 https://doi.org/10.24191/jeesr.v23i1.004

Feature Selection Methods Application Towards a New Dataset based on Online Student Activities

Muhammad Hareez Mohd Zaki, Mohd Azri Abdul Aziz*, Suhana Sulaiman and Najidah Hambali

***Abstract*—The increasing usage of classification algorithms has encouraged researchers to explore many topics including academic-related topics. In addition, the availability of data from various academic information management systems in recent years had been increasing, causing classification to become a technique that is in demand by the educational institutes. Thereby, having a classification technique is important in researching the data of students' performance. The problem during the classification of students' performance is the lack of factors used to identify and evaluate their performance. Most of the articles used students' grades as the most influential factor to predict students' performance. Thus, more significant features are needed to evaluate students' performance to improve the existing method. Due to the reason, a dataset is proposed to introduce some features that can affect the students' performances. The dataset's features are based on online students' activities during e-learning. This study will perform Analysis of Variance Test (ANOVA), Chi-squared Test, Recursive Feature Elimination (RFE) and Extra Tree algorithm (ET) as feature selection methods to pre-process the proposed dataset that is considered raw data. The experimental results showed that 'Answered all questions', 'After-class notes', 'Correct 3 and above' and 'In-class notes' were the most significant features in evaluating students' performance. The study is significant towards educational data mining in analysing the students' performance during online students' activities.**

***Index Terms*—Student performance classification, feature selection, ANOVA, chi-squared, recursive feature elimination, extra tree**

I. INTRODUCTION

STUDENT performance has been a widely explored research topic in the past few years [1]–[5] due to exciting information that can assist educators and students, mainly when advanced algorithms are applied [6]. The growing research on student performance had involved variety of features in analysing the performance. Different authors used different features in conducting their student performance research. Some research used demographical features, students' personal information to study the student performance [2], [4], [7]–[9]. Other research also used psychometric features that related to students' behaviour and mental development during their studies [10]–[14]. Commonly, most of the research used academic features such as CGPA, internal assessment, External assessment, Examination final score and extra co-circular activities of the student as prediction criteria [1], [8], [15]–[18]. Thus, previous research showed that many categories of features can be used for analysing students' performance. There was no specific features to study their performance as students would be influenced by socioeconomic, psychological, and environmental factors [19].

The problem during the classification of students' performance is one of negligence factors used to identify and evaluate their performance [8].Most of the research that studied students' performance were narrowed to students' grades and GPAs as features to analyse their performance instead of learning outcomes and other main factors that influenced their performance [5][20]. This can be prevented by using variety of features during a research. The weaknesses of having the solution is the used of many features might not coordinate well with the aim of a research because of the widely exposed features. The features used in a research should be varied especially in categories. The diversity in the categories of the features in a research is new and has yet to be explored. As a result, other determining factors could be found to identify the factors that influenced students' performance.

The advantages of having variation of features are the researchers can discover many possibilities of results and they also might discover that the actual outcomes are differ from the expected outcomes. New features that are used in research usually performed feature selection process during the pre-processing stage. There are a few feature selection techniques that are used in the previous research; Analysis of Variance Test (ANOVA), Chi-squared Test, Recursive Feature Elimination (RFE) and Extra Tree algorithm (ET). ANOVA tests are widely used due to its capability that could choose high relevance features as stated in [21], [22]. This test can also be used to handle curse of dimensionality issue [23]. The advantage also applied to Chi-squared Test. Chi-squared Test are widely used in educational data mining (EDM) especially in analyzing students' performance [24]–[28]. Besides, RFE is also

This manuscript is submitted on 14th December 2022 and accepted on 17th July 2023. M.A.A.Aziz, S.Sulaiman, and N.Hambali are with the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor (azriaziz@uitm.edu.my).

Muhammad Hareez Mohd Zaki was with the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA. (email : 2020480662@isiswa.uitm.edu.my).

*Corresponding author
Email address: azriaziz@uitm.edu.my

commonly used by many researchers due its ranking feature that made them easier to differentiate between the features. Some research that used this method were students' performance [19], [29]–[31], early detection of at-risk students [32] and microarray gene expression [23].Besides, there were also research that applied Extremely randomised trees algorithm (ET), a tree based- ensemble method for feature selection method due to its generalisation ability [33]–[35].

