INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

HRMARS MANAGEMENT ACADEMIC RESEARCH SOCIETY EXPLORING INTELLECTUAL CAPITAL

www.hrmars.com

ISSN: 2226-6348

Academic Performance Prediction Model Using Classification Algorithms: Exploring the Potential Factors

Rozianiwati Yusof, Norhafizah Hashim, Normaziah Abdul Rahman, Sri Yusmawati Mohd Yunus, Nor Azlina Aziz Fadzillah

To Link this Article: http://dx.doi.org/10.6007/IJARPED/v11-i3/14753 DOI:10.6007/IJARPED/v11-i3/14753

Received: *16* June 2022*,* ***Revised:*** *17* July 2022*,* ***Accepted:*** *30* July 2022

Published Online: *20* August 2022

In-Text Citation: (Yusof et al., 2022)

To Cite this Article: Yusof, R., Hashim, N., Rahman, N. A., Yunus, S. Y. M., & Fadzillah, N. A. A. (2022). Academic Performance Prediction Model Using Classification Algorithms: Exploring the Potential Factors. *International Journal of Academic Research in Progressive Education and Development*, *11*(3), 706–724.

Vol. 11(3) 2022, Pg. 706 - 724

http://hrmars.com/index.php/pages/detail/IJARPED | **JOURNAL HOMEPAGE**

Full Terms & Conditions of access and use can be found at
http://hrmars.com/index.php/pages/detail/publication-ethics

INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

HRMARS MANAGEMENT ACADEMIC RESEARCH SOCIETY
EXPLORING INTELLECTUAL CAPITAL

www.hrmars.com
ISSN: 2226-6348

Academic Performance Prediction Model Using Classification Algorithms: Exploring the Potential Factors

Rozianiwati Yusof, Norhafizah Hashim, Normaziah Abdul Rahman, Sri Yusmawati Mohd Yunus, Nor Azlina Aziz Fadzillah

Faculty of Computer and Mathematical Sciences, UiTM Cawangan Negeri Sembilan, Kampus Seremban, Malaysia

Email: rozian696@uitm.edu.my

Abstract

The student's performance has become the focus in higher education institutions. The ability to predict students' performance is beneficial to improve their achievement and the learning process. However, producing a prediction model for academic performance becomes challenging when an educational dataset contains various data. Many researchers have widely explored this kind of research, but many features should be investigated to affect students' achievement. Finding the potential factors influencing students' performance helps enhance students' quality. These factors will assist an institution plan a strategy for improving students' performance. This research proposes a classifier model to predict students' academic performance and define the factors influencing the performance by considering 14 attributes from demographics, learning styles, and educational background. The model development employs seven machine learning algorithms, and the best model will be selected. The factors that influence academic performance will be revealed from that model. The dataset was collected by conducting a survey at UiTM Seremban involving 233 students from Science and Technology and Social Science Streams. The Random Forest Tree produced an accurate result with the simple rules to be interpreted. The model also showed four attributes: qualification before tertiary education, SPM result, Seniority and gender positively impacting academic performance. Some factors that did not influence their performance were their parents' academic background and hometown.

Keywords: Data Mining, Academic, Performance, Prediction, Classification

Introduction

In people's lives, education is the most potent weapon. It gives people the tools they need to improve and change their quality of life. It is possible to learn it through the learning process. Formal learning is education offered deliberately and intentionally by trained instructors in higher education or university classrooms. The learning level of students can be determined by their performance. As a result, it is crucial to properly monitor students' progress

throughout their higher education learning process. Students' performance has been measured using the Cumulative Grade Point Average (CGPA) system in higher education level in Malaysia. It is an average of grade points obtained for all finished semesters. Thus, CGPAs can serve as predictors of success for the students before getting an actual working experience in the future. Higher institutions measure using the CGPA system as shown in the table below:

Table 1
CGPA Pointer Status

Pointer	Status
3.50 - 4.00	First Class
3.00 - 3.49	Second Class Upper
2.50 - 2.99	Second Class Lower
2.00 - 2.49	Third Class
0.00-1.99	Fail

The students who manage to complete their studies with a CGPA between 3.0 to 4.0 will be awarded Second Class Upper and First-Class who are marketable and successful. These grades can be used as students' performance measurements; for example, the CGPA between 3.5 to 4.0 can be considered excellent, and their understanding is between 'A-' to 'A+' grades. Below is the table of grades used in calculating the CGPA grade.

Table 2
Grading Scheme and Status

Marks	Grades	Points	Status
90-100	A+	4.00	Passed
80-89	A	4.00	Passed
75-79	A-	3.67	Passed
70-74	B+	3.33	Passed
65-69	B	3.00	Passed
60-64	B-	2.67	Passed
55-59	C+	2.33	Passed
50-54	C	2.00	Passed
47-49	C-	1.67	Fail
44-46	D+	1.33	Fail
40-43	D	1.00	Fail
30-39	E	0.67	Fail
0-29	F	0	Fail

In the COVID-19 era, it is very important to investigate the students' backgrounds that will influence their academic performance. In this study, three criteria have been chosen to find the correlation with students' academic performance. These three criteria are demographics, student learning styles, and educational background. Demographic and education criteria are essential to identify whether the student's performance relates to their lifestyle and educational background. Meanwhile, the learning style criterion has been considered to know whether the way they learn will influence students' results.

The learning styles refer to various competing and contested theories aiming to explain differences in individual learning. The learning style will influence the acceptance of students in the learning process (Ilçin et al., 2018). Regarding technology changing rapidly, the learning pattern from face to face is not the only one that should be considered (Nuankaew et al., 2019). The popular methods in teaching right now are blended learning and e-learning. These methods need students to become self-learners. Encouraging students to recognise the process of finding information and knowledge is essential (Nuankaew et al., 2019). Many models have been discussed on learning styles, such as VAK, KOLB, Felder Silverman, 4Mat, Gregorc and Honey Mumford. This study chooses the most popular and common widely used, which is the VAK model. The VAK model provides a simple way to identify the learning styles among learners. It consists of three categories of learners such as Visualize, Auditory and Kinesthetics. Visualised learners prefer to learn by seeing. Meanwhile, auditory learners learn by listening, and kinesthetic learners learn by experience, which is touching, doing, and moving (Melo, 2018).

