

JELTec (Journal of Learning Technology)
JELTec, Volume 04, Number 02, September 2026
E ISSN: 2986-3538; P ISSN: 2986-4364, pp. 101-114
https://journal.yhmm.or.id/index.php/JELTec
DOI: 10.56327/**JELTec**.v13i2.1297

Predicting University Student Academic Achievement Using the C4.5 Decision Tree Based on Socioeconomic Factors

Muhamad Muslihudin[1], Agustina[2], Tuti Puspitasari[3]
[1,2]Information System of Study Program, Institut Bakti Nusantara, Lampung
[3]Islamic Education Management of Study Program, STIT Pringsewu, Lampung
[1,2]Wisma Rini Street, No.09 Pringsewu, Lampung, Indonesia
[3]Irigasi Street, Wonodadi, Gadingrejo, Pringsewu, Lampung, Indonesia
E-Mail: mmuslihudin415@gmail.com[1], tinajeon05@gmail.com[2], tutipuspita12@gmail.com[3]

Article history:
Received: July 26, 2026
Revised: August 08, 2026
Accepted: August 28, 2026

Corresponding authors
***authors@email.ac.id**

Keywords:
Data Mining;
C4.5 Algorithm;
Learning Achievement;
Students;
Socioeconomics.

Abstract
Student academic achievement is an important indicator in assessing the success of higher education; however, it is often influenced by external factors such as socio-economic background, including parents' occupation and income. Differences in these conditions can lead to disparities in academic performance if not properly identified. This study aims to identify student academic achievement based on socio-economic background using a data mining approach with the C4.5 algorithm. The research process includes data collection through observation, interviews, questionnaires, and documentation, followed by data preprocessing, attribute selection, and classification model development using the Orange application on 510 student records with variables such as gender, Grade Point Average (GPA), parents' occupation, and parents' income. The results indicate that the C4.5 algorithm is capable of generating a decision tree model that identifies patterns between socio-economic factors and academic achievement, with GPA as the most dominant attribute. This model can be utilized as a basis for academic policy-making, such as providing educational assistance and academic support programs, and may also serve as a reference for future research to achieve more accurate and comprehensive analyses.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Higher education plays a fundamental role in strengthening human capital, promoting social mobility, and supporting sustainable economic development. In Indonesia, the implementation of higher education is governed by Law Number 20 of 2003 concerning the National Education System and Law Number 12 of 2012 concerning Higher Education, which emphasize equitable access, educational quality, and the development of students' intellectual and professional competencies[1]. Furthermore, Article 31 of the 1945

Constitution of the Republic of Indonesia establishes education as a fundamental right, reinforcing the government's responsibility to provide inclusive educational opportunities. Within this regulatory framework, academic achievement, commonly measured through the Grade Point Average (GPA), represents an important indicator of students' learning outcomes and academic progression. Nevertheless, academic achievement is influenced not only by intellectual abilities and individual learning motivation but also by socioeconomic circumstances that shape students' educational opportunities. The socioeconomic dimension is particularly relevant in Indonesia, where disparities in household income, parental employment, educational resources, and access to digital technologies may contribute to unequal learning conditions. These challenges are consistent with international research demonstrating that socioeconomic inequalities remain closely associated with educational achievement and the reproduction of educational disadvantages across generations [2]. Consequently, understanding socioeconomic disparities in higher education is essential for developing evidence-based academic interventions and strengthening educational equity.

Lampung Province provides a relevant regional context for investigating the relationship between socioeconomic characteristics and university students' academic achievement. The province accommodates various public and private higher education institutions, including universities, institutes, polytechnics, and specialized colleges, serving students from diverse geographical and socioeconomic backgrounds. This institutional diversity creates opportunities for educational participation while simultaneously introducing challenges associated with unequal household financial resources and academic support systems. Students from economically disadvantaged households may experience difficulties in obtaining learning materials, maintaining reliable internet connectivity, accessing appropriate digital devices, and meeting transportation and living expenses. Such limitations may affect learning engagement, academic persistence, and ultimately GPA. Conversely, students from financially stable households may benefit from greater access to educational facilities and supportive learning environments, although economic advantages do not necessarily guarantee superior academic performance. Research by Sirin [1], [3] established a significant relationship between socioeconomic status and academic achievement, while Richardson et al. [4] demonstrated that university academic performance is shaped by multiple psychological and contextual determinants. These findings suggest that socioeconomic conditions should be examined alongside the broader complexity of student achievement rather than treated as isolated or deterministic predictors. Accordingly, a systematic analysis of socioeconomic characteristics in the Lampung higher education context is necessary to identify potentially vulnerable student groups and support more equitable institutional decision-making.

Socioeconomic status is a multidimensional construct generally represented by household income, parental occupation, parental educational attainment, household composition, and access to material resources. In Indonesia, household economic conditions are commonly examined through income and expenditure indicators, poverty thresholds, and household welfare characteristics reported by Statistics Indonesia (BPS). However, socioeconomic classification requires careful operationalization because national poverty thresholds and expenditure-based welfare measures cannot automatically be equated with fixed parental income categories. Within higher education, parental income and employment status may serve as measurable indicators of financial capacity, particularly in relation to tuition affordability, learning facilities, and the ability to sustain educational participation. Indonesian higher education policy also recognizes economic accessibility through tuition-related mechanisms and financial assistance programs, including the Kartu Indonesia Pintar Kuliah (KIP Kuliah) initiative. International evidence

further highlights the relationship between family socioeconomic background and educational outcomes. Rodríguez-Hernández et al. [5], through a systematic review and meta-analysis, identified socioeconomic status as a relevant determinant of academic performance in higher education, while Musso et al. [6] demonstrated the potential of predictive analytics for identifying patterns associated with university academic performance. Nevertheless, the relationship between socioeconomic characteristics and GPA may vary across institutional environments, demographic groups, and educational contexts. Therefore, an empirical classification approach is needed to determine which socioeconomic attributes are most informative in distinguishing academic achievement categories among university students.

