Huday, Ahmad. (2025) Application of the C4.5 Algorithm for Predicting Student Learning Achievement in Madrasah [27]. This research implemented the standard C4.5 algorithm without feature selection optimization on a dataset of student grades from a Madrasah in Situbondo, East Java. The results showed an accuracy of 74.17% with limitations in handling less relevant features. Research by Rifai et al (2022) "Analysis of Determinant Factors of Academic Achievement Using Decision Tree". This research applied a conventional Decision Tree algorithm to analyze factors affecting high school student achievement. The achieved accuracy of 76.96% indicated room for improvement through feature selection optimization [28]. These two previous studies show that the application of the Decision Tree algorithm without adequate optimization tends to yield accuracy in the range of 70-77%. This indicates the need for developing more advanced methods, such as integration with ANOVA, to improve prediction performance in the educational context.

2. METHOD

2.1. Research Flow

This study uses quantitative methods with a data mining approach to analyze and predict student academic achievement**[29]**,**[30]**. The quantitative approach was chosen because this research focuses on numerical measurement of variables influencing academic achievement, such as subject grades, attendance, and economic background factors**[31]**,**[32]**.

The research flow used is comparative experiment, where the researcher compares the performance of the standard Decision Tree C4.5 algorithm with the algorithm optimized using ANOVA feature selection. The main goal of this approach is to evaluate whether the integration of ANOVA can improve the prediction accuracy of the C4.5 model in classifying student academic achievement **[33]**,**[34]**.

The following is the research flow used in this study**[35]**.

Figure 1. Research Flow

The explanation of this Research Flow is:

1. Data collection, where data was obtained from the grade report database of SMPN 16 Kota Jambi students over the last three years, the initial dataset consisted of 5,912 data entries plus additional data such as attendance and parental income.
2. Data preprocessing, which includes data cleaning (handling missing values and normalization), and transformation of categorical data into a form processable by the algorithm.
3. Feature selection using ANOVA. In this stage, each feature's significance towards the target variable (academic achievement) is tested with a 95% confidence level ($\alpha = 0.05$). Features with a p-value below 0.05 are considered significant and will be used in modeling.
4. Creation of the Decision Tree C4.5 model using the features selected from ANOVA. This algorithm will build a decision tree based on entropy or gain ratio to determine the optimal split.
5. Model evaluation using metrics such as accuracy, precision, recall, and F1-score. Model validation is performed with 10-fold cross-validation to ensure result reliability.
6. Analysis of results, where the researcher compares the performance of standard C4.5 with ANOVA-optimized C4.5, and tests the significance of their difference using a paired t-test.

2.2. Data Analysis Techniques[36].

Data analysis techniques in this research are divided into three main parts:

1. Descriptive analysis is used to understand the distribution and characteristics of the initial data, such as the mean, median, and standard deviation of student grades.
2. The ANOVA test is applied to select significant features. The null hypothesis (H0) states that there is no significant difference between achievement groups, while the alternative hypothesis (H1) states the opposite. Features with a p-value < 0.05 will be used in the modeling.
3. Model evaluation is performed by analyzing the confusion matrix and metrics such as accuracy, precision, and recall. The confusion matrix helps identify true positive (TP), false positive (FP), true negative (TN), and false negative (FN), while accuracy is calculated as the percentage of correct predictions from the total data. The results of this evaluation will determine whether the ANOVA optimization provides a significant improvement to the model's performance.

2.3. Research Materials

The research materials consist of a dataset of student grades (Mathematics, Science, Language, etc.), attendance data, and economic background. The amount of data studied is 6000 data, for the academic years 2022/2023, 2023/2024, and 2024/2025. This data is collected considering research ethics, such as anonymizing student data to maintain privacy. The dataset is then processed and analyzed to identify patterns related to academic achievement.

