ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
**Indian Journal of Science and Technology,** *Vol 12(5), DOI: 10.17485/ijst/2019/v12i5/140987, February 2019*

# Using Decision Tree Algorithm to Predict Student Performance

**Micheline Apolinar -Gotardo***

Leyte Normal University, Tacloban City, Philippines;
micheline.gotardo@lnu.edu.ph

## Abstract

**Objective**: Everyone has the right to education. For Higher Educational Institutions, students serve as its best asset. The prediction of students' success in their academic performance is then vital for it will benefit both students and professors, enabling the latter to do proactive measures and find ways in helping students learn, ultimately improving their academic performance. **Methods**: This study utilized the Data mining technique, specifically; the J48 algorithm was used to create the Decision Tree Model in predicting the Student Performance in Data Structures and Algorithms. For model accuracy, K-fold cross-validation and Receiving Operating Characteristics Curve (ROC) was used. The datasets used were collected from the grades of 2nd year BSIT students enrolled during the school year 2015-2016. **Findings**: The generated Decision Tree Model results showed that Finals had the highest instance and in predicting student performance in the Data Structures and Algorithms subject. It also shows that Finals has the highest factor to receive either of the following remarks: Pass, Failed or Conditional. The model was also able to identify 85.31% accuracy for the attribute Pass, 79.41% accuracy for the attribute Conditional and 91.67% accuracy for the attribute Failed. Further, the Decision Tree Model likewise revealed that for the student to pass the Data Structures and Algorithms subject they should have a grade higher than 66.12% in Midterms and a grade higher that 72.30% in Finals. **Application/Improvements:** The use of the data driven system can be used by institutions to track student performance. Data analysis is a key component to further strengthen their policies and do intervention programs where it is highly needed. Further, for more improvement of this study additional data mining techniques can be applied.

**Keywords:** Data Mining, Data Structures and Algorithms, Decision Tree Algorithm, Information Technology, Student Performance

## 1. Introduction

Everyone has the right to education[1]. Education in the Philippine is prioritized by parents; it is indispensable, a national legacy which should be instilled in every generation[2]. The Philippine Educational System had undergone various development and changes to equip its graduates with the necessary skills to be competitive with other graduates from other countries. In fact, the Commission on Higher Education (CHED) had issued CMO 46, series of 2012, known as the Philippine Higher Education through an Outcomes-based and Typology-based QA. This mandate translates to multiple missions for the Philippine Higher Education System, one of which is producing graduates with high levels of academic thinking, behavioral and technical skills/competencies that are aligned with national academic and industry standards and needs and international standards, when applicable. However, despite its efforts, it is still observed that the academic performance of students is low. Although, uni-

*Author for correspondence

versities collect an enormous amount of students' data, this remains unutilized and does not help in any decisions or policy making to improve the performance of students[3].

Earlier identifaction of factors contributing to the low performance of students is important. Students' are the asset of a university. Students' performance (academic achievement) plays an important role in producing the best quality graduates who will become a great leader and manpower for the country thus responsible for the country's economic and social development[4]. Performance is an observable or measurable behavior of a person in a particular situation[5,6]. On the other hand, Academic Performance or Academic Achievement, represents the performance outcomes by a person indicating how far that person accomplished specific goals that were the focus on the different activities in the learning process[7]. Typically, student academic performance is measured by the grades acquired by completing requirements set by their professors. Meanwhile, results of a study views that student's performance is linked with the student's profile: attitude towards class attendance, time allocation for studies, parent's level of income, mother's age and mother's education[8].

