International Journal of Engineering & Technology, 7 (2.15) (2018) 61-63

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

A Student Performance Prediction Model Using Data Mining Technique

Rohaila Abdul Razak[1*], Mazni Omar[2], Mazida Ahmad[3]

[1]*Kolej Poly-Tech MARA*
[2]*Universiti Utara Malaysia*
[3]*Universiti Utara Malaysia*
**Email: rohaila@gapps.kptm.edu.my*

Abstract

Predicting performance is very significant in the education world nowadays. This paper will describe the process of doing a prediction of student performance by using data mining technique. 257 data sets were taken from the student of semester 6 KPTM that involved four (4) academic programs which are Diploma in Computer System and Networking, Diploma in Information Technology, Diploma in Business Management and Diploma in Accountancy. Knowledge Discovery in Database (KDD) was used as a guide to the process of finding and extracting a knowledge from the dataset. A decision tree and linear regression were used to analyze the dataset based on variables selected. The variables used are Gender, Financing, SPM, GPASem1, GPASem2, GPASem3, GPASem4, GPASem5 and CGPA as a dependent variable. The result from this indicate the significant variable that contribute most to the students' performance. Based on the analysis, the decision tree shows that GPASem1 has a strong significant to the CGPA final semester of the student and the prediction accuracy is 82%. The linear regression shows that the GPA for each semester has a highly significant with the dependent variable with 96.2% prediction accuracy. By having this information, the management of KPTM can make a plan to ensure that the student can maintain a good result and at the same time to make a strategic plans for those without a good result.

Keywords: *Data mining, Knowledge Discovery in Database, Prediction, Student performance, Decision tree*

1. Introduction

Data mining can be defined as a process of analyzing data form a big data repository, compile it and turn it into useful information. This data mining technique can be used in industry, science, engineering, government and education. It can be concluded that data mining is a process of finding the relationships and patterns among the data in large relational database[1].
Recently, tutors and researches have been utilized data mining techniques in analyzing students learning in order to get a broad view of it and at the same time can improve the quality of the educational procedures [1]. Data mining also can be described as a combination of tools taken from artificial intelligence area and statistic together with database management in computer science field [2].
KPTM is an institution that practices in maintaining the quality improvement in providing the best teaching and learning service in a conducive environment in order to fulfill clients' satisfaction with excellent reputation and at the same time will lead a holistic education through global acknowledgement and acceptance.

YEAR	TOTAL DROPOUT
2014	160
2015	172
2016	99

Figure 1. Data KPTM student in 2014-2016

Figure 1 shows that the total number of students that dropout from year 2014-2016. The number increase from year 2014 to 2015. Student performance is the major concern to the management of KPTM. Among the student that dropout, there are a number of students with a good grades. This prediction is very important for the management to come out with a solution on how to manage the student based on their performance.
The performance prediction can predict the student performance at the early stage of their study duration. The management of KPTM need to identify the student that has the potential in getting good grades in order to prevent them from dropout. They also need to manage the student that do not achieved a good grades accordingly.

2. Background and Related Works

Data mining a technique of computational process of analyzing data which involve methods that involved with artificial intelligence, statistics, machine learning and database systems from different perspective and after that, summarizing into useful information in discovery patterns in large dataset. The main goal of data mining is to select or get the relevant information from a data set and convert it to become pattern for future use.

Devasia, T. P, and Hegde [3] stated that nowadays, students that were dropping out has been increasing and it has been affecting the reputation of the institute; not just only the stu-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

dents' career. The current system has no capability to analyze the data but only can maintain the information on student in the form of numerical values and it can only store and access the information.

Ktona, Xhaja and Ninka in [2] stated that data mining is an area in computer science that can be used in education and at the same time can provide a findings that will help to increase the education quality. It's a combination of tools of statistics with database management and artificial intelligence.

Different classification techniques were used by Mishra, Kumar, and Gupta in [4] to build performance model based on students' social integration, academic integration, and various emotional skills. Two algorithm were used applied which are Decision tree-J48 (Implementation of C4.5) Random Tree to predict third semester performance.
Buniyamin, Mat, and Arshad in [5] used data includes Students GCPA scores of every semester, GPA scores in their courses such as Mathematics, Signal and system, Digital System, CGPA of every semester and GPA prediction and classification of engineering students achievement. Buniyamin, Mat, and Arshad in [5] also describe a tool to identify, predict and classify students depends on their academic performance that was calculated using Cumulative Grade point average (CGPA) grades. Buniyamin, Mat, and Arshad in [5] also used student data to improve in education planning and presents techniques to obtain knowledge from databases to predict student's performance.

Arsad, Buniyamin, and Manan in [6] used Cumulative Grade Point Average (CGPA) at semester eight to identify the accomplishment in academic. Variables used are the results in the first semester for the fundamental subjects as an independent variables or input predictor variables and the output or the dependent variable was used CGPA at semester eight. Arsad, Buniyamin, and Manan [6] has also divided the variables into two which are dependent variable and independent variables or input predictor variables. The CGPA taken from semester eight was used as the output or the dependent variable and the results for the fundamental subjects in the first semester were used as independent variables or input predictor variables.

