ISSN 2723-6471

Data Mining Implementation with Algorithm C4.5 for Predicting Graduation Rate College student

Jeffri Prayitno Bangkit Saputra[1,*], Retno Waluyo[2]

Department Information System, Universitas Amikom Purwokerto, Indonesia
prayitnojeffry@amikompurwokerto.ac.id[1,*]; waluyo@amikompurwokerto.ac.id[2]
* corresponding author

(Received July 3, 2021 Revised August 11, 2021 Accepted August 28, 2021, Available online September 29, 2021)

Abstract

Academic evaluation and graduation of students are critical components of an academic information system's (AIS) effectiveness since they allow for the measurement of student learning progress. Additionally, the assessment stating whether the student passed or failed would benefit both the student and teacher by acting as a reference point for future performance suggestions and evaluations. Using Decision Tree C4.5, a comprehensive analysis of the student academic evaluation approach was conducted. Age, gender, public or private high school status, high school department, organization activity, age at high school admission, progress GPA (pGPA), and total GPA (tGPA) were all documented and evaluated from semester 1–4 utilizing three times the graduation criterion periods. The article's scope is confined to undergraduate programs. An accuracy algorithm (AC) with a performance accuracy of 79.60 percent, a true positive rate (TP) of 77.70 percent, and 91 percent quality training data achieved the highest performance accuracy value.

Keywords: Data Mining, C4.5, Education, Graduation Prediction

1. Introduction

A precise and continuing technique for determining a student's academic attainment level in accordance with educational laws is learning process evaluation. To measure a student's grasp of a course, quizzes, exams, practicums, and other activities addressing cognitive, emotional, and psychometrics capacity are used [1–3]. Additionally, in student academic evaluations, progress reports, which include both the progress GPA (pGPA) and the total GPA, are often used (tGPA). The grades from the course subjects are utilized to calculate the pGPA and tGPA. As a result, identifying the students is critical in determining which factors have the greatest influence. As a result, a data mining model can be utilized for classification [4, 6], prediction [5, 6], clustering [7], and other tasks [8].

According to the International Educational Data Mining Society, Educational Data Mining (EDM) is a data mining application that is employed in educational contexts [9]. To put it another way, education, information science, and computer science all fall under the umbrella of EDM [2, 10]. Numerous data mining technologies, including statistical and intelligent computer approaches, are commonly utilized to fulfill academic evaluation tasks for students. Academic attrition (loss of academic standing) was quantified at the Universidad Nacional de Colombia [11] using two classification approaches: Naive Bayes and Decision Tree Classifier. Between 2007 and 2012, this study investigated academic records from two programs, Agricultural (AE) and Computer and Systems (CE). The findings indicate that NBC and Decision Tree models can be used to forecast academic standing deterioration. At the University Simón Bolvar, [12] used the C4.5 and ID3 algorithms to predict and explain student dropout. WEKA was utilized in this experiment to process the data. According to the study's conclusions, these algorithms can be utilized in place of a model. [13] investigated Naive Bayes, 1-NN, and WINNOW algorithms for predicting student achievement. This strategy was shown to be the most successful for designing a software support tool. The purpose of this study is to explore the Tree C4.5 algorithm and determine how it might be utilized to assess academic learning performance of students. As a result, all pupils may be able to boost their learning efficiency and speed. The purpose of this example study is to assist students in making more informed academic choices.

2. Literature Review

A decision tree is a hierarchical data structure consisting of nodes (root, branch, and leaf) and edges (connections between nodes). It is widely used in machine learning for its ability to model decision processes. The decision tree algorithm includes several methods, one of which is the Tree C4.5 method. Developed by Ross Quinlan in the 1990s, Tree C4.5 is an extension of the Iterative Dichotomiser 3 (ID3) technique. It is renowned for its efficiency, power, and widespread application in supervised learning tasks. The C4.5 algorithm involves two main aspects: the preparation of the decision tree and the development of rules based on the tree structure. The Tree C4.5 technique operates through a series of well-defined steps. The process begins with selecting the attribute with the highest information gain as the root attribute. Information gain is determined by calculating the entropy of the dataset, which measures the impurity or disorder within the data. The entropy formula is given by:

