INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 12, Issue 4, (2023) E-ISSN: 2226-6348

Utilizing Educational Data Mining for Enhanced Student Performance Analysis in Malaysian STEM Education

Mohammad Izzuan Termedi, Aini Marina Ma'rof, Habibah binti Ab. Jalil, Iskandar Ishak

Faculty of Educational Studies

Universiti Putra Malaysia, Jalan Universiti 1, 43400 Serdang, Selangor, Malaysia

Corresponding Author's Email: izzues66@gmail.com

To Link this Article: http://dx.doi.org/10.6007/IJARPED/v12-i4/19577 DOI:10.6007/IJARPED/v12-i4/19577

Published Online: 28 November 2023

Abstract

Educational Data Mining (EDM) applies data mining in education, aiding schools to enhance student learning programs by analyzing data and success factors. In the era of big data, schools must adopt data-driven approaches. However, predicting success among diverse secondary students in Malaysia remains uncertain due to dataset size and heterogeneity. This study aims to identify key predictor variables for STEM student performance and present a systematic method for analysis, benefiting academics, schools, and the education ministry. The article explores data mining via knowledge discovery (KDD) and employs classifiers like Random Forest, PART, J48, and Naive Bayes on a dataset of Malaysian upper-secondary Science students. Utilizing WEKA for analysis, the research utilizes 21 features from the Education Repository and SAPS. Notably, J48 outperforms other classifiers. The study aids educational enhancement, enabling early intervention and improved academic achievement.

Keywords: Data mining techniques, educational data mining, performance prediction, student performance

Introduction

Investing in Science, Technology, Engineering, and Mathematics (STEM) program is necessary to foster innovation, creativity, and the maintenance of employment opportunities in the context of Industry 4.0's educational initiatives (Ali, Jaaffar, & Ali, 2021). Education in the STEM fields will also help foster an increase in critical thinking abilities, essential in expanding the economy throughout the industrial revolution.

Data mining is becoming more common for evaluating students' academic achievement (Huang, Spector, & Yang, 2019). Most scholars use the techniques of classification, regression, and clustering. Some of them employ a combination of several methodologies to produce a robust mechanism with a higher prediction accuracy model (Perner, 2015). Data mining is the

utilization of a wide variety of algorithms and tools to forecast potential future occurrences based on past happenings (Mokhtar, Alshboul, & Shahin, 2019).

According to Barbosa Manhães, da Cruz, and Zimbrão (2015), a lot of research tries to predict how well students will do in school, but relatively few such studies use the data mining approach. This work uses data mining methods to construct the most accurate model for predicting students' academic achievement in STEM fields. The study aims to describe the predictor variables that best predict student performance and to provide a systematic approach employing student data that academics, schools, and the educational ministry may use to analyze student performance.

This study will examine statistical data mining methods, including Random Forest, PART, J48, and Naive Bayes. These models will be refined to suit student performance data and tested to select the best data mining model (Livieris et al., 2019). This study compares academically successful and unsuccessful students. The Education Repository's data will be mined to confirm its quality. The research will improve the effectiveness of data mining methods employed in secondary schools and inform the Ministry of Education on student academic performance.

This study uses more accurate and authentic educational data to predict academic performance among Form 4 students in Malaysian secondary schools. The research aims are:

1) To identify and select appropriate predictor or independent variables that can be used as the inputs of predictive models in predicting students' STEM performance.

2) To identify and select the best data mining techniques: Random Forest, PART, J48, and Naive Bayes for developing predictive models in predicting students' STEM performance.

3) To validate the developed models using the data collected in upper secondary schools in Malaysia.

This study contributes to the literature in several ways. First, this study will provide a methodological technique for predicting student performance utilizing Education Data Repository data (Zulkifli, Mohamed, & Azmee, 2019). Most secondary school student input data is recorded online in the Ministry of Education database. They may have access to data but not know how to interpret it. Most Education Data Repository systems and apps are used in schools for compliance and monitoring. Individual student information permits head teachers to obtain data from teachers responsible for entering data into Education Data Repository systems. The head teacher personally collected student background, accomplishments, and attendance data.

Second, this study's conclusions may assist the Ministry of Education, instructors, and parents increase academic performance and student success. These findings may impact future curriculum strategy and help encourage these variables among students before they take the topics of interest. This investigation uses four data mining techniques to estimate student performance in related disciplines, which will benefit educators and students.

Third, this research will reveal the reliability of the data mining algorithm used. Precision, Recall, F-measure, and ROC area assist in choosing the optimal data mining method for future study. The top model can discover academically at-risk students sooner, which helps forecast dropouts. This will help the education ministry generalize STEM student performance (Ramaswami et al., 2019).

Related Works

Developing student outcome prediction models is one of education's oldest and most common practices (Romero & Ventura, 2010). Predicting academic performance in many disciplines has long been regarded as an essential research subject for various reasons. First, predictive models help teachers predict students' success and take preventive measures. The teacher may classify problematic students academically using a proven statistical model. The teacher may suggest implementing different instructional techniques to those at risk in learning and the approach to minimize student drop-out levels from suitable courses or services (Lowis & Castley, 2008).

Data mining, or knowledge discovery, as defined by Witten, Frank, and Hall (2011), is the computer-assisted process of digging through and analyzing enormous data sets and extracting the data's meaning. Data mining tools predict behaviors and future trends allowing businesses to make proactive, knowledge-driven decisions. There is a rising interest in data mining due to the recently increased amount of data, sometimes called "Big Data." This study aims to explore data mining in education, known as Educational Data Mining (EDM). The study has proliferated in the EDM industry because education data seems to contain undiscovered information, necessitating extensive research to find this new information. During the last two decades, the application of data mining techniques has gained popularity in the modern educational era, spurred by the fact that it enables all educational stakeholders to discover new, engaging, and useful knowledge about students and potentially improve some aspects of the quality of education (Mueen et al., 2016). Some excellent surveys, Romero and Ventura (2007) and Romero and Ventura (2010) presented the significant trends in EDM research, describing in detail the process of mining learning data to discover new insights and how those insights impact the activity or practitioners in education.

