



Predicting students' performance in English and Mathematics using data mining techniques

Muhammad Haziq Bin Roslan¹ · Chwen Jen Chen¹ 

Received: 26 May 2022 / Accepted: 20 July 2022 / Published online: 29 July 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

This study attempts to predict secondary school students' performance in English and Mathematics subjects using data mining (DM) techniques. It aims to provide insights into predictors of students' performance in English and Mathematics, characteristics of students with different levels of performance, the most effective DM technique for students' performance prediction, and the relationship between these two subjects. The study employed the archival data of students who were 16 years old in 2019 and sat for the Malaysian Certificate of Examination (MCE) in 2021. The learning of English and Mathematics is a concern in many countries. Three main factors, namely students' past academic performance, demographics, and psychological attributes were scrutinized to identify their impact on the prediction. This study utilized the Orange software for the DM process. It employed Decision Tree (DT) rules to determine the characteristics of students with low, moderate, and high performance in English and Mathematics subjects. DT and Naïve Bayes (NB) techniques show the best predictive performance for English and Mathematics subjects, respectively. Such characteristics and predictions may cue appropriate interventions to improve students' performance in these subjects. This study revealed students' past academic performance as the most critical predictor, as well as a few demographics and psychological attributes. By examining top predictors derived using four different classifier types, this study found that students' past Mathematics performance predicts their MCE English performance and students' past English performance predicts their MCE Mathematics performance. This finding shows students' performances in both subjects are interrelated.

Keywords Educational data mining · Performance prediction · Secondary education · Data mining techniques · English · Mathematics

✉ Chwen Jen Chen
cjchen@unimas.my

¹ Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Malaysia

1 Introduction

Educational Data Mining (EDM) involves the use of data mining (DM) techniques on data collected via various educational systems (Berland et al., 2014) to improve educational quality (Baker & Yacef, 2009). Recent EDM studies primarily focus on analyzing students' learning and behavior, analyzing educational strategies and interventions, predicting students' performance and dropout, and providing personalized recommendations to students (Romero & Ventura, 2020).

Students' academic performance is one of the main priorities for educators in determining educational success at all educational levels (Tan et al., 2019). Predicting students' performance enables early interventions and support to be devised and implemented by educators to enhance their performance, particularly for students who are at risk of failure (Lu et al., 2018; Wakelam et al., 2020). Despite the increasingly growing popularity of DM in various educational settings, EDM studies are still rather minimal and are not discussed in-depth, particularly in developing countries such as Malaysia (Jamil et al., 2018; Shahiri & Husain, 2015). The handful of prediction-related EDM studies done in the Malaysian school context includes Khairy et al. (2018), Kiu (2018), and Makhtar et al. (2017). Hence, very minimal is known about the academic performance predictors for students of this region.

The most common DM methods used in EDM research are classification and clustering (Mohamad & Tasir, 2013). In previous EDM studies, the two types of data modeling techniques used are descriptive and predictive models (Dutt et al., 2017). The common techniques used in the classification method are Decision Tree (DT), Neural Networks (NN), Naïve Bayes (NB), and Support Vector Machine (SVM) while the techniques used in the clustering method are fuzzy clustering, K-means algorithms, and discrimination analysis (Salem & Parusheva, 2018). Saa et al. (2019) discovered that DT is the most utilized technique in EDM to predict students' performance. DT is often used because it is easy to understand and has high predictive accuracy (Sathe & Adamuthe, 2021).

MCE is a compulsory national examination that is equivalent to the International General Certificate of Secondary Education (IGSCE). This examination is meant for 17 years old, Form Five students in Malaysia. In recent years, students' performance in English and Mathematics has been an issue of concern in Malaysia. According to statistics released by the Malaysia Ministry of Education, in 2018 and 2019, approximately one out of five students failed these subjects (Kementerian Pendidikan Malaysia, 2020). Although studies on the relationship between English and Mathematics have emerged in literature over the years such as Awan et al. (2011), Bagececi et al. (2014), Perez and Alieto (2018), and Stoffelsma and Spooren (2019), no known study is harnessing on data mining techniques to gain insights into it. Hence, this study aims to identify factors that predict students' MCE performance in English and Mathematics subjects, identify characteristics of students at different levels of performance, and gain more insights into the relationship between these subjects using DM techniques.

1.1 Factors that predict students' performance

Dutt et al. (2017) and Saa et al. (2019) conducted reviews of EDM literature particularly on predicting students' performance. These reviews reported that numerous studies identify factors that predict students' academic performance and the best DM method to be utilized in each context. The recent literature review by Alyahyan and Düşteğör (2020) concluded that there are five factors that researchers frequently examined when predicting students' performance. These include past academic performance (Adekitan & Salau, 2019), demographics (Atlay et al., 2019), learning environment (Mueen et al., 2016), psychological attributes (Khairy et al., 2018), and e-learning activities (Almeda et al., 2018).

Past academic performance is the most used factor to predict students' performance (Saa et al., 2019). Shahiri and Husain (2015) noted that final grades or GPA is the most significant attribute used to predict students' performance. Based on 30 studies from 2003 to 2014, Shahiri and Husain (2015) reported that one-third of them used GPA as the key factor in student performance prediction. Similarly, based on the review of 36 research papers from 2009 to 2018, Saa et al. (2019) identified the two most common factors in predicting students' performance as their previous grades and internal assessments. Therefore, there is a notable justification to include past academic performance in the present study.

Next, demographic factors are identified to be important elements for predicting students' performance (Alyahyan & Düşteğör, 2020). Hoe et al. (2013) reported a strong connection between students' demographics and their academic performance. Other studies have shown that gender is one of the student demographics that significantly affect students' performance (Saa et al., 2019). Nevertheless, some studies reported no critical effect of gender (Martínez & Gil, 2019) and students' social condition (Maghari & Mousa, 2017) on academic performance. In short, it is crucial to further examine demographic factors as there is a gap in existing knowledge on how demographics affect students' performance.

Prior studies have shown that psychological attributes are significant predictors of students' performance (Gray et al., 2014). Khairy et al. (2018) found that students' cognitive abilities influence their academic performance during secondary school. Other studies (Garg, 2018; Putpuek et al., 2018) focus on psychological attributes such as students' interest, study behavior, stress, anxiety, time concern, self-regulation, and motivation. These studies found that students' performance is affected by these factors. Although some attempts have been made to deal with these attributes, this psychological dimension has received limited attention in the literature (Saa et al., 2019). Hence, it is important to include psychological attributes in the present study to discover more about how this factor affects students' performance.

Many EDM studies examine the influence of one single factor such as past examination marks studies (Asif et al., 2017) or demographics (Atlay et al., 2019) on students' academic performance (Saa et al., 2019). Although combining many factors in a study has the potential to increase the predictive accuracy of DM algorithms (Mishra et al., 2014), the review study by Shahiri and Husain (2015) discovered very limited studies that examine how the combination of such factors affects the prediction of students' academic performance.

Moreover, most past studies are limited to factors that are frequently studied (i.e., past academic performance, demographics) (Namoun & Alshantqi, 2021). This creates an interesting gap to be examined in this study in which past academic performance, demographics, and psychological attributes are examined simultaneously to gain insights into their relevance to students’ performance prediction. However, this study excludes the learning environment factor. According to Alyahyan and Düşteğör (2020), some attributes related to this factor include program type, class type, and semester period. In this regard, the present study’s sample only involved some Malaysian public schools of similar locality and physical classroom settings. Moreover, the students involved were all in the same grade and took the MCE national examination in the same year. Hence, the present study did not examine the learning environment factor. It also excludes the e-Learning activities factor as e-Learning is not compulsory in Malaysian secondary school education before the Covid-19 pandemic struck.

