IJHEP | INTERNATIONAL JOURNAL OF HIGHER EDUCATION PEDAGOGIES
ISSN: 2669-2333
VOLUME 3, ISSUE 4

Decision Tree Algorithm Use in Predicting Students' Academic Performance in Advanced Programming Course

Ismail O. Muraina[1*], Edward A. Aiyegbusi[1] and Solomon O. Abam[2]

[1] Department of Computer Science, College of Information and Technology Education, Lagos State University of Education, Lagos, Nigeria
[2] Department of Science and Technology Education, Faculty of Education, Lagos State University, Lagos, Nigeria
niyi2all@yahoo.com

ABSTRACT

Students' academic performance or achievement has from time to time been a subject of discourse to academicians, scholars, researchers, and educational institutions all over the globe. In this regard, schools are expected to play major and active roles in ensuring that students have good performance at end of their programs. Academic performance is normally used to classify or predict how students would be ultimately capable to withstand and face future challenges after graduation. Students' academic performance/achievement in any course of study plays a vital role in contributing to and producing outstanding students who will be future viable leaders. The use of algorithms to classify and predict students' academic performance/achievement is not new in machine learning using different techniques like neural networks, logistic regression, decision trees, and many more. This study classifies and predicts with the use of a graphical technique called a Decision Tree. The dataset was built from students' attendance, practical assessment, assignment, ability to complete a free related course on the internet, test score, and examination grade; a large data set was used to construct and validate the decision tree algorithm (CHAID) for the first time. CHAID's overall accuracy, sensitivity, and specificity were assessed using a two-part dataset. The training test and testing set were compared to see how each stage of the algorithm compares to the other. Results show that the decision tree algorithm makes classification and prediction visible and clear with the use of graphics to display the results. Hence, the model built produces 96% accuracy.

Keywords: decision tree, academic achievement, classification, prediction, algorithms

Cite this article as: Muraina, I. O., Aiyegbusi, E. A., & Abam, S. O. (2022). Decision Tree Algorithm Use in Predicting Students' Academic Performance in Advanced Programming Course. *International Journal of Higher Education Pedagogies*, 3(4), 13-23. https://doi.org/10.33422/ijhep.v3i4.274

1. Introduction

For many years, researchers have used a range of techniques to categorize and forecast students' academic performance/achievement at various levels using various metrics, such as test scores, grade points, teacher rating scales, assignments, and even dropout tests (Ampofo & Osei-Owusu, 2015). Further, the authors said that academic instruction is the primary aim of education and schools are in the position to influence students toward learning, socialization, and vocational preparedness. Therefore, students' academic success is a key outcome of education, which is why every nation would strive to include the provision of quality education among its national goals for education. Educational institutions are being forced to use education as a vehicle for social change in this trend. As a result, the quality of pupils a school produces determines its success, just as the success of any educational institution is determined by how well its students perform on both academic and non-academic tests. Yusuf (2012) argued that performance should be measured not only in terms

of test and examination results but also in whether students have acquired survival skills that allow them to compete with their peers in the labor markets. According to MolokoMphale & Mhlauli, (2014), education is seen as a promoter of human development and should be at the center of any society's life and concern. It is a social artifact that embodies aspirations for the well-being and development of the society it is meant to serve. The major concern nowadays is about the type of students schools produce. Academic achievement must be given high priority because it refers to performance outcomes in intellectual domains taught at school, college, and university; it serves as an indicator of intellectual education. It is the most important prerequisite for individual and societal prosperity. This makes academic achievement a vital issue both for students, teachers, and school managers (Spinath, 2012). Hence, Martín, (2017) asserted that the academic performance of students is not only associated with an intellectual quotient (IQ) but there are other multiple variables and dimensions to which a certain predictive value can be attributed to capture cognitive, psychomotor, and affective domains of the students.

