Malaysian Journal of Social Sciences and Humanities (MJSSH) (e-ISSN : 2504-8562)
2024, Volume 9, Issue 6, e002867
DOI: https://doi.org/10.47405/mjssh.v9i6.2867

RESEARCH ARTICLE

Prediction Model of School Drop Out Factors Using Classification Techniques in Selangor

Siti Rafidah binti Sariman[1*], Habibah binti Ab Jalil[2], Erzam bin Marlisah[3]

[1]Fakulti Pengajian Pendidikan, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.
Email: rafidah.sariman@gmail.com
[2]Fakulti Pengajian Pendidikan, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.
Email: habibahjalil@upm.edu.my
[3]Fakulti Sains Komputer Dan Teknologi Maklumat, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.
Email: erzam@upm.edu.my

CORRESPONDING AUTHOR (*):
Siti Rafidah binti Sariman
(rafidah.sariman@gmail.com)

KEYWORDS:
Predicting drop out
Classification
Data mining

CITATION:
Siti Rafidah Sariman, Habibah Ab Jalil, & Erzam Marlisah. (2024). Prediction Model of School Drop Out Factors Using Classification Techniques in Selangor. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 9*(6), e002867. https://doi.org/10.47405/mjssh.v9i6.2867

ABSTRACT

Malaysia has been struggling to sustain the number of students graduating from school, with an increasing number leaving early, which poses significant concerns for the future generation. Several factors contribute to this issue, such as economic constraints, geographical challenges, transportation problems, and sociocultural norms. This paper uses a data mining approach to identify the attributes that lead to school dropouts and to determine the predictive model with the highest accuracy for forecasting dropout rates. The application of data mining approaches has proven effective in predicting students at risk of dropping out in general education. Nevertheless, there is a shortage of data mining-related studies on student attrition in public schools in Malaysia. The study utilized student and school datasets with consent from the respective departments in the Ministry of Education. This data includes information from 2,482 students across various primary schools in Selangor with initially 22 attributes collected from the dataset. After the attributes undergone feature selection process by using InfoGainAttributeEval, there are 12 features left including class attribute Status_DO. The collected data encompasses student demographics, academic performance and socioeconomic background. The experiments for this study used Decision Trees (J48), Naïve Bayes and Random Forest. By using classification techniques that were made available in WEKA, all attributes from the dataset were tested. The results of the analysis shown that Random Forest with the highest accuracy of 79.5729% in term of predicting student drop out hence indicate the reliability of this research as a decision support tool.

Contribution/Originality: The results of this study enable educators and policymakers to identify the factors that lead to school dropout and to implement targeted interventions. The model's accuracy makes it a useful tool for making informed decisions, leading to more effective education management.

1. Introduction

School dropout, often viewed as a critical challenge within educational systems worldwide, refers to the phenomenon where students disengage prematurely from formal schooling before completing their intended educational program. This issue carries significant implications, not only for individual students but also for communities and societies at large. High dropout rates can hinder economic development, perpetuate cycles of poverty, and limit individuals' opportunities for personal and professional advancement. Understanding the complex factors contributing to dropout and implementing targeted interventions is essential for fostering inclusive, equitable, and effective education systems. This issue affects not just students but also whole communities. By understanding why students drop out, a better and fairer education system can be built to find ways to help them graduate and finish their study.

According to McFarland et al. (2019), in The National Center for Education Statistics (NCES) report, a dropout is defined as someone who was enrolled in school at any point during the previous academic year but is not enrolled at the start of the current school year, has not graduated from high school or completed an education program, and does not meet any exclusionary conditions such as transferring to another school, temporary absence due to suspension or school-approved illness, or death. Meanwhile, UNICEF (2016) emphasizes the importance of defining school dropout rates with clear criteria, considering factors like compulsory school age, dates for enrollment and dropout data, types of excused absenteeism inclusion/exclusion criteria for educational programs and conditions for excluding students. These criteria ensure accurate measurement, helping stakeholders identify at-risk students, assess dropout challenges, and implement targeted interventions to improve educational outcomes globally.

Dupere et al. (2015) in his research, refers school dropout as a process by which students disengage from the educational system before obtaining a diploma or credential. However, Rumberger and Lim (2008) defined school dropout as the termination of educational enrollment before completing a level of education, typically high school or its equivalent, without obtaining a diploma or credential. On the other hand, White and Kelly (2010) refer dropping out of school as the final outcome in a sequence of disengagement from school, driven by adverse influences (risk factors) and insufficient support systems (protective factors).

Detecting students at risk of dropping out of school early is vital for several reasons. Firstly, it enables educators and policymakers to intervene promptly, providing tailored support to prevent dropouts. By identifying and assisting these students, schools can enhance overall academic achievement and graduation rates. Moreover, preventing dropout reduces long-term economic costs associated with unemployment and limited career opportunities for individuals and society. Additionally, addressing dropout risk contributes to social equity by ensuring equal access to educational opportunities for all students. Finally, data-informed decision-making allows schools and districts to allocate resources effectively and implement targeted interventions to address underlying factors contributing to dropout.

The success of any education system in their school retention programs is determined in identifying those targeted students in earlier phase. It is necessary to detect this high-risk group of students earlier to tackle the future issues thus to provide designated

prevention strategy so that the students stay in school until they graduated (Heppen & Therriault, 2008).

In recent years, advancements in machine learning algorithms have played a significant role in enhancing the accuracy and effectiveness of data mining projects. Specifically, within the field of educational data mining, these algorithms offer valuable tools for educational institutions to identify key factors influencing dropout rates. By leveraging machine learning techniques, educational researchers and administrators can analyze large datasets to uncover patterns and predictors of student attrition, enabling proactive interventions to support at-risk students and improve overall retention rates.

This study aims to identify the primary factors contributing to student dropout rates in public schools in Selangor. Additionally, it seeks to assess the efficacy of various data mining techniques in developing accurate prediction models for this purpose. The outcome of this research will be a predictive model designed to assist educational institutions in implementing effective student retention programs tailored to their specific needs.

Therefore, it's crucial to take a dynamic approach to pinpoint students at high risk of dropping out. An accurate and effective model can be built using machine learning techniques and to analyze the datasets. The decision tree (J48), random forest (RF), and Naïve Bayes classification techniques were employed in constructing the models. Regardless with many prediction models and findings on the factors that effects the dropout rate at schools in numerous literatures, there is limited research related to student's dropout especially in public schools in Malaysia.