Arsad, Buniyamin and Manan in [6] stated that the outcomes shows that basic subjects at semester 1 and 3 can give a strong influence in the final CGPA for graduation. There is direct correlation between students' strong achievement on fundamental subjects at early semester with the overall academic performance on graduation. It was concluded that the fundamental subject must be fully dominated because it can effects other subjects at the higher semester.

Mishra, Kumar and Gupta in [4] found that the result of second semester will influence the third semester result and the programming subject in second semester will form the foundation of programming subject in third semester. A good academic performance indicates a good performance in third semester. From the emotional attributes, leadership and the spirit to success can effected the student performance. The performance of both the algorithm used is sufficient; Random Tree implementation earned a higher overall accuracy (94.418%) compared to J48 with 88.372% accuracy. Also J48 is lower than the True Positive Rate, Precision and Recall measures of Random tree in determining the accuracy.

Jacob, et al. in [7] stated that the C5.0 classification algorithm has 100% classification accuracy and the comparison will be done between proposed algorithm and the bench mark algorithms such as Decision tree induction and Naïve Bayes. This is very helpful for the educator to find a specific student that need a special consideration.

3. Methodology

Methodology is a process of collecting information and data which used several concepts and theory beneath it. The methodology used in this research is Knowledge Discovery in Database (KDD) that involved with the process of finding and extracting a knowledge from a collection of data in the context of large database. This is done by using algorithm according to the specification by using a database together with any required preprocessing, subsampling and transformation of that database.

A. Data Selection

Selection of data involved with the process of determine what data that will be used for the knowledge discovery such as what kind of data available, need for any additional data and the integration between all of the data. This process is very critical in order to apply the data mining techniques and if some data and attribute is missing, it can effects the entire process.

In this study, there were 257 data set taken from the student of semester 6 which involved with 4 programs – Diploma in Computer System and Networking, Diploma in Information Technology, Diploma in Business Management and Diploma in Accountancy from KPTM Alor Setar.

There are 11 (eleven) attributes chose such as Gender, Program, State, Sponsorship, SPM, CGPA, GPA Sem 1, Sem 2, Sem 3, Sem 4, and Sem 5. This data has been taken from Campus Management System (CMS), KPTM Alor Setar.

B. Data Preprocessing

Data Preprocessing is very crucial in order to ensure that the data set used is a quality and reliable data. This steps involved with the preprocessing and cleaning data that will remove the irrelevant data and deciding strategy to handle missing fields and altering data based on the requirement. At the same time, simplifying the data set by removing the unwanted variables and analyzing the features that can be used to represent the data set depends on the goal and the task of doing data mining.

Box plot and cross tabulation in SPSS are used in this process to compare the relationship between variables in order to find out which variable do not have any relation with other variable. The variables that do not have any relation, will be taken out and the outcome is a clean data set.

C. Data transformation reduction

This step involved with changing or combining into an appropriate form in data mining. It is a statistical measure that will evaluate the degree of changes of one variable to the value of one another. The tools that will be used is SPSS. The correlation test has been done to identify the strength of the relationship between the variables and how strong the relationship is. The outcome from this phase will be data required for modelling.

D. Implementation of Data Mining

One of the technique used in this project is linear regression in SPSS. Linear regression is a statistical method that can be used to summarize and identify the relationship between two (2) or more variables. The main purpose of linear regression analysis is to find out the associations between dependent variable and independent variables when the dependent and independent variables relationships are almost linear. It can show an optimum result but it is not suitable for non-linear relationship.

The J48 algorithm were also implemented to the dataset to develop the decision tree. For the analysis, the 10 fold cross validation approached has been used, which used to evaluate the dataset by dividing them into 10 equally sized parts to test and evaluate the dataset.

4. Result and Analysis

Linear regression has been applied to the preprocessed data in order to find the relationship between dependent variables such as Gender, Sponsorship, PreviousSPM, GPA Sem 1, Sem 2, Sem 3, Sem 4, and Sem 5 and dependent variable such as CGPA. The significance level has been set to <0.05 (where p value <0.05 indicates that there is a correlation between the dependent variable and predictor).
The variables such as GPASem1, GPASem2, GPASem3, GPASem4 and GPASem5 have a highly significant with the dependent variable. Variable Financing and PreviousSPM has correlation with the dependent variable and Gender has no correlation with the dependent variable. The prediction accuracy in linear regression is 96.2% which indicates that the predictors has a strong correlation with the dependent variable.