Machine learning algorithms are considered quite powerful in classification problems and they are the focus of many studies. The prediction capability of these algorithms is very important and relevant in decision-making (Mienye, Sun & Zenghui, 2019). According to Sharma, Himani & Kumar, (2016) decision tree algorithm is regarded as one of the most popular classification techniques. A decision tree is considered a structure that includes a root node, branches, and leaf (child) nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. One of the most well-liked categorization strategies is the use of decision trees to anticipate outcomes. Consequently, a decision tree algorithm creates a categorization and predictive model that can handle both numerical and categorical data and is straightforward to comprehend, interpret, and display graphically. Decision tree analysis aids in group characteristic identification, examines correlations between independent variables and the dependent variable, and simply presents this data. The procedure can be used to determine classification guidelines for upcoming occurrences, for instance, to identify students who are likely to perform well at the end of his/her program at the institution. Jorda & Raqueno, (2019) asserted that a decision tree develops classification systems that predict or classify future observations based on a set of decision rules such algorithms include: The Classification and Regression Tree (C&R) Tree, Chi-squared Automatic Interaction Detection (CHAID), C 5.0 and, Quick, Unbiased, Efficient, Statistical Tree (QUEST). Raut &Nichat, (2017) added that a decision tree is used to sort educational problems by measuring the students' performance.

This paper is organized as follows: Section I represents the introduction/background of the study. Section II reviews related literature concerning the factors that affect the performance of students in learning. Section III represents the materials and methods used. Section IV represents the results of the findings. Section V represents the discussion of the results gotten from the previous section. Finally, Section VI shows the Conclusions.

2. Related Literature

It was observed from past studies that students' academic gain and learning performance have been affected by various factors including gender, age, teaching methodology, student's grades, parents' social economic status, environmental influence, tuition trend, study pattern, and, the time reserved for study, accommodation (Ali et al., 2013). Numerous academics have carried out in-depth analyses of the variables influencing student success at various study levels. Higher education institutions have been interested in students' academic

performance and graduation rates (Shahzadi & Ahmad, 2011). In the higher education community, research on the variables that affect university students' academic success is becoming increasingly popular. Recently, an investigation revealed how variables like learning styles, gender, and race affect student performance.

Jijo and Abdulazeez (2021) conducted a study on classification using the decision tree method for machine learning in 2021. The paper's contents, including the algorithms/approaches employed, datasets, and results obtained, are thoroughly assessed and presented. To further highlight the authors' topics and determine the most precise classifiers, all of the methodologies examined were also discussed. As a result, the applications of various dataset kinds are explored, and the results are examined.

Academic achievement, according to Steinmayr et al. (2015), embodies performance outcomes that indicate how well a person has achieved particular objectives that were the focus of activities in instructional environments, specifically in school, college, and university. Therefore, academic success should be viewed as a complex construct that includes several learning domains like cognitive, emotional, and psychomotor. There are a variety of indicators of academic achievement, including extremely basic ones like grades or results on a test of academic achievement, as well as cumulative ones like educational degrees and certificates. According to Kapur (2018), factors that affect students' academic success include their attendance in class, their homework, tests, and exams, as well as their participation in competitions and other events.

In a study by Jorda and Raqueno (2019), the dataset was split into a training set and a testing set. Using IBM SPSS Modeler Version 18.0, the training data was utilized to develop and validate two decision tree algorithms, C5.0 and Chi-squared Automatic Interaction Detection (CHAID), based on their overall accuracy and ten-fold cross-validation. The best early warning system for TUPM to identify students who are at risk for academic failure was therefore CHAID. As a result, the study concludes that the CHAID modeling algorithm worked best as a predictive model for identifying students who were likely to be retained in the COE program as well as those who were academically at risk.

To create classifier models, Mustafa (2016) used four distinct classification methods, including decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis. Accuracy, precision, recall, and specificity performance indicators are used to compare their results over a dataset made up of student replies to an actual course review questionnaire. All of the classifier models displayed comparable high classification performances, it was discovered.

Tripti, Dharminder, & Sangeeta (2014) built a performance prediction model based on students' social integration, academic integration, and a variety of emotional skills using multiple classification methodologies. On the dataset, the two algorithms used—Random Tree and J48 (Implementation of C4.5)—performed well.

The enhancement of prediction/classification approaches, which are used to evaluate skill proficiency based on academic achievement by the breadth of knowledge, was the focus of a study done by Mayilvaganan & Kalpanadevi (2014). The C4.5 method, AODE, Naive Bayesian classifier algorithm, Multi-Label K-Nearest Neighbor technique, and decision tree algorithm were used to identify the best classification accuracy and analyze student performance using Weka.