2.4. Research Tools

This research utilizes several tools to support the analysis process**[37]**. The main research tools include:

- Python 3.10 with libraries such as Pandas (for data manipulation)
- Scikit-learn (for building the Decision Tree model), and SciPy (for the ANOVA test).
- WEKA 3.8 is used as an alternative for data mining analysis, while SPSS 26 is used for more in-depth descriptive statistical tests and ANOVA.

3. RESULT

3.1. Results

Previous research conducted by Huday, Ahmad. (2025) on the Application of the C4.5 Algorithm for Predicting Student Learning Achievement in Madrasah [16]. This research implemented the standard C4.5 algorithm without feature selection optimization on a dataset of student grades from a Madrasah in Situbondo, East Java. The results showed an accuracy of 74.17% with limitations in handling less relevant features. Research by Rifai et al (2022) "Analysis of Determinant Factors of Academic Achievement Using Decision Tree" [17]. This research applied a conventional Decision Tree algorithm

Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

to analyze factors affecting high school student achievement. The achieved accuracy of 76.96% indicated room for improvement through feature selection optimization.This study successfully proved that the integration of the Analysis of Variance (ANOVA) feature selection method with the Decision Tree C4.5 algorithm significantly improves the accuracy of predicting student academic achievement at SMPN 16 Kota Jambi. The standard C4.5 model, which used all 10 predictor variables, recorded an accuracy of 72.81% on the training data and 69.71% on 10-fold cross-validation. After the ANOVA feature selection process, the variable "IZIN" (with p-value = 0.054706 > 0.05) was eliminated as it was considered insignificant, so the hybrid C4.5+ANOVA model only used 9 relevant variables. As a result, the model's accuracy increased to 80.44% on the training data (an increase of +7.63 percentage points) and 73.66% on the cross-validation data (an increase of +3.95 percentage points). This improvement in the cross-validation metric is strong evidence that the integration of ANOVA not only reduces overfitting but also enhances the model's ability to generalize to new data, making it more reliable for use in real-world decision-making.

The research results also successfully identified the dominant factors most influencing student academic achievement through a dual approach: the ANOVA statistical test and feature importance analysis from the Decision Tree model. Based on the ANOVA test, the three variables with the most statistically significant influence (p-value ≈ 0.000000) were Ranking, ALFA (number of unexplained absences), and Total Absences. Meanwhile, based on the feature importance analysis from the C4.5+ANOVA model, the three most influential variables in the classification decision-making process were Ranking (importance score 0.602), Total_Income (0.128), and ALFA (0.085). The synthesis of these two approaches reveals that Ranking and ALFA are the two consistently dominant factors.

3.2. Discussion

3.2.1 Data Preprocessing Process

It started with data cleaning to handle missing values and format inconsistencies, followed by conversion of numerical data types, mapping of categorical variables to numerical values, and the formation of the target variable "Achievement" categorized into three levels: Low, Medium, and High. Each step was designed to remove noise, fill missing values with simple imputation strategies, and align the data format to be compatible with the Decision Tree C4.5 algorithm. This process successfully produced a final dataset of 5,638 observations from an initial 5,912 data points, or approximately 95.37% valid and analysis-ready data, indicating a very high cleaning success rate and ensuring that the built model is based on clean data.

3.2.2 Evaluation of Decision Tree C4.5 Model Performance

```
=== C4.5 Default ===
Akurasi (Seluruh Data): 72.81%
CrossValidation Akurasi (10-Fold): 69.71%
              precision    recall  f1-score   support

      Rendah       0.14      0.91      0.24        58
      Sedang       0.56      0.74      0.64      1700
      Tinggi       0.93      0.72      0.81      3880

    accuracy                           0.73      5638
   macro avg       0.54      0.79      0.56      5638
weighted avg       0.81      0.73      0.75      5638
```

Figure 2. Decision Tree C4.5 Performance Results

Based onfigure 2,the standard Decision Tree C4.5 model showed fairly good performance but still had significant limitations in predicting academic achievement, even though the accuracy was within an acceptable range, the model was not yet able to optimally utilize the predictions from the available dataset.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