This study will be divided into 3 sections. In the first section we examined previous research articles related to classification techniques in machine learning to predict school dropouts in public schools. Next, we will outline the methodology employed in predicting school dropouts and the results of the prediction model. Lastly, section 3 delves into the presentation and discussion of the results and suggesting avenues for future research.

1.1. Research Objectives

The objective of this paper is to develop a prediction model and identify the factors contributing to school dropout rates in public schools in Selangor using classification techniques.

2. Literature Review

The earliest prediction model of school dropout was pioneered by Tinto (1975) which identified the causes of student dropout. The results showed that demographics, cultural, family background, socioeconomic status, academic and psychological profile were the characteristics that can influence students' decision to drop out of schools (Nicoletti, 2019). The model suggested that the major determinant of student's completion in their study was caused by the student's social and academic integration. Other integrations were also found as the key factors in the dropout model such as family background, personality, previous schooling, academic performances and the interaction between students and the faculty.

Over the past few years, there has been a notable increase in research dedicated to predicting student performance, specifically concentrating on course drop-out and retention through the application of classification techniques in supervised learning. Al-Radaideh et al. (2006) conducted an analysis of student academic data, which included factors such as student gender, age, department, high school grade, lecturer degree, lecturer gender, among others. They employed the decision tree method to construct a classification model aimed at improving the quality of the higher education system. The study found that high school grades showed the highest gain ratio and were identified as the primary node in the decision tree.

Marquez-Vera et al. (2016) proposed a method and a classification algorithm to unveil distinct and comprehensible models for predicting student dropout at the earliest possible stage. They conducted various experiments to forecast dropouts at different points in the course, identify the most effective dropout indicators, and compare the suggested algorithm with several widely recognized classical and unbalanced classification algorithms. Results showed that the algorithm was able to accurately predict student dropouts during the first four-six weeks of the program and was efficient enough to be used in the early warning system.

Viloria et al. (2019) applied neural networks (NN), decision trees (DT), and Bayesian networks (BN) to predict student dropout in India. The results revealed that both academic performance and socioeconomic status significantly influence students to leave school before graduation. The findings suggest that effectively managing these variables could lead to a reduction in the dropout rate.

Gil, Delima, and Vilchez (2020) utilized Decision Trees (DT) and Naive Bayes (NB) techniques to discern the key factors that contribute to student drop-out in a Philippines public school. They employed the Weka toolkit to implement the classifier algorithm on the dataset, and enabled them to conduct a comprehensive comparative analysis, evaluating the performance of each algorithm based on metrics such as recall, precision, and accuracy.

Meanwhile Mardolkar and Kumaran (2020) employed data mining techniques to develop thorough predictive models for student drop-out at the earliest stage. The accurate model, with high predictive accuracy, is intended for use in an early warning system to promptly identify students at a heightened risk of drop-out. Their exploration encompassed academic variables at universities and schools, sociodemographic factors, behavior, and extracurricular activities that could impact student dropout. However, only a subset of attributes with exceptionally high predictive significance was considered.

Roslan (2021) in their study utilized two popular classifiers, decision tree and logistic regression, to predict student dropout. It analyzed 7706 student records from a Malaysian private university's database from 2018 to 2019. Classifier performance was evaluated using accuracy and misclassification rates. Results showed that the decision tree with chi-square (2 branches) achieved the highest classification accuracy of 89.49% with an 80/20 data partition.

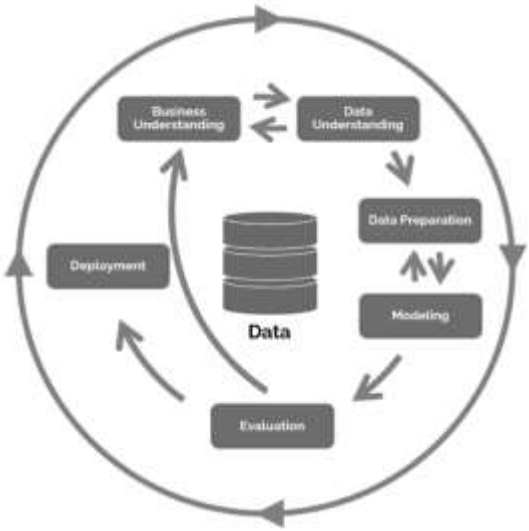
In the prediction of student dropouts, the process of selecting variables stands out as one of the most crucial stages since variables constitute the fundamental constructs in a study. As indicated in the preceding research, the factors influencing dropout rates can

vary from one country to another. According to Vijayakumaran et al. (2023) student dropout intention is influenced by parental involvement and student engagement.

3. Research Methods

CRISP-DM methodology is used as a basis of data mining process to produce the prediction model of student dropout. Figure 1 presents 6 phases from CRISP-DM that will be performed in this research. The phases are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. The phases were described in the subsections below. The prediction model will be constructed using WEKA as well as to perform the performance evaluation and for features selection process.

Figure 1: CRoss Industry Standard Process for Data Mining (CRISP-DM) Framework



Source: Provost & Fawcett (2013)

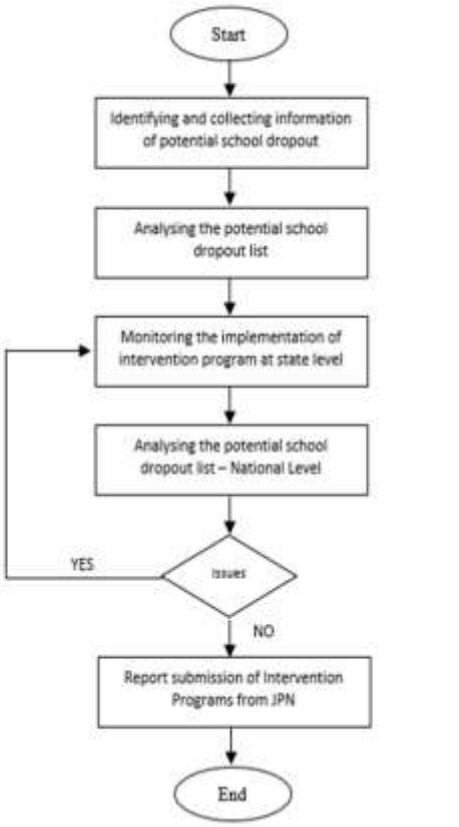
3.1. Business Understanding

The study will start with the understanding of the problems and goals and transform it into data mining problem definitions. Once the goals are constructed then the preliminary plan is designed to achieve the objectives. The objectives were ensembled from the literature reviews from previous research and related to the education scenarios exclusively in Malaysia. In the "Guidelines for Managing Students at Risk of Dropping Out" published by the Ministry of Education (2018), there are several steps that were needed to identify and process school dropouts.

