International Journal of Advanced Technology and Engineering Exploration, Vol 9(89)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.875311

Research Article

Comparative analysis of classification algorithm evaluations to predict secondary school students' achievement in core and elective subjects

Hasnah Nawang[1*], Mokhairi Makhtar[2] and Wan Mohd Amir Fazamin Wan Hamzah[3]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia[1]
School of Computer Science, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Tembila Campus, Terengganu, Malaysia[2]
School of Information Technology, Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Tembila Campus, Terengganu, Malaysia[3]

Received: 30-December-2021; Revised: 16-April-2022; Accepted: 19-April-2022

Abstract
Many researchers in educational data mining (EDM) have explored various machine learning techniques in order to predict students' performance. However, the most daunting challenge in classification modelling is selecting the most effective algorithm with the highest accuracy. A study was conducted using datasets from two Malaysian premier secondary schools, Maktab Rendah Sains Mara (MRSM) Kuala Berang and Kuala Terengganu. The purpose of this study is to respond to two key questions; the first is to examine which algorithm is the best in predicting secondary students' achievement in core and elective subjects, while the second is to study whether the same features and algorithms are capable of predicting academic performance based on students' first semester achievement. To do so, this study analysed the effectiveness of six different classification algorithms, which are naïve Bayes (NB), random forest (RF), k-nearest neighbour (kNN), support vector machine (SVM), sequential minimal optimization (SMO), and logistic regression (LGR). Each model's prediction accuracy was evaluated using 10-fold cross validation in order to identify the best model. The results showed that the RF model outperformed other models in terms of accuracy, precision, recall, and F1-Measure. With most algorithms achieving significant accuracy levels for both core and elective subjects' dataset. It is concluded that the prediction of secondary school students' achievement can begin as early as the first semester using RF for core and elective subjects with biology dataset. The accuracy obtained was 96.7% and 97.5%, respectively for the core and elective subjects.

Keywords
Classification, Prediction, Educational data mining, Students' performance predictions.

1.Introduction
Data mining is a mathematical way of data processing that is broadly used in many fields to extract useful insights from data [1]. The implementation of data mining techniques in educational datasets has rapidly evolved into one of the most important factors in the analysis of students' data. As this research area has evolved significantly, various definitions such as educational data mining, academic analytics, learning analytics, teaching analytics, data-driven education, and educational data science are now used to describe the implementation of data mining in educational data [2].

Some of the objectives of educational data mining (EDM) are to investigate educational data to determine the effectiveness of learning systems [3], analyse students' achievement [4−6], and design early warning systems for dropout or failure cases [7, 8].

Traditionally, educational institution datasets capture data from year to year as the number of students increase. The dataset for students may include demographic information, course enrolment, scholarship status, education history, grades and marks, and so on [9]. Due to the growth of data in educational databases, predicting students' achievement has now become a significant challenge for educators. Aside from that, data complexity and machine learning are not areas of expertise of

*Author for correspondence

educators, thus leading to difficulties in data processing for prediction analysis [10]. The massive growth of databases has formed a need for technology development in order to use knowledge and information constructively. However, educators rarely use this information to predict academic achievement, instead they only use it to generate basic reports on students' current achievement based on their grade point average (GPA) or current marks and grades. As a result, the extent to which available and collected data are used is not significant in contributing to the institution's decision-making process [11].

Every subject taught in school or educational institution is important whether it is core or elective since it affects the GPA; therefore, it is critical to ensure that all students are on track with their studies in order to perform well in their high school examinations. Students in Malaysia's upper secondary schools must enrol in five core subjects and four elective subjects, depending on their stream. Students at Maktab Rendah Sains Mara (MRSM) Premier School were solely provided scientific streams. Malay language, English language, History, Islamic education, and Mathematics are among the core subjects covered. Additionally, Additional Mathematics, Chemistry, and Physics are required electives for the scientific stream. Students must also choose whether to enrol in Biology or Accounting based on their preferences.

Many researchers in the field of EDM have previously focused solely on predicting students' success and failure [12] or graduate or not graduate [13] which is covered students' overall performance rather than delving deeper into predictions of how core and elective subjects affects the performance of students. Students' capabilities in some subjects differ as some may tend to excel in core subjects such as language and history and some students may have strength in science subjects such as Chemistry or Biology. The limitations of previous research focused on which algorithm may better predict secondary students' achievement, specifically in core and elective subjects. Motivated researchers to explore this scenario, so that their teachers can cater to their needs by grouping them in different groups based on their profiling [14]. This study also aims to look at whether the same variables and algorithms used in the classification of core and elective subjects could be able to predict academic performance based on students' first semester achievement.

The ability to predict students' achievement will help educators in designing different approaches to teaching, thus catering to efforts to correct the misleading belief on certain subjects' difficulty [15]. With the ability of early prediction for both core and elective subjects, teachers and educators can therefore improve their pedagogy techniques [16] and try to design and deliver content in a more effective way [17]. While school administrators may assist in administering programmes or academic workshops that can improve students' weakness in identified subjects. This knowledge also aids students in their preparation to further their education to the tertiary level based on their strengths in specific subjects. Therefore, this research provides insights to the literature through the main components of the work done, as follows:

1. To identify features that are significant to students' excellency in core and elective subjects.
2. To investigate which machine learning classifiers, perform the best among random forest (RF), naive Bayes (NB), support vector machine (SVM), k-nearest neighbor (kNN), sequential minimal optimization (SMO), and logistic regression (LGR).

This study contributes to the literature in several ways. Firstly, we develop and test six predictive models using data mining and classification techniques to predict upper-secondary students' achievement in core and elective subjects based on their achieved grades. Finally, this study differs from other studies in the field of predicting student performance because this research compares best algorithm apply to the core and elective subject separately.

2.Literature review

EDM techniques comprise classification, clustering, regression, outlier analysis, and association rules. The process of predicting the class of given data points is known as classification. Classification is the most common technique of machine learning algorithms as it is used to recognize and categorize objects. This technique can classify and assign classes to a set of data which then allows for precise estimations. Classes are also known as targets, labels, or categories. The task of estimating a mapping function (f) from discrete input variables (X) to discrete output variables is known as classification predictive modelling (y) [18]. Supervised learning is another phrase for classification. Learning and classification are both involved in the classification task. Training data are analysed by classification algorithms during

the learning phase, while test data are used to approximate a model's accuracy during the classification stage. A classification method's accuracy can be measured in terms of the rule's trustworthiness with the test dataset.

Computational techniques such as decision trees, linear programming, neural networks, and statistics are often used as classification techniques [19]. Previous studies have shown the growth of classification technique's application in analysing students' performance. Recently, many researchers have conducted studies for performance prediction using machine learning in order to gain various outcomes such as identifying students at risk, leaners' behaviour towards e-learning, students' academic achievement, and many more. A systematic review conducted has shown an increase in the number of research on performance prediction within the period from 2015 to 2019. The rising popularity of machine learning is due to the benefits of adapting machine learning to educational datasets, such as the ability to handle large amounts of data, gain new insights from raw data, and adaptability of analytical algorithms to different types of datasets [20].

The two most important factors in predicting student performance are qualities features and prediction methods [21]. In the other words, data mining techniques have been proven to be very effective in predicting students' academic achievement, based on variables used and the selection of appropriate algorithms [22]. Thus, an extensive literature review was conducted to investigate the various algorithms utilized for predicting student performance features that have been used during students' performance forecasting. For instance, [23] performed a multiple regression (MR) algorithm analysis on 478 Moroccan high school students enrolled in the Physics stream from 2015 to 2018 to predict their grades and help the teachers in making choices on whether a student needs reinforcement courses, support, or has difficulty in passing their exams. The attributes used in their research are included demographic features and academic performance features such as marks obtained for each subject and also students' attendance to the class. Their findings demonstrated that the proposed model can make better predictions of students' performance. Another study was conducted by [24] on an educational dataset consisting of 225 instances and ten attributes. Five different classifiers, which are NB, Bayesian network, ID3, J48, and neural network (NN), were used for analysis. The features employed in their study are previous academic performance, students' attendance, students' participation, seminar, lab experiments and final mark. The objective of the study is to predict students' performance using the five different classifiers, and the Bayesian network was shown to have the highest accuracy as compared to the other models, however the most significant features that impacted the performance of the classifier were not highlighted.