Shahiri, Husain, and Rashid (2015) concluded in their systematic literature review in predicting students' performance that the classification method is frequently used in educational data mining. Under the classification techniques, Neural Network and Decision Tree are the two methods researchers use to predict students' performance. Ramaswami et al. (2019) also used classification data mining techniques to predict students' academic performance. They argued that the model developed from their research can provide value to institutions in making process- and data-driven predictions on students' academic performances.

In their systematic review, Roslan and Chen (2022) found that the methods most often used by previous researchers can be categorized into classification, cluster, and regression. Among those methods, the most widely used methods of estimating students' academic performance are the classification and regression methods. In conjunction with the systematic review conducted by Roslan and Chen (2022), there was also the same kind of review conducted by Asiah et al. (2019). They proposed that the prediction method is one of the critical components for analyzing student performance. Most researchers use classification, regression, and clustering methods such as Bayesian Network, Decision Tree, Artificial Neural Network, Support Vector Machine, K-Nearest Neighbour and others. Some use mixed methods to provide robust mechanisms with a better predictive accuracy model.

Al-Barrak and Al-Razgan (2016) presented a case study in educational data mining. They discovered it was predominantly used to improve students' performance and detect early predictors of their final GPA. They utilized the classification technique, particularly in decision tree, to predict students' final GPA based on their grades on previous courses. They discovered classification rules to predict students' final GPA based on their grades in

mandatory courses. They also evaluated the most critical courses in the study plan that greatly impact the student's final GPA.

Devasia, Vinushree, and Hegde (2016) also used the data mining technique to predict students' performance. The algorithm used was classification employed in student information to predict the students' division based on previous information. Several area unit approaches are used for knowledge classification, so Naive Bayes is employed here. Information like group action, class tests, seminars, and assignment marks was collected from the students' previous information to predict their performance at the top of the semester.

Methodology

Introduction

The preceding paragraphs make specific reference, on several occasions, to the goals of the study that will be carried out. The first step is to investigate and locate characteristics that may accurately predict students' success in STEM subjects in secondary schools. Second, to develop the most accurate prediction model based on the information provided by the students. To accomplish what must be done, we will use the data mining technique (Abu Saa, Al-Emran, & Shaalan, 2019).

Analyzing the data and information of students to classify students, create decision trees or association rules, make better decisions, or enhance students' performance is an exciting field of research. This field of research primarily focuses on analyzing and understanding students' educational data, which indicates their educational performance. It generates rules, classifications, and predictions to assist students' future educational performance (Asif, Merceron, & Pathan, 2014).

The overall framework of this research is based on the knowledge discovery (KDD) paradigm through imposing data mining (DM) (See Figure 3.1). Data mining may be described as the processing or compiling of knowledge or valuable information from broad data stores (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Many researchers often mark data mining as a knowledge discovery (KDD), knowledge mining, data analysis, and computer architecture. Data mining may help identify and discover data similarities in nearly all fields of study (Sumathi & Sivanandam, 2006).

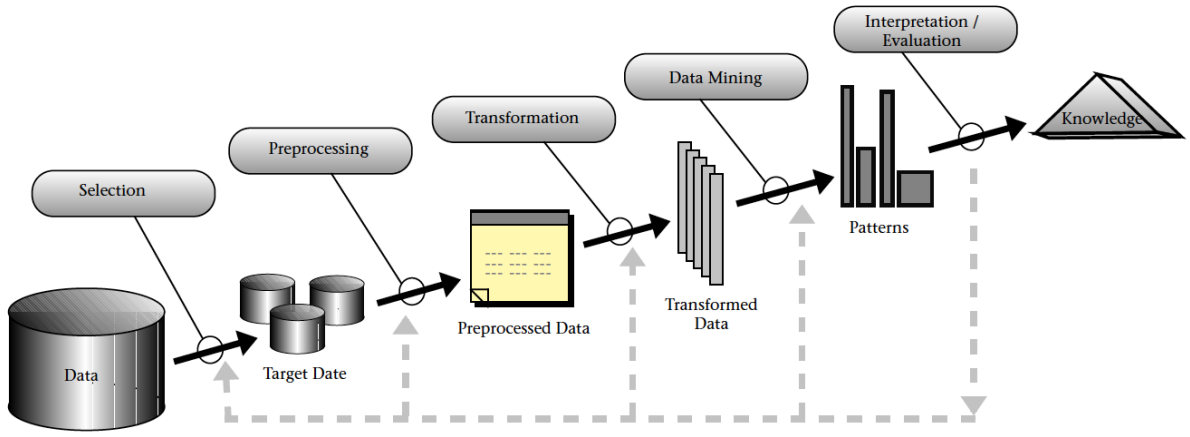

Figure 3.1. An Overview of the Steps That Compose the KDD Process (Fayyad et al., 1996)

Collecting data entails amassing all information on students, considering the elements influencing student performance (Mueen, Zafar, & Manzoor, 2016). These pieces of information may be compiled into the dataset by gathering them from the many data sources that are currently accessible. We will identify the elements that are thought to impact the

students' performance by thoroughly studying past research (Al-Barrak & Al-Razgan, 2016) and engaging in conversation with individuals who are believed to be experts on student performance (Lowis & Castley, 2008). These elements of influence will be classified as input variables in the analysis.

On the other hand, students' grades will serve as the dependent variables for this investigation (Asif, Merceron, & Pathan, 2014).

In the next step, known as "pre-processing," the data will be "cleaned," "attributes selected," "dimensionality reduced," and "data partitioned" so that a more accurate forecast can be made (Romero & Ventura, 2010).

The next step, the categorization stage, will be carried out (Shahiri, Husain, & Rashid, 2015). During the classification step, many data mining algorithms are tested using various factors before being compared with one another to see which algorithm performs the best (Roslan & Chen, 2022).

In the last step, we will analyze the models we developed in the previous stage to see how well we can forecast the students' success in STEM-related classes (Al-Barrak & Al-Razgan, 2016).