Ali et al. (2013) started a study to look into the variables influencing graduate students' academic performance at the Islamia University of Bahawalpur Rahim Yar Khan Campus. Students' grade was taken into account as a dependent variable, and the gender, age, faculty

of study, education father's or guardian's socioeconomic position, residential location, and medium of instruction were independent variables. The research found that graduate student's academic performance was highly influenced by their age, their fathers' or guardians' socioeconomic level, and their daily study time.

3. Materials and Methods

The dataset was built from students' attendance, practical assessment, assignment, ability to complete a free related course on the internet, test score, and examination grade; a training test set and a testing set were created from the dataset. The decision tree algorithm (CHAID) was developed and validated using training data, and its overall accuracy, sensitivity, and specificity were assessed using testing data.

This method is broken down into four key steps: data collection, classification, building a predictive model, and evaluation. In order to define the student's achievement level dynamically, the weights of the decision tree were concretized. Additionally, the variables of assessment are determined by the student's attendance, practical assessment, assignment, ability to complete a free related course online, test score, and examination grade.

Individual responses from 210 advanced programming course students were categorized using all the aforementioned factors, including attendance, practical assessment, assignment, ability to finish a complimentary related course online, test score, and examination grade. To evaluate the student's academic achievement, the DT-CHAID Algorithm was "conformed" with the experimental process' parameters, which used the decision tree's weights for the dynamic selection of the exercises. Because they are straightforward to comprehend and interpret, simple to display graphically, and able to handle both numerical and categorical data, decision trees have an advantage over alternative prediction models.

The assessment findings were compared to the grades and overall scores of the student's performance in the evaluation, and this helped to draw important conclusions.

4. Results

4.1. The use of Cross-Validation

Table 1 displays the specifications and results in information of the model. At specifications, the growing method used was CHAID, the dependent variable was the decision taken, gender, test scores, attendance, assignment, online free course, practical test, and final examination formed the independent variables used, validation type was cross-validation, the maximum tree depth was 3, the minimum cases in parent node was 10, while the minimum cases in child node were 5. Also, at results section of Table 1, it shows the independent variable included (Final Examination, Online Free Course, Practical Test, Assignment, and Gender), the number of nodes was 11, the number of terminal nodes was 6 while the depth of the results was 3.

Table 1.
Model Summary for cross Validation

Specifications	Growing Method	CHAID
	Dependent Variable	Decision Taken
	Independent Variables	Gender, Test Scores, Attendance, Assignment, Online Free Course, Practical Test, Final Examination
	Validation	Cross Validation
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	Final Examination, Online Free Course, Practical Test, Assignment, Gender
	Number of Nodes	11
	Number of Terminal Nodes	6
	Depth	3