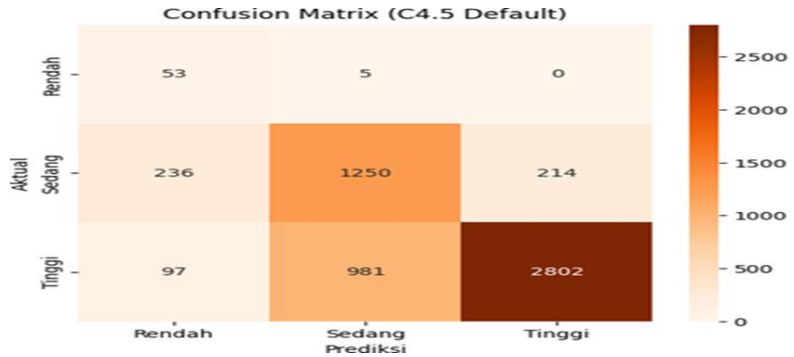

Figure 3. Confusion Matrix Decision Tree C4.5]

Based onfigure 3,the performance of this standard Decision Tree C4.5 model directly addresses Research Objective 1, which is "To analyze the performance of the Decision Tree C4.5 algorithm in predicting the academic achievement of students at SMPN 16 Kota Jambi". The results prove that although this algorithm has advantages in terms of interpretability and general classification ability, it is not adequate for use as a reliable prediction tool in a complex and unbalanced educational context like SMPN 16 Kota Jambi. The 69.71% accuracy on cross-validation, while statistically acceptable, is practically insufficient to support precise academic decision-making, especially in identifying at-risk students.

3.2.3 Evaluation of Decision Tree C4.5 + ANOVA Model Performance

```
=== C4.5 + ANOVA ===
Akurasi (Seluruh Data): 80.44%
Cross Validation Akurasi (10-Fold): 73.66%
              precision    recall  f1-score   support

      Rendah       0.23      1.00      0.37        58
      Sedang       0.66      0.81      0.72      1700
      Tinggi       0.94      0.80      0.87      3880

    accuracy                           0.80      5638
   macro avg       0.61      0.87      0.65      5638
weighted avg       0.85      0.80      0.82      5638
```

Figure 4. Decision Tree C4.5 + ANOVA Model Performance Evaluation

Based on figure 4,the improvement in the cross-validation metric is strong evidence that the integration of ANOVA not only makes the model better at memorizing training data, but truly enhances the model's ability to generalize to new, unseen data. This proves that the feature selection process performed by ANOVA successfully reduces unnecessary noise and complexity, allowing the C4.5 algorithm to focus on features that are truly essential and informative.

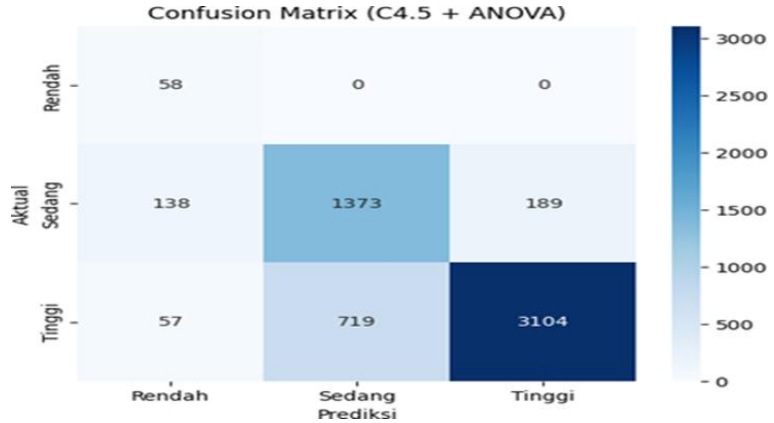

Figure 5. Confusion Matrix Decision Tree C4.5 + ANOVA

Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

Based on figure 5 the confusion matrix for the C4.5 + ANOVA model shown in the image, the following conclusions can be drawn:The model demonstrates excellent performance in predicting the "Tinggi" (High) class, with a high true positive count of 3104. Most instances actually belonging to the "High" class are correctly identified. However, there is some misclassification, with 57 "High" instances predicted as "Rendah" (Low) and 719 as "Sedang" (Medium).

