SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Prediction of Student Dropout in Malaysian's Private Higher Education Institute using Data Mining Application

Nurhana Roslan[1,*], Jastini Mohd Jamil[1], Izwan Nizal Mohd Shaharanee[1], Sultan Juma Sultan Alawi[2]

[1] School of Quantitative Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
[2] Ministry of Education, South Alsharqia, Sultane of Oman

ARTICLE INFO

Article history:
Received 21 April 2023
Received in revised form 1 March 2024
Accepted 14 April 2024
Available online 22 May 2024

Keywords:

Educational data mining; Student dropout prediction; Data mining; Student attrition

ABSTRACT

Student dropout issue is a major concern among the academics and management of the university. The higher rate of student dropout impacted the university reputation such as reducing student enrolment, affecting the revenue of the university, financial losses for the country, and increase the existence of a social problem among the students. In this study, 2 popular classifiers were utilized to predict the student dropout namely decision tree and logistic regression model respectively. Several sets of experimental setting were employed which include three set of data partitioning - along with different types of decision tree and regression model. As for the logistic regression model, different data imputation and transformation method was tested to ensure that the model built is valid. A total of 7706 student data extracted from one of the private universities in Malaysia database (between year 2018-2019) to assess the capability of the classifier. The classifier performance is evaluated using machine learning performance measure of accuracy and misclassification rate. The result indicates that, decision tree - chi-square (2 branches) achieved slightly better classification performance of 89.49% on 80/20 data partitioning. The chosen model also identified the most important variable for accurate prediction of student dropout. Application of this model has the potential to accurately predict at risk student and to reduce student dropout rates.

1. Introduction

Managing student dropout issue is a major concern among the management of the university. Dropout might be caused by many factors such as academic performance, health, family and personal reasons and varies depending on the nature of the study and the higher education provider [2,8]. The reputation of the higher education provider will be impacted if a large number of students dropped out from their respective institutions. A reduced number of student enrolments will cause many issues such as less revenue for the university. Private higher education provider in Malaysia depending more on the tuition fees compared to the public university. Thus, it is important for private higher education to get the right student that can complete their study. Therefore, there is a need

[*] *Corresponding author.*
E-mail address: hanaroslan1996@gmail.com

https://doi.org/10.37934/araset.45.2.168176

for student dropout analysis to recognize students with a high probability of dropout [1]. This analysis can offer a way for the higher education management to take precaution steps by understanding students at risk to dropout. Thus, reducing the probability of these student to dropout and quit their study.

Data mining is an important and helpful tool in the decision-making process. New knowledge can be extracted through analysing hidden patterns of data using various data mining techniques [5]. Through the process of data mining application, management of university can monitor the status of student dropout based on attributes obtained from their repository [12,13]. Hence, the tendency of student dropout can be detected. Therefore, this data mining application is useful for the university to conduct any plan in reducing and hence overcoming the issue regarding the student dropout of the university [20].

In this research work, 2 classifiers namely the decision tree and logistic regression model were developed and tested. Several experimental settings were utilized. This includes different data partitioning strategies in order to reduce bias in the data set. The performance of the models was evaluated using accuracy.

2. Methodology

2.1 Description of Method

In this study, 2 classification methods - decision tree and logistic regressions – are used to build the prediction models. Each model is compared to each other using 3 different data partition settings of training and testing - 80/30, 70/30 and 60/40 respectively. As a result, in this research works 36 different classification models setting were built as depicted in Table 3. The software used to perform the prediction of student dropout analysis is SAS Enterprise Miner. Figure 1 shows an overview of the research design for predicting student dropout using data mining technique. The research processes were conducted based on the activities designed in each phase to accomplish the respective objectives stated in this study. As depicted in Figure 1, the phases consist of phase 1, phase 2 and phase 3 which corresponds to the study objectives respectively conducted in this study. Phase 1 to phase 3 are developed based on the tool called SEMMA. SEMMA consists of sample, explore, modify, model and assess.

In achieving the first objective in this work, previous works from prominent researchers were reviewed and the data from the database of the private university were extracted. During this phase, the descriptive statistics was performed for instance, the exploratory analysis for the data set was conducted and the outliers which refer to the missing values that existed in the data set were handled using imputation methods such as replacing the missing values using mean value [19]. Phase 2 was conducted to achieve the second objective in this study by developing the classification techniques in SAS Enterprise Miner which refer to logistic regression model and decision tree model. In phase 3, all models were evaluated and compared, and the best model is selected based on the highest accuracy rate obtained.