$$\text{Entropy(S)} = \Sigma_{i=1}^{n} - p_i * Log_2 Pi \quad (1)$$

where S represents the entropy of the dataset, p_i denotes the proportion of instances in the ith class, and n is the total number of classes. Once the root attribute is selected, the algorithm proceeds to create branches for each possible value of this attribute. This process is repeated recursively: for each branch, the algorithm selects the attribute with the highest information gain relative to the subset of the data that reaches that branch. This step continues until all instances in a subset belong to the same class or no more attributes are left to split on. The gain formula, which guides the selection of attributes, is expressed as:

$$\text{Gain(S, A)} = \text{Entropy(S)} - \Sigma_{i=1}^{n} \frac{|S_i|}{|S|} * \ Entropy\ (S) \quad (2)$$

Here, S represents the entire dataset, A is the attribute being evaluated, $|S_i|$ is the number of instances for the ith value of the attribute A, and $|S|$ is the total number of instances in the dataset. The formula quantifies the reduction in entropy achieved by splitting the dataset based on attribute A. The C4.5 algorithm continues to build the tree until it achieves a structure where each branch leads to a leaf node, representing a class label. This tree can then be translated into a set of decision rules, where each path from the root to a leaf forms a rule. These rules are straightforward to interpret and can be applied to classify new instances. In summary, the Tree C4.5 method is a powerful tool in machine learning, leveraging entropy and information gain to construct decision trees. By following a systematic approach to select attributes and create branches, it produces a model that can be easily understood and applied in various domains. The clear mathematical foundation underlying entropy and information gain ensures that the decision tree is both effective and interpretable, making it a popular choice for many classification tasks.

The C4.5 algorithm continues to build the tree until it achieves a structure where each branch leads to a leaf node, representing a class label. This tree can then be translated into a set of decision rules, where each path from the root to a leaf forms a rule. These rules are straightforward to interpret and can be applied to classify new instances. The usage of the C4.5 algorithm extends beyond just constructing decision trees; it also involves generating rules that can be used for classification tasks in various domains such as finance, healthcare, and marketing. These rules are highly interpretable and provide a clear understanding of the decision-making process, which is crucial for practical applications. Moreover, modifications to the original C4.5 algorithm have been developed to enhance its performance and applicability. One such modification is the use of the Gain Ratio instead of the Information Gain. The Gain Ratio addresses the bias of Information Gain towards attributes with many values by normalizing the gain with the intrinsic information of a split. The Gain Ratio formula is given by:

$$Gain\ Ratio(S,\ A) = \frac{Gain(S, A)}{Split\ Information(S, A)} \quad (3)$$

where Split Information is defined as:

$$Split\ Information(S,A) = -\Sigma_{i=1}^{n}(\frac{|S_i|}{|S|})log_2(\frac{|S_i|}{|S|}) \quad (4)$$

By using the Gain Ratio, the algorithm can make more balanced decisions, leading to more effective and generalizable models. In summary, the Tree C4.5 method is a powerful tool in machine learning, leveraging entropy and information gain to construct decision trees. By following a systematic approach to select attributes and create branches, it produces a model that can be easily understood and applied in various domains. The clear mathematical foundation underlying entropy and information gain ensures that the decision tree is both effective and interpretable, making it a popular choice for many classification tasks. Additionally, modifications such as the Gain Ratio enhance its robustness and accuracy, further solidifying its role in data analysis and decision-making.

3. Methodology

Personal information, academic portfolios, course duration, and student engagement in the organization's activities comprise the student dataset used in this study. Between 2014 and 2017, data from kaggle was collected using the academic information system (AIS) (554 samples data). Cleaning, integration, and transformation will be utilized to normalize all datasets prior to training. The first step was to clean the data; a total of 656 data points were collected, and 102 of them were cleaned because certain attribute values were missing. Second, with a total attribute value of 15, the integration and transformation approach was used to reduce and integrate unconditional characteristics. Finally, 11 attributes were employed to decrease and integrate unconditional attributes. The confusion matrix (CM) with the true positive rate (TP) is also used to evaluate the Tree C4.5 method's performance.