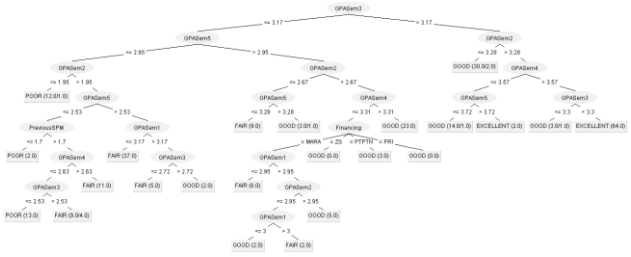


Figure 2 . J48 decision tree used in the analysis

The decision tree in Figure 4 uses J48 algorithm with 10 fold cross validation approached and has shown that the prediction accuracy is 82.5%. It indicates that the data set are classified as positive value. There is a strong significant relation between the GPASem1 and CGPA. Another variable that has significant value with CGPA is GPASem2, GPASem3, GPASem4 and GPASem5. The decision tree also indicates that financing and SPM also has significant value with the CGPA. It also shows that Gender does not contribute in the student performance.

The prediction accuracy is the method in ensuring that the techniques used, reached the right measurement of accuracy. Table II shows the prediction accuracy between Linear Regression and J48 Decision Tree. The prediction accuracy of linear regression is 96.2% which is higher compared to J48 Decision Tree.

Table II. Prediction Accuracy

Linear Regression	J48 Decision Tree
96.2%	82.5

The result is to find out which variables has a strong relation and contribute more to CGPA and it shows that variables such as GPASem1, GPASem2, GPASem3, GPASem4 and GPASem5 has a highly significant with CGPA.
The reason why variables such as GPASem1, GPASem2, GPASem3, GPASem4 and GPASem5 has a highly significant with the CGPA is mainly because these variables is a part of the CGPA.
Financing and SPM is the variables that is not a part of the CGPA so it shows less significant. It also shows that Gender doesn't have any strong effect on the CGPA of the student.

5. Conclusion and Future Works

The main goal of this paper is to discuss about predicting the student performance by using data mining technique. Phases in KDD has been used in order to find which variables that can strongly contribute to the student performance. The tools such as SPSS and WEKA has been used to assist the predicting process and to prepare the data set for modelling, to produce a clean data set, to prepare data for modelling process and to prepare a patterns of data to reach the prediction accuracy.
Predicting the performance and identifying the potential students at the early stage is very important for the management of KPTM Alor Setar to manage that particular student, to increase and sustain the number of student and at the same time can prevent them from leaving the college earlier than expected. It will help the management to find a way to reduce the number of potential student that dropout and also can help the low achiever student as early as possible.
The result shows that Linear Regression has a higher prediction accuracy compared to J48 Decision Tree. The analysis also shows factors contribute most to the performance of the students.
This paper will help the management of KPTM to manage the student performance. Predicting the performance and identifying the potential students at the early stage is very important for the management of KPTM Alor Setar to manage that particular student accordingly. Predicting the student performance is important in order for the management to increase and sustain the number of student and at the same time can prevent them from leaving the college earlier than expected. It will help the management to find a way to reduce the number of potential student that dropout and also can help the low achiever student as early as possible in getting good grades.
In the future, a different set of variables such as students' emotion, parents' occupation, parents' salary, parents' occupation and students' curriculum activity can be used as the predictor. It is to extend the factors that can contribute the students' performance and to find which factor can tribute most to the students' CGPA. This will make the research done more comprehensive and the result will be more accurate.

References

[1] F. Grivokostopoulou, I. Perikos, and I. Hatzilygeroudis, "Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance," *Proc. IEEE Int. Conf. Teaching, Assess. Learn. Eng. Learn. Futur. Now, TALE 2014*, no. December, pp. 488–494, 2015.

[2] A. Ktona, D. Xhaja, and I. Ninka, "Extracting Relationships between Students' Academic Performance and Their Area of Interest Using Data Mining Techniques," *2014 Sixth Int. Conf. Comput. Intell. Commun. Syst. Networks*, pp. 6–11, 2014.

[3] M. T. Devasia, M. V. T. P, and V. Hegde, "Prediction of Students Performance using Educational Data Mining."

[4] T. Mishra, D. Kumar, and S. Gupta, "Mining students' data for prediction performance," *Int. Conf. Adv. Comput. Commun. Technol. ACCT*, pp. 255–262, 2014.

[5] N. Buniyamin, U. Bin Mat, and P. M. Arshad, "Educational data mining for prediction and classification of engineering students achievement," *2015 IEEE 7th Int. Conf. Eng. Educ. ICEED 2015*, pp. 49–53, 2016.

[6] P. M. Arsad, N. Buniyamin, and J. A. Manan, "A neural network students' performance prediction model (NNSPPM)," *Smart Instrumentation, Meas. Appl. (ICSIMA), 2013 IEEE Int. Conf.*, no. November, pp. 1–5, 2013.

[7] J. Jacob, K. Jha, P. Kotak, and S. Puthran, "Educational Data Mining techniques and their applications," *2015 Int. Conf. Green Comput. Internet Things, ICGCIoT 2015*, pp. 1344–1348, 2016