For the "Sedang" (Medium) class, the model shows reasonably good predictive ability, correctly identifying 1373 instances. Nonetheless, a significant number of "Sedang" instances are misclassified as "Tinggi" (719 cases), indicating a tendency to over-predict the High class for actual Medium instances. Additionally, 189 "Sedang" instances are incorrectly predicted as "Rendah".

The "Rendah" (Low) class presents a different picture. While all 58 predicted "Rendah" instances are correct (resulting in no false positives for this class), the model struggles with false negatives. Specifically, 138 actual "Sedang" and 57 actual "Tinggi" instances are misclassified as "Rendah", suggesting the model has difficulty accurately identifying the Low class from other categories.

A clear class imbalance is evident in the data distribution. The "Tinggi" class dominates with 3880 instances, followed by "Sedang" at 1700, and "Rendah" with only 58 instances. This imbalance likely contributes to the model's high overall accuracy while potentially compromising its performance on minority classes.The most frequent confusion occurs between the "Sedang" and "Tinggi" classes, with 719 "Sedang" instances predicted as "Tinggi" and 189 "Tinggi" instances predicted as "Sedang". This pattern suggests the model finds distinguishing between Medium and High categories challenging.

For comprehensive evaluation, it is recommended to calculate precision, recall, and F1-score for each class, along with overall accuracy and balanced accuracy metrics. Given the class imbalance, additional metrics such as Matthews Correlation Coefficient (MCC) or multiclass ROC AUC would provide more insightful performance assessment.

3.2.4 Accuracy Comparison C4.5 VS C4.5 + ANOVA

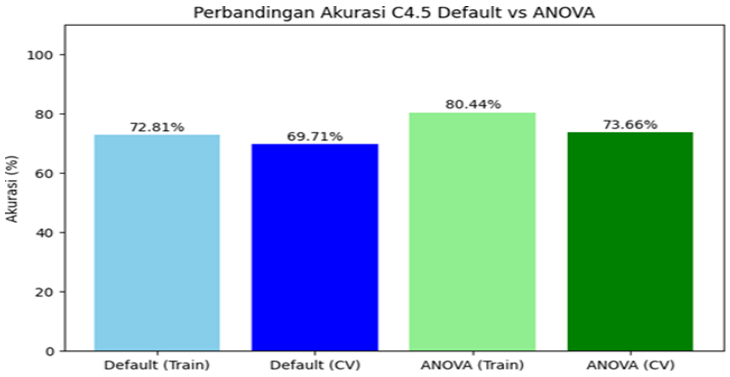

Figure 6. Accuracy Comparison C4.5 vs C4.5 + ANOVA

Based on figure 6,the visualization graph, the C4.5 model with default configuration produced an accuracy of 72.81% on the training set and 69.71% on cross-validation. On the other hand, the C4.5 model integrated with feature selection technique based on analysis of variance (ANOVA) showed improved performance, with an accuracy of 80.44% on the training data and 73.66% on cross-validation. This improvement indicates that the application of ANOVA as a feature selection method is able to enhance the model's generalization capacity, as reflected by the 3.95% increase in cross-validation accuracy, which is a crucial indicator of the model's ability to generalize patterns to previously unseen data.

3.2.5 Feature Importance Visualization

Based on figure 7,the horizontal bar chart in the Figure represents the feature importance of each predictor variable in the classification model based on the C4.5 decision tree that has been combined with the analysis of variance (ANOVA) based feature selection technique. The vertical axis lists the feature identities, while the horizontal axis shows the relative importance score in the range of 0 to 0.6. The Ranking feature stands out as the most dominant variable with an importance score approaching

Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

0.6, indicating that this feature contributes the most to the decision rule formation process in the model. The Total_Income and ALFA features occupy the next positions with scores of approximately 0.13 and 0.09 respectively, showing a secondary but still relevant influence.