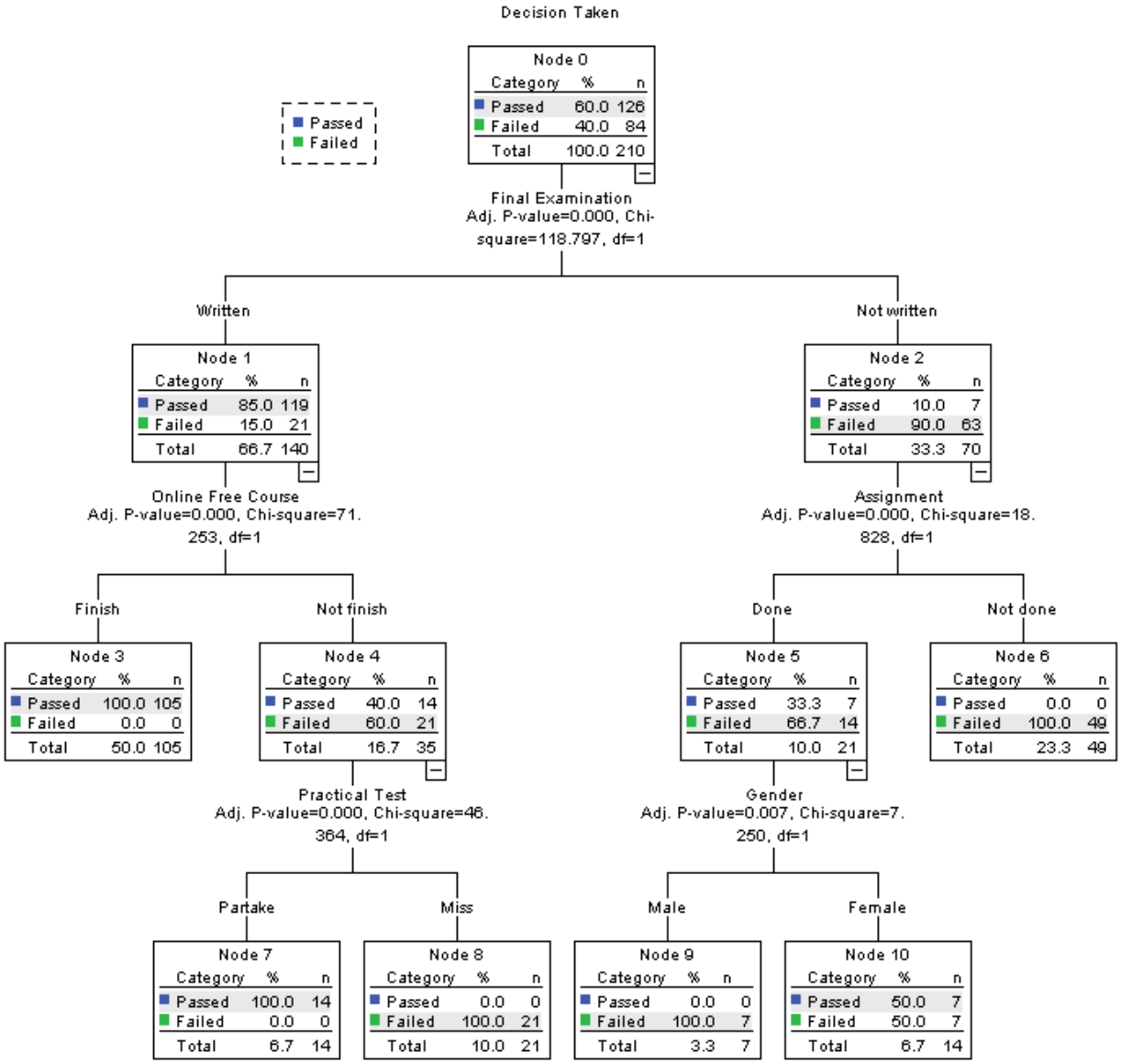

Figure 1. Decision Tree with CHAID and 10-fold validation

The number of students and the percentage of students who passed or failed is provided for each node. The splits happen in priority order. The final test was the most important consideration, in this case, therefore the 'parent' node, which had all 210 students, split into two 'child' nodes, one of which contained the students who took the exam and the other of which contained those who did not. Assignment and Online Free Courses were the next two important nodes. It was shown that all 105 students who participated in the online free course finished and got certified; among those that did not finish the online free course, 40 students passed while 21 failed. Of those that did not finish the online free course but still partook in the practical test were 14 and they all passed while those who did not finish the online free course and also missed the practical test were 21. Similarly, students that did not eventually write the final examination but did the assignment and passed were 7 while those who failed were 14. Those students that did not do the final examination or assignment failed were 40. So, 7 were male students that did the assignment but failed while 7 female students' passed as well as 7 females failed. Hence, the Practical test and gender formed the terminal nodes for the classification. The correct classifications were highlighted in grey.