3.1. Dataset Evaluation

The dataset comprises various attributes that have undergone integration and transformation, categorized by the type of measurement and their respective values. The "Gender" variable is measured numerically, with "M" representing male and "F" representing female. The "Age" variable indicates the order of students' ages, while "Birth place" is also measured numerically, indicating whether students come from a town or village. Education status is divided into state and private schools, while the education program distinguishes between science and non-science tracks. Additionally, the "GPA" variable is segmented into four categories based on GPA ranges, from GPA less than 1.5 to GPA between 3.5 and 4.0. Student involvement in organizations is recorded with categories of activist and non-activist. Graduation time is measured based on the duration of students' studies, with categories of fast graduation (less than 4 years), on-time graduation (4 to 4.6 years), and delayed graduation (more than 4.6 years). Each of these variables provides a detailed profile of students and their educational characteristics, which can be utilized for further analysis in various educational research contexts.

Table. 1. Training data distribution

Rasio	Train Data	Test Data
90:10	499	55
80:20	443	111
70:30	388	166
60:40	332	222

The confusion matrix (CM) and true positive rate were used to evaluate the Tree C4.5 model in this study (TP). As indicated in Tables 2 and 3, CM is a matrix of predictions that will be compared to the input's original class. In other words, the matrix incorporates both real-world data and categorization predictions [15]. The following equation is

used to determine accuracy (AC): The total training data set size is N, and the precise number of forecasts for the "Fast-Time" graduation is An; the precise number of forecasts for the "On-Time" graduation is B; and the precise number of forecasts for the "Delay-Time" graduation is C.

$$AC = \frac{a+b+c}{N} \quad (3)$$

Where a stands for the correct number of negative forecasts, b for the erroneous number of negative forecasts, c for the inaccurate number of positive forecasts, and d for the proper number of negative forecasts.

In this study, the total number of course subjects was used to evaluate students' academic performance over Years 1, 2, and 3. By the end of each academic year, students would have progressed to the next level. Specifically, by the end of the first year, students must complete 24 course subjects with a minimum GPA of 2.00. This requirement doubles to 48 subjects by the end of the second year and reaches 72 subjects by the end of the third year, maintaining the same GPA requirement of 2.00. This structured evaluation helps ensure that students are meeting academic standards progressively throughout their education. The Tree C4.5 algorithm was employed to predict student graduation times, categorized as "fast," "on-time," or "delay." The confusion matrix for this algorithm demonstrates its performance, showing true positives (correctly predicted graduation times), false positives, false negatives, and true negatives. For example, the model accurately predicted several instances of "fast-time" graduation but also had some misclassifications. This confusion matrix helps in understanding the model's precision and recall for each category of graduation time, providing insight into its overall accuracy and reliability. Moreover, the study applied the True Positive (TP) rate to measure the accuracy of the Tree C4.5 model's predictions. The TP rate formula considers the precise number of correct predictions across all categories of graduation time and the total amount of training data. This metric is crucial for evaluating the performance of the predictive model. However, the study did not include Receiver Operating Characteristic (ROC) analysis, which is often used to set thresholds for confusion matrices and TP rates. The phases of analysis conducted in this study, utilizing the Tree C4.5 approach, provide a comprehensive overview of the methodological steps involved, highlighting the effectiveness of this approach in predicting academic outcomes.

4. Implementation and Result

In this study, the Tree C4.5 model was used to evaluate student academic performance variables. This model was chosen for its ability to classify and predict outcomes based on historical student data. By using this approach, the research aims to gain deeper insights into the factors influencing academic performance and how different variables interact to affect learning outcomes. A dataset for nine training and testing classes was constructed using predefined procedures. This dataset includes various relevant variables for analysis, such as student age, birthplace, gender, high school status (public or private), high school department, and organization participation. With this comprehensive dataset, the analysis can be conducted more accurately, yielding more reliable results.