Studying and analysing educational data, especially students' performance, is important. Educational Data Mining (EDM) in the field of study is concerned with mining educational data to find interesting patterns and knowledge in educational organisations. This study is significant and beneficial for predicting students' performance to improve their learning process. They will be able to achieve successful academic results if they are aware of the indicators and factors that can influence the outcomes. However, creating a prediction model for academic success becomes difficult when an educational dataset comprises various data. This study explores multiple factors theoretically assumed to affect students' performance in higher education and finds a predictive model which best classifies and predicts the students' performance based on related personal and social factors. There is less research on predicting students' performance based on their educational background and learning styles. This study also considered the background streams of students and the learning styles that will influence the students' performance. Thus, the research questions of this study are; what are the main factors that influenced the students' performance and how accurate is the model produced in predicting students' performance? This study needs to achieve a few objectives to answer the research question.

1) Investigates the selected demographic, educational background and learning styles among respondents. Identified the best learning styles for success in academics.
2) Producing an accurate prediction model of students' academic performance using classifier algorithms.
3) Identified a few factors influencing the student's academic performance based on demographics, educational background and learning styles.

The order of the paper is structured in a simple manner which is divided into five sections. The first section is an introduction to summarise the background of the research. The second section summarises the essentials and other research related to the prediction model in students' performance and classification algorithms. The third section describes the process and steps in the research methodology. The fourth section is an analysis and discussion of the research results. Finally, the last section concludes all important issues related to the research and future works.

Related Work

Data mining is a popular approach to discover new and meaningful knowledge in various domains using a large dataset. This approach offers multiple tasks such as classification,

clustering, association and sequential (Husam et al., 2017; Tomasevic et al., 2020). These tasks will produce different models based on the problem to be solved and available data. A few techniques can be used to produce the best model that gives the highest accuracy of the model.

Classification is one of the popular tasks in data mining. This task can produce a classifier model for solving problems. This model can be used in making a prediction based on the past dataset by identifying interesting patterns or useful patterns in the form of rules, trees or functions (Husam et al., 2017). A dataset will be divided into training and testing in order to create a model. A training set is used to create a model representing the whole dataset, while a testing (or validation) set is used to ensure that the model is accurate. There are a few techniques for splitting the data into training and testing, such as the hold-out method, cross-validation, and random subsampling.

To ensure the model produced is reliable, a few measurements can be used to identify the best model, such as accuracy, mean absolute error, root means squared error, f-measure, precision, recall and Kappa statistic. Table 3 shows the descriptions of measurements for the classifier model.

Table 3
Metric Measurements

Metric Measurement	Formula	Descriptions
Accuracy	(TP + TN)/ (TP + FN + FP + TN)	Accuracy is the proportion of the total number of predictions where correctly calculated
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{j=1}^{n}\lvert yj - y'j\rvert$	MAE is the average over the test sample of the absolute differences between prediction and actual observation, where all individual differences have equal weight.
Root Mean Square Error	$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(yj - y'j)^2}$	RMSE is the square root of the average squared differences between prediction and actual observation.
F-Measure	2∗((Precision∗Recall)/ (Precision + Recall))	F-Measure provides a single score that balances both the concerns of precision and recall in one number.
Precision	TP /(TP + FP)	Precision is the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases.
Recall	TP/(TP + FN)	The recall is the ratio of correctly classified samples to the total number of unclassified instances and correctly classified cases.
Kappa Statistic	(observed agreement−expected agreement)/(1−expected agreement)	Kappa statistic is defined when two measurements agree only at the chance level; the value of kappa is zero. When the two measurements agree perfectly, the value of kappa is 1.0.

Numerous algorithms, including Naive Bayes, Neural Network, Nearest Neighbor, Regression, Rough Set, and Decision tree, are included in the classification technique (Gopagoni & Lakshmi, 2020; Qian et al., 2021). Prior to experimentation and data collection, the Naive Bayes or Bayesian algorithm uses statistical methods to assign probabilities or distributions to occurrences or parameters based on prior knowledge or best assumptions. Using the Bayes theorem and strong (naive) independence assumptions, a Bayes classifier is a straightforward probabilistic classifier. The independent variables are a better description of the underlying probability model (Bhatia & Malhotra, 2021). Meanwhile, Neural Network (NN) is a computational or mathematical model based on imitating a biological neural system. Many neural network algorithms exist, including ANN, CNN, and Multilayer Perceptron. The aggregate of stimuli is typically transformed nonlinearly by the neuron to produce its output value. A few continuous functions adapt to the non-linear transformation in more complex models.

Regression is a statistical method to determine the strength and character of the relationship between one dependent variable and a series of other variables (known as independent variables). Meanwhile, rough set theory can be utilised to find structural correlations in noisy or erratic data, and it applies to characteristics with discrete values. Therefore, before use, continuous-valued properties must be discretised (Qian et al., 2021). The Decision tree algorithm will work by creating a decision tree based on data attributes. Many algorithms, including C4.5, J48, and random forest trees, are based on the decision tree technique. This algorithm finds the characteristic that most effectively distinguishes across instances. The accuracy of the model can be used to evaluate the quality of the rule, tree, or function produced by this algorithm (Amirah et al., 2016; Husam et al., 2017)

Due to the enormous growth of educational data, data mining has recently become increasingly popular in education. The development of databases allows data mining to be used to extract relevant information from this data. This led to the emergence of Education Data Mining called EDM (Alyahyan & Düştegör, 2020). EDM is crucial for uncovering hidden patterns or interesting knowledge in the educational sector, such as predicting students' academic success. Students' success is critical to higher institutions because it is considered an important criterion in accessing the quality of educational institutions. Thus, many past studies have discussed academic performance from different perspectives, such as knowledge score, grade and performance. They used different criteria or features in measuring academic performance, such as GPA Poudyal et al (2022); Ahmed et al (2018) tests Yagci (2022) quizzes Poudyal et al (2022), course grades Nabil et al (2021) IELTS score (Ghazal & Allah, 2020)and pre-admission test (Mengash, 2020). However, the CGPA and GPA of students are the most common indicators used by researchers in measuring students' academic performance.

