Drop-Out Prediction in Higher Education Among B40 Students

Nor Samsiah Sani[1], Ahmad Fikri Mohamed Nafuri[2], Zulaiha Ali Othman[3]
Mohd Zakree Ahmad Nazri[4], Khairul Nadiyah Mohamad[5]
Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[1, 2, 3, 4]
Unit Pemodenan Tadbiran dan Perancangan Pengurusan Malaysia (MAMPU), Cyberjaya, Selangor, Malaysia[5]

***Abstract*—Malaysia citizens are categorized into three different income groups which are the Top 20 Percent (T20), Middle 40 Percent (M40), and Bottom 40 Percent (B40). One of the focus areas in the Eleventh Malaysia Plan (11MP) is to elevate the B40 household group towards the middle-income society. In 2018, it was estimated that 4.1 million households belong to this group. The government of Malaysia has widened access to higher education for the B40 group in an effort to reduce the gaps in socioeconomics and to improve their living standards. Statistical data shows that since 2013, a yearly intake of students in bachelor's degree programs in Malaysia's public universities amounts to more than 85,000. Despite this huge number of enrolments, not all were able to graduate, including students from low-income family background. Data mining approach with machine learning techniques has been widely used effectively and accurately to predict students at risk of dropping out in general education. However, machine learning related works on student attrition in Malaysia's higher education is generally lacking. Therefore, in this research, three machine learning models were developed using Decision Tree, Random Forest and Artificial Neural Network algorithm in order to classify attrition among B40 students in bachelor's degree programs in Malaysia's public universities. Comparative performance analysis between the three models indicates that the Random Forest model is the best model in predicting student attrition in this study. Random Forest model outperforms the other two models in terms of accuracy, precision, recall and F-measure with the value of 95.93%, 97.10%, 81.26% and 88.50%, respectively. Nevertheless, there is a statistically significant difference in performance between the Random Forest model and Decision Tree model but no statistically significant difference between Random Forest models and Artificial Neural Network model.**

Keywords*—Machine learning; prediction; student attrition; student drop-out; B40; random forest; decision tree; artificial neural network

I. Introduction

Malaysia's household income is classified into three groups, which are Bottom 40% (B40), Middle 40% (M40) and Top 20% (T20). According to the Department of Statistic, Malaysia (2017), generally in Malaysia, B40 household income is not more than RM 4,630.00. Approximately, there were 2.7 million households belonging to B40 group in 2014. The figures increased in 2018, as the government announced that 4.1 million households will continue to benefit from Bantuan Sara Hidup (Household Living Aid) (BSH) which is specially allocated to B40 group [1].

B40 had been selected as a focus group in Rancangan Malaysia 11 2016-2020 (The Eleventh Malaysia Plan) (RMK-11). Through RMK-11, the government used education as one of the strategies to boost B40 household's income and ultimately narrowing socioeconomic gap [1]-[2]. Higher education institution and skills training institutes were encouraged to allocate more seats and allowing admission by special allocation for B40 students in an effort to ensure their access to higher education is secured. As reported in Higher Education Statistic, from the year 2011 to 2015, the total number of students intake for bachelor's degree programmes in Malaysia Public Universities is more than 85,000 students yearly. The highest number of students intake was recorded in the year 2011 with 99,862 students. Nonetheless, not all were able to graduate on time. In a worse case, some dropped out voluntarily or were expelled from by university.

Student's attrition in university will negatively affect B40 students financially. The family financial burden will increase as student's education loan has to be paid even if they fail to graduate. Furthermore, it will affect a student's chances on securing a high-income job. Students drop out would also lead to a huge loss in human capitals to the nation as fewer professionals and expert skills will be produced by public universities.

Hence, a proactive approach is desperately needed in identifying students who are at risk at dropping out. An effective prediction model using machine learning technique can be implemented for that purpose. Thus, the aim of this paper is to conduct a comparative study for machine learning models in predicting attrition among B40 students, particularly in the bachelor's degree programme in Malaysia Public Universities. Decision Tree (DT) Random Forest (RF), and Artificial Neural Networks (ANN) algorithms were adopted in constructing the models.

The remainder of this paper is organized as follows. Section 2 presents previous research articles related to classification technique in education and student drop-out prediction in higher learning institutions. Section 3 describes the methodology used in predicting student's drop out in this research. Results and discussion will be discussed in Section 4, while the conclusion of this paper and further works is outlined in Section 5.

II. Literature Review

Classification is a machine learning technique that can be used to predict students drop out rate accurately to help reducing student attrition rate. The task is crucial as the ability to predict students at risk the earliest possible is a great help to keep students from leaving their studies and overcome attrition among B40 students. Classification technique had been developed and applied successfully to a wide range of real-world domains [3] – [8]. Also, the classification is playing an important role in the education domain, especially in predicting student's academic performance, whether in school or higher education institution [9]. The research had review 30 studies carried out in between year 2002 until early 2015 and discovered that Artificial Neural Network (ANN), Decision Tree (DT), Naïve Bayes (NB), k-Nearest Neighbour (k-NN) and Support Vector Machine (SVM) were often used in building prediction models. However, findings showed that ANN and DT models produced higher accuracy results than the others.

Over recent years, there has been a significant growth of research published in predicting student performance, focusing on course drop-out/ retention using the technique of classification (supervised learning). These researches concentrate on predicting final grade or Cumulative Grade Point Average (CGPA) of students by utilizing classifier algorithm [10]-[13], predicting student's performance in *Massive Online Open Courses* (MOOC) environment [14], and predicting students at risk of not graduating high school on time [15].

