**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 16, No. 3, December 2019, pp. 1584~1592
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i3.pp1584-1592 

# Supervised data mining approach for predicting student performance

**Wan Fairos Wan Yaacob, Syerina Azlin Md Nasir, Wan Faizah Wan Yaacob, Norafefah Mohd Sobri**
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Malaysia

**Article Info**

***Article history:***

Received Apr 1, 2019
Revised Jul 22, 2019
Accepted Jul 28, 2019

***Keywords:***

Classification technique
Data mining
Predictive model
Student performance

**ABSTRACT**

Data mining approach has been successfully implemented in higher education and emerge as an interesting area in educational data mining research. The approach is intended for identification and extraction of new and potentially valuable knowledge from the data. Predictive model developed using supervised data mining approach can derive conclusion on students' academic success. The ability to predict student's performance can be beneficial for innovation in modern educational systems. The main objective of this paper is to develop predictive models using classification algorithm to predict student's performance at selected university in Malaysia. The prediction model developed can be used to identify the most important attributes in the data. Several predictive modelling techniques of K-Nearest Neighbor, Naïve Bayes, Decision Tree and Logistic Regression Model models were used to predict student's performance whether excellent or non-excellent. Based on accuracy measure, precision, recall and ROC curve, results show that the Naïve Bayes outperform other classification algorithm. The Naïve Bayes reveals that the most significant factors contributing to prediction of excellent students is when the student scores A+ and A in Multivariate Analysis; A+, A and A- in SAS Programming and A, A- and B+ in ITS 472.



***Corresponding Author:***

Wan Fairos Wan Yaacob,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA Cawangan Kelantan,
18500 Bukit Ilmu, Machang, Kelantan, Malaysia.
Email: wnfairos@kelantan.uitm.edu.my

## 1. INTRODUCTION

Currently, universities are working in a very complex and highly competitive environment. One of the high quality university criteria is based on its excellent record of academic achievements. Hence, in higher learning institutions, student's performance is an important part to be focus by the management of the university. Ability to predict the student performance using data mining (DM) has received much attention [1-10]. Though predicting the student's performance is a complex task due to the increase in the number of data available relating to student's academic results in higher learning institution, data mining application can help the academic management systems to investigate and identify group of excellent students and group of dropped out students from the university.

The data mining technique is important in higher learning institution as studies on existing prediction method is still insufficient to identify the most suitable methods to predict student's achievement in particular courses. With the accurate data mining techniques prediction algorithm, it can help to identify the most important attributes in contributing to student's performance. Higher institutions can gain deep and thorough knowledge to enhance its lesson plan, assessment, evaluation planning and decision-making based on the finding obtained. DM offers many techniques such as predictive models, classification, association and many others [8, 11].

Thus, the aim of this research paper is to develop predictive models using classification algorithm to predict student's performance into excellent or non-excellent students depending on the results of their academic performance via educational data mining. Four classifiers such as Decision Trees, Naïve Bayes, K-Nearest Neighbour, Logistic Regression are adopted in predicting the student performance and categorized them either excellent or non-excellent. In this case, RapidMiner tool is utilized for the model building process to evaluate students' performance for data classification. In reviewing literature on predicting student's performance, there are two main factors being highlighted which are attributes and prediction methods. There are lots of research investigating the attributes that have been frequently used in predicting student's performance. The commonly used attributes is CGPA and internal assessment. CGPA is the most important attribute used by researcher [12] to determine the performance of the students while internal assessment such as quizzes, test and attendance were also used by researchers [13-14].

The next important attributes used is demographic such as gender [15-16] and external assessment which identify the marks or grade obtained for a particular subject [14, 17]. There is also other researcher that use physcometric factor such as interest to predict student performance [15] and performance in examination.