Proposed a method of predicting students' final grades using the recurrent neural network (RNN) on the learning data of 108 students using features such as attendance, quiz, course view, word count, slide views and memos [25]. As RNN only works with numbers, grades A, B, C, D, and F were replaced with 95, 85, 75, 65, and 55, respectively. The results showed that RNN's prediction accuracy is greater than 90% until the sixth week, therefore providing the conclusion that RNN can be used to predict final grades since early on in their studies. Carried out a study to predict students' performance using students' public examination results information collected from the Directorate of Higher Secondary Education of Tamil Nadu [26]. Three different classification algorithms, namely NB, kNN, and RF were used to assess the data. The results were then used to examine which classification algorithm is capable of accurately forecasting students' performance. According to the findings, the most accurate algorithm is NB with an increase in accuracy from 83.96% to 98.12% with the help of the AdaBoost algorithm. Furthermore, public examinations feature was a highly useful predictor in forecasting students' performance.

Collected a sample of 635 master's students from a reputable private university in Malaysia's graduate studies college to forecast students' academic achievement, specifically their cumulative grade point average (CGPA) during the postgraduate level [13]. Between the six algorithms used for the study, which were artificial neural network (ANN), SVM, least square regression (LSR), decision tree, Gaussian regression and ensemble method, the results showed that the ANN model performed the best, accounting for 89% of the variation in the students' CGPA. Another study was undertaken by [27] with the goal of analysing the significance of many well-known predictors of academic achievement in higher education. The sample consisted of 162,030 Colombia University students. In evaluation measures like as recall and F1 score, ANN able to beat existing machine-learning methods. Prior

academic achievement (39.5%), students' socioeconomic status (22.8%), university background (15.1%), and high school characteristics (10.2%) are the categories with the greatest contribution to the student's classification in the "high performance" group, according to their findings.

This study discovered that the aforesaid features that are widely employed in students' performance prediction are separated into many groups; demographic, e-learning, social network, school design, academic performance and previous education features [14]. However, there has been little research previously conducted to study either core subjects or elective subjects that are designed to contribute to the students' performance prediction. Due to the fact that few studies focus on the classification performance of core and elective subjects, as many studies focus on GPA [28] or sole subjects such as Mathematics [29] and English [30] as their main features in study, this research will delve deeper into the most significant algorithms that are capable of classifying students' performance specifically in core and elective subjects.

The identification of the appropriate algorithm has become one of the challenges in EDM since the effectiveness of algorithms vary depending on the sets of variables and the amount of data used in the prediction [18]. Thus, this paper suggests an evaluation of a high school dataset using different data mining algorithms and the classification approach in order to forecast students' academic achievement in core and elective subjects. Several algorithms, including NB kNN, RF and SVM have been discussed in terms of their superiority in predicting student performance. The main reason of the selection of these algorithms is that they have been proven to be successful in predicting students' performance using educational datasets. Thus, this study will compare the accuracy and performance of these six algorithms in order to determine the best algorithm for forecasting student performance.

NB classifiers presume that the outcome of an attribute's value on a given class is independent of the other attributes [31]. This is referred to as class conditional independence. Conditional probabilities determine the degree of dependency. Because of its simplicity, computational economy, and excellent performance, NB classifiers are more often utilised in real-world applications [32]. kNN is a representation of the lazy learning algorithm. kNN classifiers work by comparing a given test tuple to training tuples that are close to it, or by learning through analogy. The closest neighbours are taken into account, and thus the class of test element is formed. This algorithm can be applied both in classification and regression approaches. kNN algorithm is widely used in data mining and machine learning because it is easy to use and effective [33]. The RF method is an ensemble of machine learning algorithms built using randomised decision tree algorithms. One significant advantage of Random Forest is that it can be utilised for both classification and regression problems, which are popular in most machine learning systems nowadays [34]. It is the most widely used algorithm because it is simpler to implement and understand than other classification algorithms [35]. Researchers have nevertheless found that SVM performs better than the other classifiers in classification tasks. According to [36], models developed using SVM algorithms are useful in the early forecasting of unsuccessful students with an accuracy of 83%. Simultaneously, their study found that pre-processing techniques are beneficial in enhancing the performance of prediction algorithms.

Once the dependent variable is dichotomous, LGR is the appropriate regression analysis to use. In an educational setting, the logistic model can be used to predict the likelihood of a specific class or event occurring, such as pass or fail, or graduate or not graduate. As shown in a study conducted by [37], LGR outperformed other classifiers in identifying students who might graduate with poor grades or might not graduate at all in the engineering faculty of the Nigerian University with an accuracy of 85.15%. *Table 1* summarises the research on six data mining algorithms that are NB, RF, kNN, SVM, SMO, and LGR that have been applied in educational settings.

Table 1 Literature table based on selected algorithms

Algorithm	Dataset	Accuracy	Reference
RF	Academic and personal information were gathered from three different colleges in Assam, India.	99%	[9]
	A total of 772 students enrolled in e-commerce and e-commerce technology modules at a higher education institution.	88.3%	[38]

Algorithm	Dataset	Accuracy	Reference
	A total of 1054 final research projects from the Mumbai University Science Faculty.	71.48%	[39]
kNN	Demographic and academic features of 76 second-year university students enrolled in a Computer Hardware course.	89%	[40]
	Datasets provided by courses in the bachelor study programmes of the University of Basra's College of Computer Science and Information Technology for the academic years of 2017–2018 and 2018–2019.	63.3%	[41]
SMO	Academic achievement of 2,260 students in Algebra and Geometry courses of the first two years at Lyceum.	89.4%	[42]
	Six different types of learning activities in the learning management system.	82%	[43]
LGR	1,819 medical students from five consecutive cohorts of King Saud bin Abdulaziz University of Health Sciences.	66.7%	[44]
	Two Python programming classes for first-year students at a university in Northern Taiwan.	83%	[45]
NB	Data on 120 students' performance from the spring of 2013.	85.7%	[46]
	A total of 488 second-semester students of a secondary school from 2011 to 2014.	73.4%	[31]
	From 2011 to 2014, a total of 488 students from the second semester of a secondary school was involved.	81.9%	[47].
SVM	An academic dataset of 2,039 students enrolled in the Computer Science and Information College of a Saudi public university between 2016 and 2019.	75.28%	[48]
	An academic dataset gained from the UCI machine learning website, which include academic performance attributes such as Mathematics grades, Portuguese language grades, attendance, grades, time spent studying, and a list of failed subjects.	76.3%	[49]
	Data logged by a technology-enhanced learning (TEL) system called the digital electronics education and design suite (DEEDS).	75%	[50]

3.Methodology

During the experimental phase, a real dataset on upper secondary school students in Malaysia who were enrolled during the period from 2015 to 2018 in two MRSM premier schools was collected. The MRSM Premier Schools selected are MRSM Kuala Berang and MRSM Kuala Terengganu. Each upper secondary student in MRSM premier schools has to register for five core subjects and at least four elective subjects. The students need to complete four semesters and in their fourth semester, they will sit for the trial examination of Sijil Pelajaran Malaysia. The dataset comprised of variables related to students' grades obtained for core and elective subjects during the first semester together with their gender feature. Data were collected and integrated using Microsoft access. The experimental process was divided into three main categories, which are core subjects' dataset, elective with biology dataset and elective with accounting dataset.