Figure 2 illustrate the process of managing and identifying potential school dropouts. First, the process started with collecting information from the students' database manually. This method is time-consuming and demanding due to the extensive data source and large number of students involved. Then, state and district officers will need to analyze the data and oversee the intervention programs accordingly. The process will be done using a spreadsheet application such as Microsoft Excel in which the accuracy of the data and the reliability of the results may be uncertain. The analysis uncovered uncertainty surrounding the factors influencing potential school dropout, leading to a

varied and ambiguous list. This uncertainty may impact the efficacy of intervention programs, potentially rendering them less suitable for targeted groups of students.

Figure 2: Procedures for handling potential school dropout



Source : Ministry of Education (2018)

3.1.1. Identifying and collecting information of potential school dropout

State and district education officers will access student data from the database provided by the Ministry. The data will be presented in Excel format and will include a range of variables and information about the students, including personal details, parental information, and academic performance.

3.1.2. Analyzing the potential school dropout list

The data will be carefully sorted based on how likely students are to drop out of school. This sorting will involve looking at factors like students' grades, attendance, family backgrounds, and behavior. The possibility was counted manually using spreadsheet according to the variables in the data set. The data will be given to schools to proceed with intervention programs.

3.1.3. Monitoring the implementation of intervention program at state level

Schools will receive a list of students who might be at risk of dropping out, and they'll develop intervention programs accordingly. These programs will be carried out by the schools, with reports prepared by counselors. However, the programs are general and not tailored to specific groups of students. There hasn't been any further research on how effective these intervention programs are. Ideally, the programs should target specific groups of students based on the factors contributing to their risk of dropping

out. State officers will oversee and monitor the programs to ensure they receive positive feedback and make a meaningful impact on the students.

3.1.4. Report submission and intervention programs

A national-level meeting involving officials from the Ministry and states will be held to discuss issues related to potential and high-risk school dropout rates. Intervention programs will be implemented at the school level by counselors and overseen by district officers to ensure effective monitoring.

3.2. Data understanding and data preparation

The dataset used in this study was obtained from the students' and schools' dataset with the consent of the respective departments in the Ministry of Education. The first dataset is collected from the Education Repository, *Aplikasi Pangkalan Data Murid* (APDM) an online-based platform that stores demographic information's of the students. The second dataset is collected from *Sistem Aplikasi Peperiksaan Sekolah* (SAPS) to obtain students' grades in examinations.

The dataset consists of 2482 students' data from various primary schools in Selangor. Only data from primary schools that were registered under the MOE was selected. From the dataset, the statistical data for dropout students is 1043 (42%) while 1439 (58%) students have completed their study from 411 schools. Therefore, the number of dropout students is slightly lower compared to the number of students that already completed their study. Not all data is taken from every school in Selangor. Only data that has a complete set of attributes were used in the process of data analysis.

Pre-processing of the data is a method to clean and transform the data into meaningful data set for the data mining tool to perform and to produce a high-quality model. Both students and school's data were integrated into a single data set. Then, the dataset was cleaned using the dimension reduction process. The initial dataset consists of 8651 rows and after the cleaning process, only 2482 rows of data were made available for the next phase. The incomplete, redundant, or obsolete records were discarded from the dataset. Next the data were transformed into an understandable format to fit the algorithms that were available in WEKA. Table 1 shows 22 attributes in the initial dataset. A feature selection was implemented to select relevant attributes to build the prediction model. In WEKA, several attribute selection techniques were chosen to test the association between attributes and class label.

Feature selection is also known as part of the process of selecting a subset of the relevant features in order to create an accurate prediction model. To find an accurate data model, researchers need to apply feature selection as one of the useful approaches in data pre-processing (Mduma et al., 2019). The process involving identifying and eliminating unnecessary, irrelevant and redundant attributes from the data set that might reduce the accuracy of the models (Musiliu, 2020). The complexity of a model can be reduced by using fewer attributes thus making it simpler and easier to understand and will not affect the learning performance.

A high accuracy prediction model can be generated only with relevant attributes. Eliminating the unwanted attributes is an important step to develop a good machine learning model. Therefore, the dataset will undergo attributes selection process using

WEKA automated tools in feature selection. The feature selection in WEKA is divided into two parts which are Attribute Evaluator and Search Method. The attribute evaluator is a technique whereby each attribute or column in the dataset is evaluated in the context of the class label. While the search method is the technique of using multiple combination of attributes in the dataset to select a short list of chosen features. The process was separated into two tasks. The first task is Attribute Evaluator, a method by which attribute subsets are assessed. The second task is Search Method, a method by which the space of possible subsets is searched. InfoGainAttributeEval was used in WEKA for Feature Selection process.

After the feature selection was applied on the dataset, there are only 12 attributes that were selected to be used in the data mining analysis. The initial attributes from the dataset before feature selection were listed in Table 1.

Table 1: Initial attributes from the dataset

No.	Variables	Description	Data Type
1	IDMurid	Students' ID	Nominal
2	KodJantina	Students gender	Nominal
3	KeteranganKaum	Students races	Nominal
4	StatusWarganegara	Students citizenship	Nominal
5	Yatim	Students' family status	Nominal
6	StatusOKU	Students disability status	Nominal
7	JenisOKU	Type of disability	Nominal
8	Asrama	Students' living in hostel status	Nominal
9	PekerjaanPenjaga1	Parents' employment status	Nominal
10	HubunganPenjaga1	Students' gurdianship status	Nominal
11	JumPendapatan	Parents' monthly income	Number
12	KatPendapatan	Economic status of the family	Nominal
13	Lokasi	School location	Nominal
14	GBMK	Bahasa Malaysia Final Year Exam Result	Nominal
15	GBIK	Bahasa Inggeris Final Year Exam Result	Nominal
16	GSEJR	Sejarah Final Year Exam Result	Nominal
17	IDKodPPD	Schools' district code	Number
18	KodSekolah	School code	Nominal
19	StatusDLP	School status on Dual Language Programme	Nominal
20	KeteranganJenisSekolah	School types	Nominal
21	CapaianInternet	Schools' internet status	Nominal
22	Status_DO	Students' dropout status	Nominal

JumPendapatan and student_status will be reconstructed into new attributes which are KatPendapatan and Status_DO. Status_DO is a class attribute that will determine whether the status of the students was dropout or stay in school. Attribute in student_status with 'Berhenti Sekolah', 'Dibuang Sekolah' and 'Lain-lain' were transformed into 'DO' while records with no information were labelled as 'ST' which represented as students who will manage to graduate from the school. While in JumPendapatan, the initial attributes were data input by the teacher's according to the parents' monthly income. Thus, all the data input will be group into categorical data in line with the household income classification endorsed by Department of Statistics

Malaysia (DOSM). The household income classification with the new attributes label is shown on Table 2.