New dataset was proposed to be utilized in this research. The objective of this research is to find out the significant features and to test the robustness of the new proposed dataset. The features were accumulated directly from the website that undergoes online learning, and all of them are based on their activities and learning outcomes during online classes. The proposed dataset consists of a few categories. Some categories such as accessing notes, exercises and assignments have sub-categories to determine the students' behaviour and the outcomes of the assignments during the class. According to previous research, these categories are psychometric and academic factors that contribute to the performance of the students [36]–[46]. The proposed features are also from the same categories and more detailed on students' behaviour when accessing their notes and exercises. Prior to the proposed dataset; it is a raw dataset and has never been processed. Hence, a few feature selection methods will be carried out on the dataset to extract the significant features and examine the robustness of the data. Some significant features from the combination of psychometric and academic categories, were expected to be chosen after features selection process. The chosen features could be broaden as future references to analyse students' performance.

In this paper, feature selection on the proposed dataset had been performed. Several methods of feature selection methods such as ANOVA, Chi-squared test, RFE and Extra Tree, were discussed and applied to the dataset during the pre-processing stage. This research makes contributions to researchers and practitioners in students' performance classification by introducing a new dataset that focuses on the academic and psychometric factors rather than grades to analyse the students' performance.

The rest of this paper is organised as follows. In Section II, presented the literature review of other researches on the features and methods used to study students' performance. Section III elaborated the feature selection methods used in this research. Section IV was on result of the research. The conclusion on the paper is presented in section V.

II. Related Studies

Studies on students' performance are still progressing until the present day. It is important for universities or any educational sector to adapt to the students' growing development throughout their studies. Many data are used for student performance evaluation. However, no specific criteria or features can study students' performance as each student has varieties of personalities and backgrounds with a different history that may impact their future studies [1]. Commonly, researchers used classification, a part of the data mining function, to study and gain more insights regarding student performance classification.

Previous research showed that student performance studies involved various data used for the classification process. Due to this reason, there are a few distinctive categories that the researchers use to group each data; demographical factor, psychometric factor and academic factor. Demographical factors consist of gender/sex, family size, marital status, religion, place of birth, father occupation, mother occupation, father qualification, mother qualification, parental status, parental income status, attendance, the profile of previous education, address (urban/rural areas), college or school type, type of transport use, nationality, scholarship, internet and etc. A few research such as [47]–[56] used this category features to identify the impact of those features towards students' performance. These features might have a good result in previous research. However, the features used were mostly about the students' background rather than their performance in their studies.

The psychometric factor is one of the factors that affect students' performance. Generally, these factors involve the behaviour and mental development of the students during their studies [57]–[59]. A few examples of features categorised as psychometric features; students' interest in the courses, attendance, engagement time, belief, self-esteem and more. A student's interest in studying the courses was mentioned in a few research papers [11]–[14], [56], [60], [61]. Additionally, students' attendance during classes can also define the students' determination towards their studies. Many research used the attendance attribute to become one of the factor to affect the students' performance [11], [14], [39], [44], [49], [62]–[66].

The academic factor is the most used feature to evaluate students' performance. One of the features that frequently used in previous research is students' grade or Grade Point Average (GPA) [67]–[72]. GPA is used as a leading indicator of students' performance by the researchers, and it is usually measured on a scale with a specific range depending on the academic institution [71]. Academic attributes also include internal assessment such as lab work, assignments, quizzes, materials, etc [61]. External assessment is defined as marks achieved by a student in the final examination. Research stated that by giving students internal assessment, they tended to be more successful when they thoroughly studied the material and finished the homework given by the teacher. Thus, that can be concluded that the assessment can positively affect the students' performance [66].

Students' online activities already existed in previous research. In [9], the researchers used three different features categories to analyse students' performance; demographic, engagement and past performance. The results showed that engagement and the past performance had the highest accuracy in affecting the students' performance. Other research also used a few features that belong to the academic and psychometric categories such as number of view course content [73], number of each access to learning dashboards [73], total time spent online [73], student engagement in course [8][74], online session assessment [74], students' activity log [67] and more. The researchers widely use both these categories to attain the factors that can affect the students' performance.

Feature selection is a process used in which a part of the features available from the dataset are chosen for the machine learning algorithm [75]. The application of the process occurs during the pre-processing stage. The feature selection process's objective is due to its effectiveness in reducing dimensionality, removing irrelevant data, and increasing learning accuracy and efficiency[75]–[82]. These advantages are significant when dealing with raw or high-dimensional datasets as they are prone to irrelevant and redundant data, as stated in [75] and was eventually caused the machine learning techniques to decrease performance. Research [80] stated that when the data dimensionality rises, the data required to provide a reliable analysis grows rapidly. This phenomenon is known as "the curse of dimensionality". Thus, it is crucial to have feature selection process so that the insignificant features of the data can be removed and therefore, increase the performance. It also caused the running time of the learning algorithms to reduce [78]. For this research, four feature selection methods are chosen; ANOVA test, Chi-squared Test, RFE and feature importance using ET.