3.1Dataset pre-processing and feature selection

During the first phase of data collection, 562 student records were obtained from MRSM Kuala Terengganu, while 579 student records were obtained from MRSM Kuala Berang; all of whom are in the Science stream and they are 16 years old. Details on the number of students based on gender are provided in *Figure 1*. The experimental study was divided into three components, one of which was implemented on the dataset on core subjects and second on the dataset on elective subjects with Biology third is the elective subjects with accounting, as students were given the choice to select either one for their elective subject collection. Malay language, English language, Mathematics, History, and Islamic Education are the core subjects assigned to students in these premier secondary schools. In terms of elective subjects, students must enrol themselves in Additional Mathematics, Physics, and Chemistry. Additionally,

these students have the option of choosing either Biology or Accounting as their fourth elective subject. The grades obtained for the subjects are the grades for semester 1 examination. The dataset was pre-processed before being used in the first stage of the study, prior to applying the EDM techniques. Due to missing and irrelevant data in the dataset, the total of 1141 samples were decreased to 1127 during the pre-processing step. However, when the data were split into elective with Biology and elective with Accounting, more missing data had to be deleted in order to remove noise during the analysis process, reducing the total number of elective data available before splitting into elective with Biology and elective with Accounting to 1094. After the preprocessing stage, the total number of students is 1127 for the core dataset, 891 for elective subjects with Biology dataset, and 203 for elective subjects with accounting dataset. *Figure 1* depicted the gender distribution of each dataset.

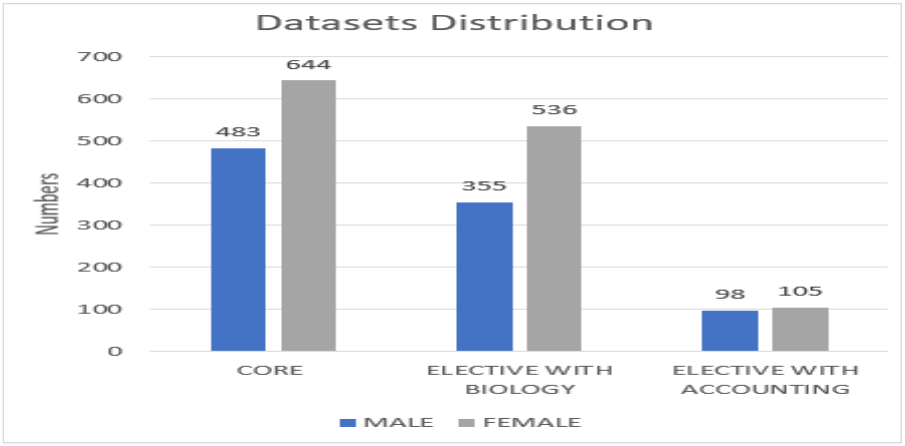

Figure 1 Gender-based distribution of datasets

The next phase conducted was the feature selection process. The goal of feature selection is to decide which attributes are more important in the prediction analysis. As a result, instead of using all features, predictive analysis can be performed with a smaller number of features. Various techniques are available to perform feature selection, and this research chose an algorithm to prioritise the attributes from most significant to least significant. The infogainattributeeval algorithm measures information gained in relation to class to determine the significance of an attribute. Irrelevant attributes such as students' ID, name, and matrix number, students' class, students' homeroom, enrolment year, and MRSM's name were eliminated during the feature selection process. The new features are summarized in *Table 2* for the core subjects' dataset, *Table 3* for elective subjects with accounting, and *Table 4* for elective subjects with biology. The correlation matrix was used to determine the relationship between students' grades and their class. *Figure 2* depicts the plot of the correlation matrix for the core subjects' dataset, which demonstrates a positive association between two grade features and the target variable, grade for Mathematics with a value of 0.73 and grade for History with a value of 0.5. In terms of elective subjects with accounting dataset, the Chemistry subject has a high relationship with the class, with a correlation value of 0.72. Additional Mathematics, Physics, and Accounting each have a positive correlation with the target of 0.62, 0.6, and 0.5, respectively. Chemistry and additional Mathematics both demonstrated a strong correlation once more with the class target in the elective with biology dataset, with the correlation values of 0.73 and 0.72, respectively. While Biology and Physics demonstrated a positive association with correlation values greater than 0.5, at 0.68 and 0.66 in both.

Table 2 List of variables for core subjects' dataset

Feature	Type	Description
Jantina	Nominal	Students' gender (male or female)
BM	Nominal	Grade for the subjects Malay Language (A+, A, B+B, C+, D, E, F)
BI	Nominal	Grade for the subjects English Language (A+, A, B+B, C+, D, E, F)
PI	Nominal	Grade for the subjects Islamic Education (A+, A, B+B, C+, D, E, F)
SJ	Nominal	Grade for the subjects History (A+, A, B+B, C+, D, E, F)

Feature	Type	Description
MM	Nominal	Grade for the subjects Mathematics (A+, A, B+B, C+, D, E, F)
GPA	Float	Students' GPA for the first semester
Class	Nominal	Class Sijil Kelas Pertama (SKP) (Excellent) Class Sijil Kelas Dua Atas (SKDA) (Good) Class Sijil Kelas Dua Bawah 1 (SKDB1) (Satisfactory) Class Sijil Kelas Dua Bawah 2 (SKDB2) (Poor)

Table 3 List of variables for elective subjects with accounting dataset

Feature	Type	Description
Jantina	Nominal	Students' gender (male or female)
MT	Nominal	Grade for the subjects Additional Mathematics (A+, A, B+B, C+, D, E, F)
FZ	Nominal	Grade for the subjects Physics (A+, A, B+B, C+, D, E, F)
KM	Nominal	Grade for the subjects Chemistry (A+, A, B+B, C+, D, E, F)
PA	Nominal	Grade for the subjects Accounting (A+, A, B+B, C+, D, E, F)
GPA	Float	Students' GPA for the first semester
Class	Nominal	Class SKP (Excellent) Class SKDA (Good) Class SKDB1(Satisfactory) Class SKDB2 (Poor)

Table 4 List of variables for elective subjects with biology dataset

Feature	Type	Description
Jantina	Nominal	Students' gender (male or female)
MT	Nominal	Grade for the subjects Additional Mathematics (A+,A,B+B,C+,D,E,F)
FZ	Nominal	Grade for the subjects Physics (A+,A,B+B,C+,D,E,F)
KM	Nominal	Grade for the subjects Chemistry (A+,A,B+B,C+,D,E,F)
BO	Nominal	Grade for the subjects Biology (A+,A,B+B,C+,D,E,F)
GPA	Float	Students' GPA for the first semester
Class	Nominal	Class SKP (Excellent) Class SKDA (Good) Class SKDB1(Satisfactory) Class SKDB2 (Poor)