Table 2: The household income classification

Income classification	Household Group	New attributes
RM 0 – RM 604	B40	A
RM 605 – RM 960		B
RM 961 – RM 3140		C
RM 3141 – RM 4850	M40	D
RM 4851 – RM 10,970		E
> RM10,971	T20	F

To improve the efficiency and accuracy of our predictive model, we removed insignificant parameters from the dataset. Specifically, attributes such as IDMurid, KodSekolah, and IDKodPPD were discarded because their association with the class label (student drop-out) was minimal and could be ignored. In addition to eliminating these non-informative features, we also took steps to ensure compliance with data protection regulations. All private particulars of the students were excluded from the dataset to adhere to the Data Protection Act and to safeguard the privacy of the individuals involved.

After this refinement process, we focused on the remaining 12 variables that showed a strong correlation with student demographics and academic performance. These variables as shown in Table 3 —StatusDLP, KeteranganJenisSekolah, KodJantina, StatusWarganegara, Yatim, StatusOKU, HubunganPenjaga1, KatPendapatan, Lokasi, GBMK, GSEJR, and Status_DO—were identified as significant predictors of student drop-out rates. By concentrating on these key attributes, we can more effectively analyze the factors that influence whether a student is likely to drop out, leading to more targeted interventions and support strategies.

Table 3: Selected attributes after Feature Selection

No.	Variables	Description	Data Type
1	StatusDLP	School status on Dual Language Programme	Nominal
2	KeteranganJenisSekolah	School types	Nominal
3	KodJantina	Students gender	Nominal
4	StatusWarganegara	Students citizenship	Nominal
5	Yatim	Students' family status	Nominal
6	StatusOKU	Students disability status	Nominal
7	HubunganPenjaga1	Students' gurdianship status	Nominal
8	KatPendapatan	Economic status of the family	Nominal
9	Lokasi	School location	Nominal
10	GBMK	Bahasa Malaysia Final Year Exam Result	Nominal
11	GSEJR	Sejarah Final Year Exam Result	Nominal
12	Status_DO	Students dropout status	Nominal

3.3. Modelling and evaluation

All prediction models will be tested using three algorithms in WEKA, which are Naïve Bayes, Random Forest and Decision Tree (J48) to get the highest performance prediction model by identifying suitable validation methods and algorithm parameters to use. All attributes that were tested in the feature selection were used in validation method testing and parameter tuning. Prediction model validation was assessed using holdout (70% - 30% and 60% - 40%) and 10 folds cross validation methods. It was found that the 10-fold cross-validation method provided the highest accuracy results for most models. Subsequently, different parameters will be tested in each prediction model to achieve the highest accuracy.

From the obtained data, there are 1075 female compared to 1407 male students. The largest number of students drop out came from Household Income in category E. Which cover Most of the students came from category C of KatPendapatan which is from B40 Household Income group. 2442 rows of the data are citizen of Malaysia whereas 40 rows of data are not the citizen of Malaysia. In the data, 1043 out of 2482 students drop out of schools which represents 42% of the overall data. Meanwhile, 1439 out of 2482 students were able to finish and graduate from schools which is 53% from the overall data. The fractions of data are almost balanced between the two categories.

Table 4 summarizes dropout (DO) and graduation (ST) rates across different household income groups (KatPendapatan) for both female and male students. Overall, there are 1043 dropouts, with males (637) outnumbering females (406). The highest dropout rate is seen in the E income group, with 339 students (133 females and 206 males), followed by the F income group with 218 dropouts (99 females and 119 males). Conversely, the total number of graduates is 1439, with males (770) again surpassing females (669). The C income group has the highest number of graduates at 616 (269 females and 347 males), followed by the E income group with 389 graduates (188 females and 201 males). This data indicates that lower household income groups (E and F) have higher dropout rates, while the C income group shows the highest graduation rates, suggesting a correlation between household income and educational outcomes.

Table 4: Statistical data of dropout rates according to household income

KatPendapatan	Drop Out (DO)		Total DO	Graduate (ST)		Total ST
	Female	Male		Female	Male	
A	16	29	45	14	19	33
B	7	21	28	16	15	31
C	102	202	304	269	347	*616*
D	49	60	109	82	86	168
E	133	206	*339*	188	201	389
F	99	119	218	100	102	202
Grand Total	406	637	1043	669	770	1439

From the Final Year Exam result, most of the students from the dataset of 2482 or 621 students get grade C from Bahasa Malaysia, 823 of the students obtained grade E in Bahasa Inggeris and 669 students get grade C in Sejarah. This shows that the majority of the students were able to pass all three subjects with the lowest grade E. Surprisingly the number of dropout students with grade A in the listed subjects are quite alarming. The drop out percentage of grade TH or absence students are the highest. Absence students did not attend or took the exam and there was no grading given by the teacher.

The reasons for the absenteeism were unknown. However, grade TH was not seen as the main cause of dropping out since there are multiple variables to be tested in the prediction model.

Meanwhile Table 5 reveals significant variations in dropout rates across different subjects and grades, with the "TH" grade representing student absences and showing the highest dropout rates: 78.46% in Bahasa Malaysia, 74.14% in Bahasa Inggeris, and 59.64% in Sejarah. Generally, lower grades (D and E) exhibit higher dropout rates compared to higher grades (A, B, and C), although an exception is seen in Bahasa Inggeris where grade E has a higher dropout rate (41.31%) than grades A and B. These findings highlight that students with absences ("TH" grade) and those with lower grades are at a greater risk of dropping out, suggesting a need for targeted interventions to support these groups.

