

# An Analysis on Student Academic Performance by Using Decision Tree Models

**Jastini Mohd. Jamil\***

School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

**Nurul Farahin Mohd Pauzi**

School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

**Izwan Nizal Mohd. Shahara Nee**

School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

## Abstract

Large volume of educational data has led to more challenging in predicting student's performance. In Malaysia currently, study about the performance of students in Malaysia institutions is very little being addressed. The previous studies are still insufficient to identify what factors contribute to student's achievements and lack of investigations on exploring pattern of student's behaviour that affecting their academic performance within Malaysia context. Therefore, predicting student's academic performance by using decision trees is proposed to improve student's achievements more effectively. The main objective of this paper is to provide an overview on predicting student's academic performance using by using data mining techniques. This paper also focuses on identifying the pattern of student's behaviour and the most important attributes that impact to the student's achievement. By using educational data mining techniques, the students, lecturers and academic institution are able to have a better understanding on the student's achievement.

**Keywords:** Academic performance; Data mining; Decision tree.



CC BY: [Creative Commons Attribution License 4.0](https://creativecommons.org/licenses/by/4.0/)

## 1. Introduction

Student academic performance is important either to student itself or to a country development. This is because one of the component for a high ranking university is based on its excellent record of academic performances. Generally, students who get a better result in higher learning institution are more likely to be employed and have high salaries. Based on the previous literature, there are a lot of definitions on student's performance. [Mat et al. \(2013\)](#) stated that student's performance can be obtained by measuring the learning assessment and co-curriculum. However, most of the studies mentioned about graduation being the measure of student's success.

Currently, there are many techniques being utilized to analyse student's performance. Data mining is one of the most popular techniques to evaluate student's performance in educational area recently. Data mining is a process used to find hidden interesting information and patterns from a huge database [Romero and Ventura \(2010\)](#). As a result, the useful information and patterns would help the students, lecturers and universities in improving an effective teaching and learning approach. This study will focus on:

1. Identifying the significant variables used in analysing student performance
2. Developing decision tree models for predicting student performance
3. Identifying the best decision tree model for predicting student performance

The next section will focus on previous works related to decision tree and student academic performance. Then, a discussion on research methodology will be described in Section 3. Lastly, in Section 4, the detail results on the decision tree prediction methods are discussed and future work are outlined.

## 2. Literature Review

This study was aiming to compare and evaluate nine different type of decision tree model as to classify the student's Cumulative Grade Point Average (CGPA) into groups either below or above CGPA of 3.00. However, it is better to start with reviewing previous work that related to this study.

[Han and Kamber \(2001\)](#) describe data mining as a tool that help the users to analyse data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. [Shana and Venkatachalam \(2011\)](#), utilized difference feature selection methods and have found that the influence of features give an impact on accuracy of the student performance model. [Baradwaj and Pal \(2011\)](#) analysed a data set of 50 Post Graduate students for mining students' performance using Decision tree method. They used different attributes that are not taken into consideration were economic background, technology exposure etc. [Ruby and David \(2015\)](#), conducted a research on the predicting the students' academic performance and identified that 7 factors out of 16 initial factors are high influencing factors using various feature selection techniques like Chi square, Information Gain, Correlation, Linear Regression and Gain Ratio. [Ramaswami and Bhaskaran \(2010\)](#) have presented a predictive

model using data mining approach to discover students' performance patterns in Maths, English and programming courses.

El-Hales (2008) did a study on evaluating student behaviour with identified factors that includes only personal and academic details of 151 students. Classification based on Decision tree was done followed by clustering and outlier analysis. Mohammed *et al.* (2012) applied the support vector machine and K-Nearest neighbour for discovering knowledge of graduate students' performance where the data had been collected for a period of 15 years [1993-2007]. Nguyen *et al.* (2007), compared the efficiency of decision tree and Bayesian technique for predicting the academic performance of Under Graduate and Post Graduate students. They found that the decision tree provided better accuracy than the Bayesian. Ruby and David (2015) compared difference data mining algorithms and proved that multilayer perceptron neural network shows a best prediction which is followed by decision tree.