Phase I: Literature Review and Dataset Extraction	
Tasks	Activities
To identify the most significant demographic factors on student dropout	- Performing Descriptive Statistics which include exploratory analysis. - Modify the dataset through identifying outliers and missing value, handling the missing value and checking the distribution of data.
Phase II: Model Development	
Task	Activities
To develop models in classifying the tendency of students to dropout or not from university using data mining applications.	- Conducting Data analysis using SAS Enterprise Miner: - Conducting a logistic regression model which refers to default regression without imputation, default regression with imputation, default regression with imputation & transformation, backward regression, forward regression and stepwise regression with different data partition set up (binary target: status) - Developing a decision tree (splitting rule: nominal target criterion: Gini, Entropy and Chi-square (max branch:2 and 3)
Phase III: Model Evaluation and Comparison	
Task	Activities
To compare and find the best model of the student dropout prediction.	Compare and evaluate all the models built and select the best model based on the highest accuracy.

Fig. 1. Overview of Research Design

2.2 Data

A secondary data approach is employed in this study by obtaining 7606 students' data from one of the private universities in Malaysia. There are 931 (12%) dropout students, and 6675 (78%) students completed their study. Hence, the number of students who completed their study in the university is higher as compared to the number of dropout students. The data set consists of the demographic data of the students in the private university that were obtained from the population data of the students in the university that includes the demographic variables that relates to the tendency of student dropout of the university. Table 1 below represents the list of variables with each description of the data set respectively. A summary of the data field for student data used for dropout prediction is given in Table 2.

Table 1
List of variables with each description of the data set

Variable Name	Model Role	Measurement Level	Description
Gender	Input	Binary	Students' gender which refers to male or female.
Student Grade	Input	Nominal	Achievement of students based on grades of Failed, Average or Excellent.
Place of Birth	Input	Nominal	State of birth each student.
Course	Input	Nominal	Courses taken by the students in the university.
Education Level	Input	Nominal	The last higher education of the students.
Sponsor/Fund Provider	Input	Nominal	Educational sponsored or educational loan to cover the tuition fees of the students.
Parent's income	Input	Interval	Income of the student's parents (in RM).
Location	Input	Ordinal	Student's living area that refers to rural, sub-rural, urban or sub-urban.
Number of dependents	Input	Interval	Dependency under the students' family for financial support.
Age	Input	Nominal	Age of students as they first enrol to the university (in years).

CGPA	Input	Interval	The cumulative grade point average, CGPA measured based on the student's performance in their courses enrolled. CGPA Range from: 0.00 to 4.00.
Curriculum Activity	Input	Nominal	The grade achieved by the students in their curriculum activity which starts from grade A, B, C, D and G.
Status	Target	Binary	Status of the students (Dropout, Not-dropout)

Thus, the definition of the relational student dataset format that is used in the prediction model is as follows:

Definition 1 Given a relational student database D, $I = \{i_1, i_2, ..., i_{|D|}\}$ the set of distinct items in D, $AT = \{at_1, at_2, ..., at_{|AT|}\}$ the set of input attributes in D, and $Y = \{y_1, y_2, ..., y_{|Y|}\}$ the class attribute with a set of class labels in D. Assume that D contains a set of n records $D = \{x_r, y_r\}_{r=1}^{n}$, where $x_r \subseteq I$ is an item or a set of items and $y_r \in Y$ is a class label, then $|x_r| = |AT|$ and $x_r = \{at_1val_r, at_2val_r, ..., at_{|AT|}|val_r\}$ contains the attribute names and corresponding values for record *r* in *D* for each attribute *at* in *AT*.

The student dataset is arranged in a row and column format. Each column is defined for attributes with their values, while the final column identified as the class attributes with a set of possible class labels.

Table 2
Summary of Data Field for Student Data

Variable Name	Measurement Level	Number of Values	Mean	Standard Deviation
Gender	Binary	2	-	-
Student Grade	Nominal	3	-	-
Place of Birth	Nominal	24	-	-
Course	Nominal	15	-	-
Education Level	Nominal	22	-	-
Sponsor/Fund Provider	Nominal	16	-	-
Parent's income	Interval	-	2192.31	1823.79
Location	Ordinal	4	-	-
Number of dependents	Interval	-	4.66	1.87
Age	Nominal	-	18.38	0.84
CGPA	Interval	-	2.47	1.06
Curriculum Activity	Nominal	5	-	-
Status	Binary	2	-	-