Educational data mining and machine learning have increasingly been employed to identify patterns associated with student achievement and to support early academic intervention. Among the available classification techniques, the C4.5 decision tree algorithm offers practical advantages because it generates interpretable decision rules, accommodates categorical and numerical attributes, and selects relevant predictors through information gain ratio. Such characteristics make it suitable for examining socioeconomic variables, including parental income, occupation, and household economic conditions. Previous research by Musso et al. [6] demonstrated the potential of artificial neural networks for predicting academic performance, while Romero and Ventura [7] highlighted the growing contribution of educational data mining to understanding student learning behavior and institutional decision-making. Similarly, Alyahyan and Düştegör [8] emphasized the importance of predictive analytics in identifying students at risk of academic underperformance and improving institutional support strategies. Although these studies demonstrate the usefulness of machine learning in educational prediction, predictive performance alone does not necessarily provide transparent explanations of how socioeconomic attributes contribute to academic classifications. Furthermore, models developed using extensive academic, behavioral, and psychological variables may not be directly applicable to institutions with limited data availability. Therefore, the C4.5 algorithm offers a potentially valuable alternative for developing an interpretable classification framework that connects socioeconomic characteristics with academic achievement categories while producing decision rules that educational administrators can understand and evaluate.

Despite the growing literature on socioeconomic inequalities and educational prediction, several research gaps remain. First, existing systematic reviews have established associations between socioeconomic status and academic achievement, but the strength and direction of these relationships vary across educational contexts, suggesting that findings cannot be generalized without institution-specific validation [1], [2]. Second, previous predictive studies have frequently emphasized overall classification accuracy or incorporated extensive combinations of academic and behavioral variables, leaving opportunities to investigate the predictive contribution of socioeconomic attributes independently. Third, although decision tree algorithms provide interpretable classification structures, the extent to which parental income and related socioeconomic indicators can reliably distinguish university students' GPA categories within the regional context of Lampung requires further empirical examination. Fourth, the practical usefulness of socioeconomic-based classification models for supporting equitable academic assistance and student welfare policies remains an important research consideration. These limitations establish the central research problem: to what extent can socioeconomic factors predict university students' academic achievement, which socioeconomic attributes provide the greatest discriminatory information, and how effectively can the C4.5 decision tree classify students into GPA-based achievement categories? Addressing these questions is particularly important because socioeconomic disadvantage should not automatically be

interpreted as evidence of low academic capability. A predictive framework must therefore be evaluated not only in terms of classification performance but also in relation to interpretability, potential socioeconomic bias, and the responsible use of student information.

Based on these considerations, this study aims to develop and evaluate an interpretable academic achievement prediction model using the C4.5 decision tree algorithm, with socioeconomic characteristics serving as the primary predictor variables. Specifically, the research seeks to identify relevant socioeconomic attributes, examine the predictive contribution of parental income and other available household characteristics, classify university students into predefined academic achievement categories based on their Grade Point Average (GPA), and extract decision rules describing the relationships between socioeconomic conditions and academic outcomes. The proposed model will be evaluated using appropriate classification performance measures, including accuracy, precision, recall, and F1-score, supported by a confusion matrix and a suitable validation strategy to assess predictive reliability. The anticipated contribution of this research lies in the development of a context-specific and interpretable socioeconomic-based academic classification framework for higher education institutions in Lampung Province. Unlike approaches that prioritize predictive accuracy without adequately explaining classification mechanisms, this study emphasizes the transparency of decision rules and their potential relevance to educational decision-making. The findings are expected to contribute theoretically to educational data mining by providing empirical evidence concerning the predictive utility of socioeconomic attributes and practically to university management by supporting the identification of students who may require additional financial assistance, academic mentoring, or targeted educational support. Ultimately, the proposed framework is intended to complement, rather than replace, comprehensive academic assessment and to promote more inclusive, transparent, and evidence-informed higher education policies.

II. RESEARCH METHOD

2.1. Data Collection

This study employed a quantitative research approach supported by multiple data collection techniques to investigate the relationship between students' socioeconomic backgrounds and academic achievement at Institut Bakti Nusantara (IBN), Lampung, Indonesia. The data collection process was designed to obtain reliable, relevant, and structured information for developing an academic achievement prediction model using the C4.5 Decision Tree algorithm. Five complementary techniques were employed: observation, interviews, questionnaires, documentation, and literature review. These methods were integrated to identify socioeconomic indicators, obtain students' academic records, verify data consistency, and establish the theoretical foundation for predictive modeling. The primary variables included parental income, parental occupation, and other available socioeconomic characteristics, while students' Grade Point Average (GPA) served as the basis for defining academic achievement categories. The integration of primary and secondary data sources was intended to improve data completeness and minimize inconsistencies before the classification process.