A classifier model for predicting academic achievement incorporates a variety of factors such as demographic, educational, social, and family backgrounds (Poudyal et al., 2022). Age, gender, race, and place of residence are examples of demographic attributes; quizzes, midterm and assignment marks or grades, attendance percentage and learning strategies are examples of educational attributes. Meanwhile, examples of social attribute categories include lifestyle, time spent on social media, number of close friends, etc. The family backgrounds include the number of children, the family's income, the parents' educational level, etc.

Most of the past research produced a model by comparing different machine learning algorithms and selecting the best model using different measurements. The accuracy metric

is a standard measurement metric used in evaluating the model. Yagci (2022) proposed a model to predict the final exam grades of undergraduate students by taking their midterm exam grades as the source data. The study produced a prediction model with 73% accuracy using the Random forest algorithm. This study compared a few machine learning algorithms such as Logistic Regression, Naïve Bayes, and K-Nearest Neighbour. Meanwhile, Razak et al. (2018) produce an accurate prediction model using Linear Regression with 96.2% accuracy. Hasan et al (2020) produced a model to predict students' overall performance at the end of the semester using data from the student information system, learning management system and mobile applications. The results showed that Random Forest accurately predicted successful students at the end of the class with an accuracy of 88.3%. Meanwhile, Ahmed et al (2018) defined the best model for performance prediction using a Decision Tree algorithm with 97.69 % accuracy. Besides using accuracy measurement, Ahmed et al. (2018) used four standard measures for evaluating classification quality: accuracy, precision, recall and F-measure. Regarding the four standard measures, Decision Tree techniques obtained more than 95% accuracy compared with other classifications. Most researchers get the accuracy between 70% to 98% of the best model to predict students' academic performance using different machine learning algorithms. It shows that the techniques used to produce the best model are based on the data condition, not the domain data.

EDM is now important in identifying certain factors that can influence the decision. For example, some attributes will have an impact on a prediction of academic achievement. There are several techniques for discovering interesting attributes that can affect the decision. Feature selection is one of the approaches to know which attributes are relevant to the decision. Numerous approaches, including Information Gain, Symmetrical Uncertainty, Gain Ratio, and Correlation-based Feature Subset Selection, are available to choose the optimal attributes to represent the entire dataset.

Thus, this study was conducted to develop an academic performance prediction model using different classification algorithms and define which attributes influence students' performance. The model was developed based on demographic information, educational background and students' learning styles. The contribution of this work is beneficial to see some of the factors that influence academic performance. Besides that, the model produced can be used as a guideline to predict the performance of new students registered as UiTM students.

Material and Methods

This research aims to develop a classifier model that can predict students' academic performance using classification techniques. The standard data mining methodology has been applied to achieve the objective of the study, which consists of 4 main steps; (1) Defining Business Goal, (2) Data Collection, (3) Data Preprocessing, and (4) Model Development. Figure 1 below shows a diagram of mining steps in detail of the model comparison for investigating the best model for predicting students' performance and defining the factors that influence the result.

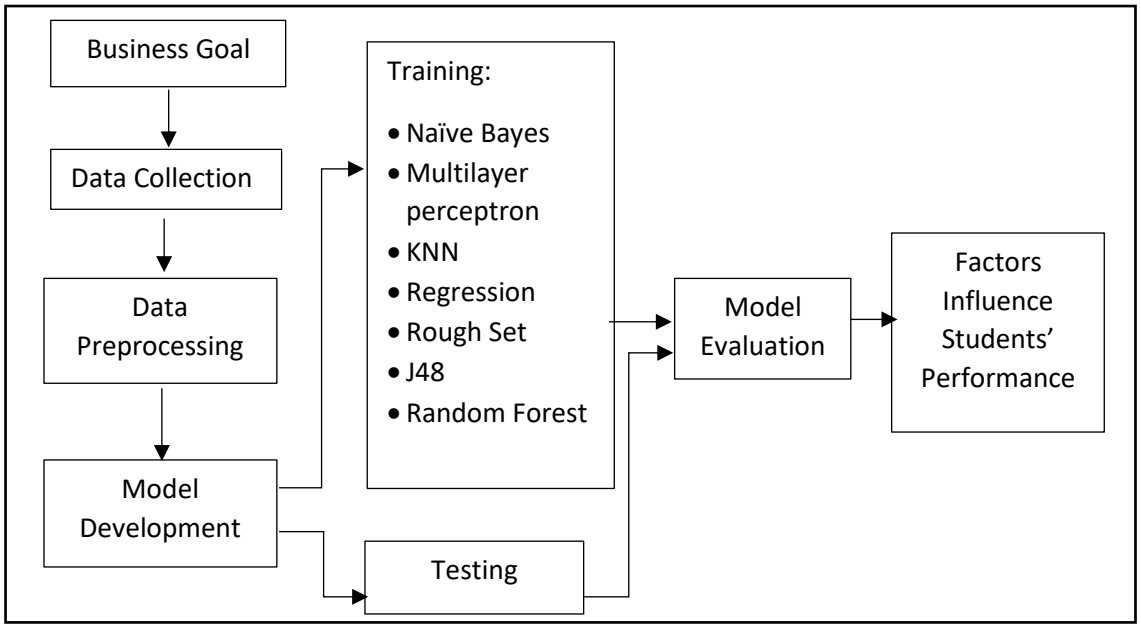

Figure 1: A Methodology for Producing a Classifier Model for Academic Students' Performance Prediction

Defining Business Goal

The study's business goals are to identify academic performance patterns among students with different backgrounds and learning characteristics. Factors that contribute to the goal will be determined through feature selection from the source data set.