Based on the descriptive information of the data, the average parent's income is RM 2192.31 which ranged from the lowest of RM 0.00 to the highest value of RM 34,355.00. The majority of the students obtained Grade A in their curriculum activity. In addition, the mean score for CGPA is 2.47 which is class moderate within the Student Grade. On average, the count of family member in the student's family are 5 people. The National Higher Education Fund Corporation (PTPTN) has funded the bulk of the students. The students pursued their study in the early age of 18. Most of the students' last Education Level is National Secondary Schools (SMK). 559 students in the dataset were found to enrol into Diploma in Islamic Studies as their preferred course. A big number of students came from Perak, which indicate that this institution is preferred among the local community as this private higher education situated in state of Perak in Malaysia. Students that come from sub-rural area contribute to the large chunk of distribution in the database. As for gender variable, it was found that male students tend to dropout more than female students. Additionally, the majority of the dropout students obtained class Failed or Moderate in their study.

2.3 Classification of Methods

This research work aims to compare the performance of 2 classification techniques within the student dropout context. A concise overview of these 2 classification methods is as follows.

2.3.1 Decision tree

Decision trees (DT) built to predict discrete-valued target functions, where a decision tree represents the learned function connecting the predictor variables to the expected variable [16]. To search the most discriminating variables and variable values, the decision tree algorithm uses a divide-and-concur approach to construct a tree-looking structure consisting of nodes and edges. Gini Index, Information Gain, Entropy, Chi-square, etc. are several information measures differentiated how DT model works to identify the most discriminating variable and variable-values. The heuristic measure of Gini Index, Entropy and Chi-Square were utilized in this research work.

2.3.2 Logistic regression

Logistic regression as defined by [17] is a nonlinear regression technique that associates a conditional probability score with each data instance. The concept of logistic regression is to examine the linear relationship between the dependent variables and independent variable [11]. The dependent variable may be binomial (as is the case in this study) or multinomial.

2.4 Evaluation Measures

To evaluate the performance of the model, a popular measurement metric known as accuracy measure was utilized. Table 3 depicted a confusion matrix for model evaluation and Eq. (1) expresses the accuracy rate measure. Accuracy rate is typically defined as the number of correctly classified instances, while the number of incorrectly classified instances is referred to as a misclassification rate.

Table 3
Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

$$\text{Accuracy} = \frac{\mathrm{TP}}{(\mathrm{TP+FP+TN+FN})} \qquad (1)$$

Figure 2 outlined the pseudo code for accuracy measure in this research work.

```
Input: Training and Testing dataset
Output: Accuracy (AR) of each classifier setting
    For each classifier, scan the training and testing dataset
        Check whether rules classify all the instances in dataset
        Calculate Misclassification Rate (MR) for each rule
      AR = (1- sum of all MRs )* 100
return AR
```

Fig. 2. Pseudo code for the Accuracy Rate

3. Results

The results of all 36 models on the accuracy measure for both training and testing are listed in Table 4. Each row is populated with their specific experimental setting. The model utilizing Decision tree with Chi-Square as nominal target criterion (2 branches) – 80/20 data partition provided the highest accuracy (89.49% - Testing dataset). In this study, the ranked importance of the predictor factors was also investigated to discover the relative contribution of each to the prediction model.

Table 4
Model comparison for prediction performance based on accuracy measure

Model	% of Data Partition	No. of Branches	Splitting Criteria		Accuracy (%)	
					Training	Testing
Decision Tree	60/40	2	Gini		91.32	88.53
			Entropy		91.34	89.19
			Chi-Square		90.53	88.96
		3	Gini		92.39	87.94
			Entropy		92.44	87.39
			Chi-Square		90.22	88.73
	70/30	2	Gini		91.30	88.13
			Entropy		90.96	89.05
			Chi-Square		90.57	88.74
		3	Gini		92.37	87.78
			Entropy		91.81	87.30
			Chi-Square		90.55	88.22
	80/20	2	Gini		90.84	89.36
			Entropy		90.84	88.90
			Chi-Square		90.45	89.49
		3	Gini		92.03	88.77
			Entropy		91.68	88.90
			Chi-Square		90.17	88.77
Logistic Regression	% of Data Partition	Regression Type	Imputation	Transform	Accuracy (%)	
					Training	Testing
	60/40	Default	No	No	88.69	87.52
		Default	Yes	No	90.05	87.98
		Default	Yes	Yes	90.09	87.98
		Backward	Yes	Yes	89.96	88.01
		Forward	Yes	Yes	89.96	88.01
		Stepwise	Yes	Yes	89.96	88.01
	70/30	Default	No	No	88.54	87.56