In educational data mining, there are several algorithms under classification task that have been applied to predict student's performance. Among the algorithms used are Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine. This study was aiming to compare and evaluate nine different type of decision tree model as to classify the student's CGPA into groups either below or above CGPA of 3.00. Decision Tree is one of a popular technique for prediction. Most of researchers have used this technique because of its simplicity and comprehensibility to uncover small or large data structure and predict the value (Kuyoro and Nnicolae, 2013; Vasile, 2007). Romero and Ventura (2007) said that the decision tree models are easily understood because of their reasoning process and can be directly converted into set of IF-THEN rules. Examples of previous studies using Decision Tree method are predicting drop out features of student's data for academic performance Agarwal *et al.* (2012), predicting third semester performance of MCA students Ajith *et al.* (2013) and also predicting the suitable career for a student through their behavioral patterns Al-Radaideh *et al.* (2006).

### 3. Methodology

The dataset used for this study was taken from Universiti Utara Malaysia (UUM). The dataset consists of 7672 cases of intake, which included gender, ethnics, entry level, program, parents' income and age as explanatory variables, whereas the

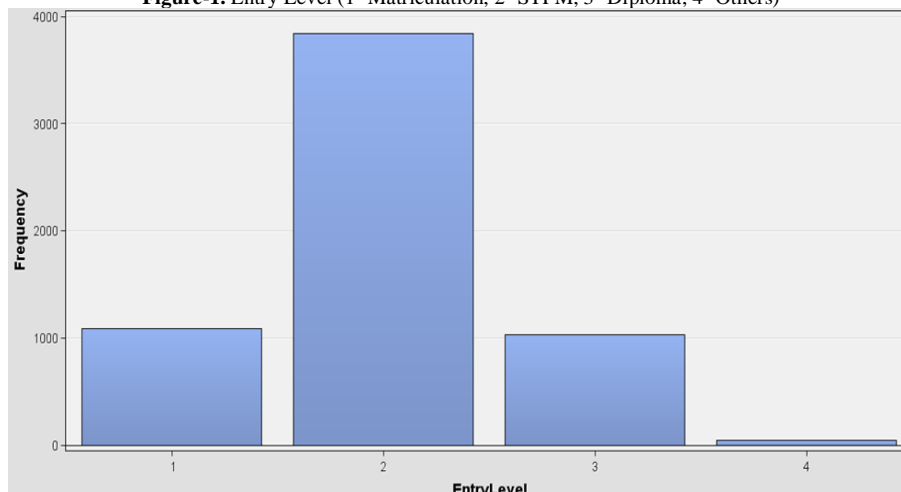
Response variable will be students' CGPA in the binary target (1 - low achiever, 0 - high achiever) as shown in Table 1. The influencing attributes are selected and are used to classify and predict the student performance using SAS data mining tool.

Table-1. The description of each response variables

Name	Model Role	Measurement Level	Description
Age	Input	Interval	Age of the students which are measured in years
CGPA	Target	Binary	A CGPA level where 0 indicates more or equal to 3.00 pointers, and 1 indicates less than 3.00 pointers
EntryLevel	Input	Nominal	The entry level of students to get into UUM
Ethnics	Input	Nominal	The ethnics of students
Gender	Input	Binary	Gender of students
ID	ID	Nominal	Number of observations
Parents Income	Input	Ordinal	Parents' income of students
ProgramCode	Input	Nominal	Categories of program study by students

From the data, it was found that more female is enrolled in degree compared to male. As for the entry level of intake, most were from STPM, followed by Matriculation, Diploma, and others.

Figure-1. Entry Level (1- Matriculation, 2- STPM, 3- Diploma, 4- Others)