Data mining is the process of sorting through large data sets[9]. Its purpose is to identify patterns and be able to establish relationships to solve problems through data analysis ultimately allowing prediction of future trends. The main functions of data mining are applying various methods and algorithms to discover and extract patterns of stored data[10]. Further, its significance to decision making makes it an essential component in various organizations. Research interest in predicting student academic performance has been increasing. A research using 1,547 datasets made use of Decision Tree (ID3) method to predict the final grades of students[11]. Predictors like the Midterm Marks, Lab Test Grade, Seminar Performance, Assignment, Measure of Student Participation, Attendance, Homework and Final Grade Marks were used. The use of ID3 resulted in the following: 292 students were "Excellent," 536 "Very Good," 477 "Good," 188 "Acceptable" and 54 "Fail". Additionally, another research made use of Educational Data Mining (EDM) from the 60 students datasets from MCA course in Pimpri Chinchwad College of Engineering at Pune University[12]. Attributes like students graduation percentage, assignment work, attendance and unit test performance were used to determine how these affect the students' university result. The findings of the research were that for the student to have good performance, a student should be good in their attendance, assignment and Unit test.

On the other hand, a research to determine the success of students in higher educational institutions was made through the use of the J48 alogrithm[13]. The researchers conducted a 60 questions survey covering the following fields: Social activity, relationships, health and academic performance. Results show that age, work, gender, stage and status has fewer effects on students' success, but students' GPA, credits, list of important notes, father work and fresh food was the most significant effect on the student success. A Research was also made on 158 students of the Information Technology Department of King Saud University, Saudi Arabia by using three classifiers: C4.5 decision tree, Naïve Bayes, and JRip[14]. The student performance of students enrolled in its Data Structures subject was the focus of the research for it has the high failure rate. Student ID, student name, grades in quiz 1, quiz 2, quiz 3, midterms 1, midterms 2, project, tutorial, final exam and total points obtained were the attributes used, from which the attirbute midterm 1 was the highest indicator in determining students' performance in the subject.

From these literatures, it can be said that predicting student academic performance is crucial in helping educators plan and strategize their lesson delivery. In conventional teaching environments, educators are able to obtain feedback on student learning experiences in face-to-face interactions with their students, enabling continual evaluation of their teaching programs[15]. But with the integration of technology in learning environments, in order to get this information, educators must find other ways to attain these. Results of the predictive model will help educators take measures to help improve students' performance. This study will make use of the

J48 algorithm a data mining technique in the prediction of the academic performance of students in their Data Structures and Algorithm subject.

## 2. Methodology

This study will make use of the Knowledge Discovery in Database (KDD) process. KDD revolves on the investigation and creation of knowledge, processes, algorithms and the mechanisms for retrieving potential knowledge from data collections[16].

### 2.1 Data Collection

A total of 108 datasets were collected from the grades of the BSIT 2nd year students enrolled in Data Structures and Algorithms during the school year 2015-2016.

### 2.2 Variables used

#### *2.2.1 Lab Exercises/Project (LEP)*

Lab Exercises are given to students after finishing the topic. These exercises are designed to challenge students with their critical thinking skills. Project is given after the Midterms Exam and serves as a completion requirement for the subject.

#### *2.2.2 Quizzes (Q)*

May come in the form of announced and pop quizzes. These are used to gauge students' understanding and comprehension of the lesson. Grades are computed as Raw score divided by the total number of items multiplied by 35 plus 60.

#### *2.2.3 Midterms (M)*

This is given during the middle of the semester. This helps the professor in determining how the students learned and fully understand the lessons.

#### *2.2.4 Finals (F)*

This is given before the end of the semester.

### 2.3 WEKA Software

The Waikato Environment for Knowledge Analysis (WEKA) software was used in the study. With GNU General Public license, WEKA is an open source software. WEKA is a collection of machine learning algorithms for data mining tasks, which contain tools for data preparation, classification, regression, clustering, association rules mining, and visualization[17].

### 2.4 Data Mining Process

Students grades are stored using the MS Excel application and then later converted into a Microsoft Excel Comma Separated Values File (.csv). Notepad++ was used to load the .csv file, and at this point, data cleaning is performed by eliminating unwanted symbols (e.g. spaces, comma and colon). As a requirement for the WEKA application, the following syntax: @Relation, @Attribute and @Data were included. Still, with the use of Notepad++, the file is then converted to Attribute-Relation File Format (ARFF). This file format was developed for use with the WEKA Software. It is an ASCII text file that describes a list of instances sharing a set of attributes. Information is then uploaded to the WEKA Application and the conversion of the pre-processed raw data to a more understandable file format.