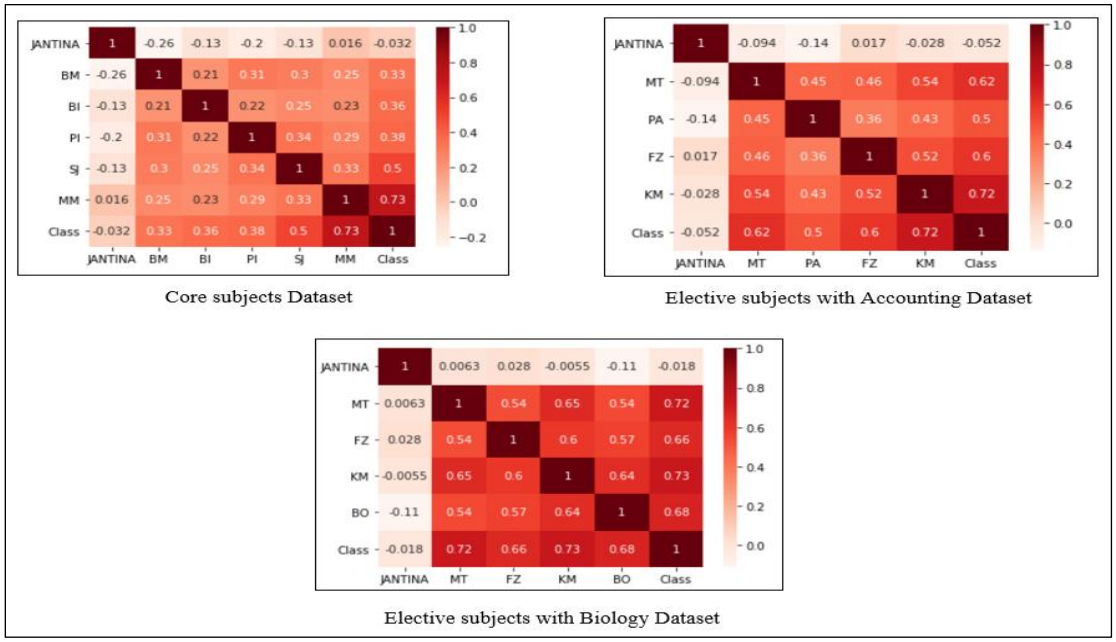

Figure 2 Correlation values for each dataset

3.2Classification

Six prediction models were developed by utilising six well-known data mining and classification algorithms which are RF, NB, SVM, kNN, SMO, and LGR. Each model was developed using 10-fold cross validation which was used for training purposes, while the rest of the folds were used for testing. The 10-fold cross-validation procedure is a resampling technique used to assess models using hidden data. This process was repeated ten times, and the models' performance can be interpreted using the ten evaluation scores. This 10-fold cross-validation technique is common since it reduces both testing bias and variance in sparse data. The classification performance of the algorithms can vary according to the characteristics of the dataset. The Waikato environment for knowledge analysis (WEKA) software and Python were used for model implementation because it is freely accessible to the public and is commonly used for data mining research work. This study was carried out in accordance with the suggested framework as shown in *Figure 3* with the number of enrolled students as mention in *Figure 1*.

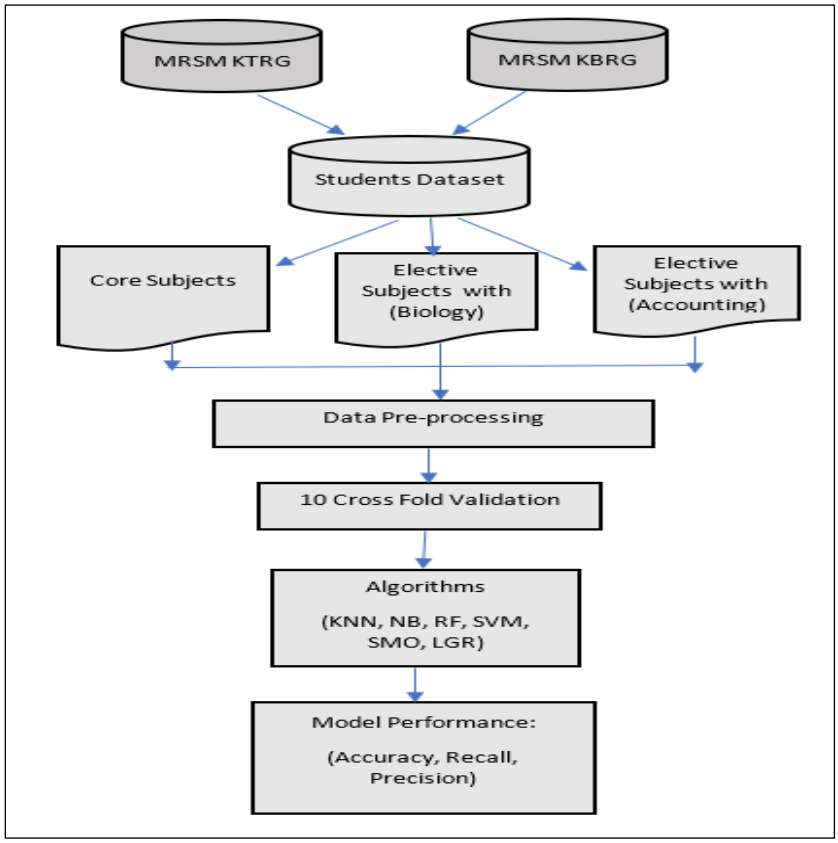

Figure 3 Research framework design

3.3Evaluation metrics

Accuracy, precision, and recall were used to measure the performance of the data mining models as formulated in Equations 1, 2, and 3, respectively. The percentage of correctly predicted results was referred to as accuracy, the percentage of positives that were correctly predicted as positive was referred to as recall, while precision was described as the percentage of correct positive findings.

- TP - the number of attributes that are correctly predicted as positive.
- FP - the number of attributes that are incorrectly predicted as positive.
- TN - the number of attributes that are correctly predicted as negative.
- FN - the number of attributes that are incorrectly predicted as negative.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$Recall = TP/(TP/TN) \quad (2)$$

$$Precision = TN/\ (TP\ /\ FN) \quad (3)$$

4.Results

To assess the models, the authors separated the integrated dataset into three CSV files that are: core subjects, elective subjects with biology, and elective subjects with accounting. All of the experiments were carried out using data from the first semester of the secondary school's upper level in order to classify their performance in the final (fourth) semester. Six analyses were conducted to determine which classifier would best forecast student performance using the collected data. The first three sets of analysis were conducted on the core subjects, elective subjects with biology, and elective subjects with accounting without the feature selection process. The effects of various feature selection techniques on classification performance were observed in the three subsequent analyses. In all analyses, the classification process used the identical classification algorithms, performance metrics, and 10-fold cross validation approach. The use of the infogainattributeval was proven to enhance the performance of each algorithm by 1.5% to 3%.

Figure 4 depicts a comparison of each algorithm's performance in the core subjects, elective subjects with accounting, and elective subjects with biology datasets following the pre-processing and feature selection phases. Once the set of data was examined, it was discovered that the RF classifier performed well in terms of accuracy as it achieved the highest accuracy for two datasets, which are the core subjects and elective subjects with biology datasets, at 96.7% and 97.5% accuracy, respectively. The second-best accuracy is achieved by the NB model, which achieved the highest accuracy for the elective subjects with accounting dataset with 91.3% accuracy. It also scored 89.58% accuracy for the core subjects' dataset and 87.65% accuracy for electives with biology dataset. For the LGR algorithm, the results demonstrated that it the method is capable of achieving good accuracy for core subjects and elective subjects with biology datasets, where both methods achieved more than 80% of accuracy at 82.54% and 81.59%, respectively. However, the accuracy of prediction for the elective subjects with accounting dataset is low at only 63.13% accuracy. The SVM model achieved the lowest accuracy for the elective subjects with accounting dataset at 60.63%, while kNN achieved the lowest accuracy for the core subjects' dataset at 65.7% and the elective subjects with biology dataset at 71.88%. Based on the performance of the algorithms, it can be concluded that RF and NB are the best algorithms that can be used to predict the three sets of data, with each algorithm achieving greater than 85% accuracy in those three datasets.

Table 3 displays the evaluation metrics for the core subjects, while *Table 4* displays the evaluation metrics for the elective subjects that include accounting, and the elective subjects that include biology. Again, RF outperformed others in terms of precision metrics, recall, and F1-Measure, as well as for the core subjects and elective subjects, including the biology dataset. SMO recorded the lowest precision, recall and F1-Measure, as the values recorded for all the three-evaluation metrics for the elective subjects including the accounting dataset did not even reach 0.5.