		Default	Yes	No	89.72	88.09
		Default	Yes	Yes	89.76	88.09
		Backward	Yes	Yes	89.69	88.00
		Forward	Yes	Yes	89.69	88.00
		Stepwise	Yes	Yes	89.46	88.39
	80/20	Default	No	No	88.31	87.85
		Default	Yes	No	89.50	88.44
		Default	Yes	Yes	89.46	88.38
		Backward	Yes	Yes	89.51	88.44
		Forward	Yes	Yes	89.51	88.44
		Stepwise	Yes	Yes	89.27	88.71

Table 5 shows 8 predictor variables. As can be seen, the most important factors came out to be CGPA, Courses, Educational Sponsorship, Educational Level, Number of Dependent, Curriculum Activity, Parent's Income and Gender.

Table 5
Variable importance – Decision tree with Chi-Square (2 branches) – 80/20 data partition

Rank	Variable Name	Feature Importance Score
1	CGPA	1.0000
2	Course	0.6814
3	Sponsor/Fund Provider	0.3490
4	Education Level	0.3131
5	Number of dependents	0.2222
6	Curriculum Activity	0.1539
7	Parent's income	0.1366
8	Gender	0.1264

4. Conclusions

In this research paper, the aim was to develop a classifier for predicting student dropout, using decision tree and logistic regression models. The study evaluated 36 different models with varying experimental settings, and despite achieving high accuracy rates, the results showed that the models needed further improvement to predict irregular/unexpected examples such as dropout students.

One of the main challenges faced in this study was the imbalance in the dataset, with a majority of non-dropout students and only a small percentage of dropout students. To address this, the researchers emphasized the importance of proper preprocessing methods to extract suitable student data and characteristics, which can help to better understand the underlying reasons for dropout and identify at-risk students who are more likely to drop out.

The results showed that the decision tree model was more effective in predicting student dropout than the logistic regression model, as it provided a more transparent model structure and clearly showed the reasoning process of different prediction outcomes. However, the researchers acknowledged the potential for further improvement by incorporating data from other sources, such as social media or survey-based data, and by exploring other predictive modelling methods like neural networks and support vector machines.

Overall, the study highlights the importance of developing accurate and transparent models for predicting student dropout, and emphasizes the need for ongoing research to improve these models and better understand the complex factors that contribute to student success or failure.

Acknowledgement

This research was supported by Universiti Utara Malaysia through University Research Grant Scheme (SO Code: 14587).

References

[1] Jamaludin, Aaishah Radziah, Wan'Atikah Wan Ibrisam Fikry, Siti Zhafirah Zainal, Fatin Shaqira Abdul Hadi, Nawal Shaharuddin, and Nurul Izzati Abd Rahman. "The effectiveness of academic advising on student performance." *International Journal of Advanced Research in Future Ready Learning and Education* 25, no. 1 (2021): 20-29.

[2] Saa, Amjad Abu. "Educational data mining & students' performance prediction." *International journal of advanced computer science and applications* 7, no. 5 (2016). https://doi.org/10.14569/IJACSA.2016.070531

[3] Alban, Mayra, and David Mauricio. "Predicting university dropout through data mining: a systematic literature." *Indian Journal of Science and Technology* 12, no. 4 (2019): 1-12. https://doi.org/10.17485/ijst/2019/v12i4/139729

[4] Amazona, Mayreen V., and Alexander A. Hernandez. "Modelling student performance using data mining techniques: Inputs for academic program development." In *Proceedings of the 2019 5th International Conference on Computing and Data Engineering*, pp. 36-40. 2019. https://doi.org/10.1145/3330530.3330544

[5] Antonenko, Pavlo D., Serkan Toy, and Dale S. Niederhauser. "Using cluster analysis for data mining in educational technology research." *Educational Technology Research and Development* 60 (2012): 383-398. https://doi.org/10.1007/s11423-012-9235-8

[6] Bilquise, Ghazala, Sherief Abdallah, and Thaeer Kobbaey. "Predicting student retention among a homogeneous population using data mining." In *International Conference on Advanced Intelligent Systems and Informatics*, pp. 35-46. Cham: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-31129-2_4