Data Collection

The sample data is collected through a survey of students at the UiTM Negeri Sembilan, Seremban Campus. The total number of respondents is 233 undergraduate students from two faculties which are Faculty of Administrative Science and Policy Studies (FSPPP) and Faculty of Computer and Mathematical Sciences (FSKM). Unfortunately, due to difficulties gathering data from students from both faculties due to the COVID-19 pandemic, just a few students were chosen as respondents. FSPPP is a social science (SS) stream, and FSKM is a science and technology (S&T) stream. About 136 respondents from FSPPP and 97 respondents from FSKM. In the sample, 48 are males, and 185 are females. Data obtained consisted of 14 attributes; three attributes are from demographic, ten attributes from educational background, and one attribute for students' learning styles. The attribute for learning style was produced from 20 questions from VAK (Visual, Auditory and Kinaesthetic) learning styles. Table 4 below shows the list of attributes for predicting academic performance.

Table 4
A List of Attributes for Predicting Academic Performance

No.	Questions	Answers
1	Age	Numeric Value
2	Gender	Female/Male
3	Hometown	City/Village
4	School Location	City/Rural
5	Type of school	Public/Private
6	SPM result	Numbers of A, B, C, D, E, G
7	Qualification before tertiary education	SPM/STPM/Matriculation/Diploma/Degree
8	Level of study	Diploma/Degree
9	Seniority	1/2/3/4/5/6/7 Above
10	Faculty	FSKM/FSPPP
11	CGPA	Numeric
12	Father's education level	SRP/SPM/STPM/Matriculation/Diploma/Degree/Master/PHD
13	Mother's education level	SRP/SPM/STPM/Matriculation/Diploma/Degree/Master/PHD
14	Learning Styles	20 Questions related to VAK learning styles

Data Preprocessing

Data preprocessing is an essential phase in the data mining step. This phase will influence the accuracy of the model. In this study, this phase involves data preparation and data cleaning. The first step is data preparation. Data preparation involves determining a class target attribute, some condition attributes, and processing learning styles output. The second step is the cleaning process which consists of data transformation and discretisation. These two processes purposely clean the dataset, including removing some attributes, generating new attributes, replacing incomplete, and handling inconsistent data. Most of the methods used in this study only work with categorical data; hence, all continuous attributes have been discretised and transformed into the appropriate format for easy implementation. Furthermore, data preprocessing tasks conducted in this study are discussed below:

a. Data Preparation

The type of school and level of study attributes have been removed from the dataset because the data sample gives a biased value which is all respondents were from public school, and the level of study is degree level. Instead, a few condition attributes have been chosen to represent the whole dataset. In addition, the current CGPA attribute has been determined as a class target to represent the students' performance. Meanwhile, the outliers from the dataset have been removed, for example, a record with age 45 and the values in parents' education attributes which are not related.

b. Data Transformation

Three new attributes have been produced by processing existing attributes. SPM result has been split into three new attributes: excellent, credit, and non-credit. Meanwhile, the learning styles attribute has been created based on the VAK model questionnaire that consists of 20 questions. There are three learning styles: Visual, Kinaesthetic, and Auditory. The questionnaire in this section has three options such as 1, 2, and 3, to be selected. Students will identify their learning styles based on the numbers 1, 2, and 3 options in the questionnaire. The rules for the VAK model questionnaire are shown in table 5 below.

Table 5
Rules for Learning Styles

If the sample of Students chose mostly 1, then Visual Learning Style
If the sample of Students chose mostly 2, then Auditory Learning Style
If the sample of Students chose mostly 3, then Kinaesthetic Learning Style

However, if the sample students choose the same number of options 1, 2, and 3, their learning styles may be mixed between two or three learning styles and labelled as multimodal (MM).

c. Discretisation

This process involves nine attributes which are age, number of excellent subjects, credit subjects and existing non-credit subjects in SPM, seniority, current CGPA, parents' education level, and learning styles. In addition, all CGPA results have been categorised into three groups: First Class, Second Class Upper, and Second Class Lower with Third Class. Meanwhile, for the learning styles attributes, the total marks produced from the VAK questionnaire will be discretised into nominal values based on the rules in Table 5.
Table 6 below shows the result after preprocessing process on dataset.

Table 6
A List of Attributes After Preprocessing Process

No.	Questions	Answers
1	Age	18-21 22-25
2	Gender	Female/Male
3	Hometown	City/Village
4	School Location	City/Rural
5	Excellent Subject (SPM)	<4, 4-6, >6
6	Credit Subject (SPM)	<4, 4-6, >6
7	Non-Credit Subject (SPM)	<4, >=4
8	Qualification before tertiary education	STPM/Matriculation/Diploma
9	Seniority	Sem 1-2, Sem 3-4, Sem >4
10	Faculty	FSKM/FSPPP
11	Current CGPA	1st (3.5-4.0) 2nd Upper (3.00-3.49) 2nd Lower & 3rd (2.00-2.99)
12	Father's education level	Level 1, 2, 3, 4
13	Mother's education level	Level 1, 2, 3, 4
14	Learning Style Preferences	Visualisation, Auditory, Kinesthetic, Multimodal

Model Development

A classification model will be created using the clean dataset, and the model will categorise the data according to the class target. The characteristics of students' performance will be determined using this model, and the factors affecting their success will be acknowledged.

The classifier model will be developed in WEKA (Waikato Environment for Knowledge Analysis) tool using different existing techniques. In this study, seven machine learning algorithms for classification task have been employed; Naïve Bayes, Multilayer perceptron, K-Nearest Neighbours (KNN), Regression, Rough Set and two Decision Tree techniques which are J48 and Random Forest algorithm. The dataset will be split into training and testing datasets using four different approaches such as 10-Folds Cross Validation (10-folds), 20-Fold Cross Validation (20-folds) and hold-out method with 80:20 and 70:30 ratios for training and testing data. The models produced using various classification and splitting techniques have been evaluated using accuracy and Kappa value. Additionally, this experiment used a reduction technique which uses symmetrical uncertainties (SU) and Classifier Attribute Evaluation (CAE). The dataset's attributes will be pruned one by one based on the value from CAE until it will produce the best model to predict the new dataset. The potential factors influencing the model produced will be identified using the ranking produced from reduction techniques. The highest ranking of attribute show that attribute highly contributes to the output.