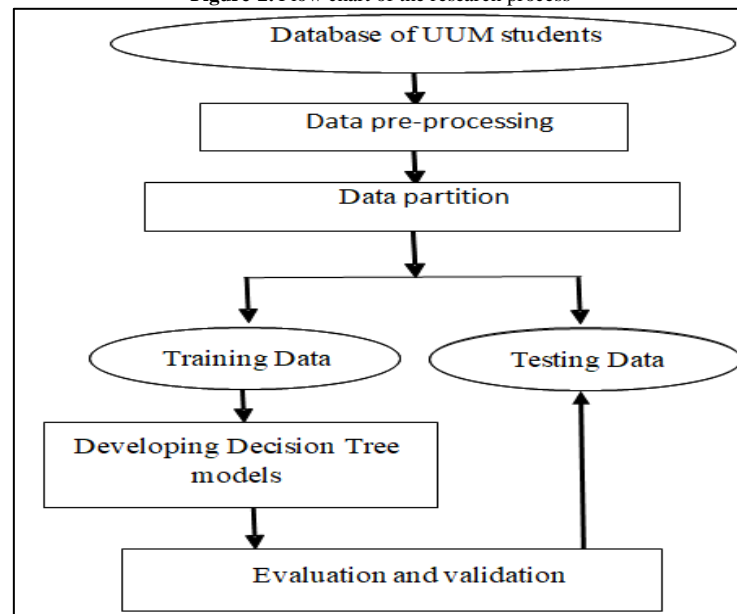
The mode of the parents' income is at the range of RM1001.00 to RM2000.00, followed by RM501.00 to RM1000.00, RM2001.00 to RM3000.00, RM3001.00 to RM4000 and so on, whereas the least income range is RM10001.00 and above.

Table-2. Parents Income

Range	Cases
1.00 - 500.00	369
501.00 - 1000.00	1553
1001.00 - 2000.00	2775
2001.00 - 3000.00	1126
3001.00 - 4000.00	593
4001.00 - 5000.00	443
5001.00 - 7500.00	436
7501.00 - 10000.00	215
10001.00 and above	162

The methodology will be described according to the SEMMA procedure where the input data will first be identified, followed by data description and pre-processing of the dataset. Next, the process used to develop the decision tree models where to classify the CGPA of the student into two groups which are below 3.00 and above 3.00.

Figure-2. Flow chart of the research process



We have tested different criterion for each of the method. For decision tree, we have tested several combinations of the number of branches (2, 3 and 4 branches) and different target splitting rules (Entropy, Gini, and Probability Chi-square). The steps required to develop the model are as shown:

Step 1: Let  $N-a$  (70%) as training sets and  $a$  (30%) as test sets, where  $N$  is the number of cases.

Step 2: Design the decision tree that consists of 2 branches with a maximum depth of 4.

Step 3: For Entropy impurity,

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t) \quad (1)$$

For Gini impurity,

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \quad (2)$$

For probability Chi-square

$$\text{Probability square}(t) = 1 - \max[p(i|t)] \quad (3)$$

where  $c$  is the number of classes.

Step 5: Calculate weighted average impurity, choose the parent's node with lowest weighted average impurity.

$$\sum_{j=1}^k \frac{N(V_j)}{N} I(V_j) \quad (4)$$

where  $I(.)$  is the impurity measure of a given node,  $N$  is the total number of records at the parent node,  $k$  is the number of attributes value (class),  $N(V_j)$  is the number of records associated with the child node  $V_j$ .

Step 6: Determine the gain of each independent variable, the highest gain will be chosen as a first child node.

$$\text{Gain}, \Delta = I(\text{Parent}) - \sum_{j=1}^k \frac{N(V_j)}{N} I(V_j) \quad (5)$$

Step 7: Repeat the steps until reaches a depth of 4 branches.

We have ran total of nine type criterion for decision tree methods. At last, we continue with the model comparison to compare the result that are obtain from these 9 models to find out the best fit model and have the best

solution for our classification problem. In the comparison, we will choose the best solution based on the misclassification rate, the lowest rate of training and testing data is desired.