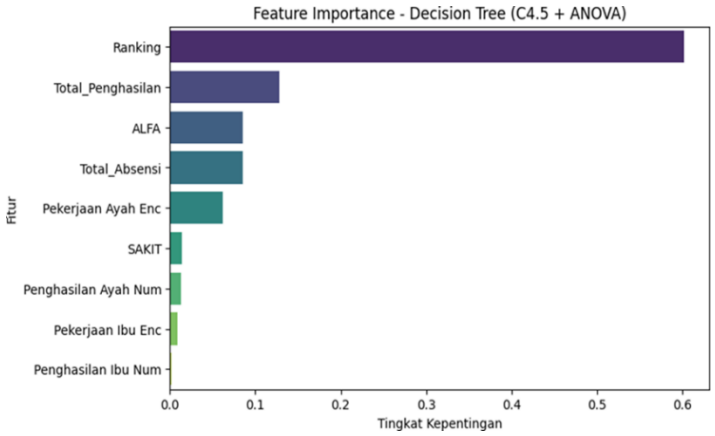

Figure 7. Feature Importance Visualization Comparison]

Kriteria Analisis	Faktor Dominan (Fitur)
ANOVA (P-Value)	Rangking, ALFA, Total_Absen
Decision Tree (Importance)	Rangkinf, ToTAL_Penghasilan, ALFA
Fitur Konsistensl	ALFA, Rangking

Figure 8. Identification of Dominant Features based on ANOVA Analysis and Decision Tree C4.5

Based on table 1,the ANOVA test measuring the statistical significance of the association between features and the target variable, the most significant factors are Ranking, ALFA, and Total_Absensi, as indicated by p-values far below the significance level $\alpha = 0.05$. On the other hand, based on the relative contribution in the decision tree model (measured through importance weight), the most influential factors are Ranking (weight 0.602), Total_Penghasilan (weight 0.128), and ALFA (weight 0.085). From the comparison of these two approaches, it was found that only two factors, namely ALFA and Ranking, consistently appear as dominant in both analyses. This shows that these two factors are not only statistically significant but also have the highest predictive weight in the model, thus truly playing an active role in determining student achievement categories. Therefore, ALFA and Ranking can be considered as key variables that significantly influence student academic achievement, making them worthy of being the main focus in data-based educational interventions or policie.

4. CONCLUSION

This research method used a data mining approach with stages including data cleaning, numeric conversion, missing value imputation, formation of derived variables, and categorization of the target variable "Achievement". The initial C4.5 model produced an accuracy of 72.81% on the training data and 69.71% on cross-validation. After feature selection was performed using ANOVA, one insignificant variable was eliminated, resulting in a hybrid C4.5+ANOVA model with nine main features. The test results showed an increase in accuracy to 80.44% on the training data and 73.66% on the cross-validation data, indicating an increase of 7.63 and 3.95 percentage points respectively. Furthermore, precision and recall for the "Low" class also increased significantly, indicating the model is more effective in detecting students at risk of low achievement.

Based on feature importance analysis and the ANOVA test, the variables most influential on student academic achievement are Ranking, Total Absences, and ALFA (number of unexplained absences), while Total Parental Income has a secondary influence. These results show that attendance discipline and student academic position are the main indicators of learning success. Overall, the integration of ANOVA with C4.5 has proven to be able to reduce overfitting, improve the model's

Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

generalization ability, and provide a scientific basis for schools in designing more effective learning interventions.The conclusion is the essence of the entire paper. Made in paragraph form, and not in list form. The conclusion does not repeat the sentences in the abstract. Of course. Here is a conclusion paragraph that incorporates the importance of the research to computer science, based on the text you provided.This study successfully demonstrates the practical and methodological value of integrating statistical feature selection with machine learning algorithms within the domain of educational data mining. By employing a hybrid C4.5+ANOVA model, the research not only identified key determinants of student achievement—namely academic ranking, attendance records, and unexplained absences—but also significantly enhanced the predictive model's performance. The marked increase in accuracy, precision, and recall, particularly for the critical "Low" achievement category, underscores the model's improved capability to identify at-risk students, thereby providing a actionable, data-driven foundation for targeted educational interventions. From a computer science perspective, this work makes a significant contribution by exemplifying how a hybrid approach can effectively mitigate the overfitting common in complex decision trees, thereby bolstering the model's generalizability and robustness on unseen data. It validates the principle that a pre-processing step with a statistically rigorous feature selection method like ANOVA can streamline models, enhance computational efficiency, and lead to more interpretable and reliable results. Ultimately, this research underscores the transformative potential of interdisciplinary methodology, where statistical analysis and machine learning converge to create intelligent systems that are not only more accurate but also more trustworthy and applicable to solving real-world problems.tolong terjemahkan ke dalam bahasa indonesia

ACKNOWLEDGEMENT

Thank you to my supervisor who has provided support in guiding me in completing this journal and also to my husband who has contributed his time and thoughts in completing this journal. I would like to convey my sincerest thanks for your mentorship, time, and invaluable insights throughout the process of completing this research journal. Your guidance has been incredibly meaningful and was a great help to me. Thank you once again for everything.

REFERENCES

[1] M. Arifin, Widowati, Farikhin, A. Wibowo, and B. Warsito, "Comparative Analysis on Educational Data Mining Algorithm to Predict Academic Performance," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarangin, Indonesia: IEEE, Sept. 2021, pp. 173–178. doi: 10.1109/iSemantic52711.2021.9573185.

[2] Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India; IEEE Senior Member, Symbiosis Institute of Technology, Pune, India, and A. Sharma, "Predicting Student Performance Using Educational Data Mining and Learning Analytics Technique," *J. Intell. Syst. Internet Things*, vol. 10, no. 2, pp. 24–37, 2023, doi: 10.54216/JISIoT.100203.

[3] C. Cui *et al.*, "Tri-Branch Convolutional Neural Networks for Top-k Focused Academic Performance Prediction," July 22, 2021, *arXiv*: arXiv:2107.10424. doi: 10.48550/arXiv.2107.10424.

[4] A. Amrein-Beardsley and J. Holloway, "Value-Added Models for Teacher Evaluation and Accountability: Commonsense Assumptions," *Educ. Policy*, vol. 33, no. 3, pp. 516–542, May 2019, doi: 10.1177/0895904817719519.

[5] V. Emslander, J. Levy, R. Scherer, and A. Fischbach, "Value-added scores show limited stability over time in primary school," *PLOS ONE*, vol. 17, no. 12, p. e0279255, Dec. 2022, doi: 10.1371/journal.pone.0279255.

[6] J. Levy, M. Brunner, U. Keller, and A. Fischbach, "Methodological issues in value-added modeling: an international review from 26 countries," *Educ. Assess. Eval. Account.*, vol. 31, no. 3, pp. 257–287, Aug. 2019, doi: 10.1007/s11092-019-09303-w.

[7] J. Liu, P. Peng, B. Zhao, and L. Luo, "Socioeconomic Status and Academic Achievement in Primary and Secondary Education: a Meta-analytic Review," *Educ. Psychol. Rev.*, vol. 34, no. 4, pp. 2867–2896, Dec. 2022, doi: 10.1007/s10648-022-09689-y.

Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

[8] X. Wang, M. Dai, and R. Mathis, "The influences of student- and school-level factors on engineering undergraduate student success outcomes: A multi-level multi-school study," *Int. J. STEM Educ.*, vol. 9, no. 1, p. 23, Dec. 2022, doi: 10.1186/s40594-022-00338-y.

[9] Y. Jang, S. Choi, H. Jung, and H. Kim, "Practical early prediction of students' performance using machine learning and eXplainable AI," *Educ. Inf. Technol.*, vol. 27, no. 9, pp. 12855–12889, Nov. 2022, doi: 10.1007/s10639-022-11120-6.