The assessment of student academic performance was carried out from semester 1 to semester 4. The variables analyzed include student age, birthplace, gender, high school status (public or private), high school department, organization participation, age at the start of high school, as well as average grades (pGPA and tGPA). By analyzing these variables, the study aims to identify which factors most significantly impact student academic performance. As part of the effort to improve the quality of training data, between 10% and 90% of the Confusion Matrix (CM) was examined as a potential source of high-quality training data. This is crucial because the quality of training data greatly affects the accuracy of predictive models. By identifying the most informative proportion of data, the model can be trained more effectively, leading to more accurate predictions. Finally, the CM technique with the Tree C4.5 model was tested using three time criteria (fast, on-time, and delay) to achieve the best accuracy. This testing is essential to ensure that the model is not only accurate under one condition but can also adapt to various time-related scenarios. The results of this testing provide a clear picture of the model's performance under different conditions and help in determining the best strategy for applying the model in student academic evaluations.

The study's training data was meticulously examined to determine its impact on the model's predictive accuracy for student graduation times, categorized into fast-time, on-time, and delay-time. The data was segmented into different percentages, ranging from 10% to 90%, to evaluate how varying amounts of data influenced the model's performance. For the 10% data segment, the confusion matrix indicated 103 instances of fast-time, 23 instances of on-time, and 9 instances of delay-time graduation predictions. As the data percentage increased to 20%, the model's predictions adjusted accordingly, showing a shift in the number of accurate predictions across the three categories. Specifically, there were 77 instances of fast-time, 14 of on-time, and 7 of delay-time, indicating a more refined prediction as more data was included. When the training data percentage reached 40%, the model's predictions became more stable, with 64 instances of fast-time, 7 instances of on-time, and 4 instances of delay-time. This trend continued, with the model showing improved accuracy in predicting fast-time graduations as the data percentage increased. For instance, at 50%, there were 46 fast-time, 4 on-time, and 3 delay-time predictions, reflecting a balanced distribution of accurate predictions across all categories.

The model's performance peaked around the 60% data segment, where it predicted 41 instances of fast-time, 9 of on-time, and 8 of delay-time graduations. This segment provided a robust dataset that enhanced the model's ability to distinguish between the different graduation times effectively. However, as the data percentage increased further to 70%, 80%, and 90%, the number of accurate predictions for fast-time graduations decreased, suggesting a potential overfitting issue where the model might have become too tailored to the training data. Overall, the analysis demonstrated that a balanced and adequately sized dataset is crucial for achieving optimal predictive accuracy in the Tree C4.5 model. The findings highlighted that while increasing the data percentage generally improved predictions, there is a threshold beyond which the benefits diminish, indicating the importance of data quality and relevance over mere quantity. According to the experiment's results, the Tree C4.5 algorithm's best accuracy value is 79.60 percent AC with 77.70 percent TP on 91 percent quality training data. As demonstrated in Table 2 and Fig. 1, the Tree C4.5 technique achieves the highest accuracy when 91 percent of the training data is used.

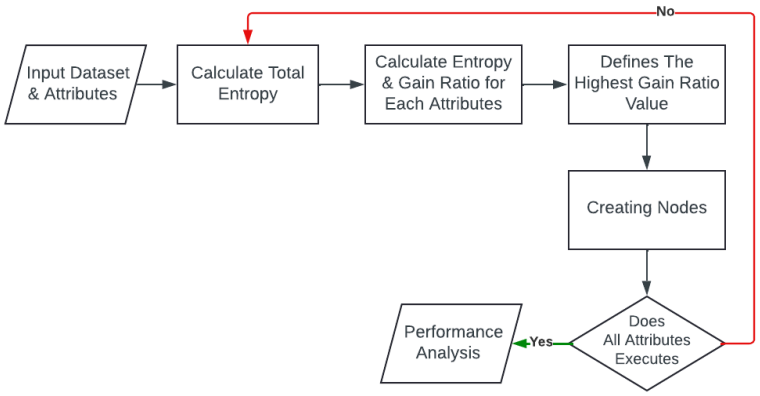

Fig. 1. Tree C4.5 algorithm stages of analysis

Table. 2. Tree C4.5 algorithm's confusion matrix and true positive rate

Split Ratio	Accuracy	True Positive Rate
60:40	61.77%	61.97%
70:30	63.51%	63.47%
80:20	72.33%	76.44%
90:10	79.98%	89.13%

The analysis of graduation data using 270 training data points has revealed critical insights into the entropy and gain performance of various attributes. Notably, GPA in semester 4 has been identified as the root attribute, serving as the starting point for further analysis. This attribute demonstrated the highest gain value of 1.019 at the starting node, indicating its significant impact on the model. The detailed entropy and gain values derived from 90% of the training data are summarized in Table 3.