Table 2.
Path Analysis

Terminal node	Path	Classification	Number correct	Number wrong
3	Written → Finish	Passed	105	0
6	Not written → Not done	Failed	49	0
7	Written → Not Finish → Partake	Passed	14	0
8	Written → Not Finish → Miss	Failed	21	0
9	Not written → Not done → Male	Failed	7	0
10	Not written → Not done → Female	Passed	7	7

Table 2 shows terminal nodes, paths, classification, number of correct classifications, and number of wrong classifications. The path analysis used helps to estimate a system of equations in which all of the variables are observed. It assumes perfect measurement of the observed variables; this implies that only the structural relationships between the observed variables are modeled. Table 2 summarized that 105 students were correctly classified that wrote the examinations and finished with passed. Likewise, 49 students were correctly classified that did not write the examination as well as did not do the assignment that eventually failed; 14 students wrote the final examination, and partook in practical but did not finish the online test, yet they passed; similarly, 21 students wrote the final examination, not finished the online test and as well missed the practical class and failed; in the same vein, 7 male students did not write the final examination, they did not do the online test and failed, while 7 female students did not write the final examination, did not do the online test and eventually failed while 7 of the female students still managed to pass the final examination. This showed 100% classification accuracy was obtained from terminal nodes 3, 6, 7, 8, 9, and 10.

Table 3.
Risk

Method	Estimate	Std. Error
Re-substitution	.033	.012
Cross-Validation	.048	.015
Growing Method: CHAID; Dependent Variable: Decision Taken		

Table 3 explains the information on the proportion of cases misclassified by the proposed classification. It shows re-submission and cross-validation estimates with standard errors. The

initial estimate was .048 (Cross-Validation) [Std. error =.015], then after re-submission, the estimated result gotten was .033 which was considered better than the initial estimate [Std. error =.012].

Table 4.
Classification

Observed	**Predicted**		
	Passed	**Failed**	**Percent Correct**
Passed	126	0	100.0%
Failed	7	77	91.7%
Overall Percentage	63.3%	36.7%	96.7%
Growing Method: CHAID; Dependent Variable: Decision Taken			

The classification table (Table 4) summarizes the percentages classified correctly. The model classified 100% of those students who passed correctly with 126 passed, 0 failed, while only 92% of those students who failed with 7 passed, 77 failed. The overall percentage of those who passed was 63.3% while the overall percentage of those who failed was 36.7%; the model accuracy level is 96.7%.

4.2. Using the Same Percentage for Both the Training Test and Testing Set

Table 5 displays the specifications and results in information of the model. At specifications, the growing method used was CHAID, the dependent variable was the decision taken, gender, test scores, attendance, assignment, online free course, practical test, and final examination formed the independent variables used, validation type was split sample of 50% - 50%, the maximum tree depth was 3, the minimum cases in parent node was 10, while the minimum cases in child node were 5. Also, it shows the independent variable included (Final Examination, Online Free Course, Practical Test, and Assignment), the number of nodes was 9, the number of terminal nodes was 5 while the depth of the results was 3.

Table 5.
Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Decision Taken
	Independent Variables	Gender, Test Scores, Attendance, Assignment, Online Free Course, Practical Test, Final Examination
	Validation	Split Sample [50,50]
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	Final Examination, Online Free Course, Practical Test, Assignment
	Number of Nodes	9
	Number of Terminal Nodes	5
	Depth	3