In Malaysia, classification techniques had been applied in education domain, but the focus was more on student performance rather than attrition. Reference [16] in their work designed a model to identify key factors that influence the drop-out rates in Computer Science course. They collected student's demographic information and transcript records which focused on the core courses offered as it gives more impact on the drop-out case. Four different classification techniques namely k-NN, DT, NN and Logistic Regression (LR) are utilized to classify the dataset. The results show that LR classifier is the most accurate (91%) as compared to other techniques used in this work. The outcome of this work reveals that there are five important courses that the student must score higher to lower the chance of dropping out.

Bedregal Alpaca et al. [17] proposed classification models based on academic information provided by university to identify a student at risk of drop-out. The student's demographic, academic performance, admission test and course information data are considered for the evaluation. From the result, it is observed that the model is able to determine the most significant variable that affects academic performance, which is the abandoned subjects.

Gil, Delima, and Vilchez [18] adapted DT and NB to identify the underlying factors of student drop-out in a public school in the Philippines. They used Weka tool kit to utilize the classifier algorithm on the selected dataset and produced a comparative result of each algorithm performance in terms of recall, precision and accuracy. Meanwhile, [19] only concentrated on k-NN to perform extensive evaluation and predict student drop-out at an early stage of study. The technique is versatile, simple and can handle different type of data. The results can help teachers to identify a student at risk of drop-out and check on their welfare.

Mardolkar and Kumaran [20] adapted data mining technique to find comprehensive prediction models of student drop-out as early as possible. The model with sufficiently high accuracy will be used in an early warning system as an effort to detect students at high risk of drop-out as soon as possible. They explored the academic variables (both at universities and former school), sociodemography, behaviour and extracurricular activities that may influence student drop-out. However, only a subset of attributes that has a very high predictive contribution on the student drop-out.

Tomasevic, Gvozdenovic and Vranes [21] conducted a research with an objective to provide a comprehensive analysis and comparison of supervised machine learning techniques for discovering students at a high risk of dropping out from the course. For this, they used various classifier such as k-NN, SVM, ANN, DT, NB and LR as the classification tool. The overall highest precision was obtained with ANN by feeding the algorithm with student engagement data in online learning and past performance data.

Viloria and Padilla [22] in their study applied NN, DT and Bayesian Network to predict drop-out among engineering students in India. As a result, it was found that academic results and socioeconomic situation have an influence on students and managing these variables helps reduce the drop-out rate.

Sangodiah et al. [23] used SVM to predict academic performance for students under probation in a private higher learning institution. The model gained 89.84% of accuracy. Likewise, [24] also used single classifier to predict postgraduate doctoral degree students that will complete their study on time by using Binary Linear Regression. The outcome revealed that only 6.8% of the students in the year 2014 were able to graduate on time.

Table I described 16 studies conducted in predicting student drop out in higher learning institution from the year 2015 until 2020. The studies indicated that academic and sociodemographic data were important features used in predicting student. Other than that, there was only one research in predicting student drop-out that uses data from server logs containing student's activities for online courses offered from various universities. All of the research reviewed here were targeting students from only one course/major in one faculty or similar institution. However, in this research, the focus will be shifted to predicting drop out among B40 students by using academic or sociodemographic data from various majors and various higher learning institutions (public universities).

TABLE I. Research in Student Drop-Out Prediction in Higher Institution Using Classification Techniques

Author, Year	**Objective**	**Data**	**Algorithm**	**Result**
Bedregal-Alpaca et al. (2020) [17]	To generate a classification model and implement them on academic information provided by the university.	Demographic, academic, admission test, course information	ANN, DT	The generated model is able to determine the most significant variable that affects academic performance, which is the abandoned subjects.
Gil et al. (2020) [18]	To identify the underlying factors of drop-out students and apply the different approach of data mining algorithms.	Academic, student attendance, sociodemographic,	DT, NB	DT model produces the best result. The model identified key factors that affect students drop-outs.
Mardolkar & Kumaran (2020) [19]	Evaluate and propose k-NN method to predict students' drop-out	Student welfare feature and academic performance	k-NN	The technique is versatile, simple and can handle different type of data.
Tomasevic et al. (2020) [21]	To provide a comprehensive analysis and comparison of supervised machine learning techniques applied for discovering students at a high risk of dropping out of the course.	Demographic, student engagement in virtual learning, academic performance	k-NN, SVM, ANN, DT, NB, LR	ANN gave the highest precision by feeding the engagement data and past performance data.
Wan Yaacob et al. (2020) [16]	To identify key factors that influence the drop-out rates in Computer Science Program and which data mining technique is the most suitable approach.	Demographic, academic (CGPA and transcript records)	k-NN, DT, NN, LR,	LR classifier was the best technique with 91% accuracy. The models identified key factors/courses that have a greater impact on drop-out.
Viloria et al. (2019) [22]	Student drop-out prediction using data mining techniques.	Academic (university), academic (school)	NN, DT, Bayesian Network	All predictive models produced similar results, but Bayesian Network has slightly higher precision.
Limsathitwong, Tieatthanont & Yatsungnoen (2018) [25]	To develop web-based system with the ability to predict students who are at risk to drop-out in Information Technology major.	Academic (First and second year students)	DT, RF	RF accuracy higher than DT
Chen, Johri & Rangwala (2018) [26]	Performance comparison between survival analysis framework and machine learning approach in predicting student attirition in Science, Technology, Engineering and Mathematic (STEM) major.	Academic and sosiodemografic	*survival analysis*, Linear Regression (LR), DT, NB, RF, Adaboosting	Survival analysis outperforms other classifiers. Important features that influence student attrition was a student's age when enrolled in university and CGPA.
Ortiz-Lozano et al. (2018) [27]	Student drop-out prediction from school of engineering in one of the universities in Spain..	Academic and sosiodemografic	DT – CART, QUEST	Model accuracy is 70%. Academic results were a good feature in predicting student drop-out.
Aulck et al. (2016) [28]	Student drop-out prediction in one of thepublic universities in United States of America.	Academic and sosiodemografic	LR, RF, k-NN	The prediction accuracy of LR model was higher than the other two models.
Liang et al. (2016) [29]	Predicting student who is at risk of leaving a course in ten days on Edx MOOC platform	Students activities record in server logs.	SVM, LR, RF, DT with *gradian boosting*	DT with the gradian boosting prediction model outperformed the others in accuracy.
Pokrajac et al. (2016) [30]	To develop student drop-out prediction model at Delaware State University.	Academic and high school data.	ANN	The accuracy of the model increased with CGPA and the number of credit hours were included in developing the model.
Márquez-Vera et al. (2016) [20]	Propose an algorithm to obtain a reliable and comprehensible classification with sufficiently high accuracy.	Academic (university), academic (school), sociodemography, behaviour and extracurricular activities	NB, SVM, k-NN, DT	Focus on early detection to be used in an Early Warning System. Classification performance is near to 100%.
S. Abu-Oda & M. El-Halees (2015) [31]	Predicting attrition for science computer students in University Al-Aqsa.	Academic , sosiodemografic and student admission.	DT, NB	DT performed better than NB with an accuracy of 98.14%.
Strecht et al. (2015) [32]	Comparing prediction models' performance for student drop-out at Porto University.	Sosiodemografic, student admission, financial aid	k-NN, RF, Adaboost, CART, SVM, NB	SVM outperformed other models based on F-measure score, but the differences were not significant.
Siri (2015) [33]	Predicting student drop-out for bachelor degree programme (healthcare) in University of Genoa, Itali.	Academic , sosiodemografic and phone conversation	ANN	Able to predict student drop-out with 76% accuracy.