ANOVA test is one of the feature selection methods used in latest research. This test is applied by comparing the 'multiple means' values of the dataset and visualising any significant difference between mean values of multiple groups (classes) [21]–[23]. However, it does not determine which group is significantly different [83]. If the test is significant, it shows that the means of at least one pair are different, but not which pair or pairs which requires additional tests [83]. ANOVA tests are widely used in various applications such as audiovisual emotion recognition [22], microarray gene expression [23], student analysis [84] and more.

Chi-squared Test is a univariate feature selection algorithm used to test independence and estimate whether the class label is independent of a feature [24]–[27]. This test has two main phases of this algorithm. In the first phase, consistency checking is performed as the stopping criteria, whereas in phase two, the results of phase one are checked. It continues until there remain no attributes for merging [26]. According to [28], Chi-squared Test is utilised to obtain the significant connection between two categorical variables. In addition, this method belongs to the filter method category. It uses a 'proxy measure' calculated from the general characteristics of the training data to score features or feature subsets as a processing step prior to modelling [85]. The advantages of having the filter method are that it can run faster, and any features chosen by the method can be passed to any modelling algorithm [85]

RFE involves the process of looping, and this method belongs to wrapper methods of feature selection. The advantage of having this method is the ability to rearrange the order of significant features through rank [29], [32]. While at each iteration of the loop structure, the less significant features will be eradicated [23]. The looping application structure is due to the variation in the significance of features at each iteration by removing less significant features [86]. However, RFE has a high computational cost. Due to the reason, a few variants are introduced to speed up the algorithm. Removing many features in each iteration can speed up the process rather than removing only one least important feature at every iteration [23]. This method is applied in many kinds of research, including educational sectors such as students' performance [19], [29]–[31], early detection of at-risk students [32] and microarray gene expression [23].

The Extra Trees algorithm (ET) is a tree-based ensemble method for supervised classification and regression problems [33]. This algorithm builds randomised trees whose structures are independent of the output values of the learning sample. According to [87] and [34], this ensemble learning technique aggregates the results of multiple decorrelated decision trees collected in a "forest" to output its classification result. ET also belongs to the embedded method group, which combines filter and wrapper methods [87]. The benefit of having this algorithm is its efficiency in computational, and the variance of the decision tree can be reduced, hence increasing the generalisation ability [33][34][35]. There are a few studies that used this method in the educational sector, such as exploring the high potential factors that affect students' academic performance [88], predicting students' performance [89], supporting students' engineering design [90] etc.

Previous research demonstrated that ANOVA, Chi-squared Test, RFE and ET are frequently used as feature selection methods during the pre-processing stage of the research [19], [29]–[32], [84], [88]–[90]. Thus, these methods are employed for the new proposed dataset to examine the robustness of the data; in addition to ensure the dataset is applicable to various studies.

In conclusion, the three categories of attributes, demographic, psychometric and academic, are easy to obtain as the previous researchers were already using the attributes widely. The proposed features are also from the similar categories; however, they are more in-depth on students' behaviour during their online classes. The similarity in the categories of attributes between the proposed features and the features used in previous research will ensure the validity of using the attributes as a dataset for the current research on student performance classification. The only difference between the features is the variety in the groups of the proposed features, which focuses on the students' psychometric effect and knowledge towards their courses. This will contribute new findings on the student performance research regarding online student activities data's effect on their performances. Thus, the proposed features are suitable for student performance classification as previous researchers were already studying the categories.

III. METHODOLOGY

This section is separated into several sections, begin with the overview of the whole process as shown in figure 1, data collection, feature selection process followed by theoretical background of the feature selection techniques used in the research. The intelligent technique was then conducted in Python in Anaconda Navigator (Anaconda3) platform. The techniques that were used in this research; ANOVA, Chi-squared Test, RFE and ET.