Data Collection

For this research, we want to gather information on the educational experiences of secondary school students, specifically Form 4 students. All the academic activities, as well as their personal or student's student demographic, socio-economic, and psychological characteristics, as well as school-related factors that are anticipated to affect student performance in STEM disciplines, will be gathered. The data collected in this study are from two sources. The first one is collected from the Education Repository in the Ministry of Education, mainly APDM (Aplikasi Pangkalan Data Murid), an online based platform that stores information on school students in Malaysia (e.g., parents' income, the number of siblings, previous school attended) and the other one is collected from SAPS (Sistem Aplikasi Peperiksaan Sekolah) to obtain students' grade in the examination. Based on extensive literature reviews conducted, the input factors and environmental factors (predictor variables) – consists of students' demographic and students' academic record as well as the target goal (academic results) chosen for this particular study from the two sources mentioned above are shown in Table 3.1.

Table:
Attributes Involved in Predicting Students' Academic Performance in Secondary Schools

Type of Factors	Attributes	Description	
Input	District	Area of the school	Students' Demographic
	Sex	Gender of the students	
	Race	Ethnicity of the students	
	Religion	Religion of the students	
	Orphan	Whether the students orphan or not	
	OKU	Whether the students disable or not	
	Nationality Of Guardian 1	Nationality of student's father	
	Job Of Guardian 1	Job of student's father	
	Income Of Guardian 1	Salary of student's father	
	No. Of Dependents	Number of siblings of the students	
	Nationality Of Guardian 2	Nationality of student's mother	
	Job Of Guardian 2	Job of student's father	
	Combined Income	Combined salary of student's parents	
	Salary Group	Type of salary group	
	DLP Status	Whether the students learn STEM subject in English	Students' Academic Record
	Dormitory	Whether the students live in dormitory	
	Mid-Year Attendance	Percentage of attendance	
	Final Year Attendance		
	Mid-Year Chemistry	Student's grades in Mid -Year Exam	
	Mid-Year Physics		
	Mid-Year Biology		
	Mid-Year Add Maths		
Environment	Type of School	Whether the school are regular, religious, or boarding	
Outcome	Final Year Chemistry	Student's grades in Final Year Exam	
	Final Year Physics		
	Final Year Biology		
	Final Year Add Maths		

The Initial Processing of the Data

Before moving on to the next stage, which is classification methods, pre-processing the data is a vital step that must be completed to get the dataset ready. It is essential to remember that the quality and dependability of the information provided directly affect the outcome of this activity. A careful and comprehensive study of the variables and the values they relate to is carried out to eliminate any idiosyncrasies. In this investigation, the primary pre-processing activities that will be carried out are the following:

Choosing Between the Features

The datasets were examined in great detail to determine the qualities that influence the output variable more.

Imbalanced data

The data is unbalanced when the number of examples in one class is much lower than the number of instances in another class. Consequently, while the classifier is in the training phase, it will accept more samples from the classes with a more significant number of instances. Because of this, while in the testing phase, classifiers are less sensitive to the classes that include fewer examples.

Data transformation

It is necessary to complete this method to carry out the required algorithms. The appropriate adjustments will be made to the datasets so that the prediction model may be satisfied. This includes identifying missing data and outliers and transforming the data to the appropriate destination for the data file.

Modeling

The fourth stage of data mining involves construction and model selection. A set of models is typically generated utilizing various statistical algorithms or data mining techniques, often called ensemble models. Another approach is to build data samples and evaluate or merge tests. Bootstrap, jackknife resampling, and validation of V-fold cross are techniques that use sample details. Upon defining essential model assumptions, it is important to set parameters in most data mining applications as the algorithm choices are often the default. Modeling algorithms such as neural networks, decision trees, and logistic regressions begin with different default settings. Technology for data mining, such as Enterprise Miner, has a tab for specifying parameter values.

Software

The study used data mining tools to analyze the most relevant student performance variables and to address the research questions. WEKA software will be used as data mining software for all analyses. WEKA stands for Waikato Environment for Knowledge Analysis, developed at the University of Waikato, New Zealand. WEKA contains visualization tools, algorithms for data analysis and predictive modeling, and graphical user interfaces for easy access to these functions. The target variable was the performance of students in STEM subjects when the data was mined, and all other variables were used as predictors.

Data Mining Techniques and Prediction Model

Techniques in data mining may identify models present but unseen in broad institutional datasets. The methodology incorporates mathematical methods, algorithms for machine learning, and visual representation to identify trends in institutional knowledge. This study used five specific predictive models in data extraction to determine the significant variables leading to a student's performance: ZeroR, Random Forest, PART, J48, and Naive Bayes. The five predictive models will classify the most critical factors impacting student performance in STEM subjects and estimate the proportion of academically at-risk students in STEM subjects. The study has also tested the accuracy of the five data mining models, which can offer further

information and insight into the most effective models in secondary schools' data mining analyses.

ZeroR

The ZeroR classification approach is the simplest one available since it just uses the target and disregards any predictors. The ZeroR classifier only estimates the category that will constitute the majority (class). ZeroR is beneficial for defining a baseline performance to serve as a standard for other classification techniques, even though it does not have any capability to predict outcomes.

Random Forest

Random Forest is a machine learning algorithm that creates many decision trees to make predictions. Each decision tree is trained on a random subset of the data and features. During the prediction phase, each tree predicts the output, and the final prediction is determined by combining the predictions of all the trees. Random Forest is practical because it can handle different data types, is less prone to overfitting, and provides feature importance measures. It can be used in various fields, such as finance, healthcare, and marketing, for fraud detection, disease diagnosis, and customer segmentation.

PART

PART is a machine learning algorithm used for classification and rule discovery. It works by recursively dividing the dataset into smaller subsets based on the values of its attributes until it creates a set of rules that can accurately classify new instances. The algorithm starts by examining the entire dataset and selecting the attribute that best divides the data into subsets with the most distinct class values. It then creates a rule based on this attribute, which divides the data into two subsets. This process is repeated for each subset until a stopping criterion is met. Once the tree is constructed, the algorithm prunes the branches that do not significantly improve the accuracy of the rule set. This results in a more straightforward and more accurate set of rules that can be used for classification.