#### 4. Results and Conclusions

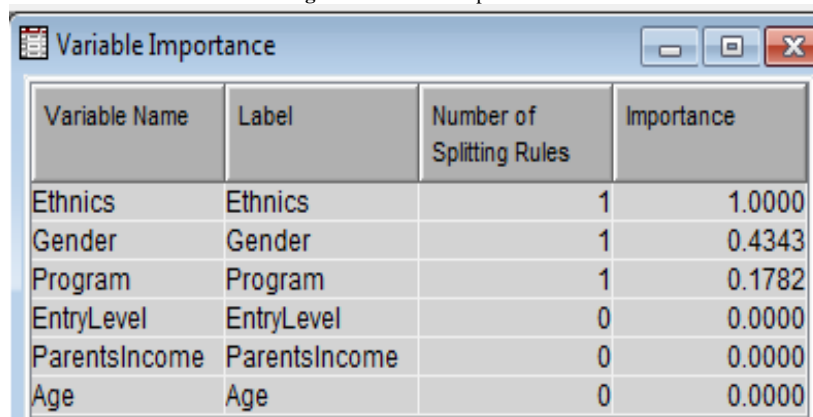
In this study, there are nine decision tree models were tested by different combination of number of branches and interval splitting rules. Thus, based on the Table 3, the decision tree model with 2 branches + Probability Chi-square splitting method, and the decision tree model with 3 branches + Probability Chi-square splitting method have the smallest misclassification rate in testing data which is 29.79%. However, decision tree model with 3 branches + Probability Chi-square splitting method have smaller misclassification rate in training data than decision tree model with 2 branches + Probability Chi-square splitting method ( $28.18\% < 28.41\%$ ). Thus, decision tree model with 3 branches + Probability Chi-square splitting method is chosen as the best model to classify student performance among decision tree models.

**Table-3.** Model Comparison Table

Model	Misclassification Rate	
	Train	Test
Decision Tree (2 branches + Entropy)	25.64%	32.45%
Decision Tree (2 branches + Gini)	24.94%	34.04%
Decision Tree (2 branches + ProbChisq)	28.41%	29.79%
Decision Tree (3 branches + Entropy)	23.09%	34.57%
Decision Tree (3 branches + Gini)	22.63%	36.17%
Decision Tree (3 branches + ProbChisq)	28.18%	29.79%
Decision Tree (4 branches + Entropy)	22.63%	32.98%
Decision Tree (4 branches + Gini)	24.94%	34.04%
Decision Tree (4 branches + ProbChisq)	30.02%	30.32%

Based on the best model chosen (3 branches + ProbChisq), we can classify the students into high achiever (3.00 pointer and above) and low achiever (3.00 pointer and below) precisely. According to the Figure 3, the most important variable in the best model is ethnics. Therefore, the parent node of this decision tree model is ethnics. This also represent the race of student is the most important variable to predict the CGPA of students either higher or lower than 3.00. However, the variable parent's income, age and entry level showed zero importance. In other words, the parent's income, student's age and admission qualification are the least important variables in this model and are excluded in this model.

**Figure-3.** Variable importance



Variable Name	Label	Number of Splitting Rules	Importance
Ethnics	Ethnics	1	1.0000
Gender	Gender	1	0.4343
Program	Program	1	0.1782
EntryLevel	EntryLevel	0	0.0000
ParentsIncome	ParentsIncome	0	0.0000
Age	Age	0	0.0000

There are seven leaf nodes generated from the decision tree model. Some of the interesting rules generated from the decision tree are shown in the Table 3.

**Table-3.** Interesting rules generated from the decision tree model

<b>Node 3</b> <b>If ethnics : Cina</b> <b>Then</b> <b>Number of Observations : 891</b> <b>Predicted class : higher achiever = 90.59%</b> <b>Predicted class : lower achiever = 9.41%</b>	<b>Node 9</b> <b>If ethnics : Melayu, India, Others</b> <b>AND Program : SARJANA MUDA PERAKAUNAN, SARJANA MUDA SAINS EKONOMI, SARJANA MUDA PEMASARAN, SARJANA MUDA KEWANGAN, SARJANA MUDA PERBANKAN, SARJANA MUDA KEWANGAN DAN PERBANKAN ISLAM</b> <b>AND Gender = Male</b> <b>Then</b> <b>Number of Observations : 479</b> <b>Predicted class : higher achiever = 11.43%</b> <b>Predicted class : lower achiever = 88.57%</b>
<b>Node 11</b> <b>if ethnics : Melayu, India, Others</b> <b>AND Program : SARJANA MUDA PENDIDIKAN PERAKAUNAN, SARJANA MUDA PENDIDIKAN TEKNOLOGI MAKLUMAT, SARJANA MUDA UNDANG-UNDANG</b> <b>AND Gender : Female</b> <b>Then</b> <b>Number of Observations : 204</b> <b>Predicted class : higher achiever = 67%</b> <b>Predicted class : lower achiever = 33%</b>	<b>Node 16</b> <b>if ethnics : Melayu, India, Others</b> <b>AND Program : SARJANA MUDA PENGURUSAN KERJA SOSIAL, SARJANA MUDA PENGURUSAN HAL EHWAL ANTARABANGSA</b> <b>AND Gender : LELAKI</b> <b>Then</b> <b>Number of Observations = 538</b> <b>Predicted class : higher achiever = 24%</b> <b>Predicted class : lower achiever = 76%</b>