A. Flowchart of the Process

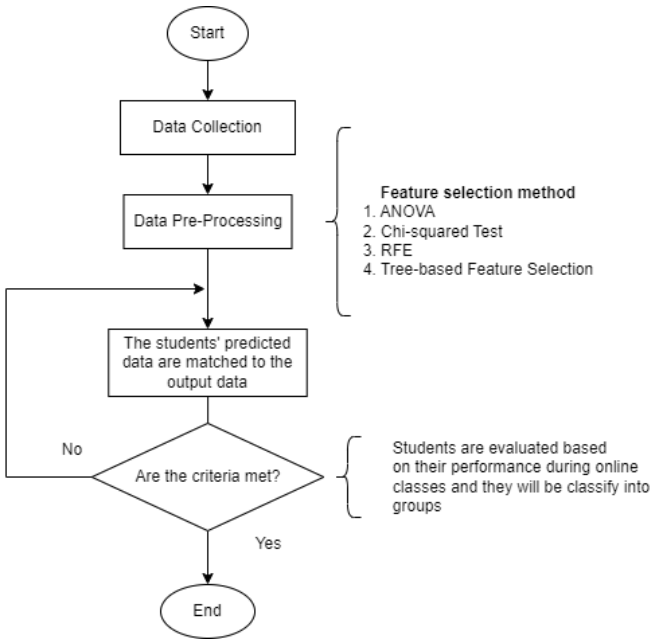

Fig. 1. The Flowchart of the Process

B. Data Collection

The new data is extracted from a website that performs online learning for computer course. The collected students' data consisted of 101 students from semester three of School of Electrical Engineering of Universiti Teknologi Mara. Most of the data were based on students' activities during online learning sessions in a course. Table I depicts the inputs as classification features for this work.

TABLE I
FEATURES USED IN THE RESEARCH

No.	Features	
1.	Accessing course materials	Before class (Before Notes)
2.		In class (In Notes)
3.		During other class (Other Notes)
4.		After class (After Notes)
5.	Note length	
6.	Accessing exercises materials	Before class (Before Exercise)
7.		In class (In Exercise)
8.		During other class (Other Exercise)
9.		After class (After Exercise)
10.	Tutorials Sections	Correct 3 and above (C3AA)
11.		Answered all questions (AAQ)
12.		Wrong before correct (WBC)
13.	Test 1 (output variable)	

The features for classification of the data is examined based on the students' online activities during their online classes. As exhibited in Table I, accessing course materials, accessing exercises materials and tutorials sections are considered as the behaviour and learning outcomes of the students. Both courses and exercises features dataset are the recorded of the students' behaviour of accessing the materials. In addition, these behaviours are divided into four groups; before class, in class, during other class and after class. These groupings are to isolate whether the students accessed or revised the materials prior in class, during other class or after the class. For note length, students' notes were recorded in the learning website. Also, there are three categories for tutorial sections: 'correct 3 and above', 'answer all questions' and 'wrong before correct'. These categories are to demonstrate the course's understanding and learning outcomes of the students' when the assignments are assign by the tutor. The output variable for this research in 'test 1'.

C. Feature Selection Process

Feature selection is performed to identify the significant features that affect the students' performance. Four feature selection methods are chosen, and each has its feature selection category. The first two methods used are Analysis of Variance test (ANOVA) and Chi-squared test. Both methods are filter method categories [85]. The second method is Recursive Feature Elimination (RFE) which belongs to the wrapper method category. Then, the third method is an embedded method category, Tree-based feature selection using Extra Tree algorithm. Fig. 1 depicts the overall flowchart for the pre-processing stage for this work.

Some methods can be set during the feature selection process to attain a specific amount of features such as ANOVA and Chi-squared test. The features chosen by the methods will be arranged in descending order so that the features will be easily recognised through their contribution in affecting the students' performance. For ANOVA and Chi-squared methods, the features are analysed through F-value and Chi-squared value respectively [85]. The higher the F-value and Chi-squared value, the higher the significance of the features towards students' performance. For RFE method, the method will rank the features according to the significance of the features [86]. Extra Tree algorithm evaluates the features by the features importance. Feature importance provides a score for each data feature between zero and one. The feature will become significant and relevant towards the output variable when the score is higher [91].