Figure 5 illustrated the confusion matrix for the highest accuracy achieved for the three datasets discuss in this study. Label (A) denotes dataset core subjects for which RF achieved 96.7% accuracy performance. According to the confusion matrix, RF can accurately classify all cases in class b (referring to SKDA: Good) and class c (referring to SKDB1: Satisfactory). This method has also misclassified 29 samples in class a (refer to SKP: good) and 9 samples in class d. (Refer to SKDB2: class poor). Meanwhile, Label (B) in *Figure 5* represents an elective with biology dataset. As presented in *Figure 4*, NB achieves the highest accuracy for this dataset, with 91.3% accuracy performance, and while there is misclassification for each class, it is less than 10% indicating that the predictions are not poor. RF was able to achieve the maximum accuracy performance for elective using accounting dataset once again, and the label (C) demonstrated how well this algorithm is able to predict the correct class for SKP, SKDA, and SKDB1. Because of the sample size for SKDB2 (poor) class is limited in these three datasets, there is a high likelihood of inaccuracy in forecasting the correct class for it.

Table 3 Evaluation metrics for core subjects

Algorithms	Precision	Recall	F1
NB	0.895	0.943	0.992
SVM	0.927	0.685	0.832
KNN	0.741	0.711	0.801
RF	1	0.993	1

Algorithms	Precision	Recall	F1
SMO	0.888	0.933	0.975
LGR	0.842	0.822	0.94

Table 4 Evaluation metrics for elective subjects

	Elective subjects include accounting			Elective subjects include biology		
Algorithms	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NB	0.8	0.899	0.842	0.897	0.863	0.88
SVM	0.613	1	0.76	0.735	0.945	0.827
KNN	0.667	0.444	0.533	0.793	0.798	0.79
RF	1	0.611	0.759	0.99	0.993	0.992
SMO	0.24	0.333	0.279	0.8909	0.819	0.862
LGR	0.474	0.5	0.486	0.83	0.872	0.851

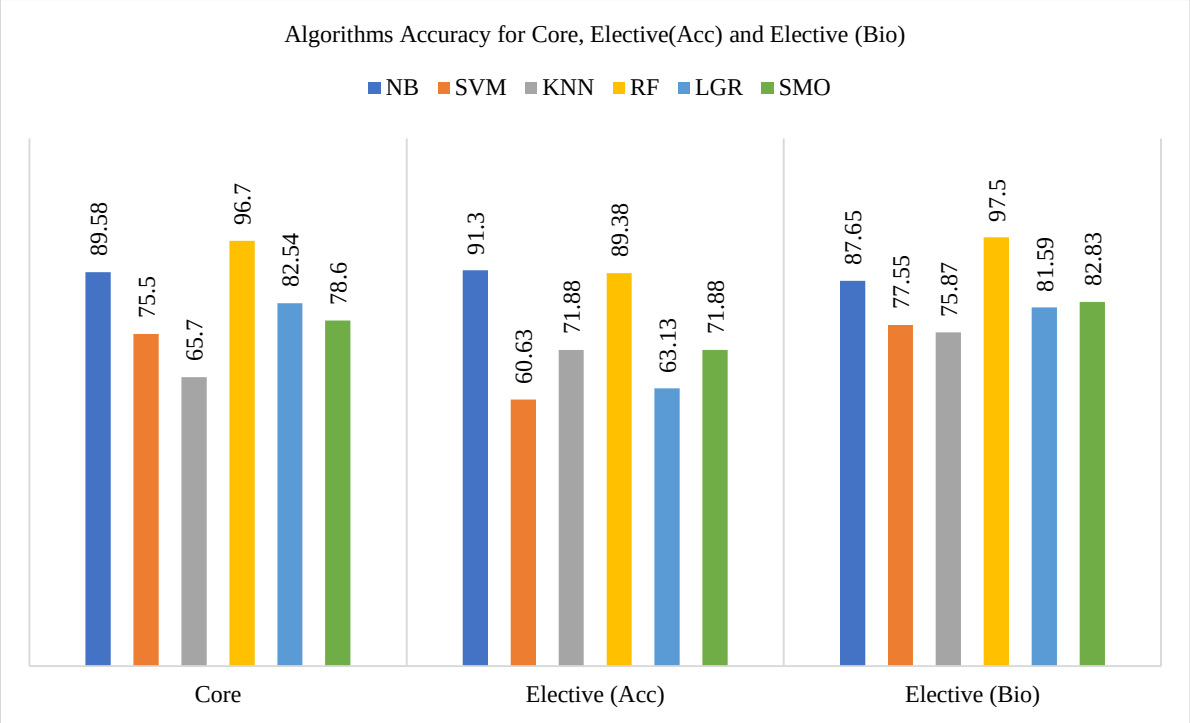

Figure 4 Accuracy of each algorithm in predicting the core, electives with accounting and electives with biology datasets

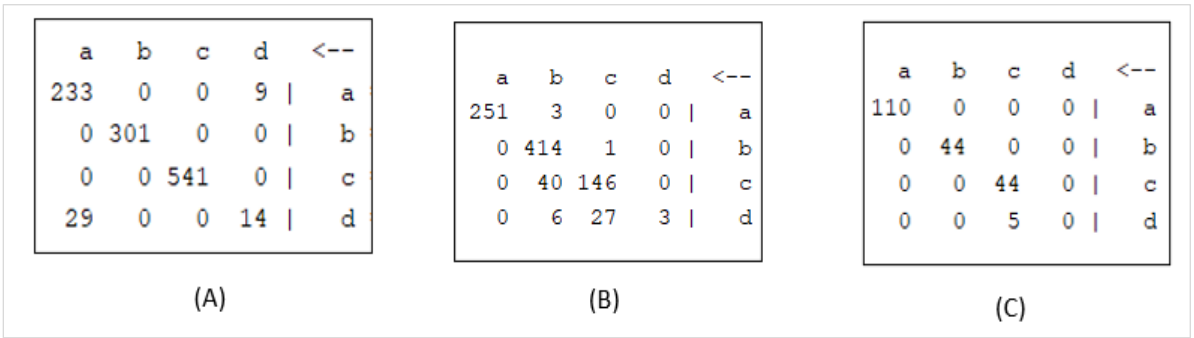

Figure 5 Confusion matrix for the best algorithms in each dataset

Figures 6 to *8* demonstrate how the variables related to the class target. Because the data was processed in both WEKA and Phyton, the study needed to convert the categorical value of some variables such as gender, grades, and the target class. The details graph plot: 0 for Male, 1 for Female, grades for each subject A+, A, B+, B, C+, C, D, E, F were converted to 0, 1, 2, 3, 4, 5, 6, 7, 8, and target class SKP (Excellent), SKDA (Good), SKDB1 (Satisfactory), SKDB2 (Poor) were transformed to 1, 2, 3 and 4. Each graph in each figure depicts the relationship between target class and gender, as well as target class and grades. Gender attributes in these three statistics indicate that male performed better than female for those three datasets. According to the core dataset depicted in *Figure 6*, as the grades increase (meaning the lower the grades as the grades are ordered from 0 to 8 represent A+ to F), the class target also increases, implying that students perform worse.