4.3. Information for Training Sample

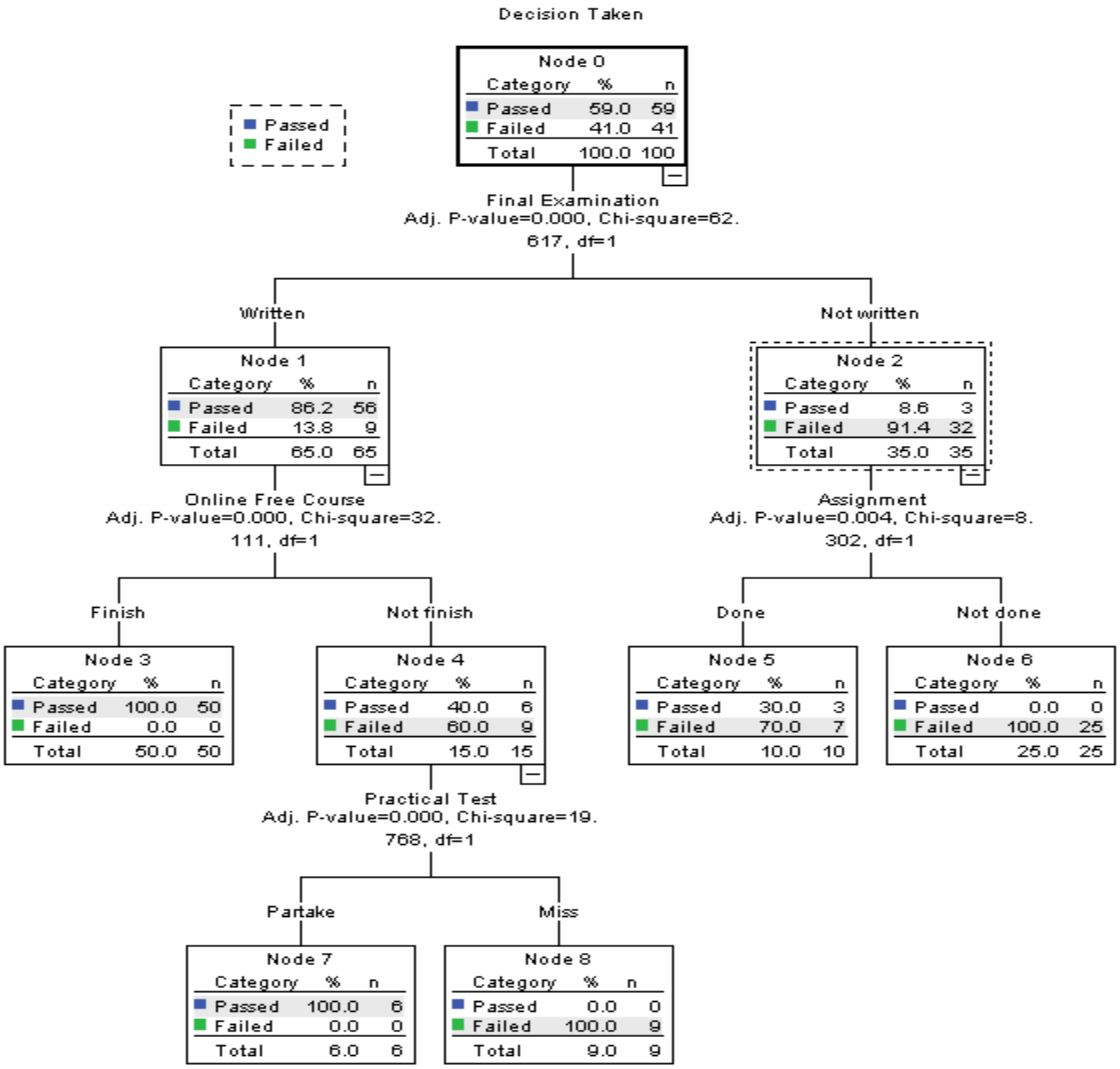

Figure 2. Decision Tree with CHAID and split sample of (50%) for the training set