Comparative studies of two or more prediction models had been the core for 13 studies while the remaining used single classifier. In comparing prediction models performance, DT was the main choice among the researchers which was used in 11 studies followed by NB/ k-NN (six studies), RF (five studies) and ANN/ SVM (four studies). Based on the review, it can be concluded that DT is the most popular choice among the researchers in predicting student drop out as it is easy to comprehend and produce high-performance prediction results. Other than that, over the recent years, classification model using ensemble learning, especially RF had been increasingly popular among researchers because the performance outcome is very high as compared to a single classifier. Thus, a comparative study between classifier, particularly DT, ANN and RF is very much needed to discover the best prediction models for student drop-out among B40 students.

III. Research Methodology

In general, this research was conducted in three phases which were Phase I – Feasibility Study, Phase II – Data Preparation ad Phase III – Modeling and Evaluation. Fig. 1 shows detailed activities for each phase in research methodology. Fig. 1 illustrated phases and details of activities for each phase for research methodology in this study. There were three softwares used in this study, which were RapidMiner for prediction models construction and performance evaluation, MariaDB database to store and pre-process data, along with SPSS for attribute selection and statistical test.

A. Data Preparation

1) Data acquisition: The dataset was provided by Bahagian Pembangunan dan Perancangan Dasar (BPPD), Kementerian Pendidikan Malaysia (Pendidikan Tinggi) which consists of 44,406 records with 23 attributes. The dataset holds student's records from 20 public universities for bachelor degree programmes, who have dropped out or graduated from the year 2014 to 2017 intake.

2) Data Pre-processing: Pre-processing of data is a method of transforming a dataset in order to better expose the information quality to the mining tool. Real world data is often incomplete, incoherent and can contain noise such as errors and outliers. Pre-processing data is therefore required to ensure that data is formatted for a given miner tool and must be adequate for a given method. Data cleaning was performed using dimension reduction process. Attributes with more than 20,000 data unavailable, redundant or obsolete were deleted from the dataset. Incomplete records or outliers were also discarded (Table II). Data cleaning also ensured that the dataset included only student records with B40 household income (not more than RM4,387.00).