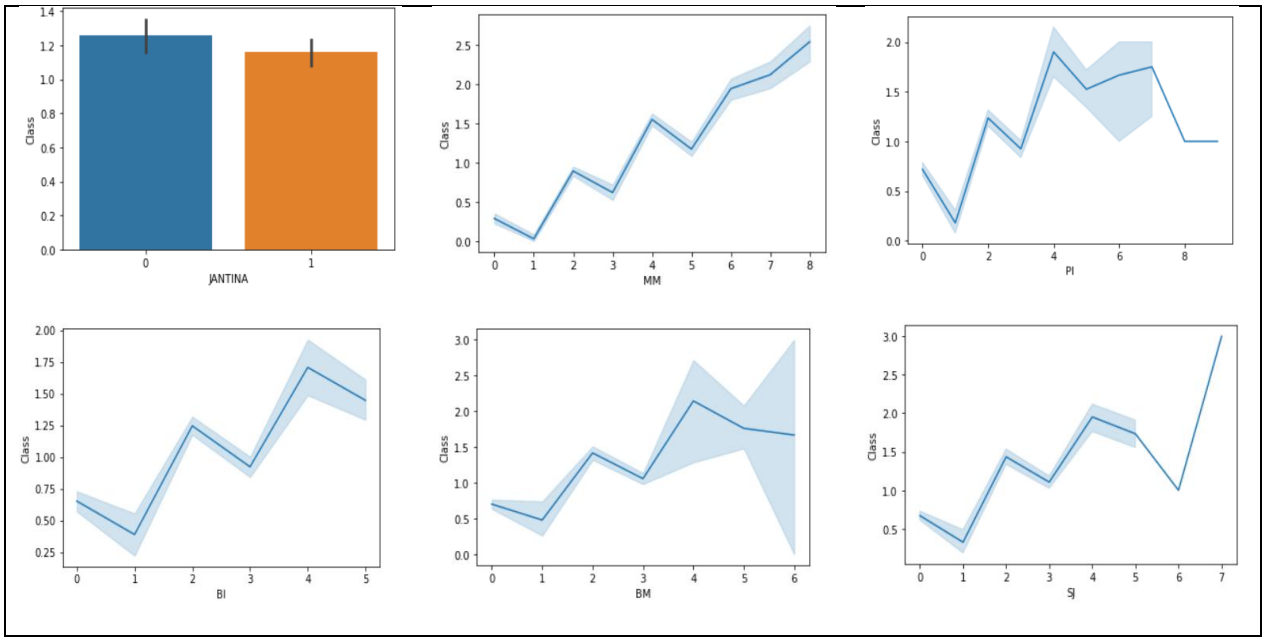

Figure 6 Graph plot describes the variables associated with the class for the core subject's dataset

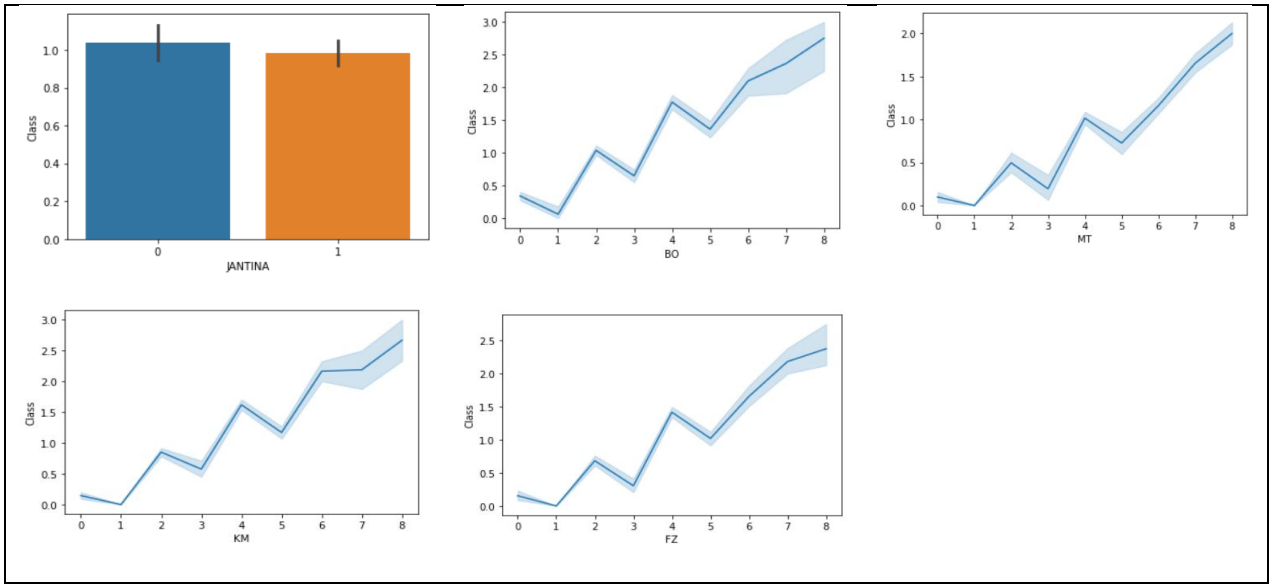

Figure 7 Graph plot describes the variables associated with the class for elective with biology subjects dataset

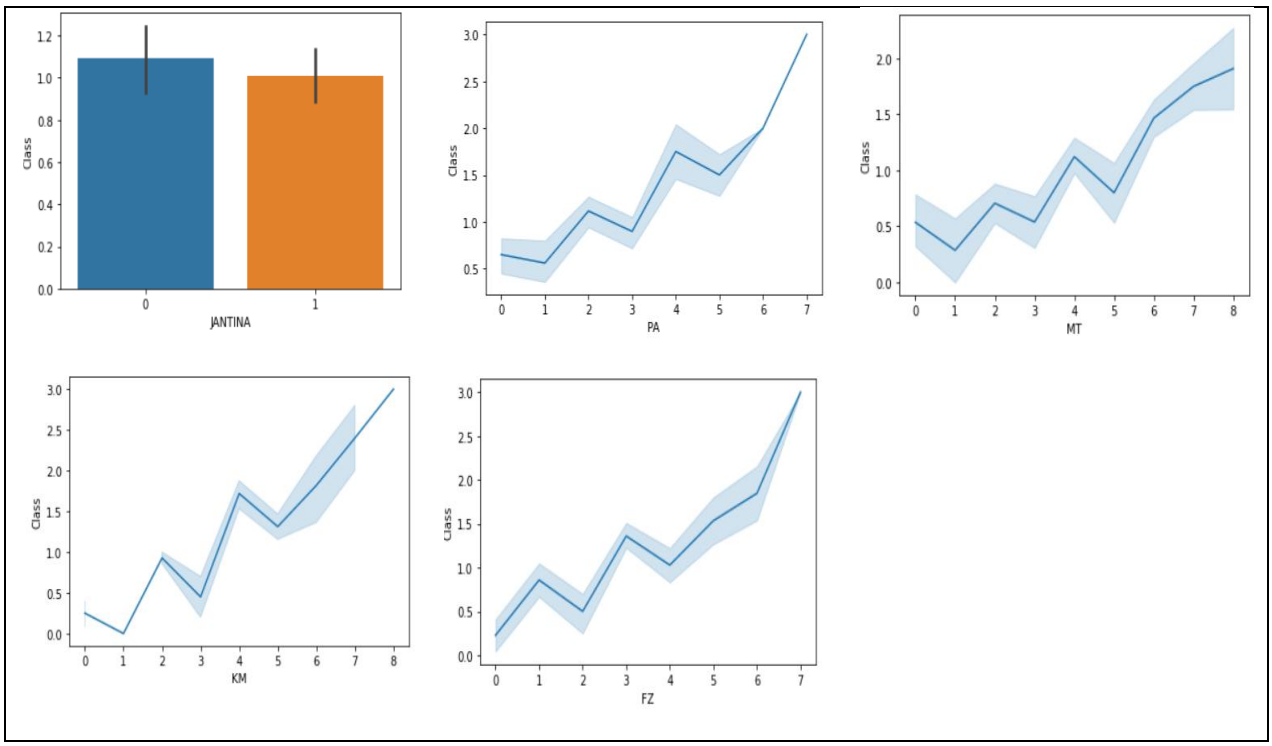

Figure 8 Graph plot describes the variables associated with the class for elective with an accounting subject's dataset

5.Discussion

The majority of students enroll in the scientific stream because they want to pursue tertiary education in the medical and engineering fields. Nevertheless, based on the correlation matrix readings and the graph plot that show the association of the subject's features to the target class, to be in science stream. It is a requirement for upper secondary students to have a strong foundation and knowledge in Mathematics. Mathematics has been shown to be closely associated with the excellent achievement in the core subjects' dataset, with the highest correlation matrix of 0.73. While additional Mathematics has achieved positive readings in the elective subject with accounting and elective subject with biology datasets, with correlation matrices of 0.72 and 0.62, respectively. Then, for those three datasets, it is reasonable to conclude that students who excel in Mathematics have a higher possibility of attaining SKP in their final exams (excellent). The gender attribute has the lowest correlation values in three datasets, with values less than 0.01. Surprisingly, gender has no substantial impact on the class target.