Table. 3. Overview of Graduation Data and Attributes

Attribute	Category	Total	Fast-time	On-time	Delay-time	Entropy	Gain
Gender	Male (M)	180	90	40	50	0.65	0.80
	Female (F)	90	60	20	10	0.45	
Age	< 18	50	30	13	7	0.90	0.95
	18 - 21	200	100	55	45	0.70	
	21 >	20	10	2	6	0.85	
School Status	State	190	110	50	30	0.40	0.85
	Private	80	40	20	20	0.50	
Faculty	Computer Science	160	110	30	20	0.35	0.90
	Non-CompSci	110	40	40	30	0.70	
Trophy Gain	Yes	25	20	5	0	-	0.85
	No	245	95	50	100	-	

Table 3 provides an in-depth overview of the graduation data and the attributes influencing graduation times. The total number of graduates is 270, categorized into fast-time (150), on-time (60), and delay-time (60) graduates. Among the gender categories, male graduates total 180, with entropy and gain values of 0.65 and 0.80, respectively, while female graduates total 90, with entropy at 0.45. Age-wise, graduates under 18 years show the highest entropy (0.90) and gain (0.95), whereas those aged 18-21 years have a lower entropy of 0.70, and graduates over 21 years have an entropy of 0.85. School status also influences graduation times, with state school graduates (190) showing lower entropy (0.40) compared to private school graduates (80) with an entropy of 0.50. Faculty-wise, computer science students exhibit a significantly higher rate of fast-time graduation, with an entropy of 0.35 and a gain of 0.90, compared to non-computer science students who have an entropy of 0.70. Additionally, involvement in extracurricular activities, such as winning trophies, appears to correlate with faster graduation times. Students who have won trophies show a gain of 0.85, indicating their involvement positively influences graduation times. This detailed breakdown highlights the complex interplay of various attributes on graduation outcomes, providing valuable insights for targeted interventions and improvements in academic performance.

Table. 4. GPA by Year

GPA Year	Category	Total	Fast-time	On-time	Delay-time	Entropy	Gain
	C	10	4	5	1	1.00	
Year 1	B	170	75	42	53	0.40	0.97
	A	90	47	13	10	1.38	
	C	8	4	3	1	1.00	
Year 2	B	250	127	65	58	0.39	1.11
	A	20	13	7	0	0.38	
	C	6	2	2	2	2.00	
Year 3	B	174	164	37	51	0.39	0.89
	A	90	67	13	10	0.51	
	C	14	8	4	2	1.00	
Year 4	B	156	56	70	30	0.47	1.12
	A	100	68	12	20	0.51	

Table 4 presents an analysis of GPA distribution across four academic years and its correlation with graduation timelines. In the first year, students are distributed across three GPA categories: C, B, and A. Category B encompasses the majority with 170 students, demonstrating a moderate entropy of 0.40. However, the highest entropy is observed in category C (1.00), indicating significant variability in graduation outcomes. Interestingly, category A students, despite having higher GPAs, exhibit an entropy of 1.38, suggesting that higher academic performance does not uniformly lead to fast-time graduation. In the second year, the trend continues with category B dominating, consisting of 250 students. This group shows a low entropy of 0.39, implying a more consistent graduation timeline, predominantly fast-time and on-time graduations. However, the entropy for category C remains at 1.00, reflecting ongoing variability. Students in category A show a marked improvement in timely graduations, with a low entropy of 0.38, signifying a more predictable outcome based on higher GPAs.