a) Observation

Observation was conducted directly at Institut Bakti Nusantara to examine the institutional environment, student academic information management, and the availability of socioeconomic data relevant to the research objectives. This technique involved systematically identifying how student information was collected, maintained, and accessed within the institution. Particular attention was given to the availability of academic performance records, socioeconomic information, and administrative procedures related to student data management. The observation

process also supported the identification of potential data quality issues, including missing information, inconsistent attribute formats, and incomplete student profiles. The findings were used to establish the research context, determine the feasibility of data collection, and support the development of a structured dataset suitable for C4.5-based classification. Observation was treated as a contextual and procedural data collection technique rather than a direct measurement of students' household economic conditions.

b) Interviews

Semi-structured interviews were conducted with relevant institutional representatives at Institut Bakti Nusantara to obtain contextual information concerning student academic records, socioeconomic data availability, institutional data management practices, and research authorization procedures. The interviews focused on identifying the sources of student information, understanding GPA assessment procedures, clarifying socioeconomic indicators, and determining the institutional requirements for accessing academic data. The interview process also aimed to identify potential limitations in the existing information system, particularly regarding data completeness, confidentiality, and consistency. Information obtained through interviews was used to support the selection of research variables, validate the interpretation of institutional records, and establish appropriate procedures for integrating questionnaire responses with academic documentation. The interview findings were not treated as direct classification inputs unless the information was converted into clearly defined, student-level variables supported by verifiable data.

c) Questionnaires

A structured questionnaire was administered to students of Institut Bakti Nusantara to collect primary data concerning their socioeconomic backgrounds. The questionnaire was designed to obtain standardized information on parental income, parental occupation, parental educational attainment, and other relevant household characteristics, subject to the availability and scope of the research instrument. Closed-ended questions and predefined response categories were used to facilitate data coding, minimize variations in interpretation, and support subsequent computational analysis. Parental income was recorded using clearly defined monetary ranges, while occupational and educational characteristics were represented through categorical attributes. Before model development, questionnaire responses were examined for completeness, consistency, duplicate records, and potentially invalid entries. The collected data were subsequently encoded and integrated with verified academic achievement information to construct the classification dataset. Participation was based on informed consent, and identifying information was excluded from the analytical dataset to protect respondents' confidentiality. The resulting socioeconomic attributes served as predictor variables in the C4.5 Decision Tree model.

d) Documentation

Document analysis was employed to obtain secondary data from institutional academic records and other relevant administrative documents. The primary purpose of this technique was to retrieve and verify students' academic achievement information, particularly their cumulative GPA, which served as the dependent variable in the classification model. Relevant documentation also included student enrollment information, institutional data management records, and supporting materials required to validate the research dataset. Academic records were matched with questionnaire responses using authorized identifiers before personal information was removed or replaced with anonymized codes. The documentation

process emphasized data accuracy, consistency, and traceability, ensuring that academic achievement categories were derived from verified institutional records rather than solely from students' self-reported information. Supporting documentation, including research authorization records and fieldwork materials, was maintained to strengthen the transparency and reproducibility of the research procedures. Access to academic records was subject to institutional approval and applicable personal data protection requirements, including Indonesia's Law Number 27 of 2022 concerning Personal Data Protection.

e) Literature Review

A structured literature review was conducted to establish the theoretical and methodological foundation of the study. Relevant scientific publications were identified through academic databases and scholarly sources, with particular emphasis on peer-reviewed international journal articles indexed in Scopus. The literature search focused on socioeconomic status, parental income, academic achievement prediction, educational data mining, decision tree classification, and the C4.5 algorithm. Priority was given to recent studies published within the preceding five years, while foundational publications were considered where necessary to explain established concepts and classification methods. The selected literature was critically examined to identify commonly used socioeconomic indicators, GPA classification strategies, data preprocessing techniques, predictive performance metrics, and methodological limitations in previous research. The review also supported the identification of the research gap, particularly the limited context-specific evidence regarding the interpretability and predictive contribution of socioeconomic characteristics to university students' academic achievement. The findings provided the basis for selecting classification attributes, determining model evaluation procedures, and positioning the proposed research within the broader field of educational data mining.

2.2. Data Integration and Preparation

Following the data collection process, the information obtained from questionnaires and institutional academic records was consolidated into a structured dataset for predictive analysis. Each observation represented an individual student, while the predictor attributes described the student's socioeconomic background. Academic achievement was represented by the student's cumulative GPA, which was subsequently transformed into predefined classification categories using explicitly documented thresholds. Data preparation involved examining missing values, removing duplicate observations, checking inconsistencies, standardizing socioeconomic categories, and encoding attributes where necessary. The prepared dataset was then used to construct the C4.5 Decision Tree model, which identifies informative classification attributes using the gain ratio criterion and generates interpretable decision rules[9], [10], [11]. To evaluate the model's generalization performance, data partitioning and validation procedures should be performed without allowing information from the evaluation subset to influence model development. Classification performance was assessed using accuracy, precision, recall, F1-score, and a confusion matrix. This integrated procedure was designed to ensure that the predictive analysis was based on systematically collected, verified, and appropriately processed student data.