4.4. Information for Testing Sample

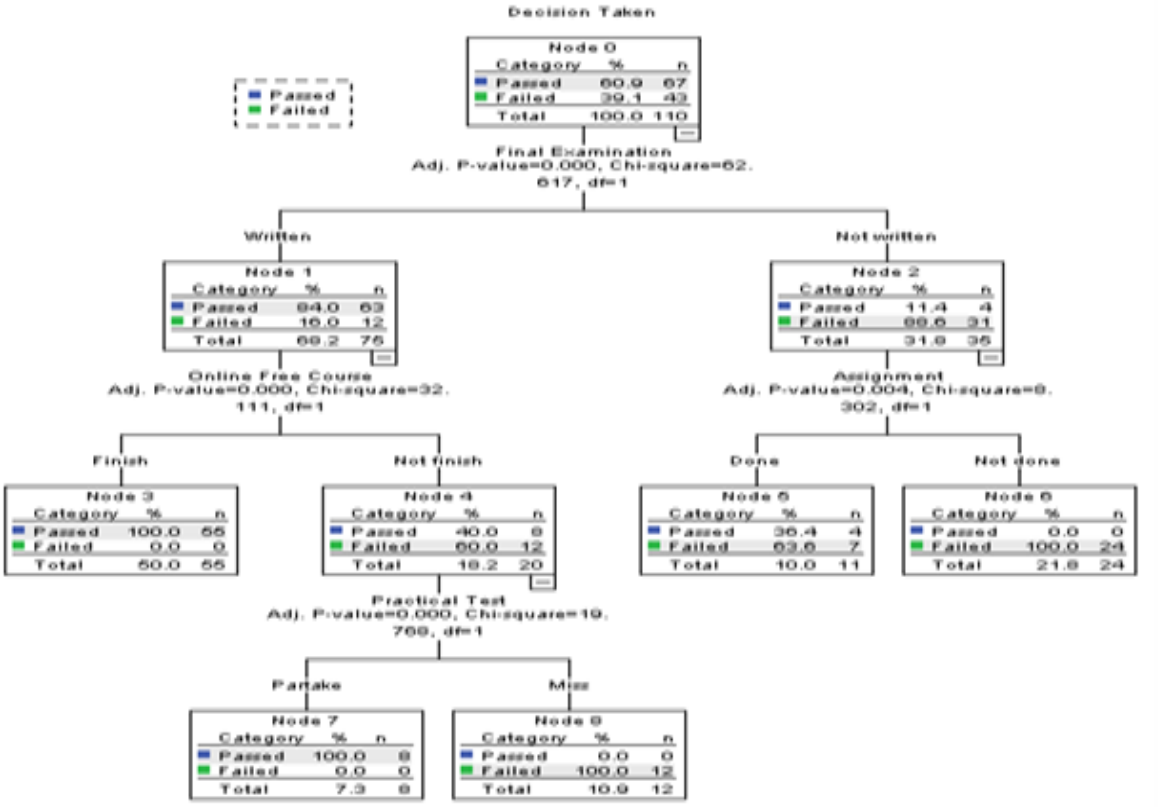

Figure 3. Decision Tree with CHAID and split sample of (50%) for the testing set

Table 6.
Risk

Sample	Estimate	Std. Error
Training	.030	.017
Test	.036	.018
Growing Method: CHAID; Dependent Variable: Decision Taken		

Table 6 explains the information on the proportion of cases misclassified by the proposed classification. It shows training and testing estimates with standard errors. The Training estimate was .030 [Std. error =.017], then the Testing estimate result gotten was .036 [Std. error =.018].

Table 7.
Classification

Sample	Observed	Predicted		
		Passed	Failed	Percent Correct
Training	Passed	56	3	94.9%
	Failed	0	41	100.0%
	Overall Percentage	56.0%	44.0%	97.0%
Test	Passed	63	4	94.0%
	Failed	0	43	100.0%
	Overall Percentage	57.3%	42.7%	96.4%
Growing Method: CHAID; Dependent Variable: Decision Taken				

The classification table (Table 7) summarizes the percentages classified correctly. The model, at the training set, classified 100% of those students who failed, and 95% of those students who passed -from which 56 students passed while 3 failed; consequently, the model overall percentage is 56% passed, 44% failed; 97% overall of accuracy. Likewise, the model at test set shows 94% of those students passed with 63 students passing, 4 students failed and 100% of those who failed with 0 passed, 43 failed. So, 57.3% passed, 42.7% failed while the overall percentage strength is 96% accuracy.

5. Discussion

A decision tree classifier is typically a statistical type of classifier that can be used to cluster datasets as well as predictions. Nodes and branches in the decision tree help to trace the classification flow easily. From the analysis, it was observed that the use of different validation still produced a reasonable result. The results showed that the use of the CHAID growing method was ideal for the classification. It was also shown that variables like Final Examination, Online Free Course, Practical Test, and Assignment contributed greatly to the classification strength of the model while gender contributed less, whereas attendance and test scores contributed nothing to the model and that is the reason they are excluded in the model. Hence, the model gave 96% and above accuracy. One of the limitations of this study is that it is restricted to the use of SPSS to implement the algorithms and covers only one algorithm out of numerous algorithms available in machine learning.