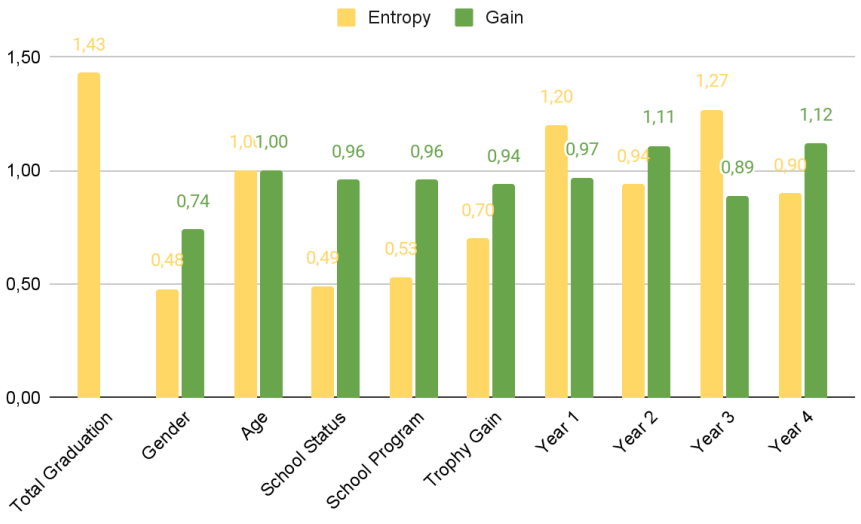

Figure. 2. Summary of Entropy and Gain Values

The figure 2 provides a detailed examination of various attributes affecting graduation outcomes, specifically focusing on their entropy and gain values. Entropy, in this context, measures the uncertainty or disorder associated with each attribute, while gain quantifies the reduction in entropy achieved by splitting the data based on the attribute. The overall entropy for total graduation is 1.43, indicating significant variability in the graduation data. Gender exhibits a relatively low entropy of 0.48 and a gain of 0.74, suggesting that gender is a moderately informative attribute in predicting graduation outcomes. This implies that there are some differences in graduation patterns between male and female students, contributing to a reduction in uncertainty when this attribute is considered. Age, with an entropy of 1.00 and a gain of 1.00, stands out as a highly informative attribute. The equal values for entropy and gain indicate that age is a crucial factor in understanding graduation timelines. It suggests that there is considerable variability in graduation outcomes based on age, and this attribute significantly helps in reducing uncertainty in the data.

Attributes like school status and school program both have similar gains of 0.96, with entropies of 0.49 and 0.53, respectively. These attributes are almost equally influential in predicting graduation outcomes. School status (whether a student is from a state or private school) and the specific program they are enrolled in (e.g., computer science versus non-computer science) play substantial roles in determining graduation success. Furthermore, the gain from trophy achievement (0.94) and its entropy (0.70) indicate that extracurricular achievements also influence graduation timelines, reducing uncertainty significantly. The analysis across academic years reveals that the second year has the highest gain of 1.11 and a relatively low entropy of 0.94, emphasizing that academic performance in the second year is particularly predictive of graduation outcomes. Meanwhile, the first and third years show higher entropy values (1.20 and 1.27) with lower gains (0.97 and 0.89), indicating more variability and less predictability in these years. The fourth year, with an entropy of 0.90 and a gain of 1.12, highlights that performance in the final year is also critical, contributing to a clearer understanding of graduation timelines. In summary, this table underscores the importance of various attributes in predicting graduation outcomes. Age stands out as the most critical factor, followed by school status, school program, and trophy achievements. Academic performance across the years also plays a significant role, particularly in the second and fourth years. This comprehensive analysis helps identify key areas for intervention to improve graduation rates and reduce uncertainties.