Next, the data modelling stage consists of five phases: Training, pattern, testing, result evaluation and knowledge representation. This is also where WEKA is used for the prediction of the Student Performance in the Data Structures and Algorithms Course. Next, cross-validation was used. Cross-validation is a model evaluation method where the entire data will not be utilized when training a learner[18]. Its most straightforward technique is called the holdout method. Here, data is divided into two, namely, the training set and test set. The training set is used to train the model, while the test set is used to evaluate it.

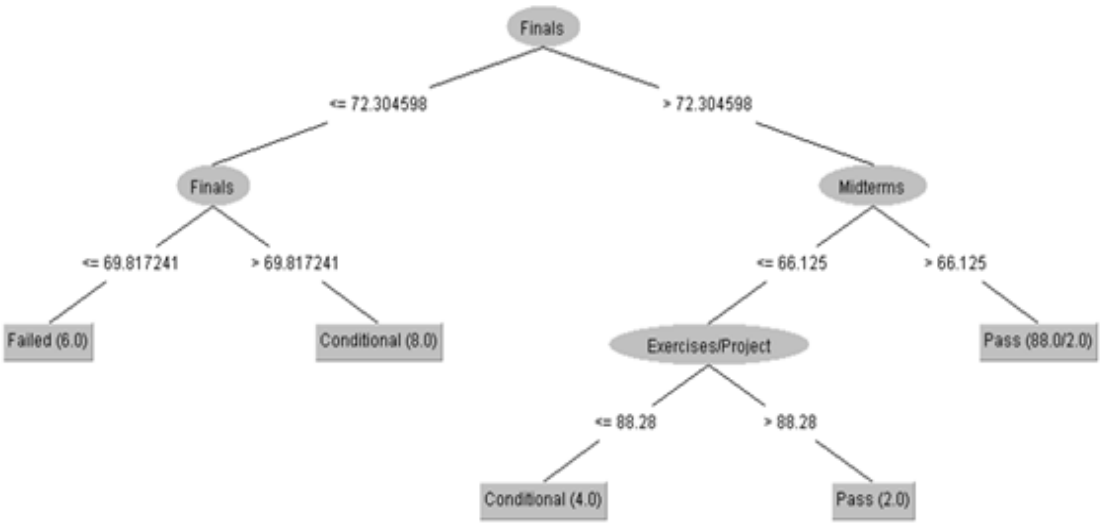
The J48 algorithm is used in the training stage and was used to build a model. The J48 classifier is a simple C4.5 decision tree for classification for the creation of a binary tree[19]. The testing stage, on the other hand, is where the K-fold cross validation is performed. This study made

use of 10-fold cross-validation. K-fold cross-validation is one way to improve the holdout method where the data set is divided into k subsets, and the holdout method is repeated k times. For each repetition, one of the k subsets is used as the test set and the other as the training set. For model accuracy, the Receiving Operating Characteristics Curve (ROC) Area under ROC Curve technique is used. ROC Area under ROC Curve techniques is a universal biostatistical tool for describing the accuracy of a model regarding predicting a phenomenon[20].

## 3. Result and Discussion

### 3.1 The Model

Figure 1 shows the graphical presentation of the pruned decision tree on Student Performance in Data Structures and Algorithms. Finals had the highest instance and became the basis for the first split between Finals <= 72.30 and Midterms > 72.30 in predicting student performance in the Data Structures and Algorithms




R1: IF (Finals <=72.30) AND (Finals <= 69.82) THEN Performance = "FAILED"
R2: IF (Finals <=72.30) AND (Finals > 69.82) THEN Performance = "CONDITIONAL"
R3: IF (Finals > 72.30) AND (Midterms <= 66.12) AND (Exercises/Project <= 88.28) THEN Performance = "CONDITIONAL"
R4: IF (Finals > 72.30) AND (Midterms <= 66.12) AND (Exercises/Project > 88.28) THEN Performance = "PASS"
R5: IF (Finals > 72.30) AND (Midterms > 66.12) THEN Performance = "PASS"