In conclusion, after we compared nine decision tree models to classify student's academic performances in term of CGPA, we found that the decision tree with 3 branches and Probability Chi-square is the best technique among other models to predict either above (higher achiever) or below (lower achiever) CGPA of 3.00. Therefore, this decision tree model able to help us to determine the important variable that affecting academic performance as to furnish the higher-level education sectors with better investment in fostering a bunch of quality undergraduates. With using decision tree model, university authority or the Kementerian Pendidikan Malaysia will have a better understanding on the factors affecting students' CGPA so that further actions can be more effective and impactful in building a better generation.

As a recommendation for further research, we would suggest that more factors that can be used in this study such as socio-economic factors, socio-cultural factors or external and internal factors since we just focus on demographic factors in this study.

## Acknowledgement

The authors would like to acknowledge the work that led to this paper, which was fully funded by the Research Generation University Grant, Universiti Utara Malaysia.

## References

- Agarwal, S., Pandey, G. N. and Tiwari, M. D. (2012). Data mining in education, Data classification and decision tree approach.
- Ajith, P., Tejaswi, B. and Sai, M. S. S. (2013). Rule mining framework for students performance evaluation. *International Journal of Soft Computing and Engineering*, 2(6): 2231 – 307.
- Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M., 2006. "Mining student data using decision trees." In *International Arab Conference on Information Technology (ACIT)*.
- Baradwaj, B. K. and Pal, S. (2011). Mining educational data to analyze students' performance.
- El-Hales, A., 2008. "Mining students data to analyze learning behavior, A case study." In *The 2008 International Arab Conference of Information Technology(ACIT2008)- Conference Proceedings, University of Sfax, Tunisia*.
- Han, J. and Kamber, M. (2001). *Data mining, Concepts and techniques*. Morgan Kaufmann Publishers: San Francisco.
- Kuyoro, S. O. and Nnicolae, G. (2013). Oludele Awodele and Samuel Okolie, Optimal algorithm for predicting students' academic performance. *International Journal of Computers and Technology*, 4(1, JAN-FEB):
- Mat, U., Buniyamin, N., Arsad, P. M. and Kassim, R., 2013. "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention." In *Engineering Education (ICEED), 2013 IEEE 5th Conference on, IEEE 2013*. pp. 126-30.
- Mohammed, M., Abu, T., Alaa, M. and El-Halees (2012). Mining educational data to improve students' performance, A case study.
- Nguyen, N., Paul, J. and Peter, H., 2007. "A comparative analysis of techniques for predicting academic performance." In *Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference*. pp. 7-12.

- Ramaswami, M. and Bhaskaran, R. (2010). CHAID based performance prediction model in educational data mining. *IJCSI International Journal of Computer Science Issues*, 7(1).
- Romero, C. and Ventura, S. (2007). Educational data mining, A survey from 1995 to 2005. *Expert Systems with Applications*, (33): 135-46.
- Romero, C. and Ventura, S. (2010). Educational data mining. *A Review of the State of the Art. IEEE*,: 601-18.
- Ruby, J. and David, K. (2015). Analysis of influencing factors in predicting students performance using MLP - A comparative study. *International Journal of Innovative Research in Computer and Communication Engineering*, 3(2).
- Shana, J. and Venkatachalam, T. (2011). Identifying key performance indicators and predicting the result from student data. *International Journal of Computer Applications*, 25(9).
- Vasile, P. B., 2007. "Analysis and predictions on students' behavior using decision trees in weka environment." In *Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, IEEE*.