Experiment and Result Analysis

This study performed two types of analysis: descriptive analysis and predictive analysis. Three descriptive analyses were performed on the dataset to describe the data's current trends or

conditions. Meanwhile, several categories of different algorithms were tested for predictive analysis by iterating the process and splitting the data into different percentages to determine the best predictive model. Furthermore, feature selection strategies were used to determine the factors influencing the students' performance.

i. Academic Performance Descriptive Analysis

Three descriptive analyses which are the distribution of demographics based on students' stream; Science and Technology (S&T) and Social Science (SS), Statistical analysis of education backgrounds, and the distribution of learning styles based on streams and Cumulative Grade Point Average (CGPA).

a. Distribution of demographics based on streams

Demographic criteria consist of age, gender, and hometown. The analysis distribution is performed by comparing the students from two streams: S&T and SS. All the students' age from the sample is between 18 to 25. However, there exists one student whose age is 31, and it is considered an outlier in data. Based on Figure 1, the SS and S&T samples are almost balanced, 58% and 42% respectively. Meanwhile, the number of students who come from rural and city areas from both faculties are 52% and 48% correspondingly. The students from S&T manage to get a high number of First-Class students compared to SS, especially students from rural areas as in Figure 2 below.

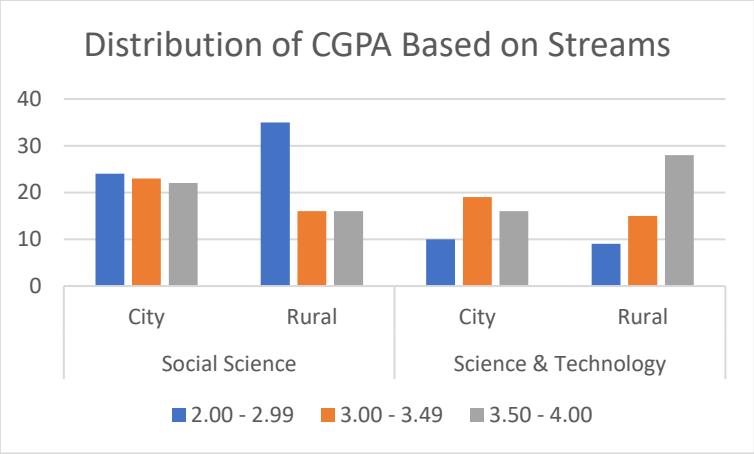

Figure 2: Distribution of Demography Criteria Based on Streams

b. Statistical analysis of education backgrounds

Table 7 below shows the description of attributes under the Education category, which represent the information on students' educational backgrounds involving their school area, education information in UiTM, Sijil Pelajaran Malaysia (SPM) results and their parents' education level.

Table 7
Statistical Analysis of Education Background Attributes

Education Backgrounds	**Statistical Value**	**Analysis**
School Location	Rural - 113 (48%) City – 120 (52%)	The respondents' school locations are rural and city, and these two values are almost balanced.
Excellent Subject (SPM)	1-3 - 79 (34%) 4-6 -126 (54%) >6 - 28 (12%)	This attribute is the SPM result which represents the number of A the respondents manage to get. Most of them tend to get 4 to 6 A's, and only 12% successfully get 6A's and above.
Credit Subject (SPM)	1-3 -144 (62%) 4-6 -71 (30%) >6 -18 (8%)	Most of the respondents get 1 to 3 passed subjects between B+ to C, and fewer get more than six credit subjects.
Non-Credit Subject (SPM)	<4 -224 (96%) >=4 -9 (4%)	Fewer students fail many subjects in SPM. About 4% of the students fail more than three subjects.
Qualification before tertiary education	STPM -47 (20%) Matriculation -90 (39%) Diploma -96 (41%)	Almost half of the students are from diploma qualification before entering bachelor degree, and just a few are from STPM qualification. However, about 39% are from matriculation.
Seniority	Sem 1-2 -102 (44%) Sem 3-4 -85 (36%) Sem >4 -46 (20%)	Most of the respondents are from part 1 and part 2, and it seems they get 1 to 2 GPA to accumulate into CGPA. Meanwhile, 36 % of them from part 3 and 4 and 20% are seniors from part 5, 6 and 7.
Faculty	FSKM -97 (42%) FSPPP -136 (58%)	The respondents are from two streams: Science and Technology for FSKM faculty and Social Science for FSPPP. These two streams are almost imbalanced in identifying students' performance from both faculties.
Father's education level	Level 1 -43 (18%) 2 -108 (46%) 3 -39 (17%) 4 -43 (18%)	About four levels of father education, level 1 represents the primary and secondary school with 18%. Meanwhile, the second level is for those who finished secondary school and got SPM certification, with 46% of them the highest. The third level, pre-university, consists of STPM, diploma or foundation with 17%. The last level is for higher education, which includes

		a bachelor's degree, master's degree and doctor of philosophy with 18%.
Mother's education level	Level 1 -27 (12%) 2 -118 (51%) 3 -47 (20%) 4 -41 (17%)	For the mother's education level, half of them, 51%, are from level 2 and less had a qualification in higher education and primary and secondary school.
Current CGPA	1st (3.5-4.0) -81 (35%) 2nd Upper (3.00-3.49) -73 (31%) 2nd Lower &3rd (2.00-2.99) -79(34%)	This attribute is a decision attribute to represent students' academic performance. The CGPA is calculated by accumulating GPA from all finished semesters. About 35% of them are successful in getting the First Class result. Meanwhile, 31% managed to get 2nd Upper, and 34% got 2nd Lower and 3rd class.

c. Learning Styles Distribution

Below is the figure that shows the learning styles among students. Most students are visual learners, and fewer are auditory learners. However, S&T students preferred the kinesthetic style compared to Social Science, which chose the visual style. Some students prefer to use more than one style called multimodal(MM), as in this study, they prefer to combine visual and kinesthetic.