**Figure 2.** Student performance decision rule.

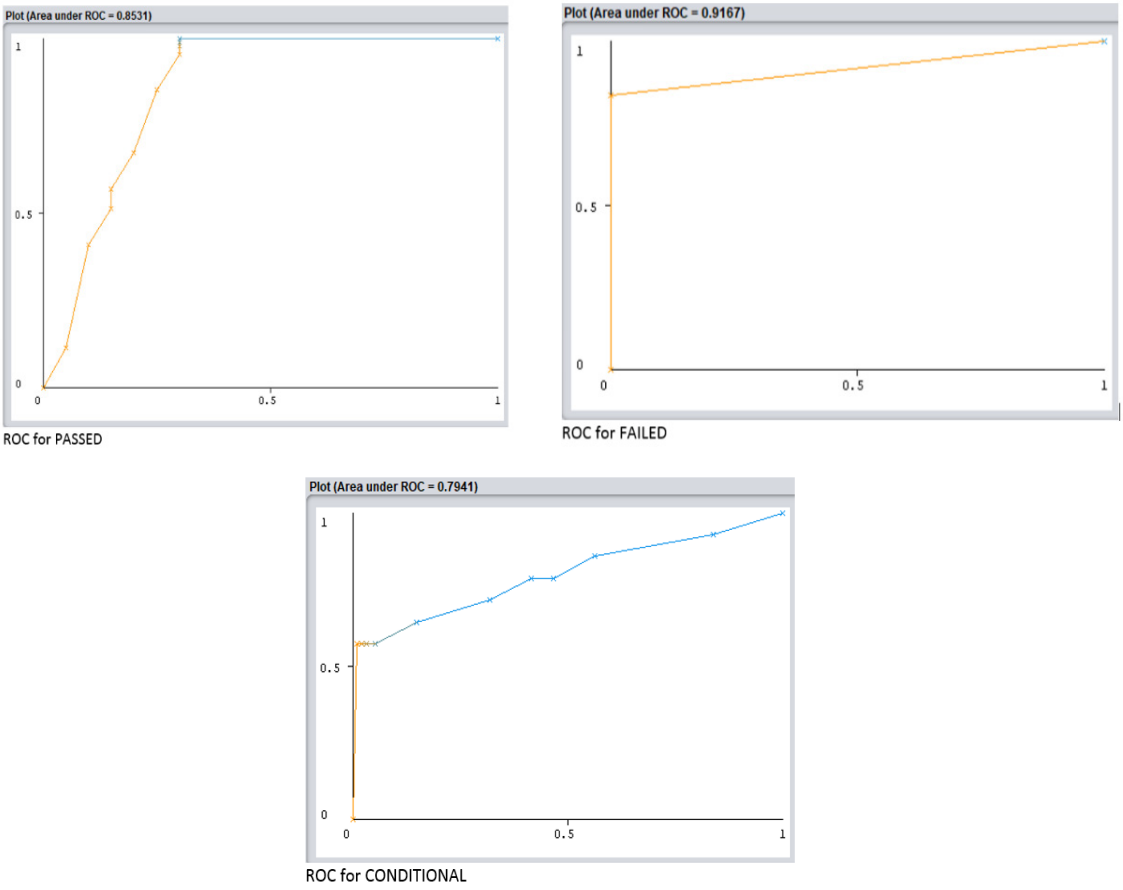



**Figure 3.** ROC curve.

**Table 1.** Confusion matrix

| a | b | c | <-- classified as |
|---|---|---|---|
| 86 | 0 | 2 | a = Pass |
| 0 | 5 | 1 | b = Conditional |
| 6 | 0 | 8 | c = Failed |

subject. Additionally, Figure 2 shows the student performance decision rule that Finals has the highest factor to receive either of the following remarks: Pass, Failed or Conditional.