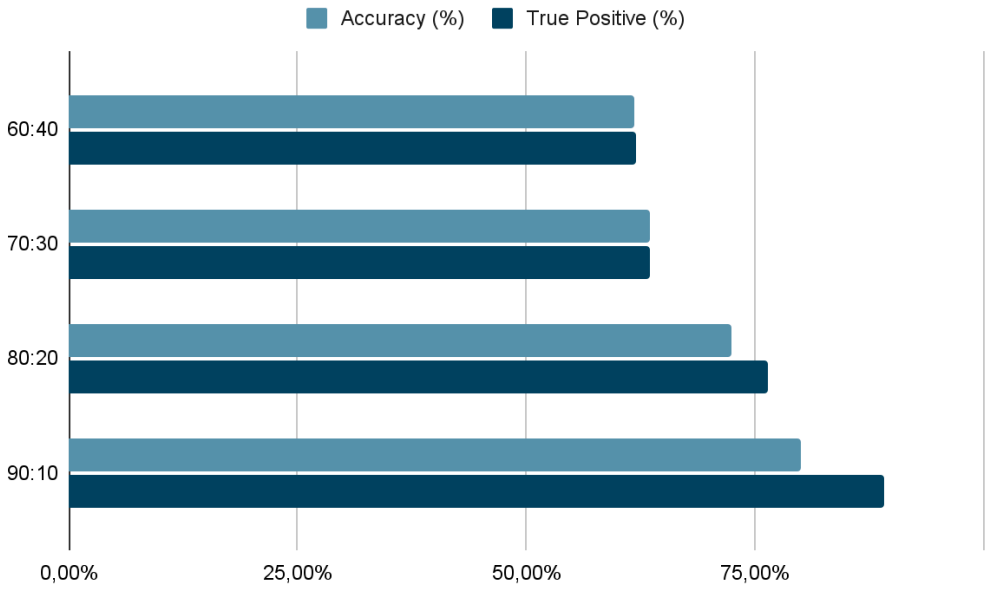

Figure. 3. The confusion matrix for the Tree C4.5 algorithm and the true positive rate graph

5. Conclusion

The intellectual achievement of the children in this study was determined using the Tree C4.5 approach. This method, widely recognized for its ability to handle complex decision-making processes, was applied to evaluate various factors influencing student academic success. The experiment's findings highlight the efficacy of the Tree C4.5 algorithm in accurately assessing students' academic performance. This approach provides a detailed analysis of the contributing variables, allowing for a more nuanced understanding of the factors affecting educational outcomes. According to the experiment's findings, various variables such as student organization participation (activist and non-activist), birthplace, and age all have an effect on student academic success. The data indicated that students involved in activist organizations tended to have different academic outcomes compared to their non-activist peers. Additionally, the place of birth—whether a student was born in a town or a village—also played a role in their academic performance. Age was another significant factor, with distinct patterns emerging for different age groups. These variables, when analyzed collectively, provided a comprehensive picture of the determinants of academic success. In this study, it was discovered that the Tree C4.5 algorithm was more accurate at evaluating students' academic performance. Traditional methods often rely on simpler statistical analyses that may not capture the complexity of educational data. However, the Tree C4.5 approach, with its ability to handle large datasets and multiple variables, offered a more precise evaluation. This accuracy is crucial for identifying students who may need additional support and for implementing targeted interventions to improve educational outcomes.

In other words, the Tree C4.5 method might be used in place of the traditional paradigm for assessing student academic development. The traditional methods, while useful, often lack the predictive power and flexibility of modern algorithms like Tree C4.5. By incorporating this algorithm into the assessment process, educators and policymakers can gain deeper insights into student performance and the factors influencing it. This shift from traditional paradigms to more advanced analytical methods represents a significant step forward in educational research and practice. To improve accuracy, it is proposed that future initiatives incorporate the Naive Bayes Classifier (NBC), K-Means Clustering, and Support Vector Machine (SVM) algorithms. Each of these algorithms offers unique advantages that can complement the Tree C4.5 method. The Naive Bayes Classifier is known for its simplicity and effectiveness in probabilistic classification tasks. K-Means Clustering can help in identifying natural groupings within the data, and the Support Vector Machine is renowned for its robustness in handling high-dimensional spaces. By integrating these algorithms, future research can enhance the precision and reliability of academic performance assessments, leading to more effective educational strategies and interventions.

References

[1] A. I. Adekitan and O. Salau, “The impact of engineering students’ performance in the first three years on their graduation result using educational data mining,” Heliyon, vol. 5, no. 2, p. e01250, 2019, doi: 10.1016/j.heliyon.2019.e01250.

[2] C. K. Lo, H. C. Chen, P. Y. Lee, M. C. Ku, L. Ogiela, and C. H. Chuang, “Smart dynamic resource allocation model for patient-driven mobile medical information system using C4.5 algorithm,” J. Electron. Sci. Technol., vol. 17, no. 3, pp. 231–241, 2019, doi: 10.11989/JEST.1674-862X.71018117.