The second part of literature in student performance prediction is about the prediction method. In data mining, prediction modelling is usually being used in predicting student's performance. The techniques can vary from classification, regression or categorization [17]. The most popular one is classification algorithms under predictive modelling techniques such as Decision Tree, Naïve Bayes, K-NN, Neural Network, Logistic Regression model and many others can be seen used by many researchers [8, 11, 18]. Several works focus on comparing these techniques particularly in predicting student performance. [17] conducted a meta-analysis and proposed a systematical literature review using data mining techniques in predicting student performance which identifies that Neural Network and Decision Tree are two highly used methods. [19] also use data mining techniques to explore ten variables that may influence dropout of students in an online program using 10-fold cross-validation technique. Work by [20] adopts classifiers such as Decision Tree, Bayesian, K-Nearest Neighbour, Rule learners to predict students' performance based on their personal and pre-university characteristics. On the other hand, [21] proposed a model of student performance predictors by using classification techniques which resulted to be satisfactory with overall accuracy of the tested classifiers is above 60%.

## 2. RESEARCH METHOD

In this study the steps involved in methodology for developing predictive model using data mining is implemented following the CRISP-DM (Cross Industry Standard Process for Data Mining) model [22]. The CRISP-DM process is a cyclic approach which consists of six steps as in Figure 1. The first step is understanding business activities and problems where the process involves transforming the business problem of predicting students' performance into data mining problem. Then, the second steps involved data analysis including collection and familiarization of raw data. Next, data preparation. The fourth step is data modeling which involved several predictive algorithms were developed including K-NN, Naïve Bayes, Decision Tree and Logistic Regression Model. After the models have been developed, next last step is model evaluation and deployment.

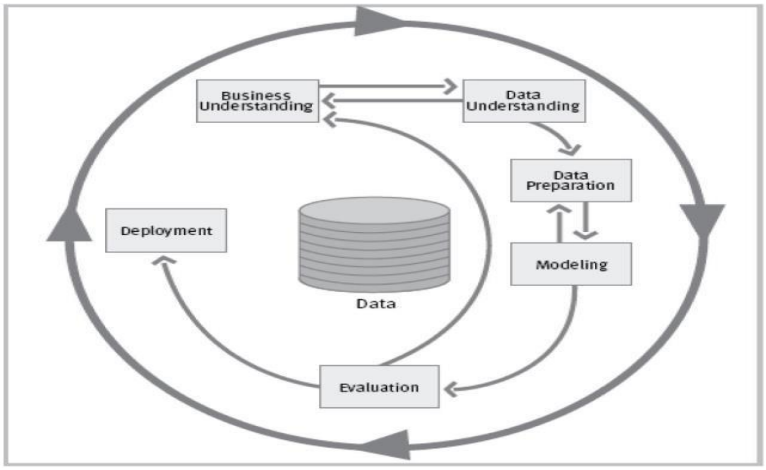


Figure 1. The CRISP-DM Model in educational data mining

### 2.1. Data Description

The data of undergraduate students of Bachelor of Science in Statistics programme has been selected from Faculty of Computer and Mathematical Sciences at Universiti Teknologi MARA Cawangan Kelantan and Universiti Teknologi MARA Cawangan Negeri Sembilan. We obtained the program study plan and courses conducted for three-year study plan with two semesters for every year which total up to six semesters. This research only focuses on the core courses offered by the programme as it dominates the study plan which in turn gives great impact to the final grade point average (CGPA). Hence, the transcript data were collected with authorize permission from Examination Department of the university. We collected 631 transcripts from year 2013 to 2016 for students who have completed their academic degrees. Each student record has the following attributes: student name, student ID, gender, final CGPA, and all the courses enrolled by the students including the course' grade. The target and input variables involved in this study are presented in Table 1.

Table 1. Description of Variables

| Attribute | Description | Possible Values |
|---|---|---|
| PERFORM | Students Performance | {Excellent, Not-Excellent} |
| STA 500 | Nonparametric Statistics | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| MAT 523 | Linear Algebra II | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| ITS 472 | Database Management Systems | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| STA 550 | Sampling Methods | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| QMT 556 | Quantitative Financial Management | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| STA 560 | Statistical Inference | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| STA 570 | Time Series Analysis & Forecasting | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| STA 600 | Intermediate Regression Analysis | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| STA 610 | SAS/R Programming | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| STA 640 | Experimental Design & Analysis of Variance | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |
| STA 680 | Applied Multivariate Analysis | {A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F} |