6. Conclusion

A decision tree is used in the educational field to classify and as well predict students' academic achievement. It was established in this study that the classification task of a decision tree is used to evaluate students' performance out of other classification algorithms. The use of a decision tree enhances the readability and visualization of the results compared to other classifiers. To predict how well the students would perform on the final exam of the semester, information about the student's attendance, practical assessment, assignments, ability to complete a free related course online, test score, and examination grade marks were directly obtained from the students. This study establishes the reliability of decision trees. Because they offer classification rules that are simpler to understand than those produced by other classification techniques, decision trees are quite popular.

References

Ali, S., Haider, Z., Munir, F., Khan, H., & Ahmed, A. (2013). Factors Contributing to the Student's Academic Performance: A Case Study of Islamia University Sub-Campus. *American Journal of Educational Research, 2013, Vol. 1, No. 8, 283-289,* https://doi.org/10.12691/education-1-8-3

Ampofo, E. T., & Osei-Owusu, B. (2015). Students' Academic Performance as Mediated by Students' Academic Ambition and Effort in the Public Senior High Schools in Ashanti Mampong Municipality of Ghana *International Journal of Academic Research and Reflection* Vol. 3, No. 5, 2015.

Jijo, B. T. & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning; Journal of Applied Science and Technology Trends *Vol. 02, No. 01, pp. 20 – 28 (2021).*

Jorda, E. R., & Raqueno, A. R. (2019). Predictive Model for the Academic Performance of the Engineering Students Using CHAID and C 5.0 Algorithm. International Journal of

Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 6 (2019), pp. 917-928.

Kapur, R. (2018). Factors Influencing the Student's Academic Performance in Secondary Schools in India. https://www.researchgate.net/publication/324819919

Martín, S. N., Rodrigo, I. G., Izquierdo, G. C., & Ajenjo, P. P. (2017). Exploring Academic Performance: Looking Beyond Numerical Grades; Universal Journal of Educational Research 5(7): 1105-1112, 2017. https://doi.org/10.13189/ujer.2017.050703

Mayilvaganan, M., & Kalpanadevi, D. (2014). Comparison of Classification Techniques for predicting the performance of Students Academic Environment, *International Conference on Communication and Network Technologies (ICCNT),* 2014. https://doi.org/10.1109/CNT.2014.7062736

Mienye, I. D., Sun, Y., & Zenghui, W. (2019). Prediction performance of improved decision tree-based algorithms: a review; 2nd International Conference on Sustainable Materials Processing and Manufacturing (SMPM 2019). https://doi.org/10.1016/j.promfg.2019.06.011

MolokoMphale, L., & Mhlauli, M. B. (2014). An Investigation on Students' Academic Performance for Junior Secondary Schools in Botswana. European Journal Of Educational Research Vol. 3, No. 3, 111-127. https://doi.org/10.12973/eu-jer.3.3.111

Mustafa, (2016). Predicting Instructor Performance Using Data Mining Techniques in HigherEducation, *IEEE Access,* Volume*:* 4,2016. https://doi.org/10.1109/ACCESS.2016.2568756

Raut, A. B., & Nichat, A. A. (2017). Students Performance Prediction Using Decision Tree Technique. International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 7 (2017), pp. 1735-1741.

Shahzadi, E., & Ahmad, Z. (2011). A Study on Academic Performance of University Students. Proc. 8th International Conference on Recent Advances in Statistics, Lahore, Pakistan – February 8-9, 2011, 255-268.

Sharma, H., & Kumar, S. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining; International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391 Volume 5 Issue 4, April 2016. https://doi.org/10.21275/v5i6.NOV164272

Spinath, B. (2012). Academic Achievement. https://doi.org/10.1016/B978-0-12-375000-6.00001-X

Steinmayr, R., Meißner, A., Weidinger, A. F., & Wirthwein, L. (2015). Academic Achievement. Education Oxford Bibliographies; https://doi.org/10.1093/obo/9780199756810-0108

Tripti M., Dharminder, K., & Sangeeta, G. (2014). Mining Students' Data for Performance Prediction, *Fourth International Conference on Advanced Computing & Communication Technologies,*2014. https://doi.org/10.1109/ACCT.2014.105

Yusuf, A. F. (2012). Influence of principals' leadership styles on students' academic achievement in secondary schools. *Journal of Innovative Research in Management and Humanities*, 3(1), 113 -121.