[7] Cohen, Anat. "Analysis of student activity in web-supported courses as a tool for predicting dropout." *Educational Technology Research and Development* 65 (2017): 1285-1304. https://doi.org/10.1007/s11423-017-9524-3

[8] Abd Rahman, Haliza, Zarina Mohd Khalid, Noraslinda Mohamed Ismail, Nur Arina Bazilah Kamisan, Siti Mariam Norrulashikin, Siti Rohani Mohd Nor, Ani Shabri, and Muhammad Fauzee Hamdan. "Statistical Analysis on Students' Evaluation and Students' Final Exam Marks in Undergraduate Mathematical Courses at Universiti Teknologi Malaysia." *International Journal of Advanced Research in Future Ready Learning and Education* 27, no. 1 (2022): 1-8.

[9] Hutagaol, Nindhia, and Suharjito Suharjito. "Predictive modelling of student dropout using ensemble classifier method in higher education." *Advances in Science, Technology and Engineering Systems Journal* 4, no. 4 (2019): 206-211. https://doi.org/10.25046/aj040425

[10] Jeevalatha, T., N. Ananthi, and D. Saravana Kumar. "Performance analysis of undergraduate students placement selection using decision tree algorithms." *International Journal of Computer Applications* 108, no. 15 (2014). https://doi.org/10.5120/18988-0436

[11] Maalouf, Maher. "Logistic regression in data analysis: an overview." *International Journal of Data Analysis Techniques and Strategies* 3, no. 3 (2011): 281-299. https://doi.org/10.1504/IJDATS.2011.041335

[12] Meedech, Phanupong, Natthakan Iam-On, and Tossapon Boongoen. "Prediction of student dropout using personal profile and data mining approach." In *Intelligent and Evolutionary Systems: The 19th Asia Pacific Symposium, IES 2015, Bangkok, Thailand, November 2015, Proceedings*, pp. 143-155. Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-27000-5_12

[13] Mduma, Neema, Khamisi Kalegele, and Dina Machuve. "A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction." *Data Science Journal* 18, no. 1 (2019). https://doi.org/10.5334/dsj-2019-014

[14] Pal, Saurabh. "Mining educational data to reduce dropout rates of engineering students." *International Journal of Information Engineering and Electronic Business* 4, no. 2 (2012): 1. https://doi.org/10.5815/ijieeb.2012.02.01

[15] Pattanaphanchai, Jarutas, Koranat Leelertpanyakul, and Napa Theppalak. "The investigation of student dropout prediction model in thai higher education using educational data mining: A case study of faculty of science, prince of Songkla Uni-versity." *Journal of University of Babylon for Pure and Applied Sciences* 27, no. 1 (2019): 356-367. https://doi.org/10.29196/jubpas.v27i1.2191

[16] Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1 (1986): 81-106. https://doi.org/10.1007/BF00116251

[17] Roiger, Richard J. "Data mining: a tutorial-based primer." (2017). https://doi.org/10.1201/9781315382586

[18] Kadoić, Nikola, and Dijna Oreški. "Analysis of student behavior and success based on logs in Moodle." In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 0654-0659. IEEE, 2018. https://doi.org/10.23919/MIPRO.2018.8400123

[19] Shaharanee, Izwan Nizal Mohd, and Fedja Hadzic. "Irrelevant Feature and Rule Removal for Structural Associative Classification Using Structure-Preserving Flat Representation." *Feature Selection for Data and Pattern Recognition* (2015): 199-228. https://doi.org/10.1007/978-3-662-45620-0_10

[20] Shahiri, Amirah Mohamed, and Wahidah Husain. "A review on predicting student's performance using data mining techniques." *Procedia Computer Science* 72 (2015): 414-422. https://doi.org/10.1016/j.procs.2015.12.157

[21] Wahyuni, S., S. KS, and M. Iswan. "The implementation of decision tree algorithm C4. 5 using RapidMiner in analyzing dropout students." In *4th International Conference on Technical and Vocation Education and Training*. 2017.

[22] Zhang, Shaoyan, Christos Tjortjis, Xiaojun Zeng, Hong Qiao, Iain Buchan, and John Keane. "Comparing data mining methods with logistic regression in childhood obesity prediction." *Information Systems Frontiers* 11 (2009): 449-460. https://doi.org/10.1007/s10796-009-9157-0

[23] Zhang, Ying, Samia Oussena, Tony Clark, and Hyeonsook Kim. "Use Data Mining to Improve Student Retention in Higher Education-A Case Study." In *ICEIS (1)*, pp. 190-197. 2010.