2.3. C4.5 Algorithm

This study employs a quantitative data mining approach using the C4.5 Decision Tree classification algorithm to predict university students' academic achievement based on socioeconomic characteristics. The C4.5 algorithm, introduced by Quinlan in 1993 as an

extension of the Iterative Dichotomiser 3 (ID3) algorithm, constructs a decision tree by recursively partitioning datasets according to the attributes that provide the highest gain ratio. Unlike its predecessor, C4.5 can handle both categorical and continuous attributes, accommodate missing values, and apply tree-pruning mechanisms to reduce overfitting and improve model generalization[12], [13], [14], [15], [16]. These capabilities make the algorithm particularly relevant for educational datasets containing heterogeneous socioeconomic information, such as parental income, parental occupation, household conditions, and access to educational resources. Previous studies have demonstrated the applicability of decision tree classification in identifying patterns associated with student performance and supporting educational decision-making [5], [6]. In this study, socioeconomic characteristics serve as independent predictor variables, while academic achievement, measured using the Grade Point Average (GPA), represents the target variable. The resulting decision tree is expected to generate interpretable classification rules that explain how combinations of socioeconomic characteristics are associated with different academic achievement categories.

The research methodology consists of several interconnected stages, including data collection, attribute identification, data preprocessing, academic achievement categorization, model development, and performance evaluation. Student socioeconomic information is obtained through structured questionnaires, while GPA records are collected from authorized institutional academic documentation at Institut Bakti Nusantara, Lampung, Indonesia. During preprocessing, the dataset is examined for missing values, duplicate observations, inconsistent records, and inappropriate attribute representations. Continuous socioeconomic attributes, such as parental income, may be processed directly through the threshold-based splitting mechanism of C4.5, while categorical attributes are organized according to their respective classifications. The algorithm subsequently evaluates candidate attributes using entropy, information gain, and gain ratio to identify the most informative decision nodes. Model performance is assessed through a suitable training and testing strategy using accuracy, precision, recall, F1-score, and a confusion matrix. The methodological design emphasizes both predictive reliability and interpretability, enabling the resulting classification rules to provide meaningful insights for institutional academic monitoring. To establish the theoretical foundation for attribute selection, Table 2.1 summarizes the variables considered in four previous studies examining student academic achievement and classification.

Table 2.1. Variables Used in Previous Studies on Student Academic Achievement

No.	Authors and Year	Research Variables
1	Nanda, Antoni, and Sarwo (2025)	Attendance, assignment scores, midterm examination scores, final examination scores, classroom participation, and academic achievement.[17]
2	Naufal Shabir and Wulan Nursafitri (2024)	Gender, age, study program, previous semester GPA, attendance, assignment scores, midterm examination scores, and final examination scores.[18]
3	Ning, Atikah, Radix, and Suhendra (2023)	Gender, residential conditions, parental socioeconomic background, social assistance or scholarship status, household facilities, and GPA.[19]
4	Daru and Gatra (2022)	GPA in the first, second, third, and fourth semesters.[19]

The comparative analysis presented in Table 2.1 indicates that previous studies have employed different combinations of academic, demographic, and socioeconomic variables to classify or predict university students' academic performance. Nanda et al. (2025) and Naufal Shabir and Nursafitri (2024) primarily emphasized academic and behavioral

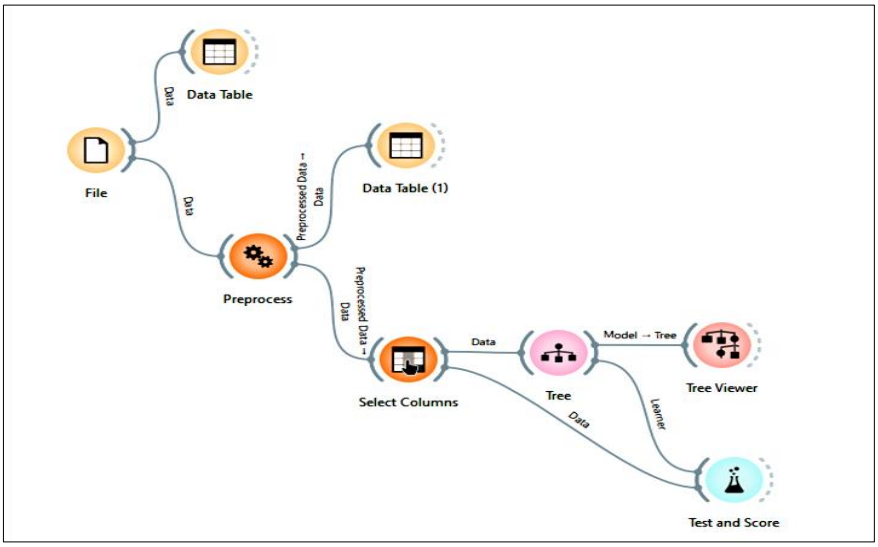
indicators, including attendance, examination results, and previous academic performance. Similarly, Daru and Gatra (2022) concentrated on historical GPA records to identify patterns of academic progression. In contrast, Ning et al. (2023) incorporated socioeconomic characteristics, including parental background, residential conditions, financial assistance, and household facilities. Although these studies provide valuable methodological foundations, they also highlight the need to investigate the predictive contribution of socioeconomic factors without relying predominantly on previous academic achievement indicators. Accordingly, the present study focuses on developing an interpretable C4.5 classification model based on socioeconomic characteristics, particularly parental income and other available household-related attributes, to predict GPA-based academic achievement categories among students at Institut Bakti Nusantara. The proposed research contributes to educational data mining by examining the discriminatory power of socioeconomic variables, identifying the most informative attributes through gain ratio analysis, and generating transparent decision rules that may support targeted financial assistance, academic mentoring, and evidence-based student support policies. However, the resulting associations will be interpreted as predictive relationships rather than evidence of direct causal effects.