1.2 Research framework

Figure 1 shows the research framework of this study. Malaysian upper secondary schools employ a grading system that comprises 10 different grades (A+, A, A-, B+, B, C+, C, D, E, and G). In this study, students’ results were divided into three performance categories: high (A+, A, A-), moderate (B+, B, C+, C), and low (D, E, G). This study examined past academic performance, demographics, and

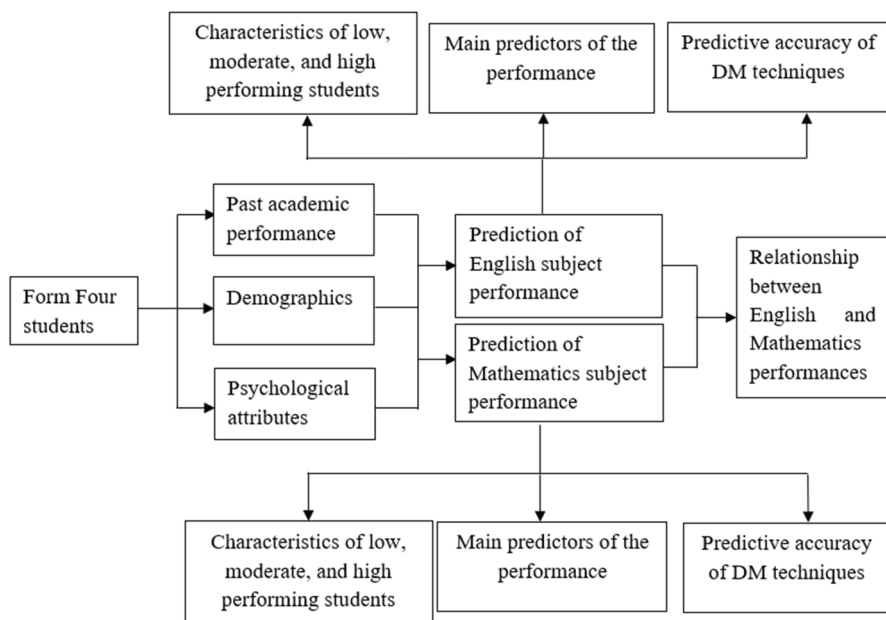


Fig. 1 Research framework

psychological attributes of Form Four students to identify prominent DT rules that define the characteristics of students with low, moderate, and high MCE performance in each of the examined subjects. The study also examined the main predictors of their MCE performance in English and Mathematics subjects respectively and determined the algorithm with the highest predictive accuracy. The study then examined the relationship between English and Mathematics performance through correlation analysis as well as examining whether students' past performance in one subject serves as a predictor for their performance in another subject.

The following lists the research questions of this study:

RQ1: What are the DT rules that define the characteristics of students with low, moderate, and high performance in the English subject respectively?

RQ2: What are the DT rules that define the characteristics of students with low, moderate, and high performance in the Mathematics subject respectively?

RQ3: What are the main predictors of students' English subject performance using DT, NN, SVM, and NB classification techniques?

RQ4: What are the main predictors of students' Mathematics subject performance using DT, NN, SVM, and NB classification techniques?

RQ5: Which predictive model has the highest predictive accuracy for students' performance in English and Mathematics subjects respectively?

RQ6: What is the relationship between students' performance in English and Mathematics subjects?

2 Methods & materials

2.1 Research design

Adapting to the Knowledge Discovery in Database (KDD) process proposed by Fayyad et al. (1996), the research process was divided into five main phases which are (i) problem identification and data selection; (ii) data preparation and pre-processing; (iii) data transformation; (iv) DM model, and (v) data interpretation.

Firstly, the problem identification was accomplished through the review of EDM literature. The researcher attempted to predict students' performance in English and Mathematics subjects as there is only a handful of EDM-based predictive studies on both subjects, specifically, and in this region, generally. For example, Lau et al. (2019) employed artificial neural networks, a DM technique, to predict students' academic performance in a Chinese university based on their entrance examination results, which include English and Mathematics. Saleh et al. (2021) used DM techniques to predict Libyan high school students' performance based on their grades in Mathematics and English subjects. Moreover, no known EDM study examines the relationship between both subjects.

The data collected were cleaned and pre-processed as there is a possibility that the data contained errors, such as missing values. This process ensured that students' information was complete. Then, students' data were transformed.

In the data transformation phase, the non-numerical data was converted to the numerical format. This is to improve the prediction effectiveness (Bergquist et al., 2019).

The fourth phase is creating a DM model. This phase explains how to predict student performance using DM techniques. The DM model was developed using training and testing with a ratio of 8:2, which means 80% of the data was used for training while 20% was used for testing. The training and testing data are divided into an 8:2 ratio. By doing so, the predictive accuracy of the algorithms reaches the highest value. If the ratio is changed to 7:3 or 6:4, the predictive accuracy will decrease. The 8:2 split draws its justification from the well-known Pareto principle (Joseph, 2022). The training set was what the DM algorithms trained on, and the test set was used to see how well classifiers performed on unseen data. Figure 2 illustrates the process of developing the DM model in Orange software to implement the prediction.

Then, in the data interpretation phase, the confusion matrix as shown in Table 1 was generated to show how the data was classified, either correctly or incorrectly (Sokolova & Lapalme, 2009). The matrix contains information about actual and predicted values. Mueen et al. (2016) define the terms as follows:

- TP: Number of True Positives (the number of instances that are positive and predicted to be positive)
- FP: Number of False Positives (the number of instances that are negative but predicted to be positive)
- FN: Number of False Negatives (the number of instances that are positive but predicted to be negative)

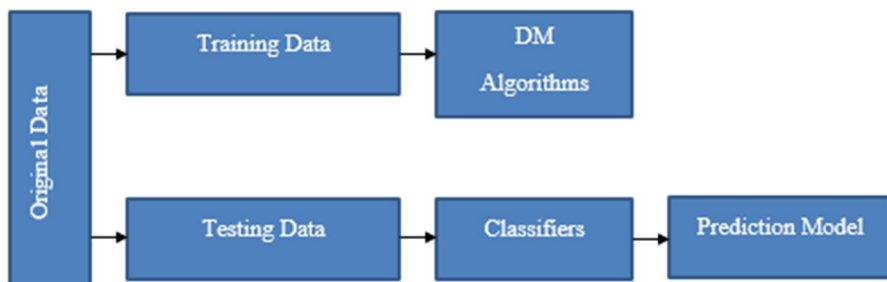


Fig. 2 DM model development process

Table 1 Confusion matrix

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

- TN: Number of True Negatives (the number of instances that are negative and predicted to be negative)

Based on the confusion matrix, the following measures (Goutte & Gaussier, 2005) were used to evaluate the prediction performance of each algorithm.

- Area Under Curve expresses how well the classifier distinguishes between positive and negative samples.
- Accuracy is the overall proportion of the total number of correct predictions.
- Accuracy (%) = $(TP+TN)/(TP+FP+FN+TN) \times 100$
- Precision is the proportion of correct positive observations.
- Precision (%) = $TP/(TP+FP) \times 100$
- Recall is the proposition of actual positive that is correctly predicted as positive.
- Recall (%) = $TP/(TP+FN) \times 100$
- F-score is a measure of a test's accuracy. F-score is the weighted harmonic mean of the precision and recall.
- F-Score (%) = $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) \times 100$

DT algorithm in Orange was used to identify prominent DT rules that define the characteristics of students with low, moderate, or high performance in English and Mathematics respectively. DT algorithm was used because it provides a tree presentation that is easy to read and interpret without requiring statistical knowledge (Berhanu & Abera, 2015). The tree diagram was confined to only four levels to incorporate only the most significant attributes affecting students'

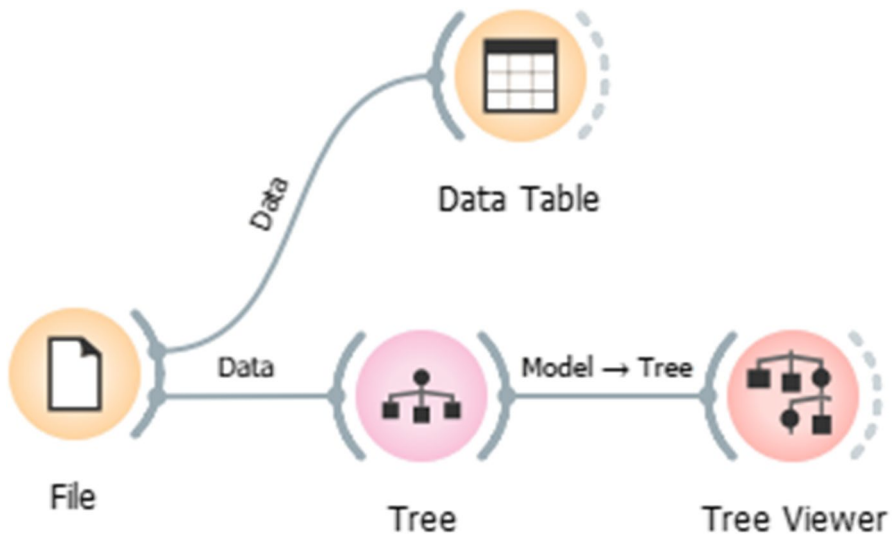


Fig. 3 Predicting students' performance process using DT

performance. Figure 3 shows the process of predicting students' performance using DT.

Orange was employed to identify the main predictors of students' performance in English and Mathematics using different DM techniques. To find out the impact of attributes on the prediction of English and Mathematics performance using different DM techniques, the 'information gain' scoring method in Orange was used. This method shows the rank of attributes that influence students' performance. The rank of the attributes provided insight into the relationship between students' performance in English and Mathematics subjects. In addition, Spearman's rank correlation, r_s was used to measure the strength of the linear relationship between these subjects using SPSS software (Bakhtiar et al., 2021).