Table 5: Statistical data according to subject grade

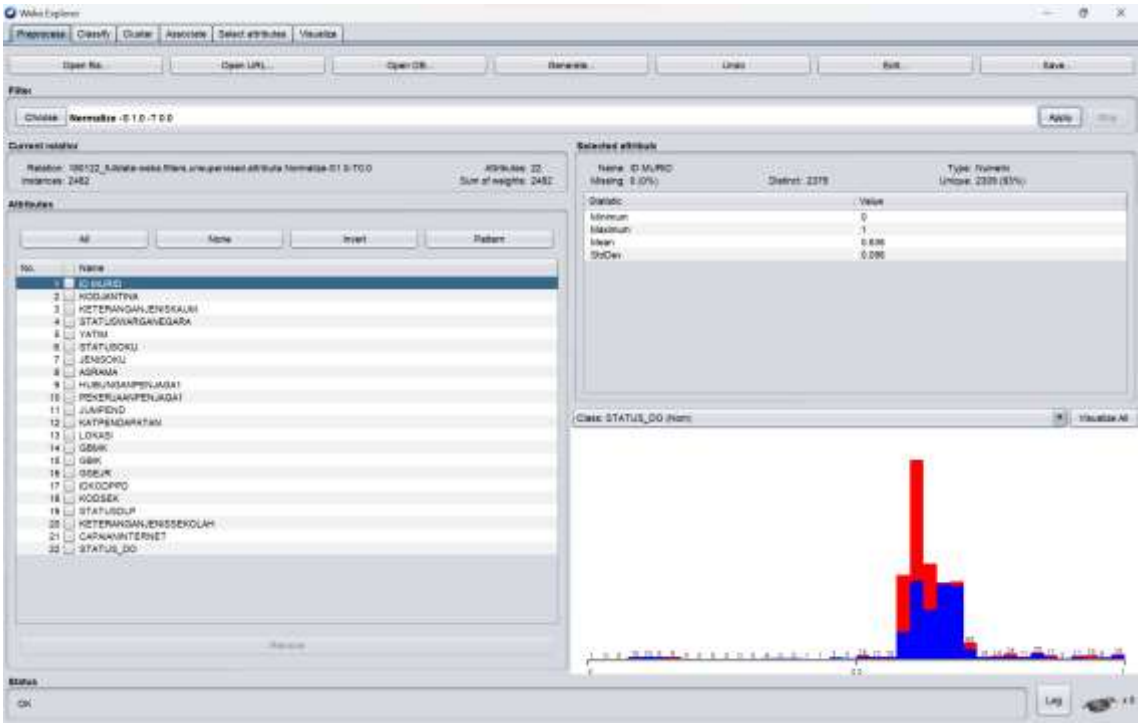
Subject	Grade	Drop Out	Graduate	Total	Drop Out Percentage
Bahasa Malaysia	A	236	212	448	52.68
	B	216	367	583	37.05
	C	221	400	621	35.59
	D	138	216	354	38.98
	E	181	230	411	44.04
	TH	51	14	65	*78.46*
Bahasa Inggeris	A	190	204	394	48.22
	B	162	219	381	42.52
	C	186	295	481	38.67
	D	122	223	345	35.36
	E	340	483	823	41.31
	TH	43	15	58	*74.14*
Sejarah	A	235	228	463	50.76
	B	237	366	603	39.30
	C	247	422	669	36.92
	D	104	209	313	33.23
	E	121	147	268	45.15
	TH	99	67	166	*59.64*

4. Results

In this study, we conduct three experiments using Naïve Bayes, Random Forest and Decision Trees (J48) to identify the factors that contribute to students' attrition. The experiments were conducted using ten-fold cross-validation, a method where the dataset is divided into ten parts. In each iteration, 90% of the data is used for training and 10% for testing. This process is repeated ten times to ensure robust evaluation. The results from these experiments were then presented and compared. The detailed accuracy of the results is presented in the following section.

The dataset was uploaded into WEKA and various preprocessing steps were conducted to ensure its quality and suitability for analysis. The initial results, as depicted in the figure, have 22 number of variables including class variable Status_DO. The value of mean and standard deviation from the preliminary data is 0.636 and 0.096 respectively. The histogram in Figure 3 illustrates the distribution of the selected attribute within the raw dataset.

Figure 3: The result of raw data when uploaded into WEKA



4.1. Naïve Bayes

The results from the WEKA analysis using the Naive Bayes classifier with 10-fold cross-validation show that the model correctly classified 1779 instances, achieving an accuracy of 71.68%, while 703 instances were incorrectly classified, resulting in an error rate of 28.32%. The Kappa statistic is 0.4124, indicating moderate agreement. The mean absolute error is 0.3374, and the root mean squared error is 0.4434, with relative absolute and root relative squared errors of 69.25% and 89.83%, respectively. For the "ST" (graduate) class, the true positive rate is 0.784, precision is 0.742, recall is 0.784, and the F-measure is 0.762. The ROC and PRC areas for this class are 0.773 and 0.784, respectively. For the "DO" (dropout) class, the true positive rate is 0.624, precision is 0.677, recall is 0.624, and the F-measure is 0.649, with ROC and PRC areas of 0.722 and 0.758, respectively. The confusion matrix shows that 1128 "ST" instances were correctly classified, while 311 were misclassified as "DO," and 651 "DO" instances were correctly classified, with 392 misclassified as "ST." Overall, the Naive Bayes classifier demonstrated reasonable performance in predicting whether a student would graduate or drop out. The result can be seen in Figure 4.

4.2. Random Forest

Figure 5 displays the results of a Random Forest classifier executed using WEKA with 10-fold cross-validation. The classifier was trained with 100 iterations and a base learner, achieving a time to build the model of 0.91 seconds. Out of 2482 instances, 1975 were correctly classified, resulting in an accuracy of 79.5729%, while 507 instances were incorrectly classified. The Kappa statistic is 0.5774, indicating moderate agreement. The Mean Absolute Error (MAE) is 0.3250, the Root Mean Squared Error (RMSE) is 0.3858, and the Relative Absolute Error (RAE) is 66.6607%. The classifier's precision, recall, F-measure, MCC, ROC Area, and PRC Area for each class (ST and DO) indicate reasonable performance, with an overall weighted average precision and recall

of 0.795. The confusion matrix shows that class 'ST' was correctly classified 1216 times and misclassified 284 times, while class 'DO' was correctly classified 759 times and misclassified 507 times. The confusion matrix shows that the classifier has a relatively high accuracy in distinguishing between the two classes. In conclusion, the Random Forest classifier demonstrates a good level of performance with an acceptable error rate, making it a reliable model for this dataset.