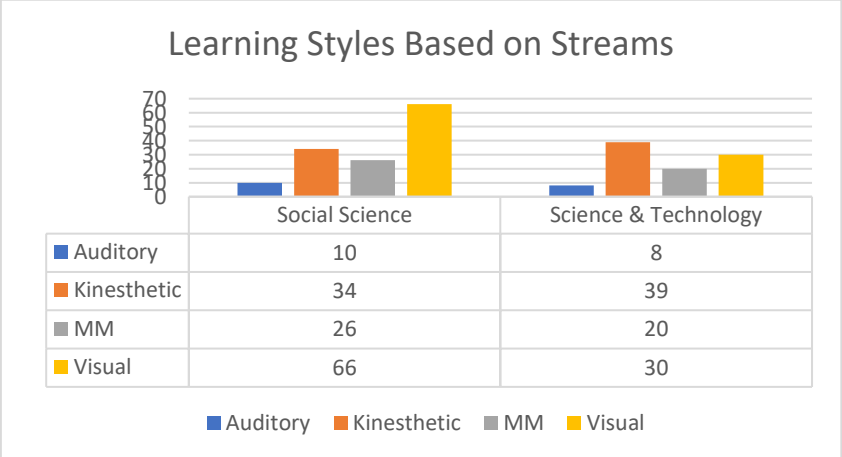

Figure 3: The Average Learning Styles

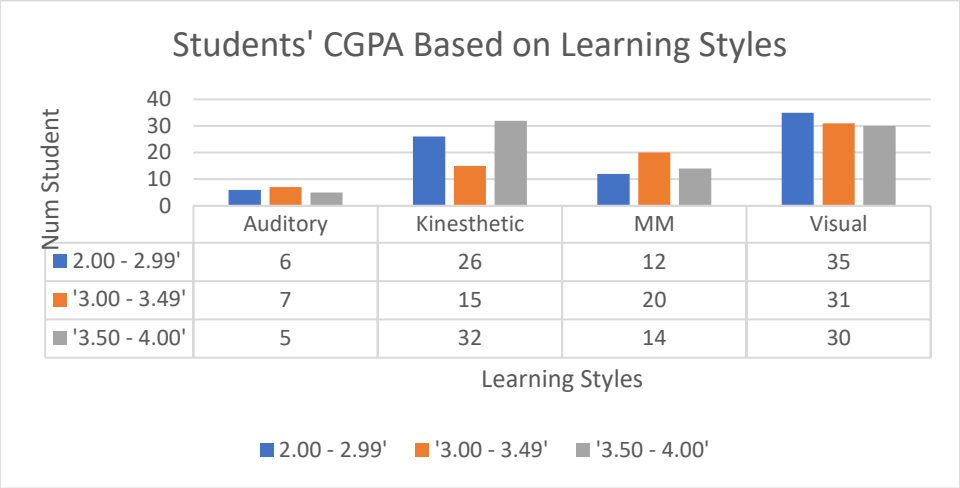

Figure 4: Academic Performance Based on Learning Styles

Based on Figure 4, the students who managed to get the First Class of CGPA are kinesthetic and visual learners. Meanwhile, for the students under Second Upper and Lower of CGPA, are visual learners. However, based on the results, learning styles do not influence students' performance, where all the learning styles give almost the same ratio of CGPA.

ii. Academic Performance Predictive Analysis

Classification Experiment Result

Table 8 demonstrates how well each model predicts students' performance, whether they managed to get First Class, Second Upper or Second Lower and Third Class. In comparing all the models produced, the Random Forest technique got an accurate model with an accuracy of 76.59%, and the Kappa test value is 0.6473 using the Hold-out validation technique. Meanwhile, K-Nearest Neighbour(KNN) manage to get the highest result for 20-Fold cross-validation with 74.25% accuracy, and the Kappa test is 0.6143. However,KNN has been selected as the best model to predict the students' performance compared to Random Forest. It is because the K-Fold technique is more accurate and efficient in splitting the data into training and testing, considering all possible values compared to the hold-out technique. The accuracy between these two techniques is slightly different. An accuracy of more than 70% is considered accurate, and the Kappa test value with more than 0.4 to 0.7 is considered a moderate value for model reliability.

The decision tree category technique is one technique that can produce an interpretability model in the form of "IF..Then" rules. Below is a set of rules in figure 5 produced from the Random Forest tree. These rules can be used in making decisions or guidelines for students' predictions.

Hence, the results showed that the performance of classification algorithms was low when the model was produced from all dataset attributes. The results become high when some attributes are reduced from the dataset. It shows that some of the attributes disturb the accuracy and reliability of the model in making a prediction. Based on the experiment, most classification algorithms produce a more accurate model when three attributes have been removed from the dataset; hometown, father and mothers' education background.