2.2 Sample

This study involved the use of archival data of Form Four students from four different urban secondary schools after obtaining consent from the Ministry of Education and respective school principals. These schools were purposively chosen and fulfilled four criteria. First, all four schools are public schools. Second, these schools are within the accessible population, which are schools located in the center of Georgetown in Penang, Malaysia. Third, these schools kept the needed archival data, and fourth, the school principals granted permission to use the archival data of their respective schools. Initially, the data of all 202 Form Four students in 2019 were gathered from all schools. These students sat for their SPM examination in early 2021. However, some of the data contained missing values. Hence, only 159 Form Four students' data were retained for analysis after data cleaning.

DM techniques are often used to extract implicit, hidden, and potentially useful information from a large database. The use of relatively small sample size is due to the limited data available from the accessible population. Hence, the results of this study are not meant to be generalized to other populations. The study by Sordo and Zeng (2005) that compared the performance of three learning algorithms (NB, SVM, and DT) using training sets with sizes ranging from 150 to 8500 cases shows that the three classifiers perform reasonably well with a small number of cases, although NB tends to perform better than SVM and DT with small datasets. Studies such as Chapman and Haug (1999) who used a dataset size of 150, Meng and Zhao (2015) who used a dataset size of 11, and Wilcox and Hripcsak (1999) who used a dataset size of 200 are evidence of the feasibility to apply DM techniques on small dataset sizes. Algarni (2016) also pointed out that as opposed to traditional research methods that require a large dataset size, DM can use a limited dataset size to build a prediction model. The data of Form Four students in 2019 were employed and these students sat for their MCE examination in early 2021.

Table 2 List of data required

Factors	Attributes
Past academic performance	Form Four English and Mathematics mid-term examinations results Form Four English and Mathematics final examinations results Actual MCE English and Mathematics results
Demographics	Gender: Male (121), Female (81) Ethnicity: Malay (186), Chinese (7), Indian (8), Others (1) Religion: Muslim (189), Buddhist (4), Hindu (6), Christian (3), Others (0) Parents' occupational status (i.e., Permanent, Temporary): Father's job—Permanent (111), Temporary (48); Mother's job – Permanent (87), Temporary (72) Parents' educational level (i.e., MCE, Diploma, Bachelor, Master, Ph.D.): Father's education – MCE (75), Diploma (36), Bachelor (35), Master (8), Ph.D. (5); Mother's education – MCE (80), Diploma (24), Bachelor (37), Master (12), Ph.D. (6) Parents' marital status: Married (148), Divorced (10), Widowed (9)
Psychological attributes	Fifteen attributes were examined via a psychometric test (Hamzah, 2011) that was administered by the Ministry of Education, Malaysia on all Form Four students. These attributes include autonomy, creativity, aggression, extrovert, achievement, diversity, intellectual, leadership, structure, resilience, help, analytical, self-criticism, vision, and transparency. Each attribute was measured on a 10-point scale. A score between 1 to 3 is considered Low, 4 to 6 is considered Moderate and 7 to 10 is considered High

2.3 Data collection

Table 2 shows the data required for each student.

3 Results & discussion

3.1 Rules defining students' of different performance levels in English and Mathematics

Referring to RQ1 and RQ2, the DT algorithm has produced key rules that define the characteristics of low, moderate, and high-performing students in English and Mathematics respectively. DT algorithm was chosen because it is more powerful than the other DM algorithms in terms of classification (Charbuty & Abdula-zeez, 2021). Furthermore, it can handle big datasets, and the visual representation of DT makes it easy for the researcher to grasp the results which are vital to determining students' characteristics (Mühlbacher et al., 2017). These rules can enlighten stakeholders, notably teachers, on the appropriate teaching approaches or interventions particularly to increase the performance of at-risk students (García-Jiménez et al., 2020).

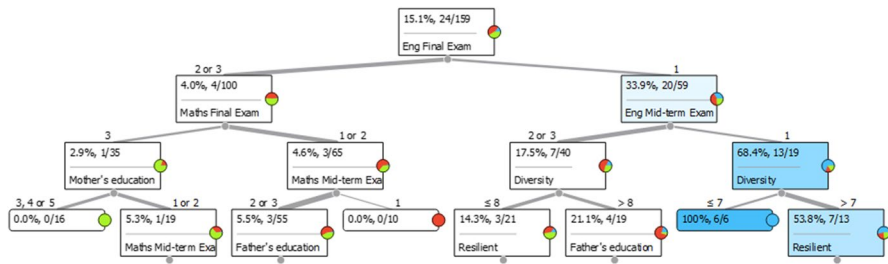


Fig. 4 Tree diagram of students with low English performance

RQ1: What are the DT rules that define the characteristics of students with low, moderate, and high performance in the English subject respectively?

Figure 4 shows the tree diagram of students who are expected to have low performance in English, involving 24 out of 159 students (15.1%). From the figure, two prominent rules that define students with the highest probability of low performance in English are identified. These rules are depicted as below:

IF (English Final Exam=Low) AND (English Mid-term Exam=Low) AND (Diversity \leq 7) THEN MCE English=Low.

IF (English Final Exam=Low) AND (English Mid-term Exam=Low) AND (Diversity $>$ 7) THEN MCE English=Low.

Based on the above DT rules, three dominant attributes are involved, which are Form Four final and mid-term examinations and diversity. In the first rule with the highest probability, students who had low performance in English and did not receive a high score of diversity (below 7 points) are more likely to have low performance in English for MCE. According to Yung and Cai (2020), past English examination results can generally predict students' future performance in the subject. Krupat et al. (2017) posited that students with low performance in previous academic examinations are at risk as there is a strong relationship between past and future academic performance. Moreover, students with low or moderate diversity levels indicate their preference toward routines, are less flexible, moderately elastic, and are not interested in trying new things. Such preference may affect their ability to answer higher-order thinking MCE questions correctly. This shows that personality traits (diversity) can be used to predict students' performance. This is supported by Miller-Matero et al. (2018) who posited a relationship between personality traits and students' performance.

Figure 5 shows the tree diagram of students who are expected to have moderate performance in English, involving 70 out of 159 students (44.0%). From the figure, three prominent rules that define students with the highest probability of moderate performance in English are identified. These rules are depicted as below:

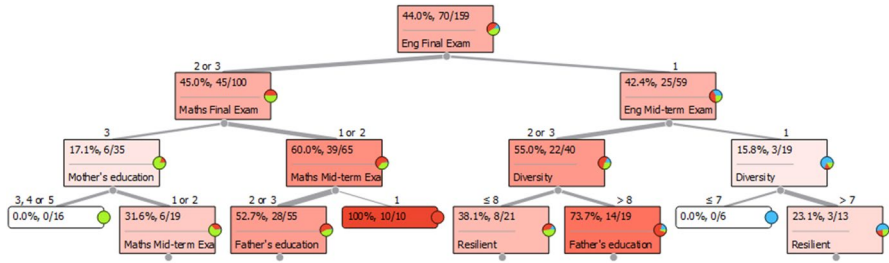


Fig. 5 Tree diagram of students with moderate English performance

IF (English Final Exam=Moderate OR High) AND (Mathematics Final Exam=Low OR Moderate) AND (Mathematics Mid-term Exam=Low) THEN MCE English=Moderate.

IF (English Final Exam=Moderate OR High) AND (Mathematics Final Exam=Low OR Moderate) AND (Mathematics Mid-term Exam=Moderate OR High) THEN MCE English=Moderate.

IF (English Final Exam=Low) AND (English Mid-term Exam=Moderate OR High) AND (Diversity > 8) THEN MCE English=Moderate.

Five attributes are involved in the rules for moderate MCE English performance: Form Four English final and mid-term examinations, Form Four Mathematics final and mid-term examinations, and diversity. According to the rule with the highest probability, students who have a moderate or high performance in the English final examination, low or moderate performance in the Form Four Mathematics final examination, and low performance in the Form Four Mathematics mid-term examination, are likely to perform moderately in English for MCE. The rule with the highest probability for moderate performance contains only past academic performance. This finding is aligned with Makhtar et al. (2017) who found that past academic performance solely can predict secondary school students’ performance.

Figure 6 shows the tree diagram of students who are expected to have high performance in English, involving 65 out of 159 students (40.9%). From the figure,

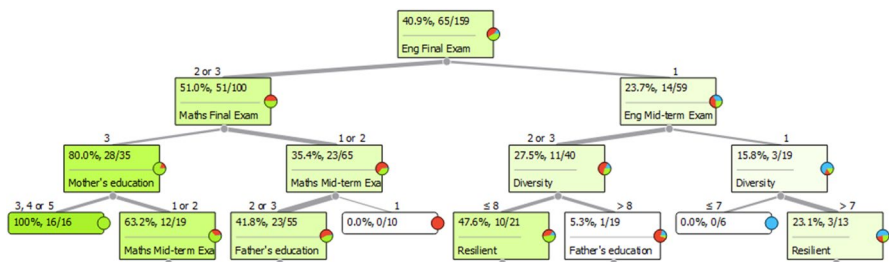


Fig. 6 Tree diagram of students with high English performance

two prominent rules that define students with the highest probability of high performance in English are identified. These rules are depicted as below:

IF (English Final Exam=Moderate OR High) AND (Mathematics Final Exam=High) AND (Mother’s educational level=Bachelor OR Master OR Ph.D.) THEN MCE English=High.