J48

The C4.5 algorithm is a classification method that, following the principles of information theory, generates decision trees. It is an adaptation of Ross Quinlan's older ID3 algorithm, also called J48 in Weka, where J stands for Java. Because of its ability to produce decision trees that can be used for classification, C4.5 is often referred to as a statistical classifier. The J48 version of the C4.5 method has many extra capabilities, such as accounting for missing data, pruning decision trees, continuous attribute value ranges, and creating rules, among other things. An open-source Java version of the C4.5 technique may be found in the WEKA data mining tool named J48. J48 allows users to classify data using either decision trees or rules derived from those trees.

Naïve Bayes

The methodology of Naive Bayes uses the probabilistic relation between classes and their attributes. Classification of the record relies on the attribute values, which can be seen as the likelihood of being registered by the specific class. The classification of Naive Bayes is based on the theorem of Bayes estimating how often x belongs to class c in (1).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

P (c|x) = probability of instance x being in class c
P (x|c) = probability of generating instance x given class c
P(c) = probability of occurrence of class c
P(x) = probability of instance x occurring

Evaluation

This phase involves an assessment of the models constructed in the model-building stage. The most popular way models are tested is to verify their output on test datasets. This study used Accuracy, Precision, Recall, F-Measure, and Receiver Operating Characteristic (ROC) to evaluate the model. If the percentage is relatively high, the model could be concluded to be a success.

To illustrate the accuracy of the model used in this study, it is essential to undergo the model comparison te A robust model evaluation methodology is important to identify the right model and to trust its efficiency (Nisbet et al., 2018). The assessment method also poses some challenges, which can be overcome by modifying those data processing or modeling activities. Such improvements can help to improve the model's predictability or implementation usability.

Classification Accuracy

The classification accuracy metric is the starting point for analyzing the model's performance. It is a measurement of how many forecasts were accurate in comparison to the total number of predictions. The model's accuracy improves in direct proportion to the ratio's value. Below is the formula for classification accuracy:

$$Accuracy = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}}$$

Precision

The model's precision may be evaluated based on the frequency of its accurate positive predictions. In other words, the likelihood that the model will predict the frequency of a positive class. The formula of precision is given below:

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall

A recall is the number of true positives divided by the number of all samples that should have been identified as positive:

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F-Measure

F-measure or F1-Score considers both precision and recall. It is the harmonic mean (average) of the precision and recall. F-measure or F1-Score is best if there is some balance between

precision and recall in the system. Oppositely, the F-measure or F1-Score is not so high if one measure is improved at the expense of the other.

$$F - measure = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Receiver Operating Characteristic (ROC)

A ROC curve is a way to graphically show how well a binary classifier algorithm performs in data mining. This algorithm classifies data into two categories, positive or negative. The ROC curve compares the proportion of true positives (correctly classified positive instances) to false positives (incorrectly classified positive instances) at different threshold values. This visualization helps to find the optimal threshold value that balances the model's classification performance. The area under the ROC curve (AUC) is a commonly used metric to measure the overall performance of the binary classifier. An AUC of 1 indicates perfect classification performance, while an AUC of 0.5 means the model performs no better than a random guess. Overall, the ROC curve is a helpful tool for evaluating and comparing the performance of binary classification algorithms in data mining.

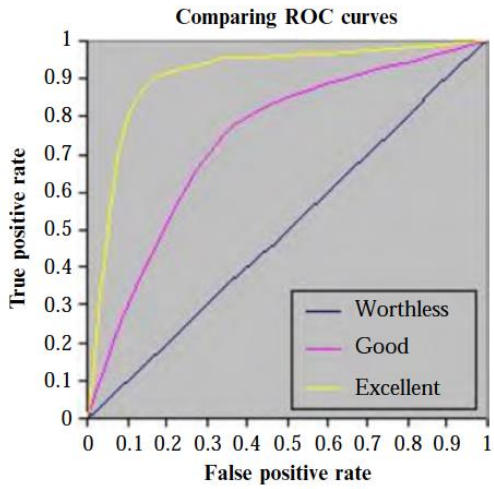

Figure 3.8. Sample ROC curves. The area under the yellow curve ("excellent" model) and the diagonal line is greater than that of the blue curve ("good" model, reflecting a greater predictive power of the yellow model than the blue model).

EXPERIMENTAL SETUP

We split the data into three categories: training, testing, and validation. So, the proportion of the data being split is 60% (8578 instances), 20% (2860 instances), and 20% (2860 instances), respectively. We trained our data using the training datasets to address the first objective. It is a crucial part of any data mining model, as it enables these models to provide reliable predictions or carry out the required functions. It demonstrates the form that the intended output should take. The dataset is analyzed several times by the model so that it may get a comprehensive understanding of its properties and improve its own performance. Figure 4.1 shows the workflow of the training datasets in WEKA environment.