In data preparation phase, we applied pre-processing technique for the collected data to prepare the data for mining purposes. The data was cleaned to ensure no missing value that above 50% and no unwanted values exist in the data set. Some irrelevant attributes were also eliminated. We also removed any data related to the preparatory year. Then we re-arranged the data so that student has the following attributes: ID, CGPA every semester and the course grade student for the duration of 2013 to 2016. In the final step, the numerical attributes were discretized to categorical ones. The target variable, CGPA group is coded into three groups: 1 = Excellent and 2=Not-Excellent. The distribution of the dataset according to the CGPA group is shown in Figure 1. Same as the student's grade in each course is coded into: A+, A, A-, B+, B, B-, C+, C, C-, D+, D, E and F. Next, the data was partitioned into training and testing set. The purpose of partitioning the data is to test on the performance of the model. In training set, the k-NN, Naives Bayes, Decision Tree and Logistic Regression model were developed to predict the students' performance. While in testing set, the models performance was measured. Rapid Miner 8.3 software was used to build the decision tree model.

### 2.2. Data Mining Method

Data mining is a business process for exploring large amountof data to discover meaningful patterns and rules [23]. It involved computational method which is has been successfully applied in many areas. Data mining techniques mostly used to build a model for prediction, classification of ndata into categories or to discover any menaingfull hidden pattern and relationships in the observed data. Hence, this can be done using algorithms of the data mining which are divided into two basic categories: i) unsupervised algorithms and ii) supervised algorithms. The task of unsupervised algorithm is to discover underlying patterns in the data without knowing the target variable. A method of clustering and association rules belongs to this group. While supervised algorithms are data mining algorithm that meant for building models with known target and are constructed to predict the class to which unknown data will belong. The most common methods of classifications are: decision trees, induction rules or classification rules, probabilistic or Bayesian networks, neural networks and hybrid procedures. In this study we investigated the impact of four algorithms: k-NN, Naive Bayes, Decision Tree and Logistics Regression.

#### 2.2.1. K-NN Algorithm

K-NN algorithm is one of well-known classification methods. This algorithm classifies objects based on closes training examples in the feature space. The closeness is defined in terms of a distance metric called Euclidean distance. Thus, the object is classified by a majority vote of its neighbor with the object

being assigned to the class most common among its k nearest neighbors. The best choice of k depends upon the data.

#### 2.2.2. Naïve Bayes Algorithm

Naïve Bayes is a method for classification based on the theory of probability [24, 25]. It is a statistical classifier which predicts class membership probabilities in which the probability of a given sample belongs to a particular class. Bayesian classifier is based on Bayes' theorem calculate the posterior probability $P(H|\boldsymbol{X})$, from $P(H)$, $P(\boldsymbol{X}|H)$ and $P(\boldsymbol{X})$ as follow [24] (Han et al., 2012): $P(H|\boldsymbol{X})=P(\boldsymbol{X}|H)P(H)P(\boldsymbol{X})$. In this study, the output is the likelihood of an excellent student (1= excellent, 0 = not excellent).

#### 2.2.3. Decision Tree

The decision tree models developed in this study are based on splitting criteria of Gini and Information Gain (Entropy). The first splitting criteria is Gini which is one of the most popular splitting criterions since it is also being used by biologist and ecologist. Gini gives probability those two items chosen at random from a population in the similar class. The measures of node in the Gini is the sum of squares of the pro-portions of the classes in the node and a perfectly pure node has a Gini score of 1 Gini index of a pure table consist of single class is zero because the probability is 1 and 1 – 12 = 0. The index also reached maximum values when all classes in the table have equal probability. The higher value of reduction in Gini Index implies that a feature is a better candidate in the classification task.

The second decision tree model is based on Information Gain (Entropy). Entropy reduction is also known as information gain splitting criterion. Entropy of a pure table (consist of single class) is zero because the probability is 1 and log (1) = 0. Entropy reaches maximum value when all classes in table have equal probability [18]. The information gain defines purity in a similar way as machine learning does. This means that if leaf is entirely pure then it is described as the classes in the leaf. On the other hand, if leaf is highly impure, then it is complicated. Perfectly pure note of entropy has lower score which is zero. The higher value Entropy indicates preference of feature for discrimination of class value. For example, if feature separates the two classes completely, it has the most Information Gain and is the best feature for classification.

#### 2.2.4. Logistic Regression

Logistic regression studies the association between dichotomous dependent variables and a set of k independent variables in which the independent variables are used to estimate the outcome of the dependent variables (Hosmer, Lemeshow and Sturdivant, 2013). The probability of event Y=1 denoted as p is obtained as follows:

$$logit = \ln\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 X$$

The goal is to estimate the probability that an event occur, p. In this study, a method called maximum likelihood is used to find the best-fit line for logistic regression.

### 2.3. Performance Evaluation

To evaluate the classifier, this study performs 10-fold cross validation by splitting the data set randomly into 10 subsets of training and testing size. Then, the performance measures were calculated. Results produced by these prediction models were compared using classification table in which it provides Accuracy, Misclassification Rate, Sensitivity, Specificity and Receiver Operating Characteristic (ROC) Chart. 10-fold cross validation is used for comparison with baseline methods for performance evaluation. Each performance evaluation parameters are defined as follows.

a. Accuracy is the closeness of a measurement of a quantity to the quantity's true.
b. Sensitivity is the true positive (TP) rate which is the probability of detecting a true outcome. It is the proportion of positive cases that are correctly identified.
c. Specificity is the true negative (TN) rate which is the probability of detecting a false outcome. It is the proportion of negative cases that are correctly identified.
d. ROC chart of a model shows the trade-off between True Positive and True Negative. For any increases of true positive rate will occur at the cost of false positive rate, ROC curve able to show the accuracy of the prediction model for every possible threshold of predicted probabilities. The vertical axis represents the TP rate whereas the horizontal axis represents the FP. As the TP rate increase, the FP rate will increase as well. However, the area under the curve (AUC) will indicate the accuracy of the model. As a bigger AUC indicates lower increment of FP rate when compare to a larger increment of TP for every increment of predicted probability threshold.

## 3. RESULTS AND ANALYSIS

In this paper, the classification of excellent student was performed using the supervised data mining predictive model based on k-NN, DT, NB and Logistic Regression Model. 10-fold cross validation was performed to validate each classifier. When the test was completed, the average performance on the test was computed to determine the accuracy of the model developed. We first present the results of k-NN model.

The resulting 10-fold cross-validation of k-NN classifier was performed at different k value to find the best k value that can measure the best accuracy. In general, cross validation has been proved tobe statistically good enough in evaluating the performance of the classifier. As referring to Figure 2, the best k that can achieved highest accuracy is with 9-NN classifier.

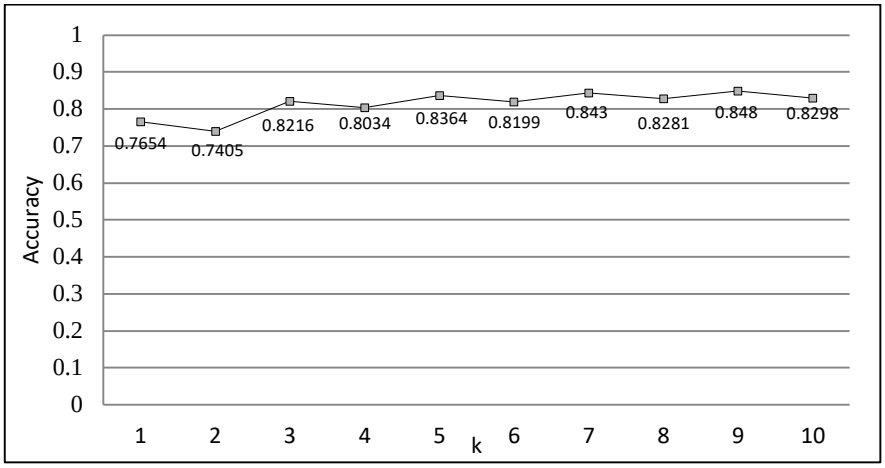


Figure 2. 10-fold cross validation accuracy levels of k-NN classifier for different k values

The decision tree model using GINI and Information Gain algorithm were alsoimplemented on the data and the results of the classification are presented in Table 2. Referring to Table 2, Information Gain algorithm has correctly classified about 82.15% for the 10-fold cross validation testing while Gini Decision Tree algorithm offer a lower rate classification of about 80.15%. The precision of Information Gain is high for Excellent class (78.61%) and Non-Excellent class (83.78%) compare to GINI with 75.25% and 83.78% for Excellent and Non-Excellent class respectively.

Table 2. Classification results for Decision Tree-algorithm

| | DT - Information Gain Precision | DT – GINI Precision |
|---|---|---|
| Excellent | 78.61 | 75.25 |
| Non-Excellent | 83.78 | 83.78 |
| Overall Accuracy | 82.15 | 80.15 |

Hence, comparing this two Decision Tree algorithm; Information Gain and GINI, Information Gain performed better. By applying pre-pruning with minimal gain of 0.01 and minimal leaf size of 3, it produces a classification tree of 19 nodes and 16 leaves. Figure 3 and Figure 4 shows the screenshot of decision tree using Information Gain results.

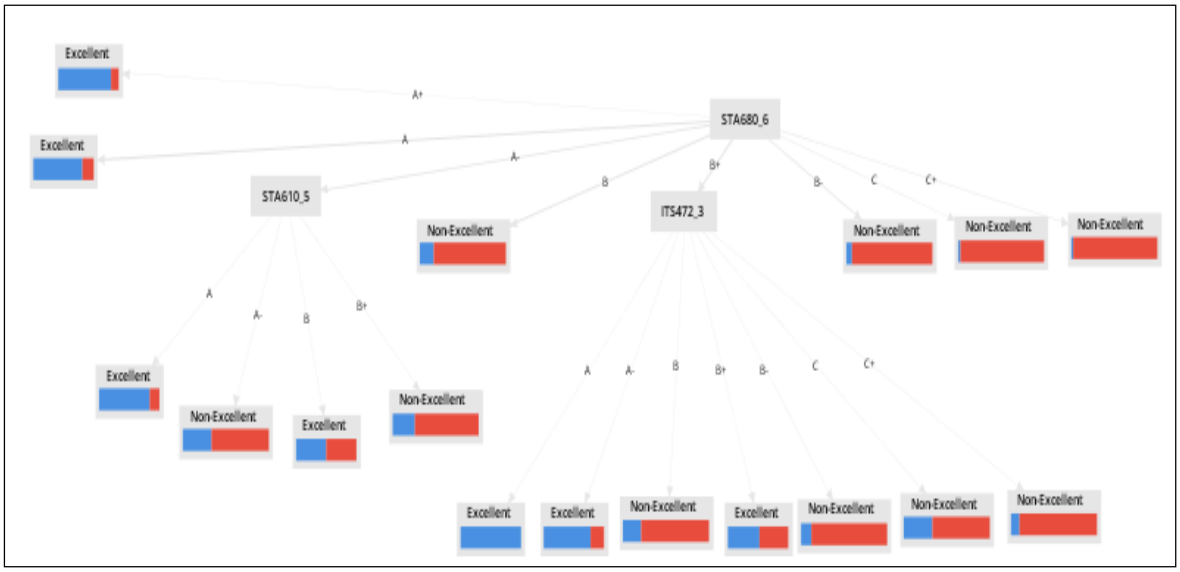


Figure 3. Decision Tree using Information Gain

**Tree**

STA680_6 = A: Excellent {Excellent=100, Non-Excellent=23}
STA680_6 = A+: Excellent {Excellent=7, Non-Excellent=1}
STA680_6 = A-
| STA610_5 = A: Excellent {Excellent=21, Non-Excellent=4}
| STA610_5 = A-: Non-Excellent {Excellent=10, Non-Excellent=20}
| STA610_5 = B: Excellent {Excellent=3, Non-Excellent=3}
| STA610_5 = B+: Non-Excellent {Excellent=8, Non-Excellent=23}
STA680_6 = B: Non-Excellent {Excellent=18, Non-Excellent=93}
STA680_6 = B+
| ITS472_3 = A: Excellent {Excellent=4, Non-Excellent=0}
| ITS472_3 = A-: Excellent {Excellent=14, Non-Excellent=4}
| ITS472_3 = B: Non-Excellent {Excellent=7, Non-Excellent=26}
| ITS472_3 = B+: Excellent {Excellent=11, Non-Excellent=10}
| ITS472_3 = B-: Non-Excellent {Excellent=3, Non-Excellent=22}
| ITS472_3 = C: Non-Excellent {Excellent=1, Non-Excellent=2}
| ITS472_3 = C+: Non-Excellent {Excellent=1, Non-Excellent=9}
STA680_6 = B-: Non-Excellent {Excellent=5, Non-Excellent=71}
STA680_6 = C: Non-Excellent {Excellent=1, Non-Excellent=35}
STA680_6 = C+: Non-Excellent {Excellent=1, Non-Excellent=44}

The results for the performance of selected classification algorithms (Accuracy, Precision and Recall) are summarized and presented in Table 3.

Table 3. Comparison of Model Accuracy

| No. | Model | Percentage Accuracy | Precision | Recall |
|---|---|---|---|---|
| 1 | K-NN | 84.80 | 73.39 | 84.65 |
| 2 | Naives Bayes | 89.26 | 85.38 | 84.19 |
| 3 | DT – Information Gain | 82.15 | 78.61 | 68.37 |
| 4 | DT – Gini | 80.99 | 75.25 | 69.30 |
| 5 | Logistic Regression (LR) | 85.28 | 79.72 | 78.60 |

The achieved results reveal that the Naïve Bayes classifier perform best (with the highest overall accuracy) followed by LR, K-NN, DT Information Gain and DT GINI. All classifier tested are performing with overall accuracy above 80% which means the error rate is low and predictions are reliable. The detection of sensitivities of Naive Bayes, LR, DT Information Gain, DT GINI and k-NN, were 85.4%, 79.7%, 78.6%, 75.3% and 73.4% respectively.

Receiver operating Characteristics (ROC) curve is also used for the evaluation of classification algorithm. The ROC Curve measure the performance of the model by plotting the true positive rates against false positive rates. A test with perfect discrimination has ROC plot that passes through the upper test corner (100% sensitivity, 100% specificity). According to ROC curve in Figure 5, it can be found that the Naïve Bayes classifier is the best classifier as the ROC curve is the approaching 1. Hence, the results indicate Naïve Bayes performs very well in predicting the performance of the student.

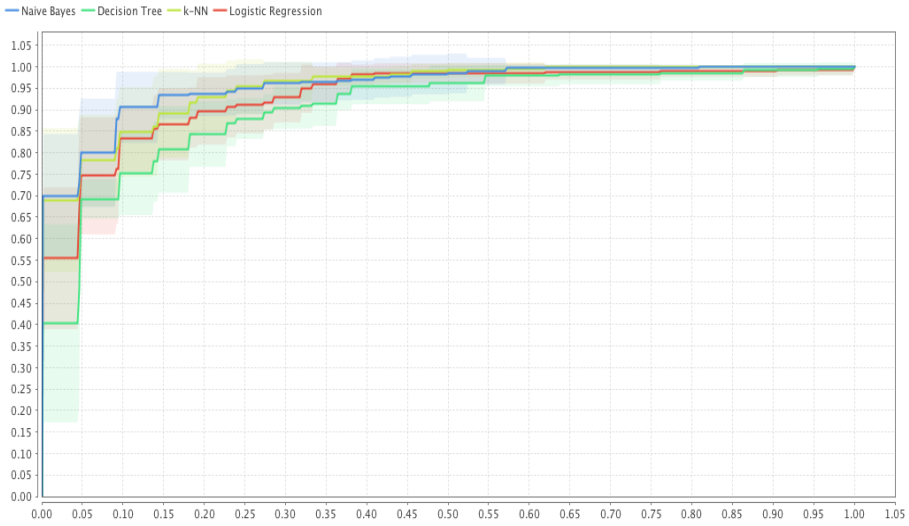


Figure 5. Comparison of ROC Curve for Naïve Bayes, Decision Tree, k-NN and Logistic Regression Model

## 4. CONCLUSION

In this paper, four supervised data mining algorithms were applied on the students performance data to predict student performance either excellent or non-excellent based on predictive accuracy. It has also been indicated that a good classifier model has to be both accurate and comprehensible [26, 27]. In this study, several predictive modelling technique of data mining approach were applied predict the student performance. The results indicate that the Naïve Bayes classifier outperformed other algorithm compared to Decision Tree, k-NN, and Logistic Regression with accurate and comprehensive classifier. This is in accordance with the findings by [20] that found Naïve Bayes model is outperforming other predictive model with higher accuracy rate. This study has proved that student performance prediction is important to be conducted for the university to improve their teaching performance. Some high influence attribute to predict student performance can be considered by the university to plan further action for improvement. The study can be further extended to predict student's performance of other courses using other attributes.

## ACKNOWLEDGEMENTS

The authors gratefully aknowledge use of the facilities of Universiti Teknologi MARA Cawangan Kelantan. This research was supported by Universiti Teknologi MARA, Malaysia through ARAS Grant 600-IRMI/DANA 5/3/ARAS (0026/2016)).

## REFERENCES

[1] A. Soni, et al., "Predicting Student Performance Using Data Mining Techniques," *International Journal of Pure and Applied Mathematics*, vol. 119, pp. 221-227, 2018.
[2] M. Bucos and B. Dragulescu, "Predicting Student Success Using Data Generated in Traditional Educational Environment," *TEM Jurnal*, vol. 7, pp. 671-625, 2018.
[3] A. Shantini, et al., "Predicting Students' Academic Performance in the University Using Meta Decision Tree Classifiers," *Jurnal of Computer Science*, vol. 14, pp. 654-662, 2018.
[4] Bendangnuksung and Prabu P., "Students' Performance Prediction Using Deep Neural Network," *International Journal of Applied Engineering Research*, vol. 13, pp. 1171-1176, 2018.
[5] A. Daud and F. Abbas, "Predicting Students Performance Using Advanced Learning Analytics," *International World Wide Conference Committee*, (FW30), 2017.

[6] A. A. Saa, "Educational Data Mining & Students' Performance Prediction," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, 2016.
[7] B. A. Pereira and A. Pai, "A Comparative Analysis of Decision Tree Algorithms for Predicting Student's Performance," *International Journal of Engineering Science and Computing*, vol. 7, 2017.
[8] A. D. Kumar, et al., "Review on Prediction Algorithms in Educational Data Mining," *International Journal of Pure and Applied Mathematics*, vol. 118, pp. 531-537, 2018.
[9] F. Widyahastuti and V. U. Tjhin, "Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron," *IEEE*, 2017.
[10] R. Asif, et al., "Predicting Student Academic Performance using Data Mining Methods," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 17, 2017.
[11] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv preprint arXiv*:1201.3417, 2012.
[12] H. E. Erdem, "A cross-sectional survey in progress on factors affecting students' academic performance at a Turkish university," *Procedia-Social and Behavioral Sciences*, vol. 70, pp. 691-695, 2013.
[13] G. Elakia and N. J. Aarthi, "Application of data mining in educational database for predicting behavioural patterns of the students," *International Journal of Computer Science and Information Technologies*, pp. 4649-4652, 2014.
[14] S. Parack and F. Z. Zahid, "Application of data mining in educational databases for predicting academic trends and patterns, in: Technology Enhanced Education (ICTEE)," *IEEE International Conference on, IEEE*, pp. 1-4, 2012.
[15] V. Aramburo, et al., "Predictive Factors Associated with Academic Performance in College Students," *Procedia-Social and Behavioral Sciences*, vol. 237, pp. 945-949, 2017.
[16] M. Garkaz, et al., "Factors affecting accounting students' performance: the case of Students at the islamic azad university," *Procedia-Social and Behavioral Sciences*, vol. 29, pp. 122-128, 2011.
[17] A. M. Shahiri, et al., "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414-422, 2015.
[18] S. K. Yadav and S. Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification," *World of Computer Science and Information Technology Journal (WCSIT)*, vol. 2, pp. 51-56, 2012.
[19] E. Yukselturk, et al., "Predicting dropout student: an application of data mining methods in an online education program," *European Journal of Open, Distance and E-learning*, vol. 17, pp. 118-133, 2014.
[20] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybernetics and information technologies*, vol. 13, pp. 61-72, 2013.
[21] C. Anuradha and T. Velmurugan, "A comparative analysis on the evaluation of classification algorithms in the prediction of students' performance," *Indian Journal of Science and Technology*, vol. 8, 2015.
[22] Ó. Marbán, et al., "A Data Mining & Knowledge Discovery Process Model, Data Mining and Knowledge Discovery in Real Life Applications," Julio Ponce and Adem Karahoca (Ed.), InTech, 2009. Available: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/a_data_mining__amp__knowledge_discovery_process_model 0
[23] Linof and Berry, "Data Mining Techniques," Wiley, Third Edition, 2011.
[24] M. Kumar and A. J. Singh, "Evaluation of Data Mining Techniques for Predicting Student's Performance," *Modern Education and Computer Science*, vol. 8, pp. 25-31, 2017.
[25] P. Galvan, "Educational Evaluation and Prediction of School Performance through Data Mining and Genetic Algorithms," *Future Technologies conference IEEE*, 2016.
[26] T. Devasia and Vinushree, "Prediction of Students Performance using Educational Data Mining," *Data Mining and Advanced Computing (SAPIENCE) IEEE*, 2016.
[27] A. U. Khasanah and Harwati, "A comparative Study to predict Student's Performance Using Educational Data Mining techniques," *IOP conf. Serie materials Science and engineering*, 2017.

## BIOGRAPHIES OF AUTHORS

Dr. Wan Fairos Wan Yaacob hold PhD in Statistics from Universiti Teknologi MARA, MSc in Statistics from Universiti Kebangsaan Malaysia and a Bachelor's degree BSsc Hons in Statistics. Dr. Wan Fairos works in Universiti Teknologi MARA as faculty member and has several years work experience in the areas of teaching, research, administrative (Head of Business Datalytics Research Group), arranging/organized research conferences, seminars, workshops, events. Dr. Wan Fairos has several research publications in well-known international Journals and conferences. She received a the Best Potential Resaerch for Commercialization Award in 2010, which was awarded by Universiti Teknologi MARA. She also won the Best Paper Award in ASST2017. Dr. Wan Fairos has also been engaged to create linkage between industry and academia while she holds position as Deputy Rector of Research and Industrial Linkages. She is a member of Institute of Statistics Malaysia and certified as Rapid Miner Data Analyst. Her research interests include data mining, predictive modelling, statistical modelling and panel count data model. She is currently completing a research on spatial and temporal random effect model on dengue disease outbreak.

Syerina Azlin Md Nasir received her Ph.D. in Information Technology from Universiti Teknologi MARA, Kelantan, Malaysia. She is a senior lecturer at Universiti Teknologi MARA, Malaysia. She has been teaching in the Faculty of Computer and Mathematical Science, Universiti Teknologi MARA since 2004. She has been engaged to research works such as conferences, workshops and become a member of Business Datalytics Group. She has leads her faculty before hold a position as Deputy Rector of Research, Industrial Linkages and Alumni. She is a Microsoft Certified Professional and Certified RapidMiner Analyst. Her earlier publications are on ontology construction and mapping such as 'Automating the Mapping Process of Traditional Malay Textile Knowledge Model with the Core Ontology, and 'Analysing the Effectiveness of COMA++ on the Mapping between Traditional Malay Textile (TMT) Knowledge Model and CIDOC CRM'. The author's current interest is on data mining, data analytics and text mining.

Wan Faizah Wan Yaacob is a Mathematics lecturer at the Universiti Teknologi MARA. She received a bachelor's degree in Actuarial Science from Universiti Teknologi MARA and a master's degree in Management Mathematics from Universiti Kebangsaan Malaysia. Her research interests include mathematical modelling, operational research and data mining.

Norafefah Mohamad Sobri hold Diploma in Statistics and Bachelor Science (BSc. Hons) Statistics from Universiti Teknologi Mara Shah Alam and Master of Applied Statistics from Universiti Putra Malaysia (UPM). Norafefah works in University Teknologi Mara as a faculty members and has several work experience in teaching and research. She is a member of Institute of Statistics Malaysia. Her research interests include data mining, time series analysis, and statistical modeling. Currently her research work is in data mining and sentiment analysis.