Figure 4: Result of the data after Naive Bayes was applied

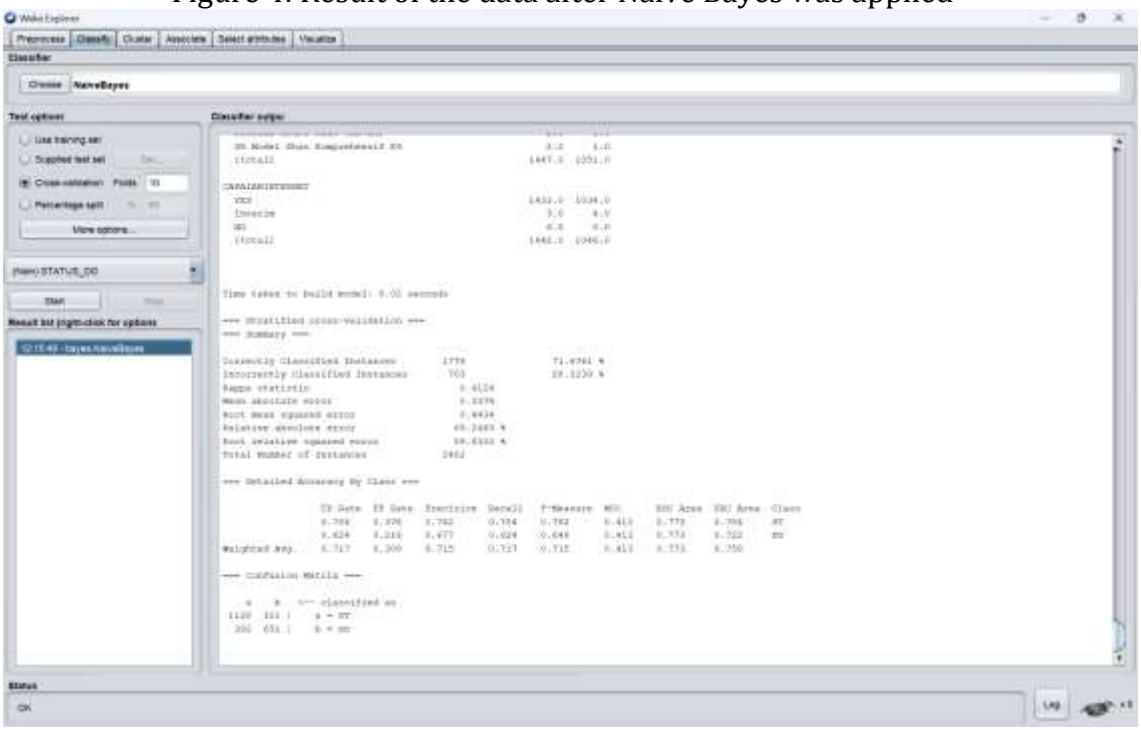
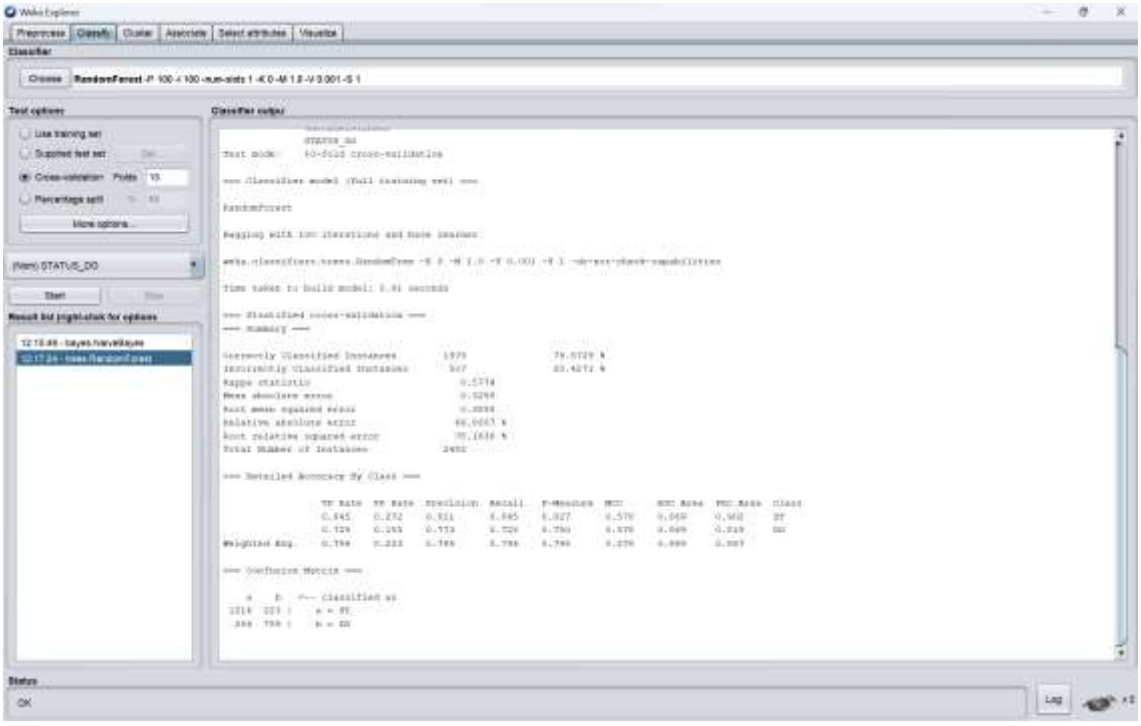