[10] W. Lastari, "Penerapan Data Mining Untuk Memprediksi Prestasi Siswa SMA Pada Dinas Pendidikan Provinsi Jambi," vol. 8, 2023, [Online]. Available: https://ejournal.unama.ac.id/index.php/jurnalmsi/article/view/864

[11] J. E. Ibarra-Esquer, B. L. Flores-Rios, M. A. Astorga-Vargas, A. C. Justo-Lopez, and G. E. Chavez-Valenzuela, "A Data-centric Approach to Tracking Student Academic Performance and Progression," *IAENG International Journal of Computer Science*, vol. 51, no. 12, pp. 1968–1979, 2024.

[12] M. Priyadharshini, S. Indra, S. Achuthan, and K. Lokesh, "Predicting Student Success: A Comparative Examination of Machine Learning Techniques," *Indian J. Comput. Sci. Technol.*, pp. 213–217, July 2024, doi: 10.59256/indjcst.20240302031.

[13] J. Ranellucci, N. C. Hall, K. R. Muis, S. P. Lajoie, and K. A. Robinson, "Mastery, Maladaptive Learning Behaviour, and Academic Achievement: An Intervention Approach," 2017.

[14] S. A. A. Kharis and A. H. A. Zili, "Learning Analytics dan Educational Data Mining pada Data Pendidikan," *J. Ris. PEMBELAJARAN Mat. Sekol.*, vol. 6, no. 1, pp. 12–20, Mar. 2022, doi: 10.21009/jrpms.061.02.

[15] T. Rahmat, "Pengaruh Kehadiran Siswa Terhadap Hasil belajar Matematika Kelas VIII MTsN 11 Agam Tahun Pelajaran 2021/2022".

[16] J. López-Zambrano, J. Lara Torralbo, and C. Romero, "Early Prediction of Student Learning Performance Through Data Mining: A Systematic Review," *Psicothema*, vol. 3, no. 33, pp. 456–465, Aug. 2021, doi: 10.7334/psicothema2021.62.

[17] L. Al-Alawi, J. Al Shaqsi, A. Tarhini, and A. S. Al-Busaidi, "Using machine learning to predict factors affecting academic performance: the case of college students on academic probation," *Educ. Inf. Technol.*, vol. 28, no. 10, pp. 12407–12432, Oct. 2023, doi: 10.1007/s10639-023-11700-0.

[18] M. Bellaj, A. Ben Dahmane, S. Boudra, and M. Lamarti Sefian, "Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance," *Int. J. Online Biomed. Eng. IJOE*, vol. 20, no. 03, pp. 55–74, Feb. 2024, doi: 10.3991/ijoe.v20i03.46287.

[19] P. M. Lyman and A. E. Olvido, "Exploring Variation in Student Academic Performance: Can Achievement in an Immersive Case Study Project Predict Exam Score in an Introductory Accounting Course?," *J. Scholarsh. Teach. Learn.*, vol. 20, no. 2, Oct. 2020, doi: 10.14434/josotl.v20i2.27648.

[20] M. Bellaj, A. Ben Dahmane, S. Boudra, and M. Lamarti Sefian, "Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance," *Int. J. Online Biomed. Eng. IJOE*, vol. 20, no. 03, pp. 55–74, Feb. 2024, doi: 10.3991/ijoe.v20i03.46287.

[21] S. Lee, C. Lee, K. G. Mun, and D. Kim, "Decision Tree Algorithm Considering Distances Between Classes," *IEEE Access*, vol. 10, pp. 69750–69756, 2022, doi: 10.1109/access.2022.3187172.

[22] V. Morosanova, T. Fomina, and I. Bondarenko, "Academic achievement: Intelligence, regulatory, and cognitive predictors," *Psychol. Russ.*, no. Query date: 2025-05-02 08:34:19, 2015, [Online]. Available: https://cyberleninka.ru/article/n/academic-achievement-intelligence-regulatory-and-cognitive-predictors

[23] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, "Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, p. 012012, Aug. 2019, doi: 10.1088/1742-6596/1255/1/012012.