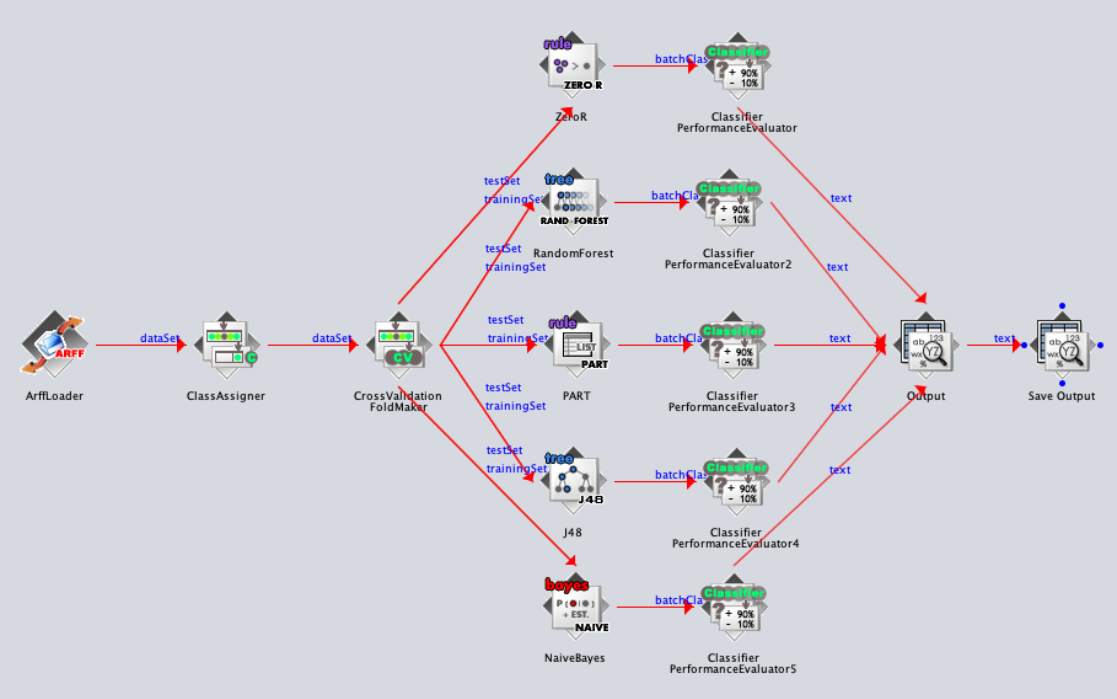

Figure 4.1. Prediction Workflow of the Training Datasets in WEKA Environment

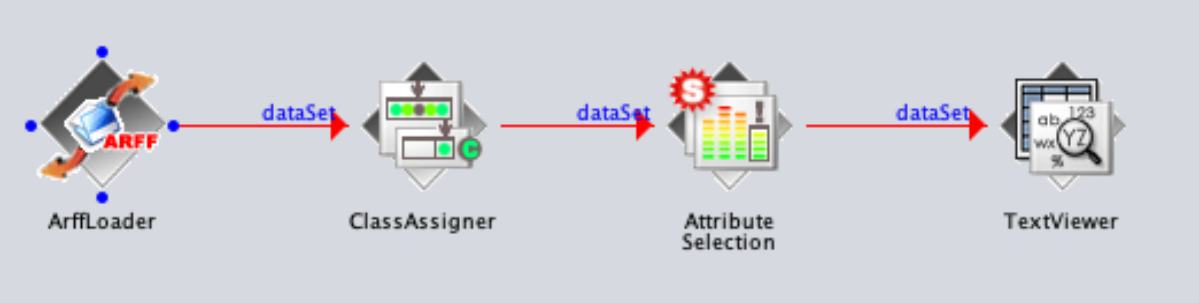

Figure 4.2. Feature Selection Workflow in WEKA Environment

A 10-fold cross-validation technique will be implemented in this training set. That is why we use CrossValidationFoldMaker as shown in Figure 4.1. Next, attribute selection was used in WEKA to answer research question 2. This step is also called feature selection. Figure 4.2 shows the workflow of the process using WEKA.

The grades used in the study are according to the actual SPM examination, which are A+ (90-100), A (80-89), A- (70-79), B+ (65-69), B (60-64), C+ (55-59), C (50-54), D (45-49), E (40-44), G (0-39) and TH (absent). The prediction group classified in this study consists of Excellent (B, B+, A-, A and A+), Good (E, D, C, and C+) and Fail (G and TH). After loading the required data, we choose the specific class to be predicted into the ClassAssigner. For example, if we want to predict the Chemistry outcome, we will assign Chemistry Final Year Examination results to the ClassAssigner. After that, the datasets will go through Attribute Selection. In WEKA, we used WrapperSubsetEval as the evaluator to determine the best attributes to be selected in our model. WrapperSubsetEval evaluates attribute sets by using a learning scheme. Cross-validation is used to estimate the accuracy of the learning scheme for a set of attributes. From Figure 4.3, we can see that we used J48 as the classifier in this attribute selection. This is because J48 is the best classifier that produced the highest prediction accuracy.

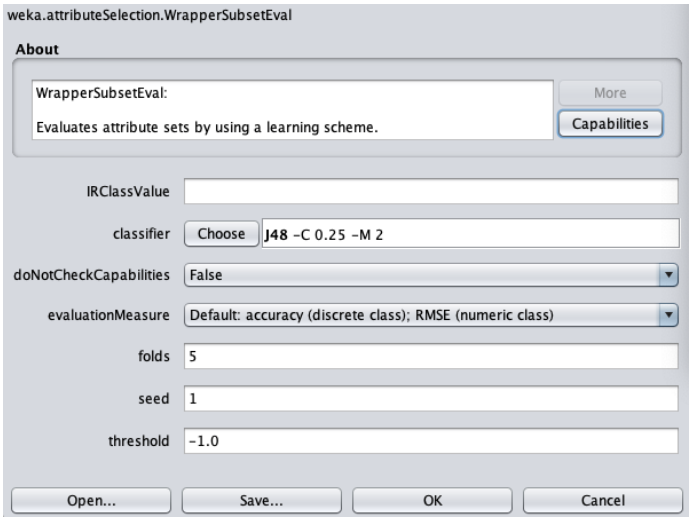

Figure 4.3. The Use of J48 in Feature Selection Workflow

Further, BestFirst searches for the best attributes to be included in our model. BestFirst searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward or start with the complete set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point). In this case, we used both directions, as in Figure 4.4.