The result of the pre-processing stage will be displayed through a few figures and graphs in results section of this paper

D. Theoretical Background

1) Analysis of Variance (ANOVA)

ANOVA test is used to compare the 'multiple means' values of the dataset and visualise whether there is any significant difference between the mean values of multiple groups (classes). The statistic for ANOVA is called the F-statistic, which can be calculated using the following steps:

a) The variation between the group is calculated as:

Between sum of squares

$$(\mathrm{BSS}) = \mathrm{n}_1(.\overline{\mathrm{X}}_1 - \overline{\mathrm{X}})^2 + \mathrm{n}_2(\overline{\mathrm{X}}_2 - \overline{\mathrm{X}})^2 + \cdots \quad (1)$$

Between mean squares

$$(\mathrm{BMS}) = \mathrm{BSS} / \mathrm{d\,f} \quad (2)$$

b) The variation within the groups is calculated as:

Within sum of squares

$$(\mathrm{WSS}) = (\mathrm{n}_1 - 1)\sigma_1^2 + (\mathrm{n}_2 - 1)\sigma_2^2 + \cdots \quad (3)$$

Within mean squares

$$(\mathrm{WMS}) = \mathrm{WSS} / \mathrm{d}\,\mathrm{f}_\omega \quad (4)$$

where *d f* = degree of freedom, *d* f_ω= *(N − k)*, σ = standard deviation *N* = Number of samples, *k* = Number of groups, and n_k= no. of samples in group *k*.

c) F-test statistic is calculated as:

$$F = BMS / WMS \quad (5)$$

The input to the algorithm is a matrix of the form N x M, where N is the total number of feature sets and M is the number of samples in the dataset.

2) Chi-squared Test

Pearson's chi-square test of independence is a statistical method used to identify the degree of association between variables [24]. This technique is applied to analyse the dependency of all attributes (factors) on the outcome attribute. So chi-square method proves helpful here. For a contingency table with 'r' rows and 'c' columns, the formula for finding the chi-square is given in equation (6).

$$x^2 = \frac{\Sigma(observed-expected)^2}{expected} \quad (6)$$

The predetermined level of significance is taken as 5% and P-values are identified using the chi-square values for each attribute.

According to [27], Chi-square is used for assessing two kinds of comparing: tests of independence and tests of goodness of fit. In feature selection, the test of independence is assessed by chi-square and estimate whether the class label is independent of a feature. Chi-square score with C class and r values is defined as

$$\mathrm{x}^2 = -\sum_{\mathrm{i}=1}^{\mathrm{r}}\sum_{\mathrm{j}=1}^{\mathrm{c}}\frac{(\mathrm{n_{ij}}-\mu_{\mathrm{ij}})}{\mu_{\mathrm{ij}}} \quad (7)$$

$\mathrm{n_{ij}}$ is the amount of samples value with the i^{th} value of the feature

$$\mu_{ij} = \frac{(n_{*j}n_{i*})}{n} \quad (8)$$

n_{i*} is the amount of samples with the i^{th} the feature value.
n_{*j} is the amount of samples in class j.
n is the number for samples.

3) Recursive feature Elimination (RFE)

RFE method in this research used linear-based kernel of Support Vector Machine as a supervised learning estimator with a fit method that uses coefficients of the weight vector, w to compute the feature ranking score [92]. Any feature, *i*th with the smallest ranking score $c_i = (w_i)^2$ is eliminated, where $w_i$represents the corresponding component in the weight vector w.

The reason of using $(w_i)^2$ as the ranking criterion is from the sensitivity analysis of the objective function $J = \frac{1}{2}||w||^2$ with respect to a variable [92]. A virtual scaling factor v is introduced into the kernel function when computing the gradient and $k(x_i, x_j)$ becomes $k(v.x_i, v.x_j)$. For a linear SVM with its kernel function, $k(x_i, x_j)$= (x_i, x_j), using the fact $v_k = 1$, the sensitivity can be computed as

$$\frac{\partial J}{\partial v_k} = \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j \frac{\partial k(x_i x_j)}{\partial v_k} \quad (9)$$

$$= \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j\,(2v_k x_k^2)$$

$$= w_k^2$$

4) Extra Tree Algorithm (ET)

In this work, an Extra Trees algorithm also known as an extremely randomised trees algorithm feature selection. The working principle of ET algorithm is by building a cluster of trees iteratively on the dataset until each tree represents a sub-dataset [33], [90]. The nodes in the trees represent features of the dataset, and the leaves represent the samples that belong to a specific class. The trees are generated by splitting the nodes, and the most related features to the target will be split first based on the information gain theory [33]. In this way, after the trees are generated, the nodes will be in order from the root to the leaves (exclude the leaf nodes) to acquire the feature sequence ordered by the significances to target [90]. After the features were ordered, feature importance is calculated by weighting the proportion of the samples that reaches a node in the whole data set with the purity of the node [90].