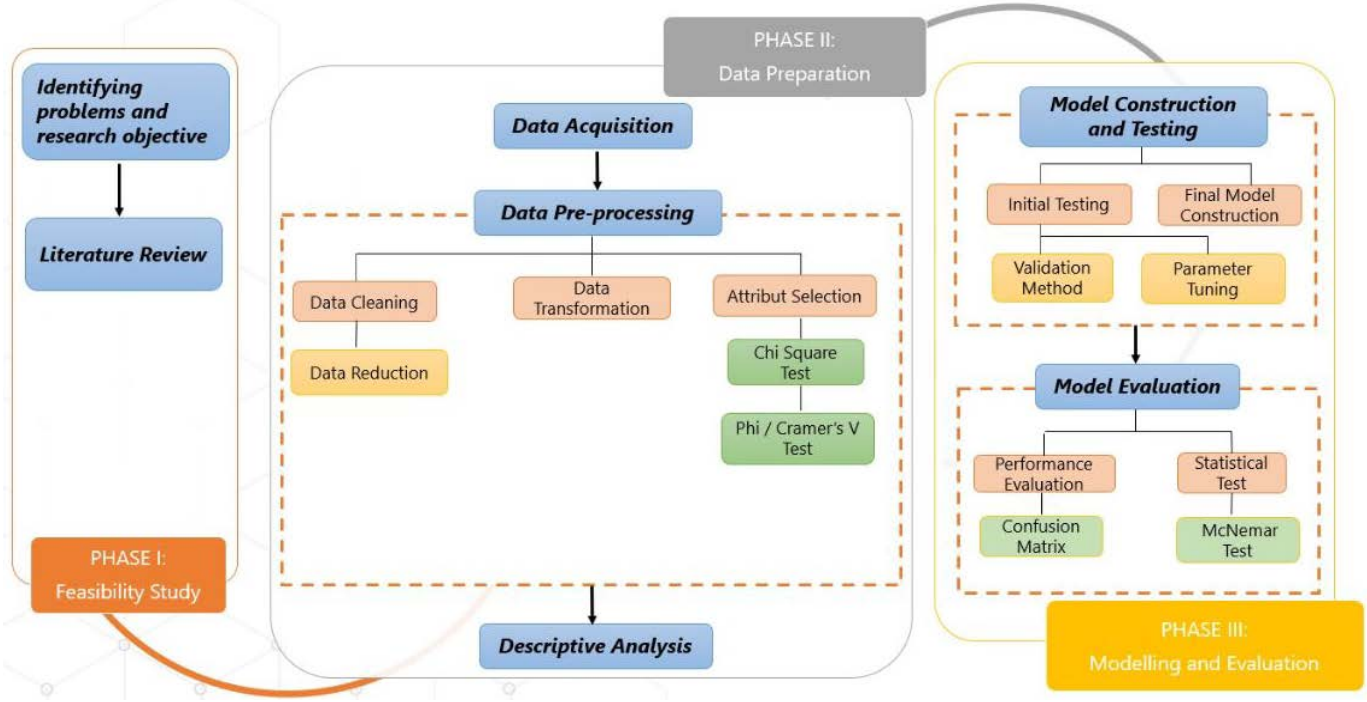

Fig. 1. Research Methodology.

TABLE II. RAW DATA ATTRIBUTES

Bil	Attribute Name	Data Type
1	student_id	nominal
2	date_of_birth	number (date)
3	gender	nominal
4	maritial_status	nominal
5	place_of_birth	nominal
6	postcode	nominal
7	state_legislative_assembly	nominal
8	parliament	nominal
9	country	nominal
10	institution_code	nominal
11	institution	nominal
12	citizenship	nominal
13	programme_code	nominal
14	programme	nominal
15	study_mode	nominal
16	study_method	nominal
17	sponsorship	nominal
18	family_income	nominal
19	student_status	nominal
20	session_intake_id	nominal
21	end_date	number (date)
22	qualification	nominal
23	cgpa_t1	number

Next, the data were transformed into a structure or understandable format befitting data mining. There were two new attributes constructed from existing attributes which were age from date_of_birth and class from student_status. Attribute class was the class label in this study. In attribute student_status, records with data' Berhenti' or 'Diberhentikan' were translated into 'C' in attribute class which represented students who dropped out while records with data' Tamat' were translated into 'G' which represented students who manage to graduate. Furthermore, attributes with varieties of data were aggregated or generalized by using the hierarchical concept.

Only relevant attributes were selected and used in building the prediction models. For this reason, the Chi-Square test was used to assess the relationship between attributes and the class label. Attributes with test result $\boldsymbol{p} < 0.05$ were considered having significant association with the class label. Afterwards, further tests were performed using Phi (ϕ) or Cramer's V (V) to each associated attribute in order to measure the strength of association. The value of ϕ or V is between 0 to 1, with 1 being the strongest and 0 being the weakest. The interpretation of the association between attributes and class label is, as shown in Table III.

Refer to Table IV, Chi-Square test results showed that all attributes had a significant association with the class label as the ***p***-value for each attribute that was less than 0.05. However, based on Cramer's V/Phi test result, *place_of_birth* and *family_income* were discarded from the dataset as their association level with the class label can be ignored. The final dataset for model construction consisted of 28,844 records with 9 regular attributes and 1 special attribute (class) (Table V).

B. Descriptive Analysis

Table VI and Table VII report the statistical analysis for each attribute after pre-processing data activity. Students who dropped out (class 'C') only represented 19.22% data in this study as compared to 80.78% of students who managed to graduate (class 'G'). More than half of the students (59.47 %) come from UiTM while UKM had the lowest number of students (0.21%).

TABLE III. PHI OR CRAMER'S V INTERPRETATION

Nilai Phi / Cramer's V	**Interpretation of Association**
0.00–0.10	Negligible
0.10–0.20	Weak
0.20–0.40	Moderate
0.40–0.60	Relatively Strong
0.60–0.80	Strong
0.80–1.00	Very Strong

Source : Lee 2016.

TABLE IV. X2 AND PHI/CRAMER'S V RESULTS (SORT BY ATTRIBUTE RANKING)

Attribute	**$\chi 2$**	***p*-Value**	**Cramer's V / Phi**	**Interpretation of Association**
cgpa_t1	18090.327	0.0000	0.79	Strong
institution	7084.965	0.0000	0.50	Relatively Strong
study_mode	2394.091	0.0000	0.29	Moderate
programme	2307.152	0.0000	0.28	Moderate
sponsorship	1838.139	0.0000	0.25	Moderate
qualification	1690.690	0.0000	0.24	Moderate
marital_status	706.977	0.0000	0.16	Weak
gender	587.057	0.0000	0.14	Weak
age	429.402	0.0000	0.12	Weak
place_of_birth	126.723	0.0000	0.07	Negligible

family_income	62.190	0.0000	0.05	Negligible

TABLE V. ATTRIBUTES IN FINAL DATASET (AFTER PRE-PROCESSING ACTIVITY)