IF (English Final Exam=Moderate OR High) AND (Mathematics Final Exam=High) AND (Mother’s educational level=MCE OR Diploma) THEN MCE English=High.

Three attributes are influential in the rules for high MCE English performance: Form Four English and Mathematics final examinations and the mother’s educational level. Based on the rule with the highest probability, students who had moderate or high performance in Form Four English final examination, high performance in Form Four Mathematics final examination, and whose mothers have a Bachelor, Master, or Ph.D. degree are probably going to perform highly in MCE English. Based on Sulphey et al. (2018), students who had excellent grades in previous examinations are more likely to perform well in future examinations. Idris et al. (2020) revealed that mothers with high educational levels positively contribute to their children’s academic performance.

RQ2: What are the DT rules that define the characteristics of students with low, moderate, and high performance in the Mathematics subject respectively?

Figure 7 shows the tree diagram of students who are expected to have low performance in Mathematics, involving 24 out of 159 students (15.1%). From the figure, one prominent rule (see below) that defines students with the highest probability of low Mathematics performance is identified.

IF (Mathematics Final Exam=Low) AND (Mathematics Mid-term Exam=Low) AND (Father’s educational level=MCE OR Bachelor OR Master OR Ph.D.) THEN MCE Mathematics=Low.

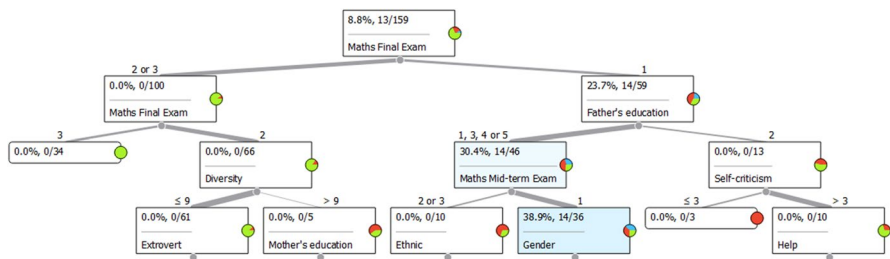


Fig. 7 Tree diagram of students with low Mathematics performance

The three dominant attributes involved in the rules are Form Four Mathematics final and mid-term examination, and the father’s educational level. Based on the rule, students who had low performance in Mathematics examinations and their father’s educational level is MCE, Bachelor, Master, or Ph.D. degree, indicating that either they either had an MCE certificate or studied post-secondary education, are more likely to have low performance in MCE Mathematics. A range of fathers’ educational levels is identified in this rule, and this may indicate the lack of involvement of fathers in their children’s education that contributes to low performance. Wong and Wong (2019) discovered that there is a significant relationship between low Mathematics performance and later examination performance. They argued that students have low performance due to a lack of enthusiasm for the subject. Muller (2018) also reported a substantial association between parents’ educational levels and their children’s academic performance.

Figure 8 shows the tree diagram of students who are expected to have moderate Mathematics performance, involving 36 out of 159 students (22.6%). From the figure, three prominent rules that define students with the highest probability of moderate performance in Mathematics are identified. These rules are depicted as below:

- IF (Mathematics Final Exam = Low) AND (Father’s educational level = MCE OR Bachelor OR Master OR Ph.D.) AND (Mathematics Mid-term Exam = Moderate or High) THEN MCE Mathematics = Moderate.
- IF (Mathematics Final Exam = Low) AND (Father’s educational level = Diploma) AND (Self-criticism \leq 3) THEN MCE Mathematics = Moderate.
- IF (Mathematics Final Exam = Moderate) AND (Diversity > 9) THEN MCE Mathematics = Moderate.

The rules involve five attributes: Form Four Mathematics final and mid-term examinations, the father’s educational level, diversity, and self-criticism. Based on the rule with the highest probability, if students have low performance in Form Four Mathematics final examination, their fathers’ educational level is MCE, Bachelor, Master, or Ph.D. degree, and have moderate or high performance in Form Four Mathematics mid-term examination, they are probably going to perform moderately in Mathematics for MCE. This finding aligns with

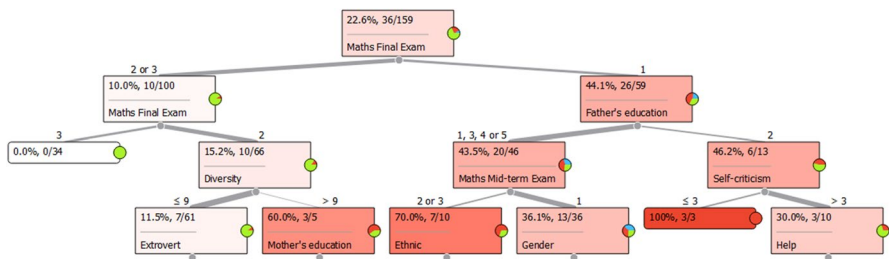


Fig. 8 Tree diagram of students with moderate Mathematics performance

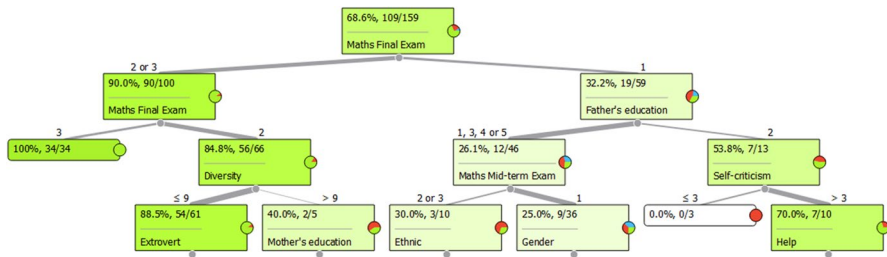


Fig. 9 Tree diagram of students with high Mathematics performance

the study by Chand et al. (2021) who discovered that past Mathematics performance of upper secondary school students can be used to predict future Mathematics performance. Li and Qiu (2018) reported that parents' educational level affects students' performance. However, the range of fathers' educational levels discovered in this rule implies equal or lack of fathers' involvement in educating their children.

Figure 9 shows the tree diagram of students who are expected to have high performance in Mathematics, involving 109 out of 159 students (68.6%). From the figure, three prominent rules are identified. These rules are depicted as below:

- IF (Mathematics Final Exam = High) THEN MCE Mathematics = High.
 IF (Mathematics Final Exam = Moderate) AND (Diversity \leq 9) THEN MCE Mathematics = High.
 IF (Mathematics Final Exam = Low) AND (Father's educational level = Diploma) AND (Self-criticism $>$ 3) THEN MCE Mathematics = High.

Four attributes are dominant in the above rules: Form Four Mathematics final examination, diversity, the father's educational level, and self-criticism. Based on the rule with the highest probability, students with high performance in the Form Four Mathematics final examination most probably will achieve high performance in MCE Mathematics. This finding is supported by Lipnevich et al. (2016) who found that previous Mathematics performance is substantially correlated with present Mathematics performance. Moreover, McKee and Caldarella (2016) concluded that early Mathematics performance greatly predicts later performance.

3.1.1 Predictors of students' performance in English and Mathematics

RQ3: What are the main predictors of students' English subject performance using DT, NN, SVM, and NB classification techniques?

Figures 10, 11, 12, and 13 show the process of identifying the main predictors of English performance using DT, NN, SVM, and NB respectively in Orange.

Table 3 shows the impact of the top five attributes on the prediction of English performance using DT, NN, SVM, and NB.