The confusion matrix in Table 1 reflects the correctly classified instances and the misclassification of the students' performance. A total of 108 classifications were made. The confusion matrix table shows the following results:

- The decision tree has classified eighty-six (86) instances as PASS and six (6) as FAILED leading to six (6) misclassifications.

**Table 2.** Cross-validation summary

| | | |
|---|---|---|
| Correctly Classified Instances | 99 | 91.6667% |
| Incorrectly Classified Instances | 9 | 8.3333% |
| Kappa statistic | 0.7128 | |
| Mean absolute error | 0.0642 | |
| Root mean squared error | 0.2241 | |
| Relative absolute error | 29.4072 % | |
| Root relative squared error | 68.831 % | |
| Total Number of Instances | 108 | |
| Cross-Validation 10-folds | | |

**Table 3.** Detailed accuracy by class

| | TP | FP | | | | | ROC Area | PRC Area | |
|---|---|---|---|---|---|---|---|---|---|
| | Rate | Rate | Precision | Recall | F-Measure | MCC | | | Class |
| | 0,977 | 0.300 | 0.935 | 0.977 | 0.956 | 0.741 | 0.853 | 0.937 | Pass |
| | 0.833 | 0.000 | 1.000 | 0.833 | 0.909 | 0.908 | 0.917 | 0.843 | Failed |
| | 0.571 | 0.032 | 0.727 | 0.571 | 0.640 | 0.599 | 0.794 | 0.602 | Conditional |
| Weighted Avg. | 0.917 | 0.249 | 0.912 | 0.917 | 0.912 | 0.732 | 0.849 | 0.888 | |

- The decision tree has classified five (5) instances as CONDITIONAL leading to zero (0) misclassifications; and

- The decision tree has classified two (2) instances as PASS, one (1) instance as CONDITIONAL and eight (8) instances as FAILED leading to three (3) misclassifications.

Table 2 shows the Cross-Validation Summary, wherein 91.67% instances were correctly classified as compared to 8.3% instances incorrectly classified. Results from Table 2 are supported by the results shown in Table 3 where it shows the complete accuracy by the class which the Precision weighted average of the student performance in Data Structures and Algorithms is 91.2%. The study also utilized the Receiving Operating Characteristics Curve (ROC) and the Area under ROC Curve (AUC) for model accuracy. Figure 3 shows that the attribute Pass has 85.31% accuracy, Conditional has 79.41% accuracy and Failed has 91.67% accuracy in Predicting the Student Performance in Data Structures and Algorithms.

### 3.2 Student Performance

For higher education institutions whose goal is to contribute to the improvement of the quality of higher education, the success of creation of human capital is the subject of a continuous analysis[21]. Result of the study made in Cordoba University, using 438 datasets in 7 Moodle courses, showed that Quizzes was the main determiner for the final marks of the students[22]. Though, Quizzes was the main determiner for the good performance of the students, the researchers also mentioned that the result could help teachers decide to promote the use of some activities to obtain higher marks or eliminate some activities because they are related to low marks.

The prediction of students' success in their academic performance is then vital for it will benefit not only the students but its professors as well. Professors in their part, will be able to proactive measures in helping students and find ways to help students learn ultimately improving their academic performance. The Decision Tree Model was able to predict 85.31% accuracy for Pass, 79.41% accuracy for Conditional and 91.67% accuracy for Failed based on the ROC curve shown of Figure 3. The Decision Tree Model likewise revealed that for the student to pass the Data Structures and Algorithms subject they should have a grade higher than 66.12% in Midterms and a grade higher than 72.30% in Finals. The Finals attribute serves as the highest indicator that can affect the student. Data Structures and Algorithms is essential for BS Information Technology course. Data structures refer to the way information is organized, while algorithms refer to the step-by-step procedure used to solve a problem. To be a good programmer, these two should be mastered by the students.