Table 8
Comparison of Classifier Model using Using Classification Algorithms

Algorithms	Splitting Techniques							
	Hold Out Validation (80:20)		**Hold Out Validation (70:30)**		**10-Fold Cross Validation**		**20-Fold Cross Validation**	
	Before Reduction (%)	**After Reduction (%)**	**Before Reduction (%)**	**After Reduction (%)**	**Before Reduction (%)**	**After Reduction (%)**	**Before Reduction (%)**	**After Reduction (%)**
Naïve Bayes	Acc-57.45 Kappa – 0.3548	Acc-59.57 Kappa – 0.3871	Acc-61.43 Kappa – 0.4092	Acc-62.85 Kappa – 0.4302	Acc-64.81 Kappa – 0.4725	Acc-68.24 Kappa – 0.5243	Acc-66.09 Kappa – 0.4915	**Acc-69.10 Kappa – 0.5366**
Multilayer perceptron	Acc-59.57 Kappa – 03946	Acc-65.71 Kappa – 0.4807	Acc-67.14 Kappa – 0.4939	Acc-67.14 Kappa – 0.4939	Acc-68.67 Kappa – 0.5287	Acc-72.10 Kappa – 0.5814	**Acc-73.39 Kappa – 0.65**	Acc-72.10 Kappa – 0.5804
KNN	Acc-68.09 Kappa – 0.5201	Acc-68.09 Kappa – 0.5201	Acc-64.29 Kappa – 0.4609	Acc-62.86 Kappa – 0.4466	Acc-66.52 Kappa – 0.4965	Acc-73.82 Kappa – 0.6071	Acc-70.39 Kappa – 0.5544	**Acc-74.25 Kappa – 0.6134**
Regression	Acc-63.83 Kappa – 0.4539	Acc-65.96 Kappa – 0.4846	Acc-65.71 Kappa – 0.4771	Acc-68.57 Kappa – 0.5214	**Acc-71.67 Kappa – 0.5735**	Acc-68.24 Kappa – 0.5218	Acc-72.53 Kappa – 0.5868	Acc-69.53 Kappa – 0.5415
Rough Set	Acc-59.57 Kappa – 0.3917	Acc-63.83 Kappa – 0.4509	Acc-61.43 Kappa – 0.4062	Acc-61.43 Kappa – 0.4008	**Acc-67.38 Kappa – 0.5085**	Acc-65.67 Kappa – 0.4829	Acc-67.81 Kappa – 0.5149	Acc-64.51 Kappa – 0.4507
J48	Acc-68.08 Kappa – 0.5201	Acc-68.08 Kappa – 0.5201	Acc-70.00 Kappa – 0.5429	Acc-70.00 Kappa – 0.5429	Acc-70.00 Kappa – 0.5482	Acc-72.03 Kappa – 0.5808	Acc-72 Kappa – 0.6114	**Acc-72.03 Kappa – 0.5808**
Random Forest	Acc-72.34 Kapa – 0.5824	**Acc-76.59 Kappa – 0.6473**	Acc-65.71 Kappa – 0.4784	Acc-74.29 Kappa – 0.4784	Acc-70.82 Kappa – 0.5612	Acc-75.96 Kappa – 0.6392	Acc-73.10 Kappa – 0.6000	Acc-74.25 Kappa – 0.6131

```
4.      What is your qualification before continue your current study? = Matriculation
|   1.  Gender. = Male
|   |   1.      Where is the location of your last/upper secondary school? = Rural: 3.00 - 3.49 (6.0/1.0)
|   |   1.      Where is the location of your last/upper secondary school? = City
|   |   |   9.  What is your father's education level? = Level 2: 2.20 - 2.99 (1.0)
|   |   |   9.  What is your father's education level? = Level 4: 3.00 - 3.49 (4.0)
|   |   |   9.  What is your father's education level? = Level 3: 2.20 - 2.99 (7.0/2.0)
|   |   |   9.  What is your father's education level? = Level 1: 2.20 - 2.99 (0.0)
|   1.  Gender. = Female
|   |   2.      Age. = 22-25
|   |   |   7.  What is your faculty? = Faculty of Administrative Science and Policy Studies: 2.20 - 2.99 (3.0/1.0)
|   |   |   7.  What is your faculty? = Faculty of Computer and Mathematical Sciences
|   |   |   |   1.      Where is the location of your last/upper secondary school? = Rural
|   |   |   |   |   Excellent = '(-inf-3.333]': 3.50 - 4.00 (0.0)
|   |   |   |   |   Excellent = '(3.333-6.667]': 3.00 - 3.49 (3.0/1.0)
|   |   |   |   |   Excellent = '(6.667-inf)': 3.50 - 4.00 (2.0)
|   |   |   |   1.      Where is the location of your last/upper secondary school? = City: 3.00 - 3.49 (6.0)
|   |   2.      Age. = 18 - 21
|   |   |   1.  Where is the location of your last/upper secondary school? = Rural: 3.50 - 4.00 (35.0/1.0)
|   |   |   1.  Where is the location of your last/upper secondary school? = City
|   |   |   |   7.      What is your faculty? = Faculty of Administrative Science and Policy Studies: 3.00 - 3.49 (7.0/2.0)
|   |   |   |   7.      What is your faculty? = Faculty of Computer and Mathematical Sciences: 3.50 - 4.00 (16.0/3.0)
```

Figure 5: A Set "If. Then" Rules Produced using Decision Tree

Table 9 below shows the correlation value of the CGPA attribute, a class target representing students' performance with other attributes in the dataset. The results show that students' performance is correlated with other attributes in the form of symmetrical uncertainty value ranking. The first ranking correlated with the class target is the qualification before tertiary education with a 0.19959 value. It demonstrates how crucial this criterion is in determining whether students may succeed in higher education. This criterion consists of qualifications from matriculation, diploma, and STPM. The findings indicate that most matriculation students achieve higher results than their peers. The gender attribute comes in second. Compared to male students, most female students achieve the highest results, and most are visual learners. Meanwhile, the third-ranking is seniority, with a value of 0.09587. To sustain their performance until the final semester, most students must ensure they receive the highest grade possible beginning in the first semester. Most students are encouraged to work

more or maintain their achievements after beginning to achieve high CGPAs in the first semester. Additionally, the number of excellent subjects in SPM also influences the performance of high institution students.

In contrast, the three factors with the lowest rankings, hometown and the educational degree of the parents, did not significantly affect children's performance. Despite applying seven different classification algorithms, the CAE approach shows that these three criteria are not crucial. All methods place these factors at the very bottom when evaluating student performance. However, other factors, including school location, SPM failure, school location, streams, and learning styles, are still crucial in determining a student's achievement.