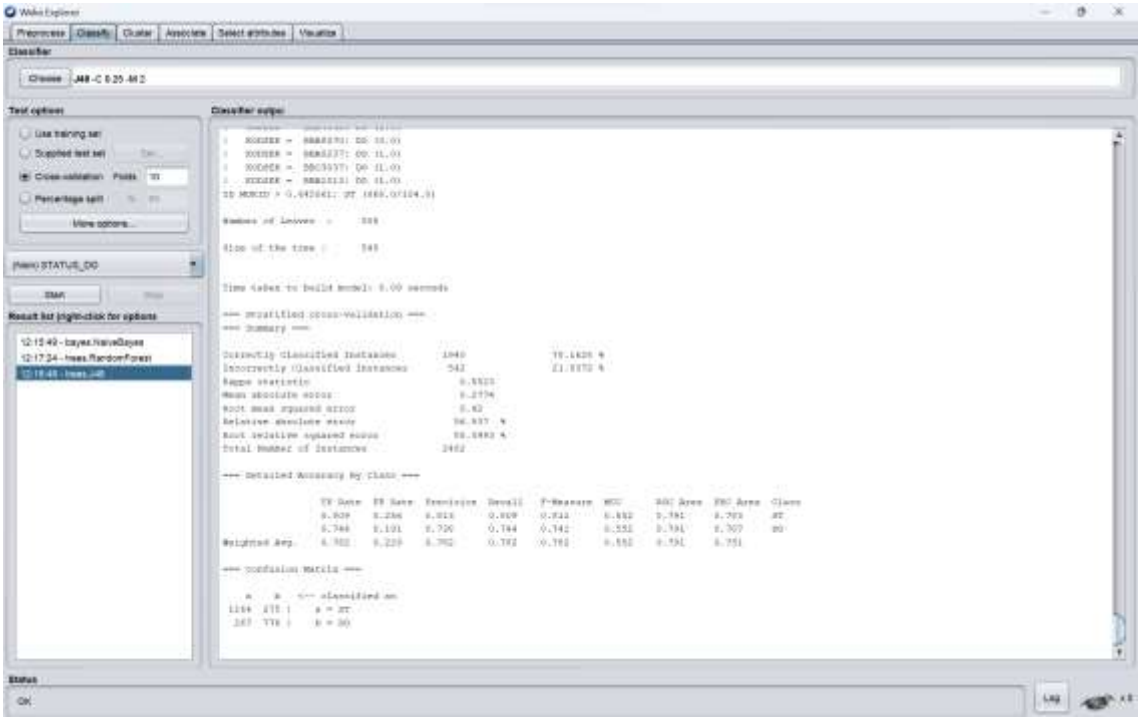
Figure 5: Result of the data after Random Forest was applied



4.3. Decision trees (J48)

Figure 6 shows the results of a Decision Tree (J48) classifier executed using WEKA with 10-fold cross-validation. The classifier was built with a confidence factor of 0.25 and a minimum number of instances per leaf of 2, resulting in a tree with 505 leaves and a size of 548. The time taken to build the model was 0.09 seconds. Out of 2482 instances, 1940 were correctly classified, yielding an accuracy of 78.1628%, while 542 instances were incorrectly classified. The Kappa statistic is 0.5523, indicating moderate agreement. The Mean Absolute Error (MAE) is 0.2774, the Root Mean Squared Error (RMSE) is 0.42, and the Relative Absolute Error (RAE) is 56.937%. The precision, recall, F-measure, MCC, ROC Area, and PRC Area for each class (ST and DO) indicate reasonable performance, with an overall weighted average precision and recall of 0.782. The confusion matrix highlights the classifier's accuracy in distinguishing between the two classes. In conclusion, the Decision Tree classifier demonstrates satisfactory performance with a moderate error rate, making it a viable model for this dataset.

Figure 6: Result of the data after Decision Tress (J48) was applied



4.4. Confusion Matrix for Naïve Bayes

The confusion matrix for the Naïve Bayes model in Table 6 shows that it correctly classified 1128 instances as 'ST' and 651 instances as 'DO'. However, it misclassified 311 'ST' instances as 'DO' and 392 'DO' instances as 'ST'. This model demonstrates a tendency to misclassify 'DO' instances as 'ST', which is evident from the higher number of false positives (392) compared to false negatives (311). Overall, the Naïve Bayes model has a moderate performance with a balanced number of errors in both classes, indicating that while it can predict both classes, there is room for improvement in reducing misclassification rates.

Table 6: Confusion Matrix of Naive Bayes

		Prediction Class	
		ST	**DO**
Actual Class	**ST**	1128	311
	DO	392	651

4.5. Confusion Matrix for Random Forest

The Random Forest model's confusion matrix in Table 7 reveals a higher accuracy compared to the Naïve Bayes model. It correctly classified 1216 instances as 'ST' and 759 instances as 'DO', with misclassifications of 284 'ST' instances as 'DO' and 759 'DO' instances as 'ST'. Although the model shows a significant number of misclassifications for the 'DO' class, it performs better overall in terms of correctly identifying 'ST' instances. The Random Forest model's strength lies in its ensemble nature, which generally enhances its predictive power and reduces overfitting, contributing to its superior performance.

Table 7: Confusion Matrix of Random Forest

		Prediction Class	
		ST	**DO**
Actual Class	**ST**	1216	284
	DO	759	759

4.6. Confusion Matrix for Decision Tree (J48)

Table 8 presents the Decision Tree (J48) model's confusion matrix indicates a solid performance with 1164 'ST' instances and 776 'DO' instances correctly classified. It misclassified 275 'ST' instances as 'DO' and 267 'DO' instances as 'ST'. The misclassification rates are lower compared to the Naïve Bayes model and somewhat comparable to the Random Forest model, particularly in distinguishing the 'DO' class. The J48 model shows a good balance in performance, providing a reliable classification with a slight edge over the Naïve Bayes model in terms of overall accuracy and precision.

Table 8: Confusion Matrix of Decision Tree (J48)

		Prediction Class	
		ST	**DO**
Actual Class	**ST**	1164	275
	DO	267	776

Based on the provided data in Table 9, the Random Forest model achieves the highest accuracy, correctly classifying 79.5729% of the instances. The Decision Tree (J48) model follows closely with an accuracy of 78.1628%, while the Naïve Bayes model has the lowest accuracy at 71.6761%.

Table 9: Comparison of prediction models with highest accuracy

Prediction Models	**Evaluation Parameter**

	Correctly Classified Instances (%)	ROC	Precision (%)	F Measure
Naïve Bayes	71.6761	0.773	74.2	0.715
Random Forest	79.5729	0.869	79.5	0.795
Decision Tree (J48)	78.1628	0.791	78.2	0.782

Random Forest is an ensemble learning method that combines the predictions of multiple decision trees, which helps reduce the risk of overfitting compared to a single decision tree that contributes to the higher accuracy of the model. By averaging the results of many trees, Random Forest reduces the variance of the model, leading to more stable and accurate predictions.