## 4. Conclusion

Research interest in predicting student academic performance has been increasing. Knowing beforehand the attributes that significantly affects the performance of student greatly helps professors in doing proactive measures for the students' benefit. This study focused on the attributes of the Data Structures and Algorithms course that will affect students' performance. J48 algorithm was used for the creation of the decision tree model, therefore, identifying that the Finals attribute gained the highest indicator that is crucial for the students passing the subject. Most importantly, a model was established in determining the Students Performance in Data Structures and Algorithms as shown in the Decision Tree, Confusion Matrix, ROC, and the Area under ROC Curve. Further, the use of the data driven system can be used by institutions to track student performance. Data analysis is a key component to further strengthen their policies and do intervention programs where it is highly needed.

## 5. References


1. Universal declaration of human rights. 1948. http://www.un.org/en/universal-declaration-human-rights/
2. Education very important to Filipinos. 1996. http://www.scmp.com/article/149110/education-very-important-filipinos
3. Performance analysis and prediction in educational data mining: A research travelogue. 2015. https://arxiv.org/abs/1509.05176
4. Ali NJ, Ali N, Jusof K, Ali S, Mokhtar N, Salamat A. The factors influencing students' performance at Universiti Teknologi Mara Kedah, Malaysia. Management Science and Engineering. 2009; 3(4):81–90.
5. Yusuf A. Interrelationships among academic performance, academic achievement and learning outcomes. Journal of Curriculum and Instruction. 2002; 1(2):87–96.
6. Simpson J, Weiner ES. Oxford English dictionary online. Oxford: Clarendon Press; 2008. p. 1–10. PMCid: PMC2235879.
7. Academic Achievement. 2017. http://www.oxfordbibliographies.com/view/document/obo-9780199756810/obo-9780199756810-0108.xml#backToTop
8. Factors affecting Students' performance. 2006. https://ideas.repec.org/p/pra/mprapa/13621.html
9. Twongyirwe T, Lubega J. Evaluation of a knowledge management dashboard for manufacturing SMEs in resource constrained areas. International Conference on Intellectual Capital and Knowledge Management and Organisational Learning; 2018. p. 1–397.
10. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996; 17(3):1–37.
11. Ahmed AD, Elaraby IS. Data Mining: A prediction for student's performance using classification method. World

Journal of Computer Application and Technology. 2014; 2(2):43–7.
12. Borkar S, Rajeswari K. Predicting student's academic performance using education data mining. International Journal of Computer Science and Mobile Computing. 2013; 2(7):273–9.
13. Hamuod A, Hashim A, Awadh W. Predicting student performance in higher education institutions using Decision Tree Analysis. International Journal of Interactive Multimedia and Artificial Intelligence. 2018; 5(2):26–31. https://doi.org/10.9781/ijimai.2018.02.004.
14. Al-barrak MA, Al-razgan MS. Predicting students' performance through classification: A case study. Journal of Theoretical and Applied Information Technology. 2015; 75(2):1–8.
15. Sheard J, Ceddia J, Hurst J, Tuovinen J. Inferring student learning behaviour from website interactions: A usage analysis. Education and Information Technologies. 2003; 8(3):245–66. https://doi.org/10.1023/A:1026360026073.
16. Norton MJ. Knowledge discovery in databases. Library Trends. 1999; 48(1):1–9.
17. WEKA 3: Data Mining Software in Java. 2019. https://www.cs.waikato.ac.nz/ml/weka/
18. Caluza LB. Predicting teachers' ICT competence in a Philippine University using J48 algorithm. Indian Journal of Science and Technology. 2018; 11(7):1–7. https://doi.org/10.17485/ijst/2018/v11i7/119062.
19. Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications. 2013; 6(2):256–61.
20. Vexler A, Yu J, Zhao Y, Hutson AD, Gurevich G. Expected p-values in light of an ROC curve analysis applied to optimal multiple testing procedures. Statistical Methods in Medical Research. 2018; 27(12):3560–76. PMid: 28504080 PMCid: PMC6212326. https://doi.org/10.1177/0962280217704451.
21. Osmanbegovic E, Suljic M. Data mining approach for predicting student performance. Economic Review. 2012; 10(1):3–12.
22. Romero C, Ventura S, Espejo PG, Hervas C. Data mining algorithms to classify students. Educational Data Mining; 2008. p. 1–10.