Based on the classification accuracy, RF algorithm had indeed outperformed other algorithms in predicting students' performance in core and elective subjects with Biology. When it comes to elective subjects with accounting, the maximum classification accuracy appears to be achieved by the NB classifier. In terms of precision, recall, and F1-Measure, the performance of RF still outscored other classifiers in core and elective subjects with biology, but for elective subjects with accounting, NB showed the highest Recall and F1-Measure. According to the findings of this study, students' performance in the first semester can be utilised to predict students' class performance in the final semester at MRSM Premier Schools. Focusing on the accuracy and the classification errors, it may be concluded that the RF classification algorithm is the most suited algorithm for the core and elective subject with biology whereas NB is the best model to predict elective subjects with accounting.

The contributions of the study are threefold. First, during feature selection, this study reveals that gender features do not significant to the students' performance prediction. Second, this study able to identify that RF is the algorithm that is suitable in for predicting core subject's dataset and elective subjects with Biology dataset, whereas NB are the techniques that are significant in predicting students' performance for elective subjects with accounting.

Third, this study able to identify that Mathematics and History are the subjects most associate to the excellent performance of students for core subject's dataset through correlation matrix and for elective subject with biology and elective subject with accounting dataset, subjects such as Additional Mathematics and Chemistry showed the consistency of lower grades in those subjects will contribute to the lower achievement for final semester examination performance. Last but not least, this study differs from other studies in the field of predicting student performance because this research compares best algorithm apply to the core and elective subject separately.

Students who choose to enroll in the science stream during their upper secondary school should evaluate a few things, such as whether their mathematical ability is sufficient to complete their high school in the science stream with excellent performance. The limitation of this work is regarding the limited sample size of 203 students for elective subjects with accounting dataset, and the samples for SKDB2 class (poor) are too small to train the data, and it is concerned that this will result in a poor approximation for the prediction. A complete list of abbreviations is shown in *Appendix I.*

6.Conclusion and future work

Modeling a student's academic performance in high school as early as the first semester is a useful strategy for assessing whether a student requires reinforcing courses or extra assistance, as well as assisting teachers in creating student profiles. For this purpose, the students' academic performance was assessed using academic and demographic data captured from two different premier high schools in Terengganu, Malaysia. This dataset, which has been collected from real student data from 2015 to 2018, was submitted to six different classification methods which are NB, RF, kNN, SVM, SMO, and LGR. The experiment results indicate that the RF algorithm is the best machine learning classifier for classifying student performance in core and elective subjects with biology dataset, with 96.7% and 97.5% accuracy, accordingly. Aside from that, our study's findings revealed that there is a significant relationship between students' strength in specific subjects and their excellency in core and elective subjects.

This offers some recommendations to improved decision-making in the education sector when confronted with unpredictable conditions. In future studies, we plan to expand our dataset with larger and more diverse sample sizes to include more data from high schools, which will be mined using various classification methods. In addition, we wish to improve the accuracy score by applying machine learning and deep learning techniques (e.g., multilayer perceptron (MLP), long short-term memory (LSTM)) to the educational dataset.

Acknowledgment

This research was supported by the Research Management and Innovation Centre, Universiti Sultan Zainal Abidin (UniSZA).

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

Hasnah Nawang: Conceptualization, methodology, investigation, analysis, interpretation of results, writing original draft. **Mokhairi Makhtar:** Validation of the models, interpretation of results, framework of methodology, project administration. review and editing. **Wan Mohd Amir Fazamin Wan Hamzah:** Validation of the analytics model, project administration, reviewing and editing.

References

[1] Aziz AA, Starkey A. Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches. IEEE Access. 2019; 8:17722-33.

[2] Romero C, Ventura S. Educational data mining and learning analytics: an updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2020; 10(3).

[3] Alhassan A, Zafar B, Mueen A. Predict students' academic performance based on their assessment grades and online activity data. International Journal of Advanced Computer Science and Applications. 2020; 11(4).

[4] Aydoğdu Ş. Predicting student final performance using artificial neural networks in online learning environments. Education and Information Technologies. 2020; 25(3):1913-27.

[5] Asif R, Merceron A, Ali SA, Haider NG. Analyzing undergraduate students' performance using educational data mining. Computers & Education. 2017; 113:177-94.

[6] Yassein NA, Helali RG, Mohomad SB. Predicting student academic performance in KSA using data mining techniques. Journal of Information Technology & Software Engineering. 2017; 7(5):1-5.

[7] Oeda S, Hashimoto G. Log-data clustering analysis for dropout prediction in beginner programming classes. Procedia Computer Science. 2017; 112:614-21.

[8] Tasnim N, Paul MK, Sattar AS. Identification of drop out students using educational data mining. In