[24] "ANALISIS PENERAPAN METODE ONE WAY ANOVA MENGGUNAKAN ALAT STATISTIK SPSS," *J. Ris. Akunt. Soedirman*, 2023, doi: 10.32424/1.jras.2023.2.2.10815.

Vol. 7, No. 2, April 2026, Page. 1116-1126
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5431

[25] L. Akbay, T. Akbay, O. Erol, and M. Kilinç, "Inadvertent Use of ANOVA in Educational Research: ANOVA is not A Surrogate for MANOVA," *Eğitimde Ve Psikolojide Ölçme Ve Değerlendirme Derg.*, vol. 10, no. 3, pp. 302–314, Sept. 2019, doi: 10.21031/epod.524511.

[26] M. L. Mouritsen, J. T. Davis, and S. C. Jones, "ANOVA Analysis of Student Daily Test Scores in Multi-Day Test Periods.," *Journal of Learning in Higher Education*.

[27] A. Huday and Zaehol Fatah, "PENERAPAN DECISION TREE C4.5 DALAM MEMPREDIKSI PREDIKAT TERBAIK DI MADRASAH TA'HILIYAH IBRAHIMY," *J. Ilm. Multidisiplin Ilmu*, vol. 2, no. 1, pp. 61–68, Feb. 2025, doi: 10.69714/be4q6n31.

[28] H. Rifa'i, Ryan Hamonangan, Dian Ade Kurnia, Kaslani, and Mulyawan, "Implementasi Algoritma Decision Tree Dalam Klasifikasi Kompetensi Siswa," *KOPERTIP J. Ilm. Manaj. Inform. Dan Komput.*, vol. 6, no. 1, pp. 15–20, June 2022, doi: 10.32485/kopertip.v6i1.131.

[29] M. S. Jailani and D. A. Saksitha, "TEHNIK ANALISIS DATA KUANTITATIF DAN KUALITATIF DALAM PENELITIAN ILMIAH," vol. Volume 15, Number 2, 2024 pp., pp. 79–91.

[30] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 905–971, Jan. 2023, doi: 10.1007/s10639-022-11152-y.

[31] N. Abdillah and F. Yuniko, "Performance analysis of data mining classification methods using c4.5 algorithm for student graduation prediction (case study at syedza saintika stikes)," *PUBLIC Health*.

[32] J. H. Yam and R. Taufik, "Hipotesis Penelitian Kuantitatif," *Perspekt. J. Ilmu Adm.*, vol. 3, no. 2, pp. 96–102, Aug. 2021, doi: 10.33592/perspektif.v3i2.1540.

[33] C. Bentéjac, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, 2021, doi: 10.1007/s10462-020-09896-5.

[34] K. Anam, B. Nurhakim, and C. Juliane, "Komparasi Algoritma Klasifikasi Data Mining Menggunakan Optimize Selection untuk Peminatan Program Studi," *Build. Inform. Technol. Sci. BITS*, vol. 4, no. 2, pp. 606–613, Sept. 2022, doi: 10.47065/bits.v4i2.2160.

[35] K. M. Unertl, L. L. Novak, K. B. Johnson, and N. M. Lorenzi, "Traversing the many paths of workflow research: developing a conceptual framework of workflow terminology through a systematic literature review," *J. Am. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 265–273, May 2010, doi: 10.1136/jamia.2010.004333.

[36] M. Faridl, "PROGRAM STUDI MAGISTER PSIKOLOGI SAINS DIREKTORAT PROGRAM PASCASARJANA UNIVERSITAS MUHAMMADIYAH MALANG".

[37] A. D. Madden, "A review of basic research tools without the confusing philosophy," *High. Educ. Res. Dev.*, vol. 41, no. 5, pp. 1633–1647, July 2022, doi: 10.1080/07294360.2021.1920895.