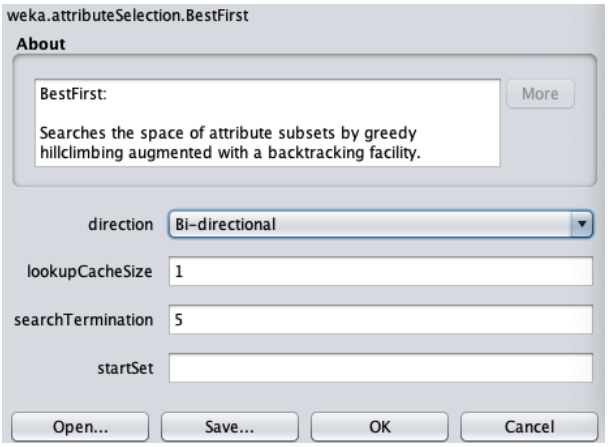

Figure 4.4. The Use of Both Directions (Forward and Backward) in Feature Selection Workflow

In comparing the accuracy of each data mining technique, such as Random Forest, PART, J48, and Naive Bayes, the following workflow in WEKA shown in Figure 5.1 has been used. Like in the attribute selection stage, we first select the required .arff file of the training data and test data into ArffLoader and then assign the specific class or prediction according to each STEM subject in ClassAssigner. Next, five different data mining techniques will be evaluated using ClassifierPerformanceEvaluator, which evaluates each classifier and produce the outcome or accuracy of the classifier. Here, we added one more classifier, which is ZeroR. ZeroR is the most basic classification approach, relying on the target and ignoring any

predictors. The ZeroR classifier predicts just the majority category (class). Although ZeroR has no prediction power, it may be used to establish a baseline performance as a standard for other classification systems.

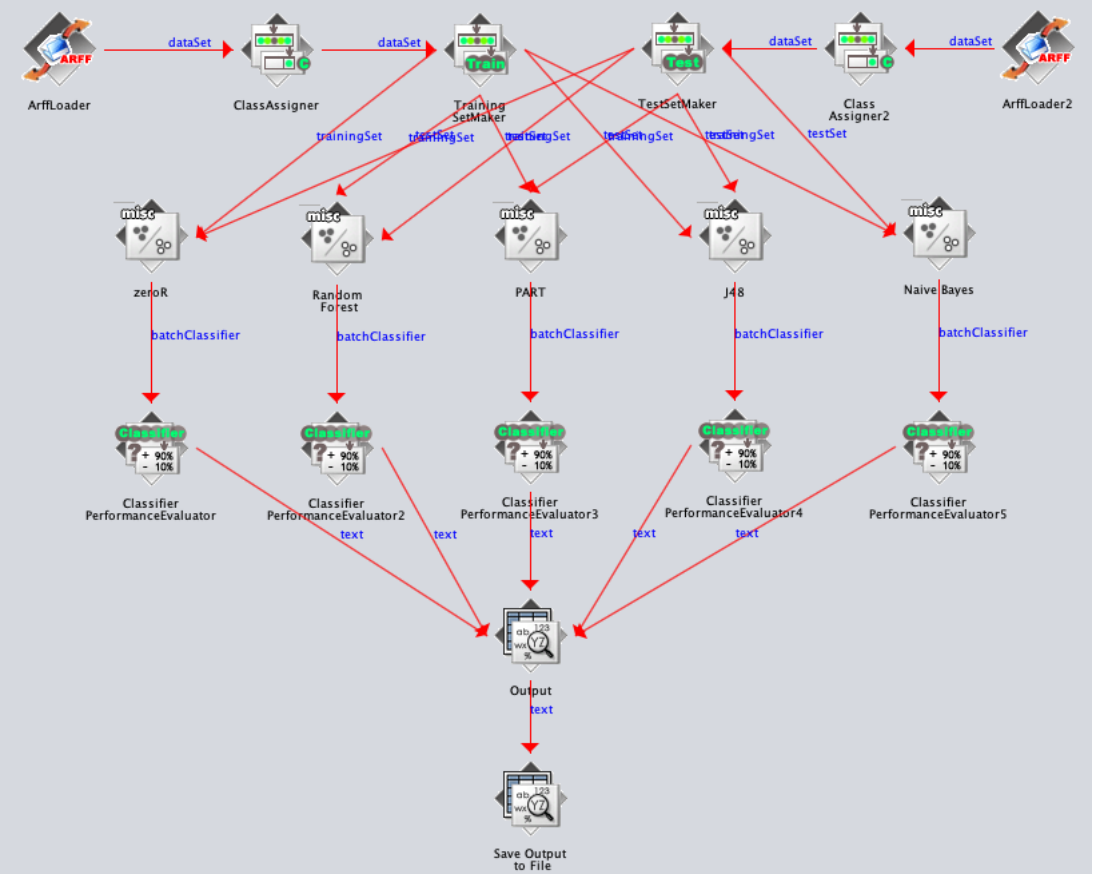

Figure 4.5. The WEKA Workflow for Evaluating Different data mining Techniques

Results And Discussions

Predictor Variables or Attribute Selection

Table 5.1 shows the selected attributes for this study (Mueen, Zafar, & Manzoor, 2016). There are 13 attributes selected for Chemistry, 13 attributes for Physics, nine attributes for Biology, and 13 attributes selected for Additional Mathematics. Chemistry has the lowest accuracy with 76.90%, and Additional Mathematics has the highest with 88.50% accuracy. Dormitory, Number of Dependents, Mid-Year Attendance, specific Mid-Year Results, and Final Year Attendance are the commonly chosen attributes for all STEM subjects (Roslan & Chen, 2022). Results were obtained using training datasets (60% of the datasets).

Using good data and data mining algorithms effectively is crucial to the effectiveness of data mining in predicting student performance (Al-Barrak & Al-Razgan, 2016). We must use the appropriate approach if we want the most significant outcomes from our data mining efforts. The best prediction results cannot be obtained just by using the algorithm (Huang, Spector, & Yang, 2019). Obtaining optimal prediction outcomes also depend on attribute or feature selection, the process of altering data for optimal data mining accuracy (Mueen et al., 2016). Attribute selection in WEKA environment was very straightforward, but the results were spectacular. However, we started with many attributes in predicting students' STEM performance, 21 attributes are exact. However, from the results we can see that not all the attributes were selected as the predictors for the prediction model. From the datasets specified in this study, the most frequent attributes used in the prediction models include

Dormitory, Number of Dependents, Mid-Year Attendance, specific Mid-Year Results, and Final Year Attendance are the commonly chosen attributes for all STEM subjects.