international conference on electrical, computer and communication engineering 2019 (pp. 1-5). IEEE.
[9] Hussain S, Dahan NA, Ba-alwib FM, Ribata N. Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science. 2018; 9(2):447-59.
[10] Flore PC, Mulder J, Wicherts JM. The influence of gender stereotype threat on mathematics test scores of Dutch high school students: a registered report. Comprehensive Results in Social Psychology. 2018; 3(2):140-74.
[11] Wise AF, Jung Y. Teaching with analytics: towards a situated model of instructional decision-making. Journal of Learning Analytics. 2019; 6(2):53-69.
[12] Mai TT, Bezbradica M, Crane M. Learning behaviours data in programming education: community analysis and outcome prediction with cleaned data. Future Generation Computer Systems. 2022; 127:42-55.
[13] Baashar Y, Hamed Y, Alkawsi G, Capretz LF, Alhussian H, Alwadain A, et al. Evaluation of postgraduate academic performance using artificial intelligence models. Alexandria Engineering Journal. 2022; 61(12):9867-78.
[14] Nawang H, Makhtar M, Hamzah WM. A systematic literature review on student performance predictions. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(84):1441-53.
[15] Cornillez JEE, Treceñe JK, De LSJR. Mining educational data in predicting the influence of mathematics on the programming performance of university students. Indian Journal of Science and Technology. 2020; 13(26):2668-77.
[16] Tsai YS, Gasevic D. Learning analytics in higher education-challenges and policies: a review of eight learning analytics policies. In proceedings of the seventh international learning analytics & knowledge conference 2017 (pp. 233-42).
[17] Joshi A, Desai P, Tewari P. Learning Analytics framework for measuring students' performance and teachers' involvement through problem based learning in engineering education. Procedia Computer Science. 2020; 172:954-9.
[18] Lang C, Siemens G, Wise A, Gasevic D. Handbook of learning analytics. New York: SOLAR, Society for Learning Analytics and Research; 2017.
[19] Hartama D, Windarto AP, Wanto A. The application of data mining in determining patterns of interest of high school graduates. In journal of physics: conference series 2019 (pp.1-6). IOP Publishing.
[20] Ndukwe IG, Daniel BK. Teaching analytics, value and tools for teacher data literacy: a systematic and tripartite approach. International Journal of Educational Technology in Higher Education. 2020; 17(1):1-31.
[21] Niyogisubizo J, Liao L, Nziyumva E, Murwanashyaka E, Nshimyumukiza PC. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization. Computers and Education: Artificial Intelligence. 2022.
[22] Karlos S, Kostopoulos G, Kotsiantis S. Predicting and interpreting students' grades in distance higher education through a semi-regression method. Applied Sciences. 2020; 10(23):1-19.
[23] Qazdar A, Er-raha B, Cherkaoui C, Mammass D. A machine learning algorithm framework for predicting students performance: a case study of baccalaureate students in Morocco. Education and Information Technologies. 2019; 24(6):3577-89.
[24] Almarabeh H. Analysis of students' performance by using different data mining classifiers. International Journal of Modern Education and Computer Science. 2017; 9(8):9-15.
[25] Okubo F, Yamashita T, Shimada A, Ogata H. A neural network approach for students' performance prediction. In proceedings of the seventh international learning analytics & knowledge conference 2017 (pp. 598-9).
[26] Navamani JM, Kannammal A. Predicting performance of schools by applying data mining techniques on public examination results. Research Journal of Applied Sciences, Engineering and Technology. 2015; 9(4):262-71.
[27] Rodríguez-hernández CF, Musso M, Kyndt E, Cascallar E. Artificial neural networks in academic performance prediction: systematic implementation and predictor evaluation. Computers and Education: Artificial Intelligence. 2021.
[28] Aiken JM, De BR, Hjorth-jensen M, Caballero MD. Predicting time to graduation at a large enrollment American university. Plos one. 2020; 15(11):1-28.
[29] Hoogland K, De KJ, Bakker A, Pepin BE, Gravemeijer K. Changing representation in contextual mathematical problems from descriptive to depictive: the effect on students' performance. Studies in Educational Evaluation. 2018; 58:122-31.
[30] Fok WW, He YS, Yeung HA, Law KY, Cheung KH, Ai YY, et al. Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine. In international conference on information management 2018 (pp. 103-6). IEEE.
[31] Mokhairi M, Nawang H, Wan SN. Analysis on students performance using naïve. Journal of Theoretical and Applied Information Technology. 2017; 31(16):3993-4000.
[32] Patil R, Tamane S. A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes. International Journal of Electrical and Computer Engineering. 2018; 8(5):3966-75.
[33] Pandey A, Jain A. Comparative analysis of KNN algorithm using various normalization techniques. International Journal of Computer Network and Information Security. 2017; 9(11):36-42.
[34] Costa-mendes R, Oliveira T, Castelli M, Cruz-jesus F. A machine learning approximation of the 2015 Portuguese high school student grades: a hybrid

approach. Education and Information Technologies. 2021; 26(2):1527-47.
[35] Priyam A, Abhijeeta GR, Rathee A, Srivastava S. Comparative analysis of decision tree classification algorithms. International Journal of Current Engineering and Technology. 2013; 3(2):334-7.
[36] Gil PD, Da CMS, Moro S, Costa JM. A data-driven approach to predict first-year students' academic success in higher education institutions. Education and Information Technologies. 2021; 26(2):2165-90.
[37] Adekitan AI, Salau O. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon. 2019; 5(2).
[38] Hasan R, Palaniappan S, Mahmood S, Abbas A, Sarker KU, Sattar MU. Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. Applied Sciences. 2020; 10(11):1-20.
[39] Viloria A, López JR, Leyva DM, Vargas-mercado C, Hernández-palma H, Llinas NO, et al. Data mining techniques and multivariate analysis to discover patterns in university final researches. Procedia Computer Science. 2019; 155:581-6.
[40] Akçapınar G, Altun A, Aşkar P. Using learning analytics to develop early-warning system for at-risk students. International Journal of Educational Technology in Higher Education. 2019; 16(1):1-20.
[41] Hashim AS, Awadh WA, Hamoud AK. Student performance prediction model based on supervised machine learning algorithms. In IOP conference series: materials science and engineering 2020 (pp. 1-18). IOP Publishing.
[42] Livieris IE, Kotsilieris T, Tampakas V, Pintelas P. Improving the evaluation process of students' performance utilizing a decision support software. Neural Computing and Applications. 2019; 31(6):1683-94.
[43] Tsiakmaki M, Kostopoulos G, Kotsiantis S, Ragos O. Implementing AutoML in educational data mining for prediction tasks. Applied Sciences. 2019; 10(1):1-27.
[44] Baars GJ, Stijnen T, Splinter TA. A model to predict student failure in the first year of the undergraduate medical curriculum. Health Professions Education. 2017; 3(1):5-14.
[45] Hung HC, Liu IF, Liang CT, Su YS. Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. Symmetry. 2020; 12(2):1-14.
[46] Marbouti F, Diefes-dux HA, Madhavan K. Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education. 2016; 103:1-15.
[47] Akçapınar G, Hasnine MN, Majumdar R, Flanagan B, Ogata H. Developing an early-warning system for spotting at-risk students by using eBook interaction logs. Smart Learning Environments. 2019; 6(1):1-15.
[48] Mengash HA. Using data mining techniques to predict student performance to support decision making in university admission systems. IEEE Access. 2020; 8:55462-70.
[49] Zohair LM. Prediction of Student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education. 2019; 16(1):1-18.
[50] Hussain M, Zhu W, Zhang W, Abidi SM, Ali S. Using machine learning to predict student difficulties from learning session data. Artificial Intelligence Review. 2019; 52(1):381-407.

Hasnah Nawang completed her BSc in Computer Science from Universiti Putra Malaysia, in 2006 and MSc in Computer Science from Universiti Sultan Zainal Abidin, Terengganu, Malaysia in 2018. Currently, she is a PhD scholar at Department of Computer Science in Faculty of Computing and Informatics, Universiti Sultan Zainal Abidin, Terengganu, Malaysia. She is also a teacher in secondary school in the Department of Mathematics and Computer Science since 2007. Her current research interests include Machine Learning, Educational Data Mining and Deep Learning.
Email: hasnah.nawang@gmail.com

Dr. Mokhairi Makhtar received his PhD from University of Bradford, United Kingdom in 2012. He is currently a Professor in the Department of Computer Science, Universiti Sultan Zainal Abidin, Terengganu, Malaysia. His current research interests include Machine Learning, Ensemble Method, Data Mining, Soft Computing, Timetabling and Optimisation, Natural Language Processing, E-Learning and Deep Learning.
Email: mokhairi@unisza.edu.my

Dr. Wan Mohd Amir Fazamin Wan Hamzah received his PhD from Universiti Malaysia Terengganu, Malaysia. He is currently a lecturer in Universiti Sultan Zainal Abidin. His research interests include Learning Analytics, Gamification, E-Learning and Cloud Computing.
Email: amirfazamin@unisza.edu.my

Appendix I

S. No.	Abbreviations	Descriptions
1	CGPA	Cumulative Grade Point Average
2	DEEDS	Digital Electronics Education and Design Suite
3	DM	Data Mining
4	EDM	Educational Data Mining
5	GPA	Grade Point Average
6	kNN	k-Nearest Neighbor
7	LGR	Logistic Regression
8	LSR	Least Square Regression

9	LSTM	Long Short-Term Memory Network
10	ML	Machine Learning
11	MLP	Multilayer Perceptron
12	MR	Multiple Regression
13	MRSM	Maktab Rendah Sains Mara
14	NB	Naïve Bayes
15	NN	Neural Network
16	RF	Random Forest
17	SMO	Sequential Minimal Optimization
18	SKDA	Sijil Kelas Dua Atas
19	SKDB1	Sijil Kelas Dua Bawah 1
20	SKDB2	Sijil Kelas Dua Bawah 2
21	SKP	Sijil Kelas Pertama
22	SVM	Support Vector Machine
23	TEL	Technology-Enhanced Learning
24	WEKA	Waikato Environment for Knowledge