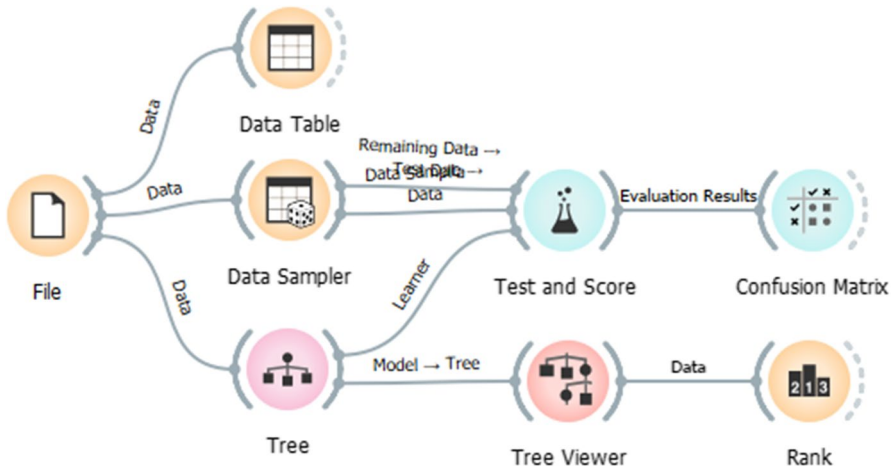


Fig. 10 Process of identifying main predictors using DT in orange

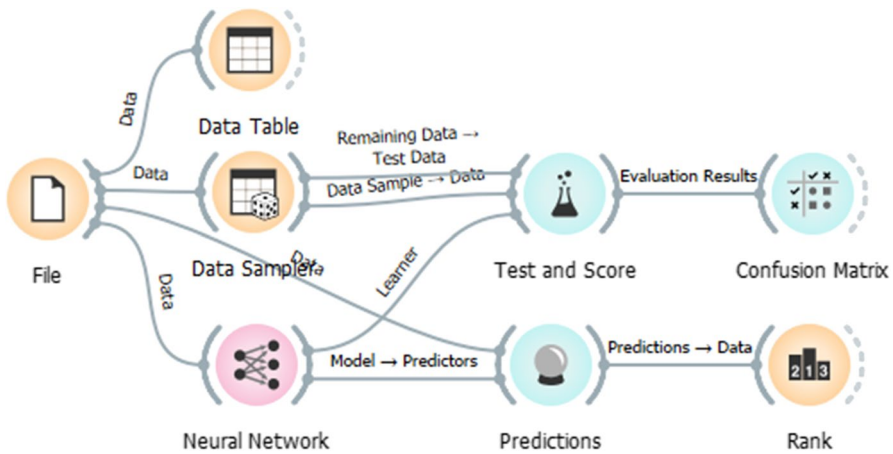


Fig. 11 Process of identifying main predictors using NN in orange

Referring to Table 3, for DT, the Form Four English final examination has the most impact on the prediction with an impact of 0.225, followed by the Form Four Mathematics final examination (0.202) and the Form Four English mid-term examination (0.165). For NN, the Form Four English final examination is the most dominant in the prediction with an impact of 0.168, followed by the Form Four Mathematics final examination (0.132) and the Form Four English mid-term examination (0.117). For SVM, the Form Four English final examination is the most influential in the prediction with an impact of 0.152, followed by the Form Four Mathematics final examination (0.121) and the Form Four English mid-term examination (0.104). For NB, the Form Four English final examination is the most prominent in the prediction with an impact of

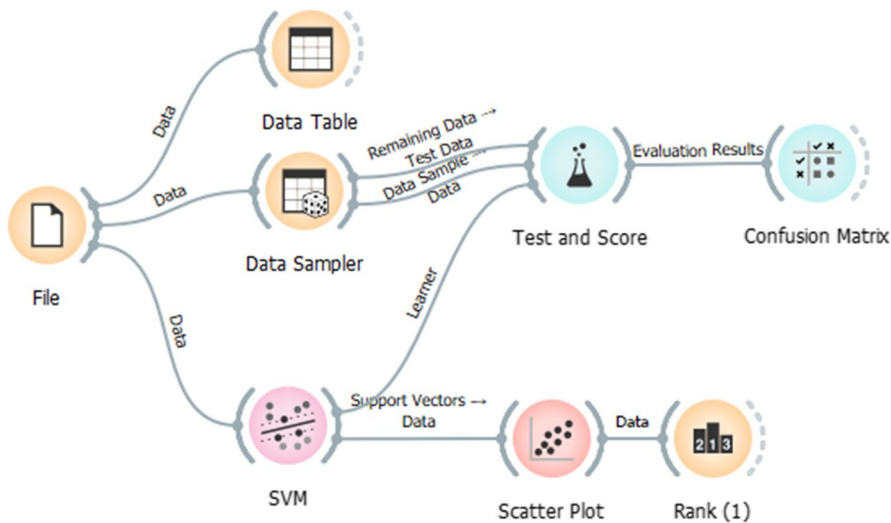


Fig. 12 Process of identifying main predictors using SVM in orange

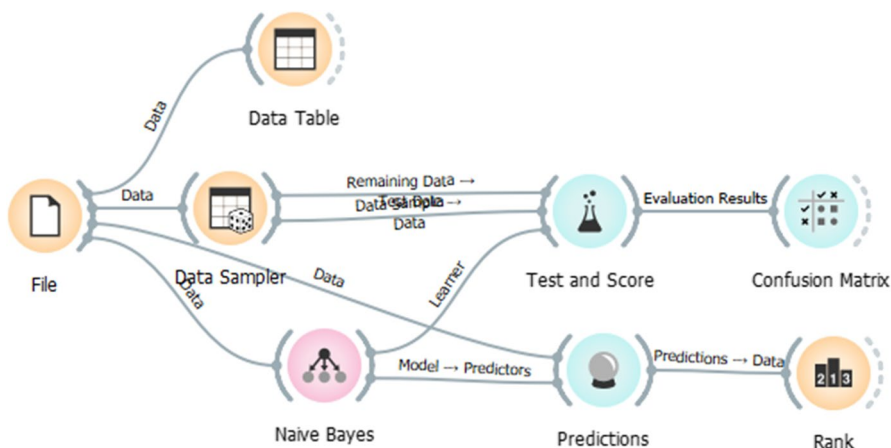


Fig. 13 Process of identifying main predictors using NB in orange

0.168, followed by the Form Four Mathematics final examination (0.132) and the Form Four English mid-term examination (0.117).

RQ4: What are the main predictors of students’ Mathematics subject performance using DT, NN, SVM, and NB classification techniques?

Similar procedures described in RQ3 were repeated for RQ4 to identify the main predictors of Mathematics performance using DT, NN, SVM, and NB in Orange. Table 4 shows the top five attributes derived from each classifier.

Table 3 Top five attributes on the prediction of English performance using DT, NN, SVM, and NB

DT	
Attributes	Impact
Form Four English final examination	0.225
Form Four Mathematics final examination	0.202
Form Four English mid-term examination	0.165
Form Four Mathematics mid-term examination	0.142
Diversity	0.090
NN	
Attributes	Impact
Form Four English final examination	0.168
Form Four Mathematics final examination	0.132
Form Four English mid-term examination	0.117
Form Four Mathematics mid-term examination	0.091
Gender	0.085
SVM	
Attributes	Impact
Form Four English final examination	0.152
Form Four Mathematics final examination	0.121
Form Four English mid-term examination	0.104
Form Four Mathematics mid-term examination	0.090
Father's educational level	0.084
NB	
Attributes	Impact
Form Four English final examination	0.168
Form Four Mathematics final examination	0.132
Form Four English mid-term examination	0.117
Form Four Mathematics mid-term examination	0.091
Gender	0.085

Referring to Table 4, for DT, the Form Four Mathematics final examination has the most impact on the prediction with an impact of 0.341, followed by the Form Four Mathematics mid-term examination (0.235) and Form Four English final examination (0.124). For NN, the Form Four Mathematics final examination is the most dominant in the prediction with an impact of 0.223, followed by the Form Four Mathematics mid-term examination (0.151) and Form Four English final examination (0.093). For SVM, the Form Four Mathematics final examination is the most influential in the prediction with an impact of 0.205, followed by the Form Four Mathematics mid-term examination (0.135) and the Form Four English final examination (0.082). For NB, the Form Four Mathematics final examination is the most prominent in the prediction with an impact of 0.223, followed by the Form Four Mathematics mid-term examination (0.151) and the Form Four English final examination (0.093).

Table 4 Top five attributes on the prediction of Mathematics performance using DT, NN, SVM, and NB

DT	
Attributes	Impact
Form Four Mathematics final examination	0.341
Form Four Mathematics mid-term examination	0.235
Form Four English final examination	0.124
Diversity	0.086
Form Four English mid-term examination	0.083
NN	
Attributes	Impact
Form Four Mathematics final examination	0.223
Form Four Mathematics mid-term examination	0.151
Form Four English final examination	0.093
Ethnicity	0.074
Gender	0.065
SVM	
Attributes	Impact
Form Four Mathematics final examination	0.205
Form Four Mathematics mid-term examination	0.135
Form Four English final examination	0.082
Parents' marital status	0.064
Father's educational level	0.048
NB	
Attributes	Impact
Form Four Mathematics final examination	0.223
Form Four Mathematics mid-term examination	0.151
Form Four English final examination	0.093
Form Four English mid-term examination	0.059
Diversity	0.044

For RQ3 and RQ4, it has been discovered that past academic performance is the most influential attribute to predict students' performance in English and Mathematics. The top three attributes that predict English performance were consistently derived from the four classification techniques, which are Form Four English final examination, Form Four Mathematics final examination, and Form Four English mid-term examination. Similarly, Form Four Mathematics final examination, Form Four Mathematics mid-term examination, and Form Four English final examination are the top three attributes for predicting Mathematics performance based on the four classification techniques.