Attribute	**Data Type**	**Details / Data**
age	nominal	Age group during enrolment in bachelor’s degree Programme (< =20, >= 21)
gender	nominal	Gender (Male, Female)
marital_status	nominal	Marital Status (Single, Married, Divorced/Widowhood)
institution	nominal	Public Universities
programmes	nominal	Student's bachelor's degree programmes group (Engineering, Humanities, Art, Health, Journalism & Information, Computing, Mathematics & Statistics, Manifacturing & Processing, Education, Security Service, Transport Service, Social Service, Environmental Protection, Business & Administration, Agriculture, Forestry & Fisheries, Sains – Broad Programme, Physical Science, Life Science, Social Science, Social & Behavioural Science, Architecture & Building, Law, Veterinary)
study_mode	nominal	Student's mode of study (Part Time, Full Time)
sponsorship	nominal	Student's education financing (Government Agencies, Private Institutions, Self Funding, Education Loan, Foundation, Others)
qualification	nominal	Student's qualification for bachelor's degree admission (Pre-universiy, Diploma, Matriculation, SPM, STPM, Others)
cgpa_t1	nominal	Students’s first year CGPA (Below 2.00, 2-2.99, 3-3.49, 3.5-4.00)
class	nominal	Class label (C, G)

TABLE VI. STATISTICAL DATA FOR ATTRIBUTE CGPA_T1, STUDY_MODE, SPONSORSHIP, QUALIFICATION, MARITAL STATUS, GENDER, AND AGE

Attribute	**Data**	**Sum**	**% (Sum)**	**Class**		**% C**
				C	**G**	
cgpa_t1	Below 2.00	3778	13.10	**3734**	44	**98.84**
	2-2.99	9573	33.19	1255	8318	13.11
	3-3.49	**11406**	**39.54**	429	10977	3.76
	3.5-4.00	4087	14.17	127	3960	3.11
study_mode	Part Time	566	1.96	563	3	**99.47**
	Full Time	**28278**	**98.04**	**4982**	23296	17.62
sponsorship	Government Agencies	1080	3.74	305	775	28.24
	Private Institutions	21	0.07	0	21	0.00
	Self Funding	**18348**	**63.61**	**3348**	15000	18.25
	Education Loan	7735	26.82	954	6781	12.33
	Foundation	24	0.08	1	23	4.17
	Others	1636	5.67	937	699	**57.27**
qualification	SPM	118	0.41	114	4	**96.61**
	STPM	5930	20.56	1320	4610	22.26
	Matriculation	3998	13.86	1449	2549	36.24
	Pre-university	502	1.74	161	341	32.07
	Diploma	**18004**	**62.42**	**2420**	15584	13.44
	Others	292	1.01	81	211	27.74
marital_status	Single	**28592**	**99.13**	**5331**	23261	18.65
	Married	238	0.83	203	35	**85.29**
	Divorced/Widowhood	14	0.05	11	3	78.57
gender	Male	8448	29.29	2362	6086	**27.96**
	Female	**20396**	**70.71**	**3183**	17213	15.61
age	< = 20	12348	42.81	**3060**	9288	**24.78**
	> = 21	**16496**	**57.19**	2485	14011	15.06

TABLE VII. NUMBER OF DROPPED OUT AND GRADUATED STUDENTS ALONG WITH DROP OUT PERCENTAGE BY PROGRAMME GROUPS

Programme Groups	**Class**		**Sum**	**Drop Out Percentage**
	C	**G**		
Sains - Broad Programme	34	0	34	100.00
Veterinary	10	0	10	100.00
Environmental Protection	2	0	2	100.00
Transport Service	1	0	1	100.00
Social Service	90	19	109	82.57
Agriculture, Forestry & Fishery	61	27	88	69.32
Security Service	53	25	78	67.95
Life Science	116	57	173	67.05
Education	67	79	146	45.89
Law	63	104	167	37.72
Architecture & Buliding	374	633	1007	37.14
Engineering	**1313**	2663	3976	33.02
Journalism & Information	230	843	1073	21.44
Personal Service	156	609	765	20.39
Social & Behavioural Science	261	1029	1290	20.23
Computing	364	1440	1804	20.18
Art	315	1418	1733	18.18
Health	117	564	681	17.18
Physical Science	168	972	1140	14.74
Manifacturing & Processing	169	1203	1372	12.32
Humanities	151	1086	1237	12.21
Business & Administration	1303	**9506**	10809	12.05
Mathematics & Statistics	127	1022	1149	11.05

Majority of B40 students managed to obtain a CGPA higher than 2.00 in the first year of their study. Nearly 40% of the students obtained their first year CGPA between 3.00 – 3.49. Even though students with first CGPA lower than 2.00 percentage is the lowest (13%), this group is most likely to drop-out as almost all of the students (98.84 %) did not continue their study. Business and Administration programme group contributed the largest number of students (10,809 students followed by Engineering (3,976 students) and the lowest was Transport Service with only one student. Further analysis also found that six out of ten programme groups with most numbers of students that dropped out and obtained CGPA lower than 2.00 were from Science, Technology, Engineering and Mathematics (STEM) major (Engineering, Computing, Manifacuturing and Processing, Mathematics and Statistics, Physical Science and Architecture).