IV. Results And Discussion

Feature selection process is performed on the dataset. The variety of feature selection methods applied to the proposed dataset will test the data consistency when producing results. Thus, four methods are used in the feature selection process; ANOVA test, Chi-squared test, RFE and ET.

1) ANOVA test

ANOVA test evaluates each of the features by the value of F. The higher F-value shows that the feature affects the students' performance significantly. The features chosen by the ANOVA test are clearly shown through the scatter plot in Fig. 2. The features chosen in descending order consisted of 'Answered all questions', 'Correct 3 and above', 'During-other-class notes' and 'Before-class exercises'.

Feature selection method such as ANOVA has a specific goal: to compare the means of the response variables for various combinations of the classification variables [22]. In [93], ANOVA are used to determine the mean difference problems by using between and within group variance differences. This method's execution results in the F-value of each feature in the proposed dataset. The higher the F-value, the higher the difference between the groups on the independent variable [83]. When there is significant differences between the groups, it implies that the variable is different from others and not identical which infers the variables are pertinent in a research. Referring to Fig. 2, there are four features chosen by ANOVA

for higher F-value namely; 'Answered All Questions', 'Correct 3 and Above', 'During-other-class Exercise' and 'Before-class Exercises'. Thus, these indicate that the findings are significant in contributing to the students' performance.

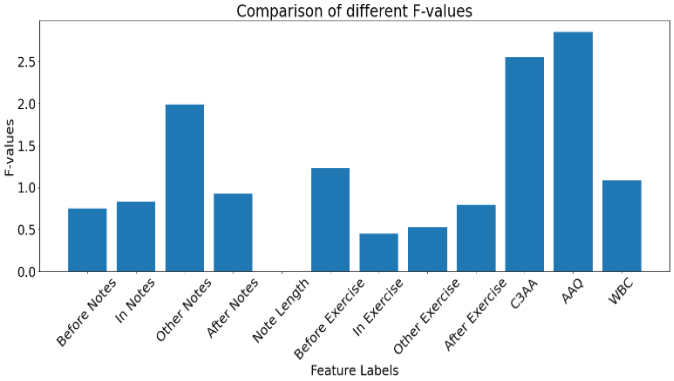

Fig. 2. The Bar Plot of Chi-squared Test

2) Chi-squared test

As mentioned previously, Chi-squared test evaluates each of the features by the value of chi-squared. The higher the chi-squared value, the higher the contribution to the students' performance. The differences in the features chosen by the Chi-squared test are presented in the bar plot in Fig. 3. As shown in Fig. 3, the features in descending order consisted of 'After-class notes', 'During-other-class notes', 'In-class notes' and 'Before-class notes'. The test selected 'After-class notes' as the most significant feature as it has the highest chi-squared value.

Chi-squared Test also analyse the dependency of all features on the outcome feature [83]. This proposed dataset is suitable for this method as the features in the dataset will be evaluated to investigate the correlation and the contribution of the features towards the outcome feature, 'Test 1'. The larger the chi-squared value, the higher the probability for a significant difference in the feature. As depicted in Fig. 3, the findings infer that the highest value of chi-squared implies highly influential factor. Thus, 'After-class Notes' is the most significant feature that affect the students' performance compared to other features.

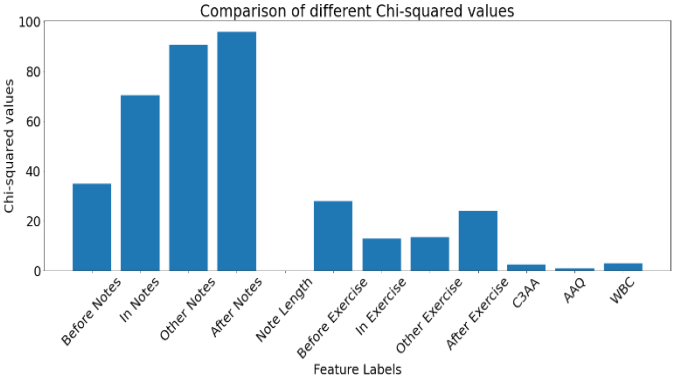

Fig. 3. The Bar Plot of Chi-squared Test

3) Recursive feature elimination (RFE)

RFE is the method that evaluates each of the features by ranks. After applying the method, each variable is ranked using numeric numbers. The best feature will be shown as number '1'. Then, other variables will be ranked in ascending order. The increase in the number shows that the features are ineffective in affecting students' performance. Table II below shows the variables with their ranks. The top four variables are 'Correct 3 and above', 'Answered all questions' 'Before-class exercise', and 'Wrong before correct'.