Additionally, the model's ability to handle missing values more effectively than other models also contributes to its higher accuracy. These characteristics make Random Forest a robust and reliable model for classification tasks, often outperforming simpler models like decision trees and Naïve Bayes.

5. Conclusion

According to the data, the Random Forest model attains the highest accuracy, accurately categorizing 79.5729% of the instances. Following closely is the Decision Tree (J48) model with an accuracy of 78.1628%, whereas the Naïve Bayes model shows the lowest accuracy at 71.6761%. Thus, Random Forest model is the best prediction model in this research because it surpassed the other models in terms of accuracy and precision.

Concurrently, from the initial dataset, which comprised 22 attributes, we conducted a feature selection process. This process identified 12 attributes that significantly contribute to predicting student drop-out rates. The selected attributes are StatusDLP, KeteranganJenisSekolah, KodJantina, StatusWarganegara, Yatim, StatusOKU, HubunganPenjaga1, KatPendapatan, Lokasi, GBMK, GSEJR, and Status_DO. These attributes were determined to have the most significant impact on predicting whether a student is likely to drop out, thereby refining our model and improving its accuracy and efficiency in making predictions. By focusing on these key features, we can better understand and address the factors influencing student drop-out rates.

The result can be benefitted the academic institutions or the ministry of education to identify the factors that contribute to school dropout as it can create specific early intervention plans to tackle the issue. In the future, this research can be furthered by using multiple data sets with many attributes by applying regression techniques.

To address potential increases in dropout rates, policymakers and educators in Malaysia need to implement targeted interventions to support at-risk students, such as providing access to technology and internet connectivity, offering financial assistance programs, enhancing mental health support services, and implementing strategies to re-engage disconnected students. Additionally, efforts should be made to address learning gaps and provide flexible learning options to accommodate diverse student needs. Ongoing monitoring and evaluation of dropout rates and related factors are essential to inform evidence-based interventions and policy decisions.

Ethics Approval and Consent to Participate

All procedures performed in this study have been approved by the Ethics Committee for Research Involving Human Subjects at Universiti Putra Malaysia (Jawatankuasa Etika Universiti Penyelidikan Melibatkan Manusia) and Educational Research Application System (ERAS) provided by the Research Ethics Committee of Education Planning and Policy ResearchDivision (EPRD), Ministry of Education Malaysia.

Acknowledgement

Part of this article was extracted from a master thesis submitted to Universiti Putra Malaysia.

Funding

This study received no funding.

Conflict of Interest

The authors reported no conflicts of interest for this work and declare that there is no potential conflict of interest with respect to the research, authorship, or publication of this article.

References

Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). *Mining Student Data Using Decision Trees*. ACIT.

Dupere, V., Leventhal, T., Dion, E., Crosnoe, R., Archambault, I., & Janosz, M. (2015). A Stress Process, Life Course Framework of Dropout. *Review of Educational Research*, *85*(4), 591–629. https://doi.org/10.3102/0034654314559845

Gil, J. S., Delima, A. J. P., & Vilchez, R. N. (2020). Predicting Students' Dropout Indicators in Public School using Data Mining Approaches Jay. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(1), 5–9. https://doi.org/https://doi.org/10.30534/ijatcse/2020/110912020

Heppen, J. B., & Therriault, S. B. (2008). *Developing early warning systems to identify potential high school dropouts*. Washington DC : The National High School Center at the American Institutes for Research.

Mardolkar, M., & Kumaran, N. (2020). Forecasting and Avoiding Student Dropout Using the K-Nearest Neighbor Approach. *SN Computer Science*, *1*(2). https://doi.org/10.1007/s42979-020-0102-0

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Forrest Cataldi, E., Bullock Mann, F., and Barmer, A. (2019). *The Condition of Education 2019*. NCES 2019144.

Mduma, N., Kalegele, K., & Machuve, D. (2019). Machine Learning Approach For Reducing Students Dropout Rates. *International Journal of Advanced Computer Research*, *9*(42), 156–169. https://doi.org/10.19101/ijacr.2018.839045

Ministry of Education. (2018). Garis Panduan Mengurus Murid Berisiko Cicir di Sekolah. *Ministry of Education*. https:// https://www.moe.gov.my/pekeliling

Musiliu, B. (2020). *Comparison of Feature Selection Techniques for Predicting Student ' s Academic Performance*. *International Journal of Research and Scientific Innovation, 7*(8), 97-101.

Nicoletti, Maria do Carmo. (2019). Revisiting the Tinto's Theoretical Dropout Model. *Higher Education Studies*, *9*(3), 52. https://doi.org/10.5539/hes.v9n3p52

Provost, P., Fawcett, T. (2013). *Data Science for Business: What You Need To Know About Data Mining and Data-Analytic Thinking.* O'Reilly Media.

Roslan, N. (2021). Prediction of Student Dropout in Malaysian's Private Higher Education Institute using Data Mining Application. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 2326–2334. https://doi.org/10.17762/turcomat.v12i3.1219

Rumberger, R. W., & Lim, S. A. (2008). Why Students Drop Out of School : A Review of 25 Years of Research. In *California Dropout Research Project*. Retrieved from https://www.issuelab.org/resources/11658/11658.pdf.

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.2307/1170024

UNICEF and UIS (2016). Monitoring Education Participation: Framework for Monitoring Children and Adolescents who are Out of School or at Risk of Dropping Out. *UNICEF Series on Education Participation and Dropout Prevention,* Vol I. UNICEF Regional Office for Central and Eastern Europe and the Commonwealth of Independent States.

Vijayakumaran, N., Mohd Yusof, H., Oulaganathan, S., & Saundra Rajan, D. K. (2023). The Impact of Parental Involvement and Student Engagement on School Dropout Intention: A Systematic Literature Review. *International Journal of Education, Psychology and Counseling*, *8*(50), 36–46. https://doi.org/10.35631/ijepc.850003

Viloria, A., Padilla, J. G., Vargas-Mercado, C., Hernández-Palma, H., Llinas, N. O., & David, M. A. (2019). Integration of data technology for analyzing university dropout. *Procedia Computer Science*, *155*(2018), 569–574. https://doi.org/10.1016/j.procs.2019.08.079

White, S., & Kelly, F. (2010). The School Counselor's Role in School Dropout Prevention. *Journal of Counseling and Development*, *88*(2), 227–235. https://doi.org/10.1002/j.1556-6678.2010.tb00014.x