III. RESULTS AND DISCUSSION

3.1. Research Results

The experimental dataset consisted of 510 student records collected from Institut Bakti Nusantara (IBN), Lampung, Indonesia, through the integration of primary data obtained from structured questionnaires and secondary data retrieved from institutional academic documentation. The dataset included five principal attributes: gender, cumulative Grade Point Average (GPA), academic achievement category, parental occupation, and parental income. The socioeconomic attributes were used to investigate potential patterns associated with students' academic performance. Data processing and classification were conducted using Orange Data Mining software through a workflow comprising data importation, preprocessing, attribute configuration, decision tree construction, and model visualization. The academic achievement categories were defined using a GPA threshold of 3.00, distinguishing students classified as achieving from those classified as non-achieving. The reported average GPA across the 510 observations was approximately 3.40, indicating that the overall academic performance of the sampled students was above the specified classification threshold. However, this descriptive average does not independently establish the predictive performance of the classification model.

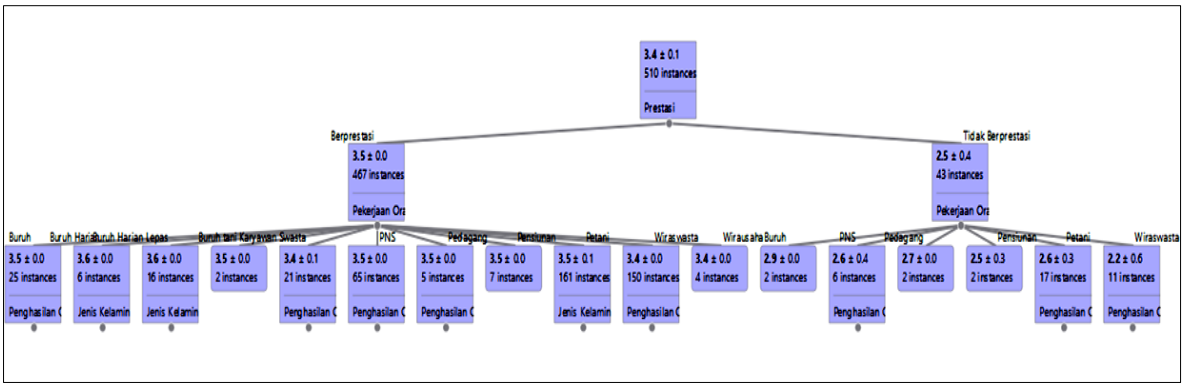
.



Gambar 3.1. *Workflow Orange*

Figure 3.1 illustrates the data processing workflow used to develop and visualize the decision tree classification model. The workflow begins with importing the student dataset into Orange Data Mining, followed by preprocessing to examine data completeness, consistency, and attribute representation. The attribute configuration stage determines the target variable and the predictors used for classification. Subsequently, the Tree widget generates a decision tree model by recursively partitioning the dataset according to the available attributes. The Tree Viewer widget provides a visual representation of the resulting decision structure, allowing the identification of splitting attributes, classification branches, and terminal nodes. This workflow facilitates the interpretation of relationships between socioeconomic characteristics and student academic achievement. However, the Orange Tree learner's splitting and pruning settings must be documented before the implementation can be described as an exact reproduction of Quinlan's C4.5 algorithm.

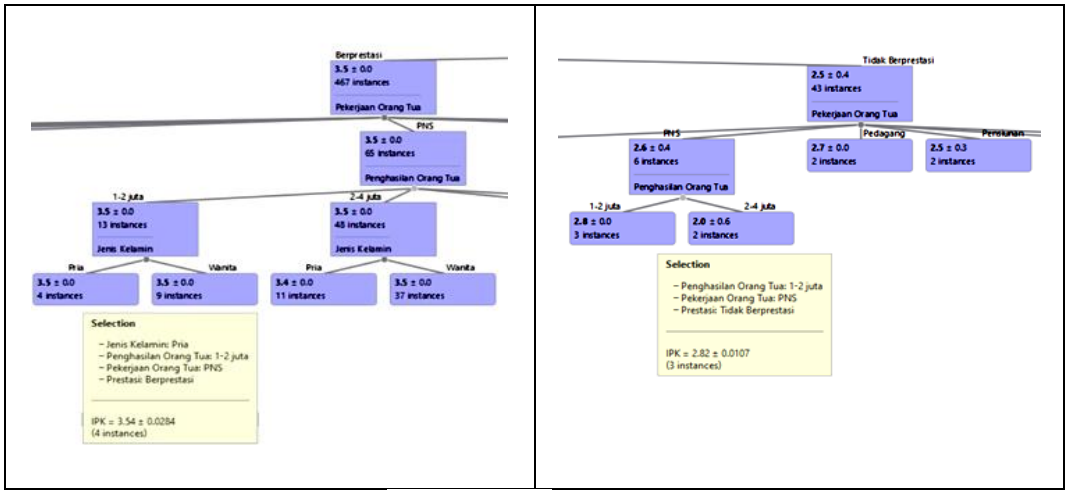
The decision tree visualization provides an overview of the classification structure generated from the 510 student records. Based on the reported model, the initial decision split was associated with the academic achievement status attribute, followed by additional partitions involving parental occupation, parental income, and gender. Parental occupations represented in the dataset included civil servants, traders, farmers, laborers, and other employment categories. These branches illustrate variations in the socioeconomic backgrounds of students across academic achievement groups. The resulting decision tree therefore offers a structured representation of how the selected variables partition the observations into increasingly specific subgroups. Nevertheless, because academic achievement status was defined from GPA, including this status as an input variable when predicting GPA introduces target leakage. Consequently, the original tree structure should be interpreted as an exploratory description of the dataset rather than valid evidence of socioeconomic-based predictive accuracy.