For RFE method, looping process is involved. Each iteration evaluated the feature importance of the proposed dataset. The recursion is required because for some measures the relative importance of each feature can change substantially when evaluated over a different subset of features during the stepwise elimination process (in particular for highly correlated features) [86]. As shown in the Table II, the features of the proposed dataset that are chosen, have ranked according to the highest feature importance using the variable ranking techniques of the RFE method. The other features of the proposed dataset are ranked lower as the features with less significance are eradicated [29].

TABLE I
THE RANKING OF EACH VARIABLE

No.	Features		Ranks
1.	Accessing course materials	Before class (Before Notes)	6
2.		In class (In Notes)	11
3.		During other class (Other Notes)	8
4.		After class (After Notes)	10
5.	Note length		12
6.	Accessing exercises materials	Before class (Before Exercise)	3
7.		In class (In Exercise)	7
8.		During other class (Other Exercise)	5
9.		After class (After Exercise)	9
10.	Tutorials Sections	Correct 3 and above (C3AA)	1
11.		Answered all questions (AAQ)	2
12.		Wrong before correct (WBC)	4

4) Extra Tree algorithm (ET)

ET evaluates each of the features by its feature importance. The feature strongly affects the students' performance when the feature importance is high. The features is elected in descending order consisted of 'In-class notes', 'During-other-class notes', 'After-class notes', and 'After-class exercises'. The differences in the ET's features are presented in line plot as depicted in Fig. 4. The algorithm chose 'In-class notes' as the most significant feature.

The fourth feature selection that are used in this research is tree-based feature selection. During the construction of the forest by extra trees, for each feature, the normalized total reduction of the Gini coefficient used to split feature decisions is calculated, which is called the importance of the Gini factor [33]. The Gini is expressed according to the feature importance through line graph and then the top four features are selected as shown in Fig. 4.

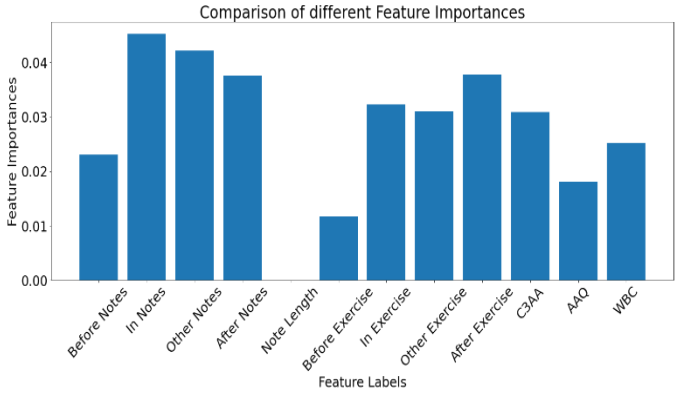

Fig. 4. The Bar Plot of ET

The top four features are taken from each method and listed in Table III. From the observation of the result, all feature selection methods have different features that influence the students' performance as shown in Table III. The first features of each method which have high contribution towards students' performance are varied from each other. For ANOVA test, 'Answered all questions' feature is consider the most significant feature that affects students' performance. For Chi-squared test, 'After class notes' feature affects student performance the most as the feature has the highest Chi-squared value. For RFE, 'Correct 3 and above' feature is ranking first to indicate the feature is the most effective feature to influence the students' performance. For Extra Trees algorithm, the most important feature to affect students' performance is 'In-class notes'. By assessing the findings, there are two methods that shared similar influential feature, 'During-other-class Notes' for which are Chi-squared Test and Extra Trees algorithm.

TABLE III
THE RANKING OF EACH VARIABLE

Methods	Features Chosen (Descending order)
Analysis of Variance (ANOVA)	Answered all questions, Correct 3 and above, During-other-class notes, Before-class exercises
Chi-squared Test	After-class notes, During-other-class notes, In-class notes, Before-class notes
Recursive Feature Elimination (RFE)	Correct 3 and above, Answered all questions, Before-class exercises, Wrong before correct
Extra Trees algorithm (ET)	In-class notes, During-other-class notes, After-class notes, After-class exercises

V. CONLUSION

In this paper, a dataset consisted of academic and psychometric feature is proposed. The dataset was based on online learning activities that are collected from a website that performs online learning for computer course. After collecting the data, the raw dataset performed data pre-processing to enhance its quality. A few feature selection techniques are applied to the dataset during the pre-processing stage consisted of ANOVA test, Chi-squared Test, RFE and Extra Trees algorithm. The application of the techniques to the dataset results in different chosen features for each method.