Table 9
Correlation of CGPA with Other Attributes using Symmetrical Uncertainty

Attribute	**Symmetrical Uncertainty**
Qualification before tertiary education	0.19959
Gender	0.10363
Seniority	0.09587
Excellent	0.06431
School Location	0.0522
Fail	0.04641
Stream	0.04045
Age	0.03319
Pass With Moderate	0.02962
Learning Styles	0.01643
Father's education level	0.01134
Mother's education level	0.01011
Hometown	0.00862

Conclusion

This paper has shown how data mining helps identify interesting knowledge from a dataset that influences students' academic performance. The model is produced based on demography, education and learning style attributes in the form of rules generation. The rules are produced based on the current CGPA of students, which consists of First Class, Second Class Upper and Second Class Lower, and Third Class. This model can be used to predict the students, whether they perform well or not in higher institutions. In this study, the prediction model from KNN and Random Forest algorithms produced accurate decisions, especially for a new dataset with 74.25% and 76.59% of accurateness. Meanwhile, the qualification before tertiary education highly impacts students' performance, where most of the students from matriculation manage to get the highest result in higher education. Besides that, the female students scored well compared to male students. Meanwhile, the students who are successful in their SPM level will achieve good result in higher education.

The descriptive analysis provided a clear picture of analytical data capabilities in the pursuit of more detailed knowledge to assist in making a decision for students' academic performance. Most students who manage to get 3.0 to 4.0 CGPA are senior students. A few Second Upper students have the same criteria as the First-Class students, but there are from part 1 and part 2 students. The result increases after they repeat the failure subjects.

Meanwhile, the learning styles do not really influence students' performance, but most of the students who are successful in their studies are kinesthetic and visual learners.

In the future, this study can be enhanced by collecting a huge number of respondents from different faculties and different backgrounds of the study; to ensure the model produced is more accurate due to the limited amount of data that reflects the students' background from the whole UiTM system. Besides that, many attributes can be considered in identifying academic performance, such as utilising the Learning Management System (LMS), students learning time (SLT), teaching approach, number of assessments and others. Thus, it can help students and lecturers identify the best approach to improving and achieving the best academic performance.

References

Ahmed, R. M., Ali, A. A., Omran, N., & Omran, N. F. (2018). Predicting and Analysis of Students' Academic Performance using Data Mining Techniques. In *International Journal of Computer Applications* (Vol. 182, Issue 32). https://www.researchgate.net/publication/347520956

Alyahyan, E., & Düştegör, D. (2020). Predicting Academic Success In Higher Education: Literature Review And Best Practices. In *International Journal of Educational Technology in Higher Education* (Vol. 17, Issue 1). Springer. https://doi.org/10.1186/s41239-020-0177-7

Amirah, W., Azlan, W., Liew, S.-H., Choo, Y.-H., Zakaria, H., & Low, Y. F. (2016). *Wavelet Feature Extraction And J48 Decision Tree Classification Of Auditory Late Response (Alr) Elicited By Transcranial Magnetic Stimulation*. *11*(10). http://users.aber.ac.uk/rkj/book/wekafull.jar.

Bhatia, S., & Malhotra, J. (2021). Naïve Bayes Classifier For Predicting The Novel Coronavirus. *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*, 880–883. https://doi.org/10.1109/ICICV50876.2021.9388410

Ghazal, A., & Allah, F. (2020). Using Machine Learning To Support Students' Academic Decisions. *Journal Of Theoretical and Applied Information Technology*, *98*(18). www.jatit.org

Gopagoni, D. R., & Lakshmi, P. v. (2020). Automated Machine Learning Tool: The First Stop For Data Science And Statistical Model Building. *International Journal of Advanced Computer Science and Applications*, *2*, 410–418. https://doi.org/10.14569/ijacsa.2020.0110253

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting Student Performance In Higher Educational Institutions Using Video Learning Analytics And Data Mining Techniques. *Applied Sciences (Switzerland)*, *10*(11). https://doi.org/10.3390/app10113894

Husam, I. S., Abuhamad, A. A. B., Zainudin, S., Sahani, M., & Ali, Z. M. (2017). Feature Selection Algorithms For Malaysian Dengue Outbreak Detection Model. *Sains Malaysiana*, *46*(2), 255–265. https://doi.org/10.17576/jsm-2017-4602-10

Ilçin, N., Tomruk, M., Yeşilyaprak, S. S., Karadibak, D., & Savcl, S. (2018). The Relationship Between Learning Styles And Academic Performance In Turkih Physiotherapy Students ,Specialist Studies In Education. *BMC Medical Education*, *18*(1). https://doi.org/10.1186/s12909-018-1400-2

Melo, G. O. (2018). *Enhancing Oral Skill Through Learning Styles VAK Enhancing Oral Skill Development Through Learning Styles VAK*. http://campus.educadium.com/newmediart/file.php/137/Thesis_Repository/recds/assets/T

Mengash, H. A. (2020). Using Data Mining Techniques To Predict Student Performance To Support Decision Making In University Admission Systems. *IEEE Access*, *8*, 55462–55470. https://doi.org/10.1109/ACCESS.2020.2981905

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks. *IEEE Access*, *9*, 140731–140746. https://doi.org/10.1109/ACCESS.2021.3119596

Nuankaew, P., Nuankaew, W., Phanniphong, K., Imwut, S., & Bussaman, S. (2019). Students Model In Different Learning Styles Of Academic Achievement At The University Of Phayao, Thailand. *International Journal of Emerging Technologies in Learning*, *14*(12), 133–157. https://doi.org/10.3991/ijet.v14i12.10352

Poudyal, S., Mohammadi-Aragh, M. J., & Ball, J. E. (2022). Prediction of Student Academic Performance Using a Hybrid 2D CNN Model. *Electronics (Switzerland)*, *11*(7). https://doi.org/10.3390/electronics11071005

Qian, W., Huang, J., Wang, Y., & Xie, Y. (2021). Label Distribution Feature Selection For Multi-Label Classification With Rough Set. *International Journal of Approximate Reasoning*, *128*, 32–55. https://doi.org/10.1016/j.ijar.2020.10.002

Razak, R. A., Omar, M., & Ahmad, M. (2018). A Student Performance Prediction Model Using Data Mining Technique. In *International Journal of Engineering & Technology* (Vol. 7, Issue 2).

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An Overview And Comparison Of Supervised Data Mining Techniques For Student Exam Performance Prediction. *Computers and Education*, *143*. https://doi.org/10.1016/j.compedu.2019.103676

Yağcı, M. (2022). Educational Data Mining: Prediction Of Students' Academic Performance Using Machine Learning Algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z