Gambar 3.2. *Tree Viewer*

Figure 3.2 presents the main classification structure reported by the Tree Viewer. The root node separates students according to their existing academic achievement category, while subsequent branches distinguish observations based on parental employment and household income characteristics. The visualization suggests that students belonging to the same achievement group may have heterogeneous socioeconomic backgrounds. This pattern is relevant because it indicates that socioeconomic conditions do not correspond to a single uniform academic outcome. However, the position of a variable in the tree should not automatically be interpreted as its causal importance. In particular, the dominant position of academic achievement status is expected when that variable is derived directly from GPA. To establish the actual predictive contribution of socioeconomic characteristics,

the model must be reconstructed using only socioeconomic and other legitimately available predictor variables, with the GPA-derived achievement category retained exclusively as the target.



Gambar 3.3. *Tree Viewer* 2

Figure 3.3 highlights the variation in academic achievement among students with comparable socioeconomic characteristics. The presence of both achieving and non-achieving students within the same parental occupation and income category indicates that household financial conditions alone are insufficient to account for all observed differences in GPA. Other variables not included in the dataset, such as learning motivation, study habits, academic preparedness, access to learning resources, and institutional support, may also be relevant. These findings are consistent with the broader understanding that university academic achievement is a multidimensional phenomenon shaped by interactions between individual and contextual factors. Nevertheless, the present analysis cannot determine the independent effects of these unmeasured variables. The principal contribution of the decision tree visualization is therefore its ability to reveal descriptive classification patterns and identify socioeconomic subgroups that may warrant further investigation.

The analysis of 510 student records demonstrates the feasibility of organizing academic and socioeconomic information into a structured dataset for decision tree analysis using Orange Data Mining. The reported visualization identified parental occupation, parental income, and gender as attributes involved in subgroup partitioning, while also revealing different academic outcomes among students with similar socioeconomic backgrounds. These observations provide preliminary evidence of heterogeneity in the relationship between socioeconomic characteristics and academic achievement. However, the original model configuration contains a methodological limitation because GPA-derived academic achievement status was included among the predictors. Therefore, its predictive validity cannot be established from the reported tree alone. A revised model that excludes GPA and its derived categories from the predictor set, together with appropriate out-of-sample evaluation, is necessary before making conclusions about classification accuracy, the relative importance of socioeconomic attributes, or the model's suitability for institutional decision-making.

Important methodological corrections for the manuscript: The reported workflow should be revised so that academic achievement category is the sole classification target and GPA is used only to derive that category, not as an input predictor. The rule for students with GPA exactly equal to 3.00 must also be specified. In addition, Orange's standard Tree widget should not be described as an exact C4.5 implementation without

verifying its splitting criterion and pruning procedure. Finally, the original Orange screenshots, confusion matrix, class distribution, and validation results should be inserted into the manuscript to substantiate the reported findings. The diagrams above are explanatory reconstructions, not the original research figures.

3.2. Model Performance Evaluation and Analysis

The performance of the Decision Tree model was evaluated using the Test and Score feature in Orange Data Mining, employing a cross-validation procedure on a dataset comprising 510 student records from Institut Bakti Nusantara, Lampung, Indonesia. The evaluation aimed to determine the model's ability to predict students' cumulative Grade Point Average (GPA) based on the attributes included in the dataset. Five regression evaluation metrics were employed: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2). These metrics provide complementary perspectives on prediction errors, model reliability, and the proportion of GPA variability accounted for by the model. The evaluation results are presented in Table 3.1.

Table 3.1. Performance Evaluation of the Decision Tree Model

Model	MSE	RMSE	MAE	MAPE	R^2
Decision Tree	0.108	0.329	0.223	0.083	0.256

The experimental results indicate that the Decision Tree model achieved an MSE of 0.108 and an RMSE of 0.329, suggesting that the model's predictions deviated from the observed GPA values by approximately 0.329 GPA points in root mean squared terms. The MAE of 0.223 indicates an average absolute prediction error of approximately 0.223 GPA points. Meanwhile, the MAPE of 0.083, equivalent to 8.3%, indicates the average relative prediction error, assuming that the metric was calculated as a proportion rather than a percentage. These findings suggest that the model captures certain patterns associated with students' academic performance, although the magnitude of the prediction errors must be interpreted relative to the GPA distribution and appropriate baseline models. In particular, the reported error values cannot independently establish that the model performs well without comparison against simpler prediction approaches, such as predicting the training-set mean GPA. Furthermore, the observed prediction errors indicate that the socioeconomic and demographic characteristics included in the model do not fully account for individual differences in academic achievement.