The attribute Number of Dependents in this study belongs to students' demographic, whereas Dormitory, Mid-Year Attendance, all Mid-Year Results, and Final Year Attendance belong to students' academic records. So, we can say that students' academic records are chosen as the attributes to be put into the prediction model most of the time. This aligns with the study conducted by Shahiri, Husain, and Rashid (2015) and Roslan and Chen (2022), which reported that almost one-third of previous studies used students' academic records such as CGPA. Similarly, Abu Saa, Al-Emran, and Shaalan (2019) found that the most common elements in predicting student success are students' past grades and internal evaluations based on a review of 36 research publications from 2009 to 2018. This is backed even further by Asif, Merceron, and Pathan (2014), who found that predicting student success only based on academic outcomes is possible, independent of any other variables.

Table 5.1.

Attribute Selection

ATTRIBUTE SELECTION J48				
SUBJECT	CHEMISTRY	PHYSICS	BIOLOGY	ADD MATHS
ACCURACY	0.769	0.799	0.820	0.885
NO. OF SELECTED ATTRIBUTES	13	13	9	13
SELECTED ATTRIBUTES	1. Dlp Status 2. Sex 3. Race 4. Religion 5. Dormitory 6. Nationality Of Guardian 1 7. Job Of Guardian 1 8. No. Of Dependents 9. Nationality Of Guardian 2 10. Salary Group 11. Mid Year Attendance 12. Mid Year Chemistry 13. Final Year Attendance	1. District 2. Type Of School 3. Dlp Status 4. Sex 5. Dormitory 6. Oku 7. Religion 8. Nationality Of Guardian 1 9. No. Of Dependents 10. Salary Group 11. Mid Year Attendance 12. Mid Year Physics 13. Final Year Attendance	1. District 2. Type Of School 3. Dlp Status 4. Dormitory 5. Job Of Guardian 1 6. No. Of Dependents 7. Mid Year Attendance 8. Mid Year Biology 9. Final Year Attendance	1. District 2. Dlp Status 3. Sex 4. Religion 5. Dormitory 6. Oku 7. Nationality Of Guardian 1 8. No. Of Dependents 9. Nationality Of Guardian 2 10. Salary Group 11. Mid Year Attendance 12. Mid Year Add Maths 13. Final Year Attendance

The Best Data Mining Techniques or Algorithms

Table 5.2 shows the accuracy of the five data mining techniques, including the number of correctly and incorrectly classified instances using test datasets (20% of the datasets). The baseline accuracy performance showed by ZeroR is in the middle value for three STEM subjects except for Additional Mathematics. The accuracy for all STEM subjects in the study can be considered reasonable, with the accuracy being more significant than 70%. However, most of the prediction accuracy using selected attributes showed better accuracy than entire attributes. The highlighted values are the best accuracy and classifier for each STEM subject.

Table 5.2.
Model Accuracy Comparison for Each STEM Subject

EXCELLENT, GOOD OR FAIL – SPECIFIC SUBJECT						
CLASSIFIERS	**ACCURACY**	**CORRECTLY CLASSIFIED INSTANCES**	**INCORRECTLY CLASSIFIED INSTANCES**	**ACCURACY**	**CORRECTLY CLASSIFIED INSTANCES**	**INCORRECTLY CLASSIFIED INSTANCES**
	CHEMISTRY – FULL 18 ATTRIBUTES			**CHEMISTRY – SELECTED 13 ATTRIBUTES**		
ZeroR	47.48%	1358	1502	47.48%	1358	1502
Random Forest	69.93%	2000	860	67.73%	1937	923
PART	67.27%	1924	936	66.89%	1913	947
J48	70.77%	2024	836	71.05%	2032	828
Naïve Bayes	69.69%	1993	867	69.65%	1992	868
	PHYSICS – FULL 18 ATTRIBUTES			**PHYSICS – SELECTED 13 ATTRIBUTES**		
ZeroR	56.85%	1626	1234	56.85%	1626	1234
Random Forest	72.97%	2087	773	74.30%	2125	735
PART	70.80%	2025	835	72.73%	2080	780
J48	75.94%	2172	688	76.43%	2186	674
Naïve Bayes	74.79%	2139	721	76.19%	2179	681
	BIOLOGY – FULL 18 ATTRIBUTES			**BIOLOGY – SELECTED 9 ATTRIBUTES**		
ZeroR	46.64%	1334	1526	46.64%	1334	1526
Random Forest	78.39%	2242	618	76.08%	2176	684
PART	76.96%	2201	659	76.61%	2191	669
J48	80.21%	2294	566	80.14%	2292	568
Naïve Bayes	78.15%	2235	625	79.79%	2282	578
	ADD MATHS – FULL 18 ATTRIBUTES			**ADD MATHS – SELECTED 13 ATTRIBUTES**		
ZeroR	79.69%	2279	581	79.69%	2279	581
Random Forest	82.27%	2352	507	83.18%	2379	481
PART	80.73%	2309	551	81.89%	2342	518
J48	83.15%	2378	482	83.71%	2394	466
Naïve Bayes	82.06%	2347	513	82.90%	2371	489

The prediction models were compared using accuracy as a means to identify the best model, including correctly and incorrectly classified instances. J48 was the best classifier and data mining algorithm for all the chosen prediction models (Roslan & Chen, 2022). J48 is one of the algorithms in the family of decision trees, and this result strengthens the use of decision trees among most of the researchers, as stated by the systematic literature review conducted by Roslan and Chen (2022).

Validate the Chosen Model

20% of validation data will be analyzed alongside training data and test data to see in detail the accuracy of our chosen prediction model. The following results will discuss the detailed accuracy for each STEM subject individually, including the Precision, Recall, F-Measure, and ROC area values.