Almost two-third of B40 students (63.61%) self-funded their study while the balance (36.39%) used education loan or received financial aid from government agencies, a private institution, foundation or other sources. Self-funded students were also presumed to drop-out as 60% of the students quit their study. Being a part-time student also can be a disadvantage, as 99.47% of them failed to graduate. Students who were single got a higher chance of finishing their degree as their percentage of dropping out was very low as compared to married or divorced/widowed students. When it comes to gender, over 70% of students in this study were female, but their drop out rate is 15.61% lower than male students. Finally, students in the age group o 20 years old and below when enrolled in a bachelor degree programme are most likely to drop-out than students in the age group 21 years and above.

C. *Modelling and Evaluation*

1) Model construction and testing: Each prediction model (DT, RF and ANN) was tested beforehand to determine the validation method and algorithms parameters that can be used to produce high performance prediction model. All nine attributes were used in validation method testing and parameter tuning. Prediction model validation was tested using holdout (70 %– 30% and 60% – 40%) and 10-folds cross-validation methods, and the latter was chosen as it gave the highest accuracy results for the majority of the models. Next, each prediction model also was constructed repeatedly by using a different parameter to achieve highest accuracy result. Parameter tuning results are as shown in Table VIII, and these parameters were used in building the final prediction models.

TABLE VIII. PARAMETER TUNING RESULTS

Prediction Models	Parameter
Decisin Tree (DT)	*criterion*: *information_gain* *maximal_depth*: 30
Randon Forest (RF)	*number_of _tree*: 90 *criterion: information_gain* *maximum_depth*: 50
Artificial Neural Network (ANN)	*Hidden layer*: 8 *Learning rate*: 0.01

Different numbers of attributes were used in building final prediction models. At first, the models were build with attributes that had moderate to strong relationship with class label. Subsequently, attributes with weak relationship were added one-by-one based on attribute ranking. The importance of weak attributes can't be neglected as they might be useful in producing high performance prediction models. Table IX shows attribute representation for final models construction.

2) Model evaluation: Prediction model's performance was evaluated by comparing the value of accuracy, precision, recall and F measure. Those values were calculated based on the confusion matrix technique, as shown in Table X. Prediction results and actual class were put in a matrix for comparison depending on a positive and negative value. Class 'C' was marked as positive value while class 'G' was negative.

In addition to performance comparison, the statistical test was performed to decide the best prediction model. This study used the McNemar test to determine if there was a significant difference statistically to the proportion of error between two prediction models with a significance level of 0.05 (α = 0.05). The significant difference between the proportion of error of two prediction models is also interpreted as a significant difference in performance between two prediction models (Dietterich, 1998).

TABLE IX. ATTRIBUTE REPRESENTATION FOR FINAL MODELS CONSTRUCTION

Attribute Representation	**Attributes**
6 Attributes	*cgpa_t1*, *institution*, *study_mode*, *programme*, *sponsorship*, *qualification*
7 Attributes	*cgpa_t1*, *institution*, *study_mode*, *programme*, *sponsorship*, *qualification*, *marital_status*
8 Attributes	*cgpa_t1*, *institution*, *study_mode*, *programme*, *sponsorship*, *qualification*, *marital_status*, *gender*
9 Attributes	*cgpa_t1*, *institution*, *study_mode*, *programme*, *sponsorship*, *qualification*, *marital_status*, *gender*, *age*

TABLE X. CONFUSION MATRIX

		Prediction Class	
		Positive (C)	**Negative (G)**
Actual Class	**Positive (C)**	True Positive (TP)	True Negative (TN)
	Negative (G)	False Negative (NP)	True Negative (TN)

IV. RESULTS AND DISCUSSION

The results indicated that RF model gives the highest accuracy in predicting student drop-out with 95.93%, followed by ANN with 95.86% and DT with 95.84%. The highest accuracy for RF model was produced with seven attributes while the others by using six attributes. However, the accuracy for RF model with six attributes was higher than ANN and DT models with the same number of attributes (refer Fig. 2). Consistently, RF also yields a higher accuracy rate than the other two models, even by applying different numbers of attributes. This showed that prediction performance could be improved with the use of ensemble learning. This result is also inline with research outcome by [20], which predicts students' drop-out in higher learning institution, revealing that the accuracy of the prediction model using RF l was higher than DT.

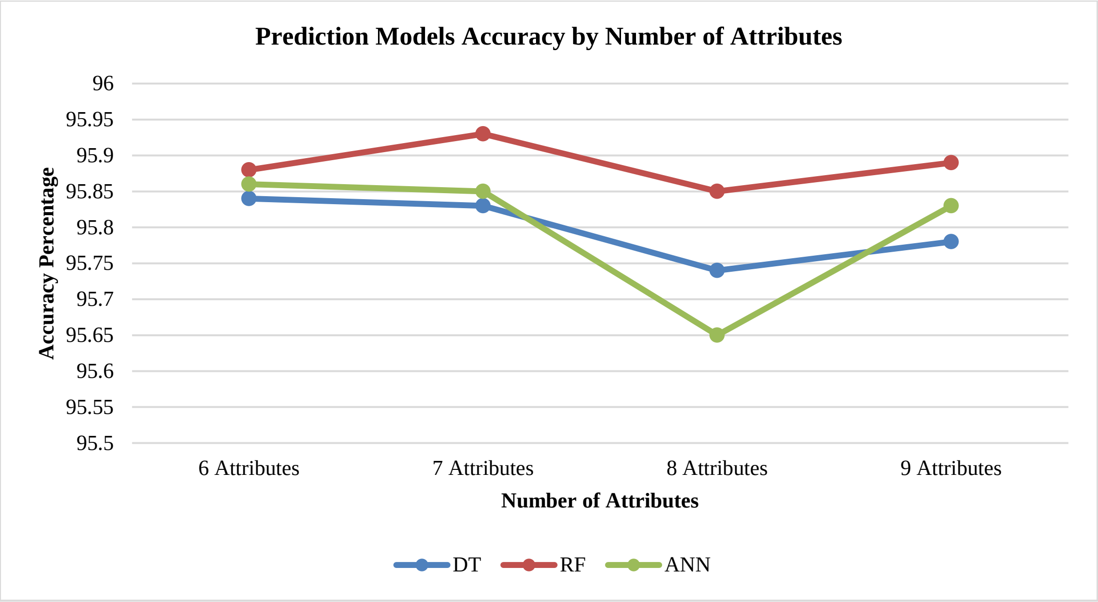

Fig. 2. Prediction Model Accuracy based on Number of Attributes.