Based on Table III, the first chosen features for each method consisted of 'Answered all questions', 'After-class notes', 'Correct 3 and above' and 'In-class notes', were considered as the most significant features to be used in evaluating students' performance. According to the results, academic and psychometric factors play an important role in affecting the students' performance. The feature selection results showed that some academic features such as 'Answered all questions' and 'Correct 3 and above' significantly impact the students' performance academically. Other psychometric features such as 'In-class exercises' and 'After-class exercises' also impact the effectiveness of students' performance by the behaviour of students in doing the exercises at different environmental condition. This showed that academic and psychometric factors can significantly influence the student's performance. Thus, the dataset can be used for further research, especially on implementing classification approaches on the students' performance.

Thus, future research by using the proposed dataset which involves the use of academic and psychometric features and these feature selection techniques are recommended to be used in the future research.

ACKNOWLEDGEMENT

The authors would also like to express the gratitude for the supports given by the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM) Shah Alam. Also, this research work is funded by Special Research Grant (GPK). Grant no 600-RMC/GPK 5/3 (029/2020).

REFERENCES

[1] M. Kumar, A. J. Singh, and D. Handa, "Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques," *Int. J. Educ. Manag. Eng.*, vol. 7, no. 6, pp. 40–49, 2017, doi: 10.5815/ijeme.2017.06.05.

[2] Y. K. Salal, S. M. Abdullaev, and M. Kumar, "Educational Data Mining : Student Performance Prediction in Academic," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 4c, 2019.

[3] H. Almarabeh, "Analysis of Students' Performance by Using Different Data Mining Classifiers," *Int. J. Mod. Educ. Comput. Sci.*, vol. 9, no. 8, pp. 9–15, 2017, doi: 10.5815/ijmecs.2017.08.02.

[4] A. A. Rimi, "Developing Classifier for the Prediction of Students' Performance Using Data Mining Classification Techniques," *AURUM J. Eng. Syst. Archit.*, vol. 4, no. 1, pp. 73–91, 2020.

[5] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Appl. Sci.*, vol. 11, no. 1, pp. 1–28, 2021, doi: 10.3390/app11010237.

[6] S. Natek and M. Zwilling, "Student data mining solution-knowledge management system related to higher education institutions," *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6400–6407, 2014, doi: 10.1016/j.eswa.2014.04.024.

[7] C. Anuradha and T. Velmurugan, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance," *Indian J. Sci. Technol.*, vol. 8, no. 15, 2015.

[8] A. Abu Saa, M. Al-Emran, and K. Shaalan, "Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques," *Technol. Knowl. Learn.*, vol. 24, pp. 567–598, 2019, doi: 10.1007/s10758-019-09408-7.

[9] B. Kumar Verma, D. N. Srivastava, and H. Kumar Singh, "Prediction of Students' Performance in e-Learning Environment using Data Mining/ Machine Learning Techniques," *J. Univ. Shanghai Sci. Technol.*, vol. 23, no. 05, pp. 596–593, 2021, doi: 10.51201/jusst/21/05179.

[10] B. Kapur, N. Ahluwalia, and S. R, "Comparative Study on Marks Prediction using Data Mining and Classification Algorithms," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 3, pp. 632–636, 2017, doi:

10.26483/ijarcs.v8i3.3066.
[11] E. Filiz and E. Öz, "Finding the best algorithms and effective factors in classification of Turkish science student success," *J. Balt. Sci. Educ.*, vol. 18, no. 2, pp. 239–253, 2019, doi: 10.33225/jbse/19.18.239.
[12] A. O. Ameen, M. A. Alarape, and K. S. Adewole, "Students' Academic Performance and Dropout Prediction," *Malaysian J. Comput.*, vol. 4, no. 2, p. 278, 2019, doi: 10.24191/mjoc.v4i2.6701.
[13] V. Vijayalakshmi and K. Venkatachalapathy, "Comparison of Predicting Student's Performance using Machine Learning Algorithms," *Int. J. Intell. Syst. Appl.*, vol. 11, no. 12, pp. 34–45, 2019, doi: 10.5815/ijisa.2019.12.04.
[14] V. G. Karthikeyan, P. Thangaraj, and S. Karthik, "Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation," *Soft Comