Journal of Physics: Conference Series

Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques

To cite this article: W F Wan Yaacob *et al* 2020 *J. Phys.: Conf. Ser.* **1496** 012005

ECS UNITED
The Electrochemical Society
Advancing solid state & electrochemical science & technology
247th ECS Meeting
Montréal, Canada
May 18-22, 2025
Palais des Congrès de Montréal
Showcase your science!
Abstracts due December 6th

Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques

W F Wan Yaacob[1,2], N Mohd Sobri[3], S A Md Nasir[4], W F Wan Yaacob[5], N D Norshahidi [6] and W Z Wan Husin[7]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Bukit Ilmu, 18500 Machang, Kelantan, Malaysia.

[2]Business Datalytics Research Group, Universiti Teknologi MARA Cawangan Kelantan, Kampus Kota Bharu, Lembah Sireh, 15050 Kota Bharu, Kelantan, Malaysia

[3,4,5,6,7]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Bukit Ilmu, 18500 Machang, Kelantan, Malaysia

Email : [1]wnfairos@uitm.edu.my

Abstract. The increasing number of students dropping out is a major concern of higher educational institutions as it gives a great impact not only cost to the students but also a waste of public funds. Thus, it is imperative to understand which students are at risk of dropping out and what are the factors that contribute to higher dropout rates. This can be done using educational data mining. In this paper, we described the uses of data mining techniques to predict student dropout of Computer Science undergraduate students after 3 years of enrolment in Universiti Teknologi MARA. The experimental results showed an achievable reliable classification accuracy from the selected algorithm in predicting dropouts. Decision tree, logistic regression, random forest, K-nearest neighbour and neural network algorithm were compared to propose the best model. The results showed that some of the machines learning algorithms are able to establish effective predictive models from student retention data. The Logistic Regression model was found to be the best learners to predict the dropout students with identified potential subject causes. In addition, we also presented some findings related to data exploration.

1. Introduction

Student retention issue at higher institution has gained wide attention and has become an important problem to the university administrators to ensure the student gains success or graduate and decrease the student's drop out from the institution. It is important for the university to understand the behaviour of successful students. Thus, many studies have been done to understand student failure or drop out issue at university level as discussed in [1-3]. Much of the work done has identified the main factors that affect the low performance of students such as failure while others focused on the characteristics that influence the students to drop out from university. The seminal work began with Tinto [2] model in student retention literature, which explored and explained the causes of students to drop out from the university. Among the characteristics of the students that can influence school failure were demographics, cultural, social, family, or educational background, socioeconomic status, psychological profile, and academic progress [2]. The model suggested that student's social and academic integration into the educational institution is the major determinant of completion and the model also identified some key influences on integration such as the student's family background, personal characteristics, previous schooling, prior academic performance, and interactions between student and the faculty. Bharadawaj and Pal [4-5] identified factors such as senior secondary exam, annual family income, residence and family status as important factors influencing student's dropout. Meanwhile, other studies used enrolment data such as in Kovacic [6], and mother's education and family income in Devasia et al.[7] to predict dropout. Recently, Perez et al.[1] used student demographics and transcripts records with courses grade performance to predict student's dropout.

Despite the existence of multiple findings on the factors that have been identified to influence student's dropout in the literature, there is limited research related to student's dropout from computer science area. Thus, there is a need to investigate the causes that lead students from Computer Science program to withdraw from the course before its completion. This can be achieved using data mining techniques by analyzing and extracting information from existing student data to predict student's dropout. The development of machine learning algorithms of data mining in recent years has enabled a large number of successful data mining projects in various applications specifically in educational data mining. Considering the factors that influence the dropout rates in universities, the aim of this study is to identify the key determinants of undergraduate student dropout rates in Computer Science program of Universiti Teknologi MARA and to determine which data mining technique is more suitable to find these key determinants. Hence, this paper demonstrates several machine learning algorithms to model student dropout using student admission data of 2016 gathered from academic databases. The best predictive model can then be used to predict new incoming students of those who are most likely to benefit from the support of the student retention program.

2. Related Work

Data mining is a business process that explores and analyses a large amount of data to extract meaningful patterns and rules [7]. This technique of data mining emerges from a combination field of statistics, artificial intelligence and data management that have been successfully applied in various fields such as marketing, banking, finance, educational research, fraud detection, and many others. Education is one of the domains where the primary concern is the evaluation and, in turn, enhancement of educational organizations. Educational Data Mining (EDM) is a discipline engaged with the development of methods and techniques not only for exploring and analyzing the data that come from educational context but also for extracting hidden information as to better understand the students. This information can be used in several educational processes such as predicting course enrollment, predicting student performance, predicting student dropout rate, detecting a typical value in students' transcripts, and improvement of student models that would predict student's characteristics or academic performances as explained in Tekin [8] and Yukselturk et al. [9].

Predicting the dropout of potential students is a workable solution for preventing dropouts. Tinto's model [2] is the most widely accepted model in student retention literature. Tinto concluded that the decision of students to persist or drop out of their studies is strongly related to their degree of academic and social integration at university. However, Brunsden et al. in [10] tested Tinto model and concluded that Tinto's model may not be the most appropriate for dropout research. Durso and Cunha [11] did a study to identify explanatory factors for undergraduate student's dropout from Accounting program of a Brazilian public university. They used survey database consisted of socioeconomic and demographic information of 371 students. The logistic regression model proposed by the study was able to accurately predict 77% of the cases dropout/completion from the sample. Five semi-structured interviews with those in the sample who dropped out of their studies were done in terms of qualitative study. The findings of the research have helped to understand the phenomenon of undergraduate student dropout from Accounting program and stresses the importance of rethinking public policies for the retention of talent and, especially, of those students who depend on their work to maintain their studies. Kim and Kim [12] conducted a study to examine the possible causes of university dropout in South Korea. They focused on four fundamental categories: resources, students, faculty and university characteristics. They used three-year balanced panel data from 2013 – 2015 and estimated them by using nonlinear panel data models. The results showed that cost and burden for student's financial resources, qualitative and quantitative features of faculty, and type/size of the university have significant effects on university dropout.

There are also several other studies that used data mining approach to predict student's dropout. Tan and Shao [13] used the method of Artificial Neural Network (ANN), Decision Tree (DT) and Bayesian Networks (BNs) as a prediction model, by selecting students' personal characteristics and academic performance as input attributions. The results showed that all the three machine learning methods were effective for student's dropout prediction, but DT presented better performance. Meanwhile, Mustafa, Chowdhury and Kamal [14] used data mining to develop a dynamic dropout prediction model for universities, institutes and colleges. Factors used were gender, financial condition and dropping year to classify the successful from unsuccessful students. Classification and Regression Tree (CART) and CHAID were used to examine the factors after applying data separation. CART was the most successful in growing the tree with an overall percentage of correct classification, then followed by CHAID.

In another study, Yukselturk, Ozekes and Türel [9] examined the prediction of dropouts through data mining approaches in an online program. The variables involved in this study were gender, age, education level, previous online experience, occupation, self-efficacy, readiness, prior knowledge, locus of control and the dropout status. Four data mining approaches were applied in order to classify dropout's students that included k-Nearest Neighbour (k-NN), Decision Tree (DT), Naïve Bayes (NB) and Neural Network (NN). 10-fold cross validation was used to train and test all the methods. The detection sensitivities of 3-NN, DT, NN and NB classifiers were 87%, 79.7%, 76.8% and 73.9% respectively. Online technologies, self-efficacy, online learning readiness and previous online experience were found the most important factor in predicting the dropout using Generic Algorithm (GA). Abu-Oda and El-Halees [15] used different data mining approach in examining and predicting students' dropout from a total of 1290 computer science students who graduated from ALAQSA University between 2005 and 2011. Different classifiers were used on the data sets such as Decision Tree and Naïve Bayes, and they were tested using 10-fold cross validation. The accuracy of the classifiers was 98.14% and 96.89% respectively. In addition, FP-growth algorithm was also used to discover hidden relationships between student's dropout status and enrolment persistence. The results showed that mastering 'digital design' and 'algorithm analysis' courses has a great effect in predicting student persistence in the major and decreases student likelihood of dropout.

Bharadwaj and Pal [4] used EDM to evaluate student performance among 300 students from five different colleges who were enrolled in an undergraduate computer course. They employed a Bayesian classification scheme of 17 attributes, of which the score in a senior secondary exam, residence, various habits, annual family income, and family status were shown to be important parameters for academic performance. In a second study, Bharadwaj and Pal [5] constructed a new data set which included student attendance, and test, seminar, and assignment grades in order to predict academic performance. A similar study was proposed by Kovacic [6], who applied EDM to identify which enrollment data could be used to predict student academic performance. In this study, he used CHAID and CART algorithms on a dataset of student enrollment.

Al-Radaideh et al. [16] analysed student's academic data (student gender, student age, student department, high school grade, lecturer degree, lecturer gender, among others) by building a classification model using the decision tree method to improve the quality of higher educational system. They found that high school grade was the attribute with the highest gain ratio and was considered the root node of the decision tree. The Holdout method and the K-Cross-Validation method (k-CV) were used to evaluate the model. However, they found that the collected samples and attributes were not sufficient to generate a classification model of high quality. Gerben et al. [17] conducted a case study in which they used machine learning techniques to predict student success using features extracted from student pre-university academic records. Their experimental results showed that simple and intuitive classifiers such as decision trees gave useful results with accuracies between 75% and 80%. One of their findings was that the strongest predictor of success was the grade for the Linear Algebra course, which was not seen as the decisive course. Despite these studies, it is not clear which data mining algorithms are preferable in this context. For example, Luan in [18] built predictors using clustering as means of data exploration and classification. In [19], Romero and Ventura presented a survey on EDM where one of their findings was that association analysis has become a popular approach. Meanwhile, Aulck et al. [20] did a study to model student dropout using the largest known database of higher education attrition through 32500 students' demographic and transcript records at one of the nation's largest public universities. The result showed that several early indicators of student attrition and student dropout can be accurately predicted even when predictions are based on a single term of academic transcript data. It highlighted the impact of student retention and success using machine learning. Finally, Herzog in [21] presented the results of a case study where Bayesian networks and neural networks were outperformed by decision tree algorithms on small educational dataset.

3. Methodology

Predictive modelling for binary classification was performed in this study where dropout (0,1) is the target variable. The Cross Industry Standard Process for Data Mining methodology (CRISP-DM) as discussed in Chapman et al. [22] was referred to as a basis of data mining methodology conducted to solve the student's dropout issues. This method is useful for planning and solving business problem. The process involves 6 phases as described in figure 1.

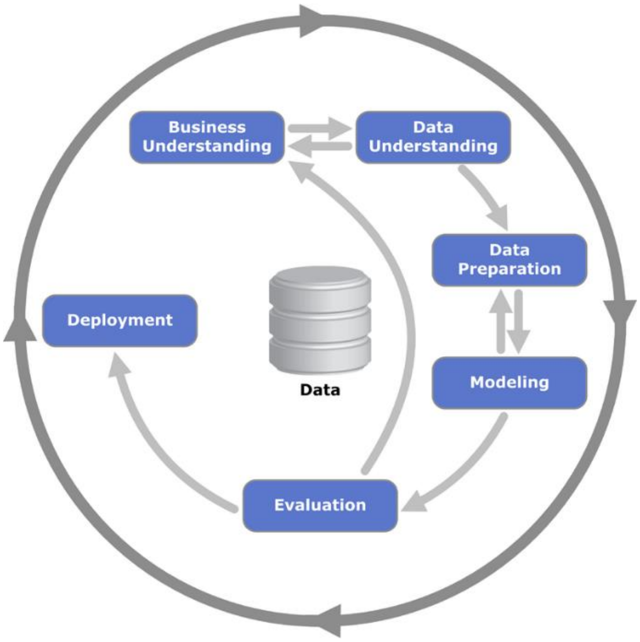


Figure 1. Cross Industry Standard Process for Data Mining Methodology

The first phase involves business understanding in which this study had defined the student's dropout issue in the previous section. The second phase requires understanding the data set, which involves data collection and measurement. Next phase is data preparation. This includes coding, feature extraction, cleaning and loading the data. Then, the modelling phase describes various applications of machine learning algorithm to build the predictive model that can predict the target using respective inputs to solve business problem. Evaluation phase focuses on assessing the performance of the model built. Final phase of deployment would be operationalization of the model using the new data set. The details of each phases are described in the following section. We used Orange 3.0 for data mining to develop the predictive model. Orange visualization and statistical modelling widget were used for data visualization and predictive models respectively.

4. The Data Set

The data set used in this study are referring to 64 students enrolled in Computer Science program Universiti Teknologi MARA Cawangan Kelantan, Malaysia.

Table 1. Description of Variables.

Attribute	Description	Possible Values
DROPOUT	Students Drop Out	{1 = Dropout, 0 = Graduated}
CGPA	Cumulative Grade Point Average	{0 – 4.00)
GENDER	Students Gender	{1 = Male, 0 = Female}
ACC106	Introduction to Financial Accounting And Reporting	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC118	Fundamentals of Algorithm Development	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ELC121	Integrated Language Skills II	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
MAT133	Intro to Mathematics	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC128	Structured Programming	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC159	Computer Organization	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ECO120	Principles of Economics	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ELC151	Integrated Language Skills II	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
MAT183	Calculus I	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC238	Object Oriented Programming	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ELC231	Integrated Language Skills III	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ITS232	Introduction to Database Management Systems	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
MAT210	Discrete Mathematics	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
MGT162	Fundamentals of Management	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC204	Practical Approach of Operating Systems	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC248	Fundamentals of Data Structures	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC253	Interactive Multimedia	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ENT300	Fundamentals of Entrepreneurship	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ITS332	Information Systems Development	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
STA116	Introduction to Probability and Statistics	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ATT270	Digital Electronics	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC301	Visual Programming	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
CSC318	Web Application Development	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}
ITT300	Introduction to Data Communication and Networking	{4.00, 3.67, 3.33, 3.00, 2.67, 2.33, 2.00, 1.67, 1.33}

The data were collected from two sources; admission information which included minimum demographic information (gender, birth date) and transcript records which included the courses taken and the grades for each of them, the academic program and the academic cumulative average. This research used computer science students entering the university from the first term of 2016 to the second term of 2016. The confidentiality of data was preserved by not using any personal data like ID number, date of birth, or name. This research only focused on the core courses offered by the programme as it dominates the study plan which in turn gives great impact to the dropout. Hence, the transcript data were collected with authorized permission from Examination Department of the university. 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 used in this study are presented in table 1.

5. Data Preparation and Exploration

In data preparation phase, we applied pre-processing technique for the collected data to prepare the data for mining purposes. The data were cleaned to ensure no missing value and no unwanted values exist in the data set. Some irrelevant attributes were also eliminated. Then the data were re-arranged so that the student has the following attributes: Dropout, CGPA, Gender and the course grade score student for the duration of 2016 - 2018. The target variable, Dropout is coded into two groups: 1 = Dropout and 2=Not Dropout (Graduated). Then, we performed a data profiling to obtain descriptive statistics to further understand the data. The distribution of dropout students segmenting by gender is displayed in figure 2. Comparison of CGPA distribution by dropout can be seen in figure 3. The boxplot in figure 4 indicates the average of the dropout students who obtained CGPA between the ranges of 2.33 to 2.81 cumulative grade point average. Further investigation on the CGPA segmenting by dropout status for selected courses is displayed in figure 5. As expected, dropout students presented lower grade averages across almost all selected subjects. Mathematics and IT courses were the most challenging for the students, followed by Computer Science Subject and Management.

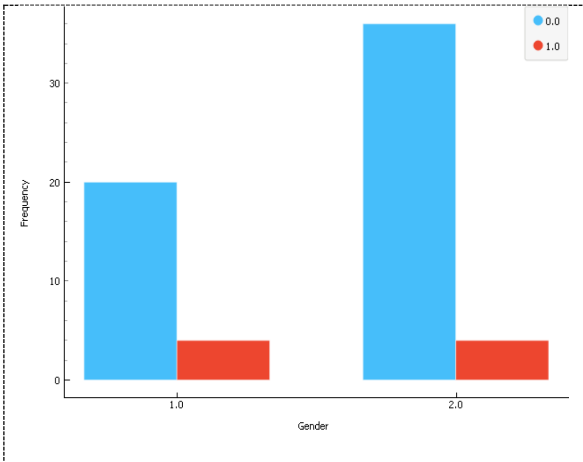


Figure 2. Percentage of drop out by gender.

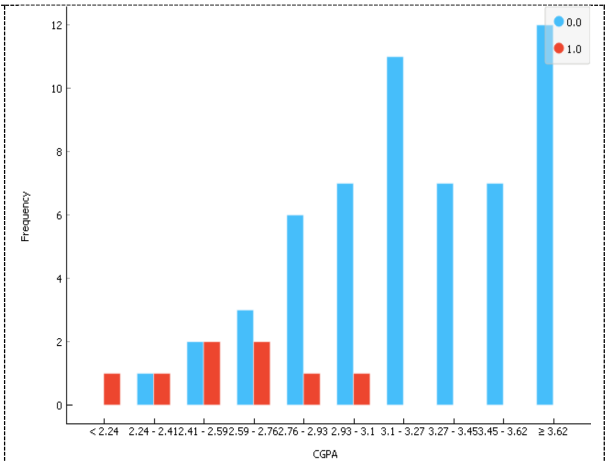


Figure 3. Comparison of CGPA by Dropout.

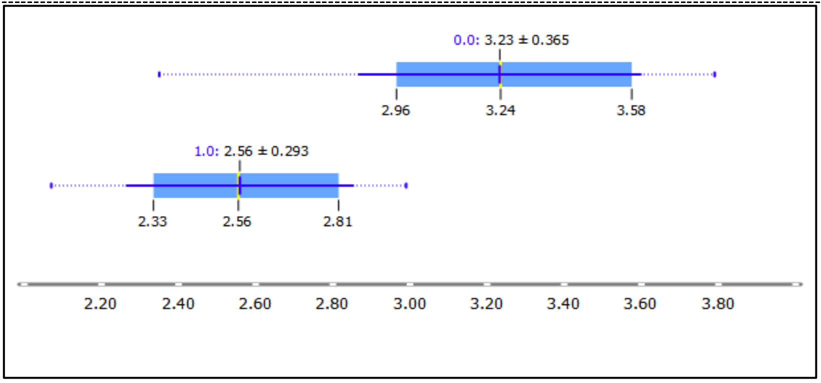


Figure 4. Box Plot of CGPA by Drop Out

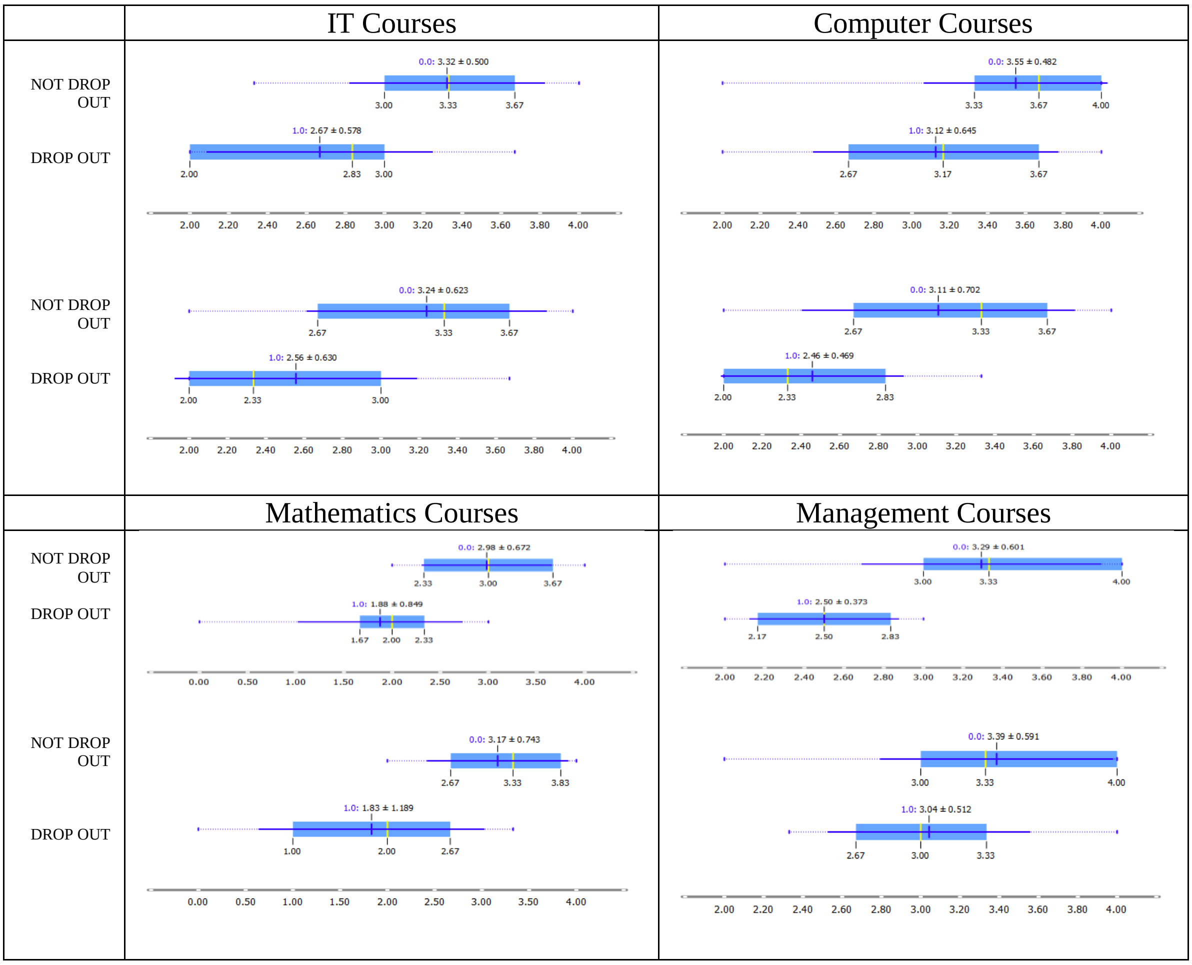
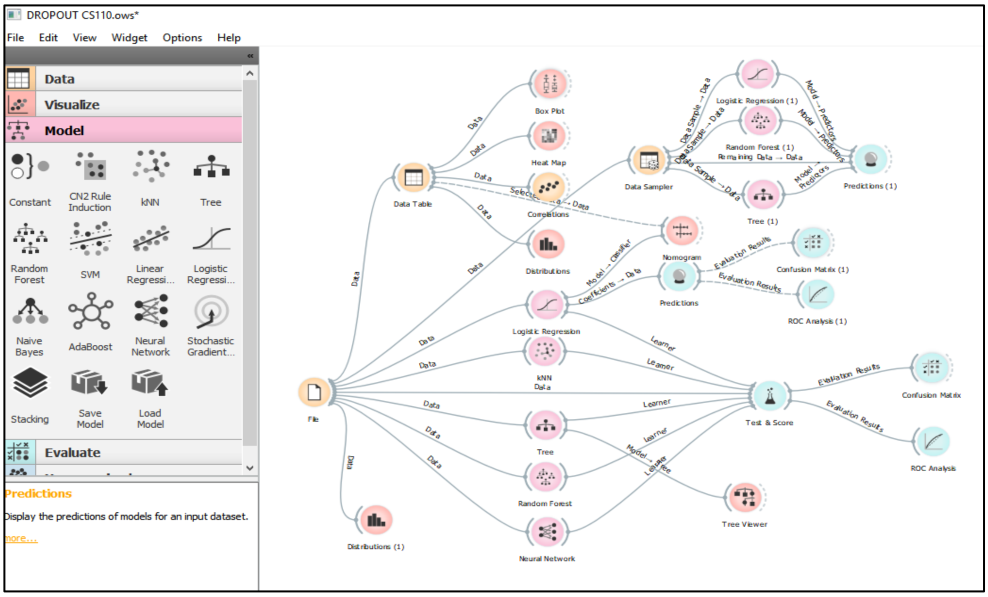


Figure 5. Box Plot of Courses Category by Drop Out

After visual exploration, the data were partitioned into training and testing set by the method of random sampling using test and score widget in Orange. The percentage proportion of training and testing data were divided into 60:40. The purpose of partitioning the data is to test the performance of the model. In training set, the k-NN, Naives Bayes, Decision Tree, Neural Network, Logistic Regression and Random Forest model were developed to predict the students' performance. In testing set, the models performance was measured. The overall flow diagram of Orange Tool is displayed in figure 6 below.



Predictions

Display the predictions of models for an input dataset.

more...

Figure 6. Flow diagram of Orange Tool

6. Modelling and Evaluation

Data mining techniques are mostly used to build a model for prediction, classification of data into categories or to discover any meaningful hidden pattern and relationships in the observed data. The most common methods of classifications are decision trees, classification rules, probabilistic or Bayesian networks, neural networks and hybrid procedures. In this study, we investigated the impact of five algorithms: k-NN, Decision Tree, Neural Network, Logistics Regression and Random Forest.

6.1. K-NN

K-NN 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 [23]. Thus, in this study, the students are classified by a majority vote of its neighbor with the object being assigned to the class most common among its k nearest neighbor's. The best choice of k depends upon the data.

6.2 Decision Tree

The decision tree models trained in this study is an induce binary tree with two minimum number of instances in leaves, a split greater than 5, with maximum 100 node levels for depth of classification tree and stop splitting the nodes after majority reach 95%.

6.3 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 as discussed in Hosmer, Lemeshow and Sturdivant [23]. The goal of this model is to estimate the probability that an event occurs, p which is the probability of dropout. In this study, maximum likelihood was used to find the best-fit line for logistic regression.

6.4 Neural Network

The Artificial Neural Network (ANN) trained in this study used multilayer perceptron (MLP) model with a maximum of 100 allowable hidden layer and maximum iteration of 200.

6.5 Random Forest

This study trained the Random Forest model with the following parameters: number of trees in the forest = 10, maximal number of considered feature= 5, and unlimited tree depth and stop splitting nodes with maximum instances= 5.

To evaluate the classifier, this study performs 10-times random sampling by splitting the data set into 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 Classification Accuracy (CA), AUC, Precision, Recall and Receiver Operating Characteristic (ROC) Chart. The results were used for comparison with baseline methods for performance evaluation.

7. Results and Discussion

As suggested based on previous works, models offering the best accuracy include Decision Tree, Logistic Regression, k-NN and Artificial Neural Network. In this study, apart from four predictive model, we also trained Random Forest model as a complemented technique because it is suitable for classification and regression. Random Forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mean prediction of the individual trees.

We first presented the results of Decision Tree model. The model produced a classification tree of 13 nodes and 7 leaves (Figure 7).

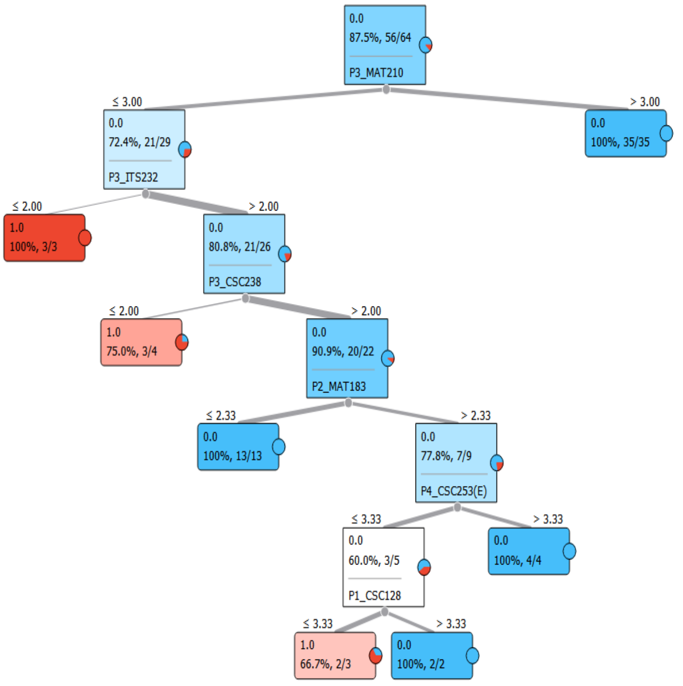


Figure 7. Decision Tree Output.

From the decision tree, we found that:

Case 1: A computer science student with Discrete Mathematics (MAT210) lower and equal to 3.00 and Failed (lower than 2.00) Introduction to Database Management Systems (ITS232) has 100% chance for dropping out from the university.

Case 2: A computer science student with Discrete Mathematics (MAT210) lower and equal to 3.00, Introduction to Database Management Systems (ITS232) of greater than 2.00 and Failed (lower than 2.00) Object Oriented Programming (CSC238) has 75% chance for dropping out from the university.

Case 3: A computer science student with Discrete Mathematics (MAT210) lower and equal to 3.00, Introduction to Database Management Systems (ITS232) of greater than 2.00, Object Oriented Programming (CSC238) greater than 2.00, Calculus I (MAT183) greater than 2.33, Interactive Multimedia (CSC253) lower than 3.33 and Structured Programming (CSC128) lower and equal than 3.33 has 66.7% chance for dropping out from the university.

For logistic regression model, the significant coefficients are displayed in table 2. As expected, it found that five important courses of Discrete Mathematics, Object Oriented Programming, Calculus I and Fundamentals of Data Structures had a negative influence on the dropping out factor. Hence, the higher the score obtained for these five important courses, the lower the risk of the student dropping out from the university.

Table 2. Significant Coefficient for logistic regression.

No.	Courses	Coefficients
1	MAT210	-0.2496
2	CSC138	-0.2447
3	MAT183	-0.5693
4	CSC238	-1.1324

Finally, table 3 presents the results for the overall performance of five selected predictive models algorithms based on Classification Accuracy, Precision and AUC. The achieved results revealed that the Logistic regression classifier performed best (with the highest overall classification accuracy) followed by k-NN, Random Forest, Neural Network and Decision Tree. All models tested performed with overall accuracy above 80% which means the error rate was low and predictions were reliable. The detection on the sensitivities of Logistic Regression, k-NN, Random Forest, Neural Network and Decision Tree were 89.9%, 87.8%, 85.2%,82.4% and 80.9% respectively.

Table 3. Comparison of Model Accuracy.

No.	Model	Classification Accuracy (CA)	Area Under Curve (AUC)
1	Logistic Regression	0.908	0.876
2	kNN	0.892	0.585
3	Random Forest	0.881	0.824
4	Artificial Neural Network	0.838	0.691
5	Decision Tree	0.812	0.619

The Receiver Operating Characteristics (ROC) curve is also used for the evaluation of classification algorithm. The ROC Curve measures 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 8, it was found that Logistic Regression model is the best classifier as the ROC curve is approaching 1. Hence, the results indicated that Logistic Regression performs very well in predicting the dropout students.

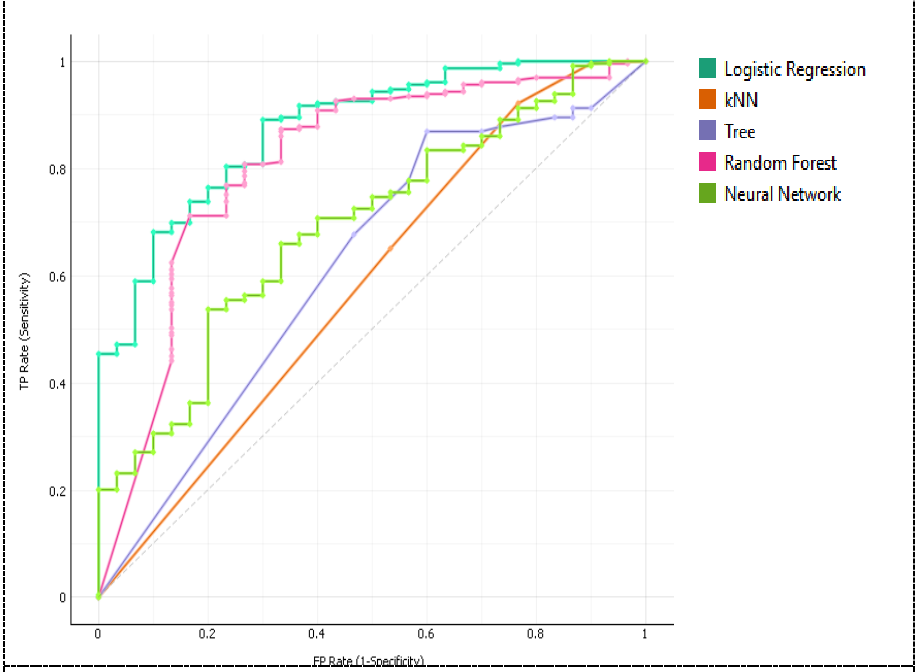


Figure 8. Models evaluation using ROC-AUC.

8. Conclusions

Data mining has played an important role in Educational research. Applying data mining in this domain can help to assess students' performance, predicting drop out and analyzing student performance. In this paper, we have applied five predictive model algorithms on the enrolment student's data to predict the student's dropout based on predictive accuracy. It has also been indicated that a good classifier model has to be both accurate and comprehensible. The results indicated that the best classification accuracy and AUC was achieved by Logistic regression classifier. It outperformed other algorithms compared to k-NN, Random Forest, Artificial Neural Network and Decision Tree with better accuracy and comprehensive classifier. Five important features were necessary to achieve this accuracy (Discrete Mathematics, Object Oriented Programming, Calculus I and Fundamentals of Data Structures). It implies that these courses have greatest impact in dropout prediction. The study can be further extended for future work by collecting a larger dataset from all campuses and applying the model using such data to show how it is generalized to other specific programs. The findings must be shared with faculty members to validate and refine the model and implement it for a better productive environment.

References

[1] Casanova, J. R., Cervero Fernández-Castañón, A., Núñez Pérez, J. C., Almeida, L. S., & Bernardo Gutiérrez, A. B. (2018). Factors that determine the persistence and dropout of university students. Factores determinantes de la permanencia y abandono de los estudiantes universitarios. Psicothema.

[2] Tinto, V.: Dropout from higher education: a theoretical synthesis of recent research. Rev. Educ. Res. 45(1), 89-125 (1975) 22. Wirth, R.: CRISP-DM: towards a standard process model for data mining. In: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, pp. 29-39 (2000)

[3] Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining Education data to predict student's retention: a comparative study. arXiv preprint arXiv:1203.2987.

[4] Baradwaj, B.K., Pal, S.: Mining educational data to analyze students' performance. arXiv preprint

arXiv:1201.3417 (2012)

[5] Bhardwaj, B.K., Pal, S. (2012) Data mining: a prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418

[6] Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

[7] Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 91-95). IEEE

[8] Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. Eurasian Journal of Educational Research, 54, 207-226.

[9] Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and e-learning, 17(1), 118-133.

[10] Brunsden, V., Davies, M., Shevlin, M., & Bracken, M. (2000). Why do HE students drop out? A test of Tinto's model. Journal of further and Higher Education, 24(3), 301-310.

[11] Durso, S. D. O., & Cunha, J. V. A. D. (2018). Determinant factors for undergraduate student's dropout in an accounting studies department of a Brazilian public university. Educação em Revista, 34.

[12] Kim, D., & Kim, S. (2018). Sustainable education: Analyzing the determinants of university student dropout by nonlinear panel data models. Sustainability, 10(4), 954.

[13] Tan, M. & Shao P. (2015). Prediction of student dropout in e-learning program through the use of machine learning method. International Journal of Emerging Technologies in Learning (iJET), 10(1), 11-17.

[14] Mustafa M. N., Chowdhury L., & Kamal M. S. (2012, May). Students dropout prediction for intelligent system from tertiary level in developing country. In 2012 International Conference on Informatics, Electronics & Vision (ICIEV) (pp. 113-118). IEEE.

[15] Abu-Oda G. S. & El-Halees A. M. (2015). Data mining in higher education : University student dropout case study. International Journal of Data Mining & Knowledge Management Process (IJDKP), 10(1), 15-27.

[16] Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006, December). Mining student data using decision trees. In International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.

[17] Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining.

[18] Jing, L.: Data mining and its applications in higher education. New Dir. Inst. Res. 2002(113), 17-36 (2002). https://doi.org/10.1002/ir.35, https://onlinelibrary. wiley.com/doi/abs/10.1002/ir.35

[19] Romero, C., Ventura, S.: Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl. 33(1), 135 - 146 (2007). https://doi.org/10.1016/j.eswa.2006. 04.005, http://www.sciencedirect.com/science/article/pii/S0957417406001266 Applying DM Techniques to Predict Student Dropout 125

[20] Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.

[21] Herzog, S. (2006). Estimating student retention and degree?completion time: Decision trees and neural networks vis?à?vis regression. New directions for institutional research, 2006(131), 17-33.

[22] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 16.Seidman, A.: Retention revisited: R= E, Id+ E & In, Iv. Coll. Univ. 71(4), 18-20 (1996)

[23] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Acknowledgments

This work was funded by Universiti Teknologi MARA through ARAS Grant (600-IRMI/DANA 5/3/ARAS (0026/2016)). We are grateful to Department of Academic Affairs and Internationalization for providing the dataset.