The coefficient of determination (R^2) provides additional insight into the model's explanatory performance. The obtained R^2 value of 0.256 indicates that the model accounts for approximately 25.6% of the variation in students' GPA relative to the mean-prediction baseline under the evaluation procedure used. This finding suggests that the selected attributes provide limited but potentially meaningful predictive information concerning academic performance. However, the remaining unexplained variation should not automatically be attributed to specific factors without additional empirical evidence. Other potentially relevant determinants include prior academic achievement, learning motivation, study habits, parental educational background, psychological characteristics, institutional learning environments, and access to educational resources. The relatively modest R^2 value highlights the complexity of academic achievement and suggests that socioeconomic indicators alone may be insufficient to develop a highly reliable prediction model. Consequently, future model development could incorporate additional theoretically justified predictors and compare Decision Tree performance with alternative machine learning techniques to determine whether predictive accuracy and generalization can be improved.

Overall, the evaluation demonstrates that the Decision Tree model provides an initial quantitative framework for examining the predictability of university students' GPA using

the available research attributes. The reported RMSE of 0.329, MAE of 0.223, MAPE of 8.3%, and R^2 of 0.256 indicate that the model has some predictive capacity but also substantial limitations in explaining differences in academic achievement. These results contribute to the research objectives by providing empirical performance measures and highlighting the need for more comprehensive predictor selection and model validation. Nevertheless, an important methodological distinction must be addressed: MSE, RMSE, MAE, MAPE, and R^2 are regression metrics, not classification accuracy measures. Therefore, the reported results support an evaluation of numerical GPA prediction rather than a validated C4.5 classification of academic achievement categories. To align the evaluation with the proposed classification research, academic achievement should be defined as the target category, GPA and its derived attributes must be excluded from the predictor set, and model performance should be assessed using classification metrics such as accuracy, precision, recall, F1-score, and a confusion matrix.

The findings of this study are consistent with previous research emphasizing the multidimensional relationship between socioeconomic background and university students' academic achievement. Rodríguez-Hernández et al. [2], through a systematic review of socioeconomic status and academic performance in higher education, demonstrated that socioeconomic characteristics are relevant to academic outcomes, although their relationships vary depending on institutional contexts, measurement approaches, and student populations. Similarly, Richardson et al. [3] identified multiple psychological and contextual factors associated with university academic performance, indicating that student achievement cannot be adequately explained by a limited set of demographic or socioeconomic indicators. These findings provide a theoretical basis for interpreting the present study's coefficient of determination ($R^2 = 0.256$), which indicates that the Decision Tree model accounted for approximately 25.6% of the observed GPA variation relative to the mean-prediction baseline. The remaining unexplained variation suggests that other potentially relevant factors, including learning motivation, prior academic performance, self-regulated learning, and institutional support, should be investigated. Therefore, the present findings reinforce the importance of considering socioeconomic conditions as one component of a broader academic achievement prediction framework rather than as the sole determinant of student success.

Furthermore, the application of the Decision Tree algorithm in this study is supported by Alyahyan and Düştegör [6], who emphasized that educational data mining techniques can facilitate early identification of students at risk of academic underperformance and support evidence-based institutional interventions. Their review highlighted the importance of selecting appropriate predictor variables, implementing systematic data preprocessing, and evaluating predictive models using reliable validation procedures. Similarly, Romero and Ventura [5] demonstrated that educational data mining and learning analytics provide valuable approaches for discovering patterns in educational datasets and improving academic decision-making. In the present study, the Decision Tree model produced an RMSE of 0.329, an MAE of 0.223, and an R^2 of 0.256, suggesting that the available attributes provide some predictive information but do not fully explain variations in students' GPA. Compared with the broader methodological recommendations of previous studies, these findings highlight the importance of improving feature selection, incorporating additional academic and behavioral variables, and conducting comparative evaluations with alternative machine learning algorithms. The principal potential contribution of this research is the development of an interpretable socioeconomic-based prediction framework for university students in Lampung, although further validation is required to establish its predictive reliability and practical effectiveness

CONCLUSION

This study investigated the prediction of university students' academic achievement based on socioeconomic characteristics using a Decision Tree algorithm and a dataset of 510 students from Institut Bakti Nusantara, Lampung, Indonesia. The research integrated primary data collected through structured questionnaires and secondary data obtained from institutional academic records, focusing on parental occupation, parental income, gender, and Grade Point Average (GPA). The experimental results obtained using Orange Data Mining demonstrated that the Decision Tree model achieved a Mean Squared Error (MSE) of 0.108, Root Mean Squared Error (RMSE) of 0.329, Mean Absolute Error (MAE) of 0.223, Mean Absolute Percentage Error (MAPE) of 8.3%, and a coefficient of determination (R^2) of 0.256. These findings indicate that the model captured certain patterns associated with students' academic performance but accounted for only 25.6% of GPA variation relative to the mean-prediction baseline. The analysis also revealed that students with similar parental occupations and income levels could exhibit different academic outcomes, suggesting that socioeconomic characteristics alone are insufficient to explain variations in academic achievement. Consequently, the findings highlight the multidimensional nature of academic performance and the potential relevance of additional academic, behavioral, and institutional factors. The study contributes to educational data mining by exploring an interpretable Decision Tree approach for examining socioeconomic patterns associated with university students' academic outcomes in a regional Indonesian higher education context. From a practical perspective, the resulting framework may provide preliminary insights for university administrators seeking to develop more targeted academic mentoring, financial assistance, and student support programs. However, the relatively modest predictive performance indicates that the model requires further refinement before being implemented as a reliable institutional decision-support tool. Moreover, the reported evaluation metrics assess numerical GPA prediction rather than categorical classification, and the inclusion of GPA-derived achievement status among the predictors presents a potential target leakage problem. Future research should therefore reconstruct the model using socioeconomic attributes exclusively as predictors, verify the implementation of the C4.5 algorithm, and evaluate categorical academic achievement using accuracy, precision, recall, F1-score, and confusion matrix analysis. Further studies should also incorporate additional predictors, apply rigorous cross-validation, and compare alternative machine learning algorithms across multiple higher education institutions. These improvements would strengthen model generalizability, interpretability, and practical relevance while supporting more equitable and evidence-based educational policies.