TABLE XI. PERFORMANCE RESULTS COMPARISON BETWEEN STUDENTS DROP-OUT PREDICTION MODELS WITH HIGHEST ACCURACY

Prediction Models	**Evaluation Parameters**			
	Accuracy (%)	**Recall (%)**	**Precision (%)**	**F Measure**
DT (6 attributes)	95.84	80.99	96.83	0.882
RF (7 attributes)	**95.93**	**81.26**	**97.10**	**0.885**
ANN (6 attributes)	95.86	81.08	96.89	0.883

Performance between prediction models was evaluated by comparing the value of accuracy, recall, precision and F measure (refer Table XI). Aside from accuracy, the results also showed that RF leads ANN and DT with regards to recall value with 81.26%, 81.03% and 80.99%. This means that RF model succeeded in predicting more students who will drop-out (class 'C') correctly from the total number of student who were actually dropped out in this study. Likewise, the highest value for precision was also recorded by RF with 97.10% which means that the model was able to predict more class 'C' precisely from the total number of students who were predicted to drop-out. ANN took second place in precision with 96.89% while DT the last place with 96.83%. When comparing F measure, RF also the highest with 0.885, followed by ANN with 0.883 and DT with 0.882.

Generally, RF is the best model in predicting drop-out among B40 students in this study as it outperformed the other two models with reference to accuracy, recall, precision and F measure, subsequently ANN and DT models. Nevertheless, the difference in accuracy and F measure value between the three models were very narrow, with 0.07% to 0.09% and 0.002% to 0.003%, respectively. Hence, the statistical test (McNemar) results were referred to in determining a significant difference in performance between the prediction models.

Based on Table XII, McNemar test results proved that statistically:

1) DT and RF models had a significant difference in proportion of error;

2) DT and ANN models had no significant difference in proportion of error; and

3) ANN and RF models had no significant difference in proportion of error.

This implied that even though RF is the best model in predicting drop-out among B40 students in this study, there is a significant difference in performance only between RF and DT, but contrarily, no significant difference in performance between RF and ANN.

TABLE XII. MCNEMAR'S *P*-VALUE (*2 SIDED*) RESULTS

Prediction Models	**DT**	**RF**	**ANN**
DT		**0.004**	0.719
RF	**0.004**		0.224
ANN	0.719	0.224	

V. CONCLUSION

Drop-out prediction among B40 students in bachelor's degree programmes can be implemented by using classification technique. Prediction model using RF was selected as the best model in this study as it outperformed ANN and DT in accuracy, recall, precision and F measure. However, statistically, the difference in performance was only significant between RF and DT, not between RF and ANN.

Results of this research are expected to benefit B40 students, public universities and the government. Early prevention steps can be deployed by public universities to avoid drop-out to produce more graduates. B40 students who are at risk to drop-out will be able to graduate with the help of their university and getting better job opportunities that will improve their socioeconomic status. These students also will become assets to the government as professionals and skilful worker that can be contributed to the nation's future development.

In future, this study can be furthered by applying regression technique to predict when attrition will happen with the additional data of students who are still studying and the exact date of drop-out. Besides, the association rule technique can be applied to discover hidden patterns that can be used to identify students at risk, and the results can be verified by the experts from the ministry or universities.

ACKNOWLEDGMENT

The authors would like to thank Universiti Kebangsaan Malaysia (UKM) and Ministry of Education, Malaysia (MOE) under the Research University Grant (project code: GUP-2019-060 and FRGS/1/2018/ICT02/UKM/02/6) for funding and supporting this research.

REFERENCES

[1] Abu, R. Hamdan, R. and N.S. Sani, "Ensemble Learning for Multidimensional Poverty Classification," Sains Malaysiana., vol. 49(2), pp.447-459 2020.

[2] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. M. Sarim, "Machine learning approach for bottom 40 percent households (B40) poverty classification," IJASEIT, vol. 8, pp. 1698-1705, 2018.

[3] J. D. Holliday, N. Sani, and P. Willett, "Calculation of substructural analysis weights using a genetic algorithm," J. Chem. Inf. Model, vol. 55, pp. 214-221, 2015

[4] J. D. Holliday, N. Sani, and P. Willett, "Ligand-based virtual screening using a genetic algorithm with data fusion," Match-Commun. Math. Co., vol. 80, pp. 623-638, 2018.

[5] N. S. Sani, I. I. S. Shamsuddin, S. Sahran, A. H. A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection," IJASEIT, vol. 8, pp. 1486-1493, 2018.

[6] S. Shabudin, N. S. Sani, K. A. Z. Ariffin and M. Aliff, "Feature Selection for Phishing Website Classification," International Journal of Advanced Computer Science and Applications, vol. 11(4), pp. 587-595, 2020.