This is in line with the review by Shahiri and Husain (2015) who found that academic results are the most important attribute to predict students' performance. The findings are supported by Saa et al. (2019) as they identified that the most common aspects in predicting students' performance are students' previous grades and internal assessments. Next, Makhtar et al. (2017) postulated that students' performance can be solely predicted using academic records. Furthermore, this attribute is the most employed in predicting students'

performance since it has a quantitative value that is both tangible and appropriate for predicting students' performance (López-Zambrano et al., 2021).

Referring to RQ1 to RQ4, the trees and predictors of students' performance in English and Mathematics incorporate some demographic attributes such as the parents' educational levels. This finding demonstrates that past academic performance is not the only attribute that influences students' performance. This finding is backed up by Kiu (2018) who concludes that parents' educational level is significant as it impacts their children's performance. The parents' marital status and ethnicity did not appear in any tree diagrams and were only derived once among the eight predictions. Hence, these attributes are considered insignificant. The parents' occupational status is found to have no bearing on the predictions. This is supported by findings by Etemadpour et al. (2011) which posited that parents' educational attainment is more pertinent than their occupational status in determining students' academic performance.

Gender is another demographic attribute that is recognized as one of the top five predictors of English performance by the NN classifier. This finding is in line with Parajuli and Thapa's findings (2017). They discovered that there were considerable gender differences in students' academic performance, with female students outperforming their male counterparts. Ioannis and Maria (2019) show that female students are more likely to score better in several subjects than male students.

Psychological attributes such as diversity and self-criticism are found in the trees and predictors. Khairy et al. (2018) utilized a similar method (the results of a psychometric test) to predict secondary school students' performance. However, according to the findings, most psychological attributes play a minor role in predicting students' performance (Verma & Yadav, 2020). In short, past academic performance is crucial to predicting students' performance in English and Mathematics and certain demographics and psychological attributes are also influencing the predictions.

3.2 Predictive model with the highest accuracy

RQ5: Which predictive model has the highest predictive accuracy for students' performance in English and Mathematics subjects respectively?

Table 5 Summary of measures on DM classifiers for English performance prediction

DM classifiers	Prediction measures (%)				
	Area under curve	Accuracy	Precision	Recall	F1-score
DT	89.3	87.1	87.4	87.1	87.2
NN	83.3	71.0	72.0	71.0	71.1
SVM	80.3	74.2	75.6	74.2	74.6
NB	81.3	71.0	76.5	71.0	72.3

Table 5 shows the summary of measures on DM classifiers for English performance prediction.

Based on Table 5, DT outperforms other DM classifiers for all measures in terms of predicting students' performance in English. DT shows the highest predictive accuracy among the DM models, with an accuracy of 87.1%, followed by SVM, NB, and NN. This finding is consistent with findings by Nawai et al. (2021) who found that DT surpassed other DM classifiers in terms of predictive performance. It may be argued that DT is the most dominant classifier since it creates classification rules that apply to both nominal and categorical data (Shetu et al., 2021). In addition, DT more closely resembles human decision-making than other DM approaches (Burkart & Huber, 2021).

Table 6 shows the summary of measures on DM classifiers for Mathematics performance prediction.

Referring to Table 6, although DT and NB have similar highest predictive accuracy (83.9%), a further examination into the Area Under Curve, Precision, and F1-Score shows NB has surpassed DT and other DM classifiers in the prediction of students' Mathematics performance. Saritas and Yasar (2019) have discovered that the NB classifier is highly effective in some applications. In Chandrasekar and Qian (2016), the prediction accuracy of the NB classifier improved by using and combining appropriate data. In comparison with other DM algorithms, Shinde and Prasad (2017) posited that NB has tremendous performance in real-time predictions, and it can also perform well in multi-class predictions (Ahuja et al., 2020). In short, NB has distinct advantages in the prediction task (Amra & Maghari, 2017).

This study reveals DT and NB as the two top classifiers while NN and SVM are the two classifiers with the lowest accuracy in predicting students' English and Mathematics performance respectively. Compared with the meta-analysis on the accuracy of classification techniques that were used to predict students' performance in studies from the year 2002 to 2015 by Shahiri and Husain (2015), the analysis reveals slightly different results. Shahiri and Husain (2015) reported that NN shows the highest prediction accuracy at 98% and DT is second at 91%. SVM and KNN show similar accuracy, which is 83%, while NB has the lowest prediction accuracy at 76%. Hence, the recent research provides new insight into the prediction accuracy of DM classifiers.

Table 6 Summary of measures on DM classifiers in Mathematics performance prediction

DM classifiers	Prediction measures (%)				
	Area under curve	Accuracy	Precision	Recall	F1-score
DT	82.9	83.9	84.3	83.9	83.2
NN	87.1	74.2	77.6	74.2	71.2
SVM	72.2	71.0	80.3	71.0	65.0
NB	92.8	83.9	84.8	83.9	83.8

3.3 Relationship between students' English and Mathematics performance

RQ6: What is the relationship between students' performance in English and Mathematics subjects?

For RQ6, it can be concluded that there is a relationship between students' performance in English and Mathematics. Form Four Mathematics final and mid-term examinations are among the most important attributes when employing four algorithms to predict students' performance in English. The Form Four English final examination is also among the top attributes in predicting students' performance in Mathematics using DT, NN, SVM, and NB. However, Form Four English mid-term examination has less impact on students' performance in Mathematics. To measure the relationship between these subjects, SPSS software was used to apply Spearman's rank correlation, r_s . Table 7 shows the correlation between English and Mathematics by using Spearman's rank correlation, r_s . There is a low positive correlation between Form Four Mathematics mid-term (0.345) and final examination (0.391) with MCE English performance. There is also a low positive correlation between the Form Four English mid-term (0.285) and final examination (0.367) with MCE Mathematics performance. The MCE English and MCE Mathematics performances are moderately correlated (0.433).

Based on the weight of attributes identified by DM classifiers and Spearman's rank correlation, it can be deduced that English and Mathematics performance are interrelated. The findings are consistent with Henry et al. (2014), which also investigated the relationship between English and Mathematics performance. They discovered that English is a statistically significant predictor of Mathematics performance, and Mathematics performance improves simultaneously with English performance. Next, Rudd and Honkiss (2020) discovered that there is an indirect mediation relationship between English proficiency and Mathematics performance. The examination of the predictors of the four DM classifiers provides interesting insights into the relationship between these two subjects in addition to insights gained via the conventional statistical methods.

4 Conclusion

The study provides a deeper understanding of attributes that influence the prediction of students' performance in English and Mathematics. It is critical to include past academic performance because it substantially impacts the prediction. Besides past academic performance, certain students' demographics such

Table 7 Results of Spearman's rank correlation

Actual MCE English performance	r_s	Actual MCE Mathematics performance	r_s
Form Four Mathematics mid-term examination	0.345	Form Four English mid-term examination	0.285
Form Four Mathematics final examination	0.391	Form Four English final examination	0.367
Actual MCE Mathematics performance	0.433	Actual MCE English performance	0.433

as parents' educational level and gender as well as psychological factors such as diversity and self-criticism also contribute to the prediction, at least in the context of this study.

This study could help academics determine which DM algorithm is most effective in predicting students' performance. According to the findings, DT and NB algorithms are the best in predicting students' performance in English and Mathematics, respectively. DT can reveal the characteristics of students with low, moderate, and high performance and provides good predictive performance for both subjects. Prediction of students' academic performance and the specific characteristics or attributes that define students of different levels of performance in these subjects enables early interventions and personalized support to be given to students for optimizing their performance thus, alleviating the issue of the high failure rate of both subjects in MCE.

The performance in English and Mathematics is found to be interrelated. This finding provides additional insights as over the decade, researchers attempt to gain insights into the relationship between English and Mathematics (Bagececi et al., 2014; Rambely et al., 2013). Limitations of this study include the small sample size employed which limits the generalizability of the results. There is also a lack of diversity in terms of chosen students' religions and ethnicities as well as the type and locality of chosen schools as all chosen schools are urban public schools. In addition, this study is confined to four DM techniques and two subjects. Future studies may include a larger number of students from various types of schools of diverse localities. Future studies may also involve other types of data mining techniques and explore other subjects as these may reveal previously undiscovered information.

Acknowledgements This work is supported by the Malaysian Ministry of Higher Education, Fundamental Research Grant Scheme, FRGS/1/2020/SS10/UNIMAS/01/1, and UNIMAS Zamalah Scholarship.

Funding This work is funded by the Malaysian Ministry of Higher Education, Fundamental Research Grant Scheme, FRGS/1/2020/SS10/UNIMAS/01/1, and UNIMAS Zamalah Scholarship.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval statement The study obtained approval from the Education Policy Research and Development Division, Ministry of Education, Malaysia to use the archival data from the schools involved.