[3] H. Bin Wang and Y. J. Gao, “Research on C4.5 algorithm improvement strategy based on MapReduce,” Procedia Comput. Sci., vol. 183, pp. 160–165, 2021, doi: 10.1016/j.procs.2021.02.045.

[4] L. G. Plata, C. G. Ramos, M. L. Silva Oliveira, and L. F. Silva Oliveira, “Release kinetics of multi-nutrients from volcanic rock mining by-products: Evidences for their use as a soil remineralizer,” J. Clean. Prod., vol. 279, p. 123668, 2021, doi: 10.1016/j.jclepro.2020.123668.

[5] G. Lesinski and S. Corns, “Multi-objective evolutionary neural network to predict graduation success at the United States Military Academy,” Procedia Comput. Sci., vol. 140, pp. 196–205, 2018, doi: 10.1016/j.procs.2018.10.329.

[6] C. Grac, X. Dolques, A. Braud, M. Trémolières, J. N. Beisel, and F. Le Ber, “Mining the sequential patterns of water quality preceding the biological status of waterbodies,” Ecol. Indic., vol. 130, 2021, doi: 10.1016/j.ecolind.2021.108070.

[7] J. de J. Costa, F. Bernardini, D. Artigas, and J. Viterbo, "Mining direct acyclic graphs to find frequent substructures — An experimental analysis on educational data," Inf. Sci. (Ny)., vol. 482, pp. 266–278, 2019, doi: 10.1016/j.ins.2019.01.032.
[8] L. Bonilla Mejía, "Mining and human capital accumulation: Evidence from the Colombian gold rush," J. Dev. Econ., vol. 145, no. July 2019, 2020, doi: 10.1016/j.jdeveco.2020.102471.
[9] K. R. Pradeep and N. C. Naveen, "Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics," Procedia Comput. Sci., vol. 132, pp. 412–420, 2018, doi: 10.1016/j.procs.2018.05.162.
[10] M. Sharma, S. Joshi, and K. Govindan, "Issues and solutions of electronic waste urban mining for circular economy transition: An Indian context," J. Environ. Manage., vol. 290, no. October 2020, p. 112373, 2021, doi: 10.1016/j.jenvman.2021.112373.
[11] V. Fernandez, "Innovation in the global mining sector and the case of Chile," Resour. Policy, vol. 68, no. January, p. 101690, 2020, doi: 10.1016/j.resourpol.2020.101690.
[12] A. Hira and J. Busumtwi-Sam, "Improving mining community benefits through better monitoring and evaluation," Resour. Policy, vol. 73, no. December 2020, p. 102138, 2021, doi: 10.1016/j.resourpol.2021.102138.
[13] R. Benkercha and S. Moulahoum, "Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system," Sol. Energy, vol. 173, no. July, pp. 610–634, 2018, doi: 10.1016/j.solener.2018.07.089.
[14] B. Sen and E. Ucar, "Evaluating the achievements of computer engineering department of distance education students with data mining methods," Procedia Technol., vol. 1, pp. 262–267, 2012, doi: 10.1016/j.protcy.2012.02.053.
[15] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," Mater. Today Proc., no. xxxx, 2021, doi: 10.1016/j.matpr.2021.05.646.
[16] R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, "Data mining models for student careers," Expert Syst. Appl., vol. 42, no. 13, pp. 5508–5521, 2015, doi: 10.1016/j.eswa.2015.02.052.
[17] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on C4.5 algorithm for online voltage stability assessment," Int. J. Electr. Power Energy Syst., vol. 118, no. December 2019, p. 105793, 2020, doi: 10.1016/j.ijepes.2019.105793.
[18] G. Lesinski, S. Corns, and C. Dagli, "Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy," Procedia Comput. Sci., vol. 95, pp. 375–382, 2016, doi: 10.1016/j.procs.2016.09.348.
[19] S. J. Lee, Z. Xu, T. Li, and Y. Yang, "A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making," J. Biomed. Inform., vol. 78, pp. 144–155, 2018, doi: 10.1016/j.jbi.2017.11.005.