.

DAFTAR PUSTAKA

[1] Republic of Indonesia, "Law Number 12 of 2012 concerning Higher Education," 2012.

[2] Republic of Indonesia, "Law Number 27 of 2022 concerning Personal Data Protection," 2022.

[3] Republic of Indonesia, "Law Number 20 of 2003 concerning the National Education System," 2003.

[4] M. Richardson, C. Abraham, and R. Bond, "Psychological correlates of university students' academic performance: A systematic review and meta-analysis," *Psychol. Bull.*, vol. 138, no. 2, pp. 353–387, 2012, doi: 10.1037/a0026838.

[5] C. F. Rodríguez-Hernández, E. Cascallar, and E. Kyndt, "Socio-economic status and academic performance in higher education: A systematic review," *Educ. Res. Rev.*, vol. 29, p. 100305, 2020, doi: 10.1016/j.edurev.2019.100305.

[6] M. F. Musso, E. Kyndt, E. C. Cascallar, and F. Dochy, "Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks," *Front. Learn. Res.*, vol. 1, no. 1, pp. 42–71, 2013, doi:

10.14786/flr.v1i1.13.
[7] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 3, p. e1355, 2020, doi: 10.1002/widm.1355.
[8] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, p. 3, 2020, doi: 10.1186/s41239-020-0177-7.
[9] P. N. Tan, M. Steinbach, V. Kumar, and A. Karpatne, *Data Mining*, vol. 11, no. 1. 2019.
[10] M. F. F. Mardianto, B. S. Pratama, M. Audilla, and E. Pusporani, "Sentiment Analysis of Suicide on X Using Support Vector Machine and Naive Bayes Classifier Algorithms," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 9, no. 1, pp. 60–75, 2025, doi: 10.29407/intensif.v9i1.23742.
[11] S. Mukodimah, M. Muslihudin, D. R. Mustofa, and D. Susianto, "Naive Bayes Classifier Method Analysis and Support Vector Machine (SVM) Student Graduation Prediction," *NEUROQUANTOLOGY*, vol. 20, no. 12, pp. 3522–3533, 2022, doi: 10.14704/NQ.2022.20.12.NQ77360.
[12] F. S. Fauzi, Rita Irviani, Andino Maseleno, "Revolutionizing Education through Technology : Big Data and Online Learning," in *CICCSE*, 2017, p. 44. doi: 10.1166/asl.2011.1261.
[13] C. J. Mantas and J. Abellán, "Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data," *Expert Syst. Appl.*, vol. 41, no. 10, 2014, doi: 10.1016/j.eswa.2014.01.017.
[14] M. Muslihudin, "Analisis Prediksi Mahasiswa Tidak Tepat Waktu Menyelesaikan Studi Dengan Menggunakan Metode Algoritma C 4.5 (Studi Kasus: STMIK Pringsewu)," IBI Darmajaya, 2015.
[15] A. H. Mahfudzin, S. Sriyanto, S. Sutedi, and W. Wasilah, "Prediction of DHF Disease Using Bagging Algorithm with Decision Tree C4.5," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 7, no. 3, pp. 747–758, 2025.
[16] N. N. Nandang Iriadi, "Kajian Penerapan Metode Klasifikasi Data Mining Algoritma C4.5 Untuk Prediksi Kelayakan Kredit Pada Bank Mayapada Jakarta," *J. Tek. Komput. AMIK BSI*, vol. 2, no. 1, pp. 132–137, 2016.
[17] S. D. Nanda Pratama Alfyandri, Hari Antoni Musril, "Implementation of the C4 . 5 Algorithm to Build A Prediction Model for Student Success in Database Courses," *Knowbase Int. J. Knowl. Database*, vol. 05, no. 02, pp. 132–144, 2025, [Online]. Available: https://ejournal.uinbukittinggi.ac.id/ijokid/article/view/10083%0Ahttps://ejournal.uinbukittinggi.ac.id/ijokid/article/download/10083/2624
[18] Wulan Nadia Nursafitri and Naufal Shabir Ramadhan, "Analisis Klasifikasi UKT Mahasiswa Berdasarkan Tingkat Penghasilan Orang Tua Menggunakan Algoritma C4.5," *Invent. J. Inov. dan Tren Pendidik. Teknol. Inf.*, vol. 2, no. 1, pp. 1–9, 2024, doi: 10.37630/inventor.v2i1.1248.
[19] L. A. R. R. dan S. Ning Ratwastuti, "Prediksi Prestasi Belajar Mahasiswa Vokasi Menggunakan Algoritma C4.5," *J. Ilm. KOMPUTASI*, vol. 18, no. 1, p. 2015, 2015.