Table 5.3.
Detailed Accuracy Comparison of the Chosen Model for Each STEM Subject

Subject	CHEMISTRY			PHYSICS			BIOLOGY			ADDITIONAL MATHEMATICS		
Classifier	J48											
Data	Training	Test	Validation	Training	Test	Validation	Training	Test	Validation	Training	Test	Validation
Accuracy	80.51%	78.39%	79.97%	83.78%	83.18%	83.15%	88.17%	88.74%	87.80%	84.95%	84.93%	85.80%
Precision	0.816	0.799	0.815	0.835	0.829	0.831	0.883	0.892	0.880	0.845	0.846	0.861
Recall	0.805	0.784	0.800	0.838	0.832	0.831	0.882	0.887	0.878	0.849	0.849	0.858
F-Measure	0.800	0.777	0.795	0.836	0.830	0.831	0.881	0.888	0.877	0.847	0.848	0.859
ROC Area	0.790	0.768	0.789	0.818	0.823	0.832	0.869	0.868	0.876	0.826	0.833	0.840

By looking into all the results of detailed accuracy in the tables, we can see that the chosen model with the selected attributes and J48 as the classifier showed a very promising accurate prediction. This is because each subject's accuracy was consistent for each training, test, and validation data and above 80% of accuracy (Roslan & Chen, 2022). Moreover, the most detailed accuracy involving Precision, Recall, F-Measure, and ROC area was above 0.8 or 80% (Witten, Frank, & Hall, 2011). The accuracy consistency throughout each training, test, and validation data shows that our chosen model was not overfitting. The model is "overfit" when it has internalized the noise and conformed too closely to the training set. This causes the model to be unable to generalize successfully to data that is not part of the training set. If a model cannot generalize successfully to new data, it cannot accomplish the classification or prediction tasks for which it was designed (Zulkifli, Mohamed, & Azmee, 2019).

Conclusion

This study used data mining to analyze factors affecting students' academic performance in STEM subjects. The study found that student's academic record, including CGPA, previous results, test scores, grades, marks, and attendance, significantly impacted their performance. Decision trees, including J48, were used as the primary algorithm for data mining. J48 was the most effective classifier for predicting student performance with high accuracy. The study suggests that this information could be used to identify at-risk students and improve academic outcomes for students. The study successfully developed a system for predicting student performance and demonstrated its usefulness in real-life academic situations. Although the design has room for improvement, it represents a significant advancement in this field.

References

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. In Technology, Knowledge and Learning (Vol. 24, Issue 4). Springer Netherlands. https://doi.org/10.1007/s10758-019-09408-7

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees: A Case Study. International Journal of Information and Education Technology, 6(7), 528–533. https://doi.org/10.7763/ijiet.2016.v6.745

Ali, G., Jaaffar, A. R., & Ali, J. (2021). STEM Education in Malaysia: Fulfilling SMEs' Expectation. In Modeling Economic Growth in Contemporary Malaysia (pp. 43–57). Emerald Group Publishing Ltd. https://doi.org/10.1108/978-1-80043-806-420211005

Asiah, M., Nik Zulkarnaen, K., Safaai, D., Nik Nurul Hafzan, M. Y., Mohd Saberi, M., & Siti Syuhaida, S. (2019). A Review on Predictive Modeling Technique for Student Academic Performance Monitoring. MATEC Web of Conferences, 255, 03004. https://doi.org/10.1051/matecconf/201925503004

Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting Student Academic Performance at Degree Level: A Case Study. International Journal of Intelligent Systems and Applications, 7(1), 49–61. https://doi.org/10.5815/ijisa.2015.01.05

Barbosa Manhães, L. M., da Cruz, S. M. S., & Zimbrão, G. (2015). Towards automatic prediction of student performance in STEM undergraduate degree programs. 247–253. https://doi.org/10.1145/2695664.2695918

Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016, 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Data Mining and Knowledge Discovery in Databases. Communications of the ACM, 39(11), 24–26. https://doi.org/10.1145/240455.240463

Huang, R., Spector, J. M., & Yang, J. (2019). Educational Technology a Primer for the 21st Century. In Journal of Educational Television (Vol. 1, Issue 2). https://doi.org/10.1080/1358165750010212

Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2019). Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach. Journal of Educational Computing Research, 57(2), 448–470. https://doi.org/10.1177/0735633117752614

Lowis, M., & Castley, A. (2008). Factors affecting student progression and achievement: Prediction and intervention. A two-year study. Innovations in Education and Teaching International, 45(4), 333–343. https://doi.org/10.1080/14703290802377232

Mokhtar, S., Alshboul, J. A. Q., & Shahin, G. O. A. (2019). Towards Data-driven Education with Learning Analytics for Educator 4.0. Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1339/1/012079

Mueen A, Zafar B, Manzoor U. Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. Int J Mod Educ Comput Sci. 2016;8(11):36–42.

Perner, P. (2015). Machine Learning and Data Mining in Pattern Recognition 11th International Conference, MLDM 2015 Hamburg, Germany, July 20-21, 2015 Proceedings. Lecture Notes in Computer Science (Including Subseries Lecture Notes in

Artificial Intelligence and Lecture Notes in Bioinformatics), 9166, 403–414. https://doi.org/10.1007/978-3-319-21024-7

Ramaswami, G., Susnjak, T., Mathrani, A., Lim, J., & Garcia, P. (2019). Using educational data mining techniques to increase the prediction accuracy of student academic performance. Information and Learning Science, 120(7–8), 451–467. https://doi.org/10.1108/ILS-03-2019-0017

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135–146. https://doi.org/10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 40(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Roslan, M. H. bin, & Chen, C. J. (2022). Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021). International Journal of Emerging Technologies in Learning, 17(5), 147–179. https://doi.org/10.3991/ijet.v17i05.27685

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Sumathi, S., & Sivanandam, S. N. (2006). Introduction to Data Mining and its Applications. In Studies in Computational Intelligence (Vol. 29). https://doi.org/10.1007/978-3-540-34351-6

UNESCO. (2017). Cracking the code: Girls' and women's education in science, technology, engineering, and mathematics (STEM). https://unesdoc.unesco.org/ark:/48223/pf0000245656

Witten, I.H., Frank, E. and Hall, M.A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition, Morgan Kaufmann Publishers, Burlington.

Zulkifli, F., Mohamed, Z., & Azmee, N. A. (2019). Systematic research on predictive models on students' academic performance in higher education. International Journal of Recent Technology and Engineering, 8(2 Special Issue 3), 357–363. https://doi.org/10.35940/ijrte.B1061.0782S319