Conflict of interest There is no potential conflict of interest in this study.

References

- Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5(2), e01250. <https://doi.org/10.1016/j.heliyon.2019.e01250>

- Ahuja, R., Chug, A., Gupta, S., Ahuja, P., & Kohli, S. (2020). Classification and clustering algorithms of machine learning with their applications. In *Nature-Inspired Computation in Data Mining and Machine Learning* (pp. 225–248). Springer, Cham. https://doi.org/10.1007/978-3-030-28553-1_11
- Algarni, A. (2016). Data mining in education. *International Journal of Advanced Computer Science and Applications*, 7. <https://doi.org/10.14569/IJACSA.2016.070659>
- Almeda, M. V., Zuech, J., Utz, C., Higgins, G., Reynolds, R., & Baker, R. S. (2018). Comparing the factors that predict completion and grades among for-credit and open/mooc students in online learning. *Online Learning Journal*, 22(1), 1–18. <https://doi.org/10.24059/olj.v22i1.1060>
- Alyahyan, E., & Düşteğör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1), 3. <https://doi.org/10.1186/s41239-020-0177-7>
- Amra, I. A. A., & Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In *2017 8th International Conference on Information Technology* (pp. 909–913). IEEE. <https://doi.org/10.1109/ICITECH.2017.8079967>
- Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate student's performance using educational data mining. *Computers & Education*, 111, 117–194. <https://doi.org/10.1016/j.compedu.2017.05.007>
- Atlay, C., Tieben, N., Hillmert, S., & Fauth, B. (2019). Instructional quality and achievement inequality: How effective is teaching in closing the social achievement gap? *Learning and Instruction*, 63, 101211. <https://doi.org/10.1016/j.learninstruc.2019.05.008>
- Awan, R. U. N., Noureen, G., & Naz, A. (2011). A study of relationship between achievement motivation, self concept and achievement in English and Mathematics at secondary level. *International Education Studies*, 4(3), 72–79. <https://doi.org/10.5539/ies.v4n3p72>
- Bagceci, B., Kutlar, E. L., & Cinkara, E. (2014). The relationship between English and Math success & some variables at freshman level. *Journal of Education and Practice*, 5(29), 6–11. <https://www.iiste.org/Journals/index.php/JEP/article/view/16180>
- Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17. <https://doi.org/10.5281/zenodo.3554657>
- Bakhtiar, A., Suliantoro, H., Ningsi, R. H., & Pitipaldi, K. (2021). Relationship of quality management system standards to industrial property rights in Indonesia using Spearman correlation analysis method. In *IOP Conference Series: Earth and Environmental Science* (Vol. 623, No. 1, p. 012092). IOP Publishing. <https://iopscience.iop.org/article/https://doi.org/10.1088/1755-1315/623/1/012092/pdf>
- Bergquist, S. L., Layton, T. J., McGuire, T. G., & Rose, S. (2019). Data transformations to improve the performance of health plan payment methods. *Journal of Health Economics*, 66, 195–207. <https://doi.org/10.1016/j.jhealeco.2019.05.005>
- Berhanu, F., & Abera, A. (2015). Students' performance prediction based on their academic record. *International Journal of Computer Applications*, 131(5), 27–35. <https://doi.org/10.5120/ijca2015907348>
- Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19, 205–220. <https://doi.org/10.1007/s10758-014-9223-7>
- Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245–317. <https://doi.org/10.1613/jair.1.12228>
- Chand, S., Chaudhary, K., Prasad, A., & Chand, V. (2021). Perceived causes of students' poor performance in mathematics: A case study at Ba and Tavua secondary schools. *Frontiers in Applied Mathematics and Statistics*. <https://doi.org/10.3389/fams.2021.614408>
- Chandrasekar, P., & Qian, K. (2016, June). The impact of data pre-processing on the performance of a Naive Bayes classifier. In *2016 IEEE 40th Annual Computer Software and Applications Conference* (Vol. 2, pp. 618–619). IEEE. <https://doi.ieeecomputersociety.org/https://doi.org/10.1109/COMPSAC.2016.205>
- Chapman, W. W., & Haug, P. J. (1999). Comparing expert systems for identifying chest x-ray reports that support pneumonia. In *Proceedings of the AMIA Symposium* (p. 216). American Medical Informatics Association. <https://doi.org/10.1006/jbin.2001.1000>
- Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20–28. <https://doi.org/10.38094/jast20165>

- Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991–16005. <https://doi.org/10.1109/ACCESS.2017.2654247>
- Etemadpour, R., Zhu, Y., Zhao, Q., Hu, Y., Chen, B., Sharier, M. A., & ... Paiva, J. G. S. (2020). Role of absence in academic success: An analysis using visualization tools. *Smart Learning Environments*, 7(1), 1–25. <https://doi.org/10.5130/AJCEB.v20i3.7056>
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–37. <https://doi.org/10.1609/aimag.v17i3.1230>
- García-Jiménez, J., Rodríguez-Santero, J., & Torres-Gordillo, J. J. (2020). Influence of contextual variables on educational performance: A study using hierarchical segmentation trees. *Sustainability*, 12(23), 9933. <https://doi.org/10.3390/su12239933>
- Garg, R. (2018). Predict student performance in different regions of Punjab. *International Journal of Advanced Research in Computer Science*, 9(1), 236–241. <https://doi.org/10.26483/ijarcs.v9i1.5234>
- Goutte, C., & Gaussier, E. (2005, March). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval* (pp. 345–359). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25
- Gray, G., McGuinness, C., & Owende, P. (2014, February). An application of classification models to predict learner progression in tertiary education. In *2014 IEEE International Advance Computing Conference (IACC)* (pp. 549–554). IEEE. <https://doi.org/10.1109/IAAdCC.2014.6779384>
- Had, M. Z. C., & Rashid, R. A. (2019). A review of digital skills of Malaysian English language teachers. *International Journal of Emerging Technologies in Learning*, 14(2). <https://doi.org/10.3991/ijet.v14i02.8732>
- Henry, D. L., Baltes, B., & Nistor, N. (2014). Examining the relationship between math scores and English language proficiency. *Journal of Educational Research and Practice*, 4(1), 2. <https://doi.org/10.5590/JERAP.2014.04.1.02>
- Hoe, A. C. K., Ahmad, M. S., Hooi, T. C., Shanmugam, M., Gunasekaran, S. S., Cob, Z. C., & Ramasamy, A. (2013, November). Analyzing students records to identify patterns of students' performance. In *2013 International Conference on Research and Innovation in Information Systems* (pp. 544–547). IEEE. <https://doi.org/10.1109/ICRIIS.2013.6716767>
- Idris, M., Hussain, S., & Ahmad, N. (2020). Relationship between parents' education and their children's academic achievement. *Journal of Arts & Social Sciences (JASS)*, 7(2), 82–92. [https://doi.org/10.46662/jass-vol7-iss2-2020\(82-92\)](https://doi.org/10.46662/jass-vol7-iss2-2020(82-92))
- Ioannis, B., & Maria, K. (2019). Gender and student course preferences and course performance in computer science departments: A case study. *Education and Information Technologies*, 24(2), 1269–1291. <https://doi.org/10.1007/s10639-018-9828-x>
- Jamil, J. M., Pauzi, N. F. M., & Nee, I. N. M. S. (2018). An analysis on student academic performance by using decision tree models. *The Journal of Social Sciences Research*, 6, 615–620. <https://doi.org/10.32861/jssr.spi6.615.620>
- Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. <https://doi.org/10.1002/sam.11583>
- Kementerian Pendidikan Malaysia. (2020). *Laporan Analisis Keputusan SPM 2019*. Retrieved from <https://www.moe.gov.my/en/kuat-turun/laporan-dan-statistik/lp/3324-laporan-analisis-keputusan-spm-2019/file>
- Khairy, A. M., Adam, A., & Yaakub, M. R. (2018). Data analytics in Malaysian education system: Revealing the success of Sijil Pelajaran Malaysia from Ujian Aptitud Sekolah Rendah. *Asia-Pacific Journal of Information Technology and Multimedia*, 7(2), 29–45. <https://doi.org/10.17576/apjitm-2018-0702-03>
- Kiu, C. C. (2018). Data mining analysis on student's academic performance through exploration of student's background and social activities. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)* (pp. 1–5). IEEE. <https://doi.org/10.1109/ICACCAF.2018.8776809>
- Krupat, E., Pelletier, S. R., & Dienstag, J. L. (2017). Academic performance on first-year medical school exams: How well does it predict later performance on knowledge-based and clinical assessments? *Teaching and Learning in Medicine*, 29(2), 181–187. <https://doi.org/10.1080/10401334.2016.1259109>
- Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, 1(9), 1–10. <https://doi.org/10.1007/s42452-019-0884-7>