[7] T. K. M. Zali, N. S. Sani, A. H. Abd Rahman, and M. Aliff, "Attractiveness Analysis of Quiz Games," International Journal of Advanced Computer Science and Applications, vol. 10(8), pp. 205-210, 2019.

[8] Z. A. Othman, A. A. Bakar, N. S. Sani, and J. Sallim, "Household Overspending Model Amongst B40, M40 and T20 using Classification Algorithm," International Journal of Advanced Computer Science and Applications, vo1. 11(7), pp. 392-399, 2019.

[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," Procedia Comput. Sci., vol. 72, pp. 414–422, January 2015.

[10] M. A. Al-Barrak and M. Al-Razgan, "Predicting students final GPA using decision trees: A case study," Int. J. Inf. Educ. Technol., vol. 6, pp. 528–533, July 2016.

[11] E. A. Amrieh, T. Hamtini and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," 2015 IEEE Conf. Appl. Electr. Eng. Comput.Technol. (AEECT), Amman, Jordan, pp. 1–5. November 2015.

[12] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, October 2017.

[13] F. Widyahastuti and V. U. Tjhin, "Predicting students performance in final examination using linear regression and multilayer perceptron," 2017 10th Int. Conf. Human Syst. Interact. (HSI), pp. 188–192, July 2017.

[14] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "Predicting student's academic performance in a MOOC environment," 11th Int. Conf. Data Mining, Comput., Commun. Ind. Appl. (DMCCIA-2017), pp. 119–124, December 2017. [Umer, R., Science, M., Susnjak, T., Mathrani, A., Science, M., & Suriadi, S].

[15] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison, "Who, when and why: A machine learning approach to prioritizing students at risk of not graduating high school on time," Proc. 5th Int. Conf. Learning Anal. Knowl., New York, pp. 93–102, March 2015.

[16] W. W. Yaacob, N. M. Sobri, S. M. Nasir, N. D. Norshahidi, and W. W. Husin, "Predicting student drop-out in higher institution using data mining techniques," J. Physics: Conf. Series 2020, vol. 1496, 012005, March 2020.

[17] N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama, and P. Yanque-Churo, "Classification models for determining types of academic risk and predicting drop-out in university students, Int. J. Adv. Comput. Sci. Appl., vol. 11, pp. 266–272, 2020.

[18] J. S. Gil, A. J. P. Delima, and R. N. Vilchez, "Predicting students' drop-out indicators in public school using data mining approaches." Int. J. Adv. Trends in Computer Sci. Eng., vol. 9, pp. 774–778, 2020.

[19] M. Mardolkar and N. Kumaran, "Forecasting and avoiding student drop-out using the K-nearest neighbor approach," SN Computer Sci., vol. 1, pp. 1–8, March 2020.

[20] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S.Ventura, "Early drop-out prediction using data mining: A case study with high school students," Expert Systems, vol. 33, pp. 107–124. February 2016.

[21] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," Comput. Educ., vol. 143, p. 103676, January 2020.

[22] A. Viloria, J. G. Padilla, C. Vargas-Mercado, H. Hernández-Palma, N. O. Llinas, and M. A. David, "Integration of data technology for analyzing university drop-out," Procedia Comput. Sci., vol. 155, pp. 569–574, January 2019.

[23] A. Sangodiah, P. Beleya, M. Muniandy, L. E. Heng, and C. Ramendran, "Minimizing student attrition in higher learning institutions in Malaysia using support vector machine," J. Theoritical Appl. Inf. Technol., vol. 71, pp. 377–385, January 2015.

[24] S. S. Shariff, N. A. Rodzi, K. A. Rahman, S. M. Zahari, and S. M. Deni, "Predicting the "graduate on time (GOT)" of PhD students using binary logistics regression model," AIP Conf. Proc. 2016, vol. 1782, p. 050015, October 2016.

[25] K. Limsathitwong, K. Tiwatthanont, and T. Yatsungnoen, "Drop-out prediction system to reduce discontinue study rate of information technology students," Proc. 2018 5th Int. Conf. Business and Industrial Research: Smart Technol. Next Generation of Information, Eng., Business and Social Sci. (ICBIR 2018), pp. 110–114, May 2018.

[26] Y. Chen, A. Johri, and H. Rangwala, "Running out of STEM: A comparative study across STEM majors of college students at-risk of dropping out early," Proc. 8th Int. Conf. Learn. Anal. Knowl., pp. 270–279, March 2018.

[27] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, "University student retention: Best time and data to identify undergraduate students at risk of drop-out," Innovations Educ. Teach. Int., vol. 57, 74–85, January 2020.

[28] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student drop-out in higher education," arXiv preprint arXiv:1606.06364., June 2016.

[29] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Drop-out prediction in edx MOOCs," Proc. - 2016 IEEE 2nd Int. Conf. Multimedia Big Data, pp. 440–443, April. 2016.

[30] D. D. Pokrajac, K. R. Sudler, P. Y. Edamatsu, and T. Hardee, "Prediction of retention at historically black college/university using artificial neural networks," 2016 13th Symp. Neural Networks and Applications (NEUREL), pp. 1–6, November 2016.

[31] G. S. Abu-Oda, and A. M. El-Halees, "Data mining in higher education: University student drop-out case study," Int. J. Data Mining & Knowl. Manage. Proc., vol. 5(1), pp. 15–27, January 2015.

[32] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, "A comparative study of classification and regression algorithms for modelling students' academic performance," Proc. 8th Int. Conf. Educ. Data Mining, 392–395, January 2015.

[33] A. Siri, "Predicting drop-out at university using Artificial Neural Networks," Italian J. Soc. Educ., vol. 7, pp. 225–247, June 2015.