- Li, Z., & Qiu, Z. (2018). How does family background affect children's educational achievement? Evidence from contemporary china. *The Journal of Chinese Sociology*, 5(1), 1–21. <https://doi.org/10.1186/s40711-018-0083-8>
- Lipnevich, A. A., Preckel, F., & Krumm, S. (2016). Mathematics attitudes and their unique contribution to achievement: Going over and above cognitive ability and personality. *Learning and Individual Differences*, 47, 70–79. <https://doi.org/10.1016/j.lindif.2015.12.027>
- López-Zambrano, J., Lara Torralbo, J. A., & Romero Morales, C. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, 33(3), 456–465. <https://hdl.handle.net/11162/211835>
- Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Journal of Educational Technology & Society*, 21(2), 220–232. <http://hdl.handle.net/2433/231307>
- Maghari, A., & Mousa, H. (2017). School students' performance prediction using data mining classification. *International Journal of Advanced Research in Computer and Communication Engineering*, 6(8), 136–141. <https://doi.org/10.17148/IJARCCCE.2017.6824>
- Makhtar, M., Nawang, H., & Wan Shamsuddin, S. N. (2017). Analysis on students' performance using Naïve Bayes classifier. *Journal of Theoretical & Applied Information Technology*, 95(16).
- Martínez, C. R., & Gil, M. G. (2019). Gender differences in school performance and attitudes toward school. *Ensaio: Avaliação e Políticas Públicas Em Educação*, 1–21. <https://doi.org/10.1590/s0104-40362019002702235>
- McKee, M. T., & Caldarella, P. (2016). Middle school predictors of high school performance: A case study of dropout risk indicators. *Education*, 136(4), 515–529. Retrieved from <https://eric.ed.gov/?id=EJ1104172>
- Meng, M., & Zhao, C. (2015). Application of support vector machines to a small-sample prediction. *Advances in Petroleum Exploration and Development*, 10(2), 72–75. <https://doi.org/10.3968/7830>
- Miller-Matero, L. R., Martinez, S., MacLean, L., Yaremchuk, K., & Ko, A. B. (2018). Grit: A predictor of medical student performance. *Education for Health*, 31(2), 109. https://doi.org/10.4103/efh.efh_152_16
- Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students' data for prediction performance. In *2014 Fourth International Conference on Advanced Computing & Communication Technologies* (pp. 255–262). IEEE. <https://doi.org/10.1109/ACCT.2014.105>
- Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97(6), 320–324. <https://doi.org/10.1016/j.sbspro.2013.10.240>
- Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, 8(11), 36–42. <https://doi.org/10.5815/ijmecs.2016.11.05>
- Mühlbacher, T., Linhardt, L., Möller, T., & Piringer, H. (2017). Treepod: Sensitivity-aware selection of pareto-optimal decision trees. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), 174–183. <https://doi.org/10.1109/tvcg.2017.2745158>
- Muller, C. (2018). Parent involvement and academic achievement: An analysis of family resources available to the child. In *Parents, their children, and schools* (pp. 77–114). Routledge. Retrieved from <https://www.taylorfrancis.com/chapters/edit/https://doi.org/10.4324/9780429498497-4/parent-involvement-academic-achievement-chandra-muller>
- Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), 237. <https://doi.org/10.3390/app11010237>
- Nawai, S. N. M., Saharan, S., & Hamzah, N. A. (2021). An analysis of students' performance using cart approach. In *AIP Conference Proceedings* (Vol. 2355, No. 1, p. 060009). AIP Publishing LLC. <https://doi.org/10.1063/5.0053388>
- Parajuli, M., & Thapa, A. (2017). Gender differences in the academic performance of students. *Journal of Development and Social Engineering*, 3(1), 39–47. <https://doi.org/10.3126/jdse.v3i1.27958>
- Perez, A. L., & Alieto, E. (2018). Change of "tongue" from English to a local language: A correlation of mother tongue proficiency and mathematics achievement. *Online Submission*, 14, 132–150. Retrieved from <https://files.eric.ed.gov/fulltext/ED597105.pdf>
- Putpuek, N., Rojanaprasert, N., Atchariyachanvanich, K., & Thamrongthanyawong, T. (2018). Comparative study of prediction models for final GPA score: a case study of Rajabhat Rajanagarindra

- university. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science* (pp. 92–97). IEEE. <https://doi.org/10.1109/ICIS.2018.8466475>
- Rambely, A. S., Ahmad, R. R., Majid, N., & Jaaman, S. H. (2013). The relationship of English proficiency and mathematics achievement. In *Recent Advances in Educational Technologies* (pp. 139–145). Retrieved from [http://www.wseas.us/elibrary/conferences/2013/Cambridge USA/EET/EET-24.pdf](http://www.wseas.us/elibrary/conferences/2013/Cambridge%20USA/EET/EET-24.pdf).
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355. <https://doi.org/10.1002/widm.1355>
- Rudd, M., & Honkiss, L. (2020). Analysing the correlation between English proficiency and academic performance among Thai university students. *Athens Journal of Education*, *7*(1), 123–137. <https://doi.org/10.30958/aje.7-1-6>
- Saa, A. A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, *24*(4), 567–598. Springer Netherlands. <https://doi.org/10.1007/s10758-019-09408-7>
- Saleh, M. A., Palaniappan, S., & Abdalla, N. A. A. (2021). Predicting student performance using data mining techniques in Libyan high schools. *Edukasi*, *15*(2), 91–100. <https://doi.org/10.15294/eduka.si.v15i2.30068>
- Salem, A. B. M., & Parusheva, S. (2018). Exploiting the knowledge engineering paradigms for designing smart learning systems. *Eastern-European Journal of Enterprise Technologies*, *2*(2), 38–44. <https://doi.org/10.15587/1729-4061.2018.128410>
- Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88–91. <https://doi.org/10.18201/ijisae.2019252786>
- Sathe, M. T., & Adamuthe, A. C. (2021). Comparative study of supervised algorithms for prediction of students' performance. *International Journal of Modern Education & Computer Science*, *13*(1). <https://doi.org/10.5815/ijmecs.2021.01.01>
- Shahiri, A. M., & Husain, W. (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>
- Shetu, S. F., Saifuzzaman, M., Moon, N. N., Sultana, S., & Yousuf, R. (2021). Student's performance prediction using data mining technique depending on overall academic status and environmental attributes. In *International Conference on Innovative Computing and Communications* (pp. 757–769). Springer, Singapore. https://doi.org/10.1007/978-981-15-5148-2_66
- Shinde, T. A., & Prasad, J. R. (2017). IoT based animal health monitoring with Naive Bayes classification. *International Journal of Emerging Trends & Technology*, *1*(2), 252–257.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Sordo, M., & Zeng, Q. (2005, November). On sample size and classification accuracy: A performance comparison. In *International Symposium on Biological and Medical Data Analysis* (pp. 193–201). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11573067_20
- Stoffelsma, L., & Spooren, W. (2019). The relationship between English reading proficiency and academic achievement of first-year science and mathematics students in a multilingual context. *International Journal of Science and Mathematics Education*, *17*(5), 905–922. <https://doi.org/10.1007/s10763-018-9905-z>
- Sulphrey, M. M., Al-Kahtani, N. S., & Syed, A. M. (2018). Relationship between admission grades and academic achievement. *Entrepreneurship and Sustainability Issues*, *5*(3), 648–658. [https://doi.org/10.9770/jesi.2018.5.3\(17\)](https://doi.org/10.9770/jesi.2018.5.3(17))
- Tan, R. Z., Wang, P. C., Lim, W. H., Ong, S. H. C., & Avnit, K. (2019). Early prediction of students' mathematics performance. *Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, 651–656. <https://doi.org/10.1109/TALE.2018.8615289>
- Verma, S., & Yadav, R. K. (2020). Effect of different attributes on the academic performance of engineering students. In *2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation* (pp. 1–4). IEEE. <https://doi.org/10.1109/ICATMRI51801.2020.9398442>
- Wakelam, E., Jefferies, A., Davey, N., & Sun, Y. (2020). The potential for student performance prediction in small cohorts with minimal available attributes. *British Journal of Educational Technology*, *51*(2), 347–370. <https://doi.org/10.1111/bjet.12836>

- Wilcox, A., & Hripcsak, G. (1999). Classification algorithms applied to narrative reports. In *Proceedings of the AMIA Symposium* (p. 455). American Medical Informatics Association. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232569/pdf/procamiasymp00004-0492.pdf>
- Wong, S. L., & Wong, S. L. (2019). Relationship between interest and mathematics performance in a technology-enhanced learning context in Malaysia. *Research and Practice in Technology Enhanced Learning*, 14(1), 1–13. <https://doi.org/10.1186/s41039-019-0114-3>
- Yung, K. W. H., & Cai, Y. (2020). Do secondary school-leaving English examination results predict university students' academic writing performance? A latent profile analysis. *Assessment & Evaluation in Higher Education*, 45(4), 629–642. <https://doi.org/10.1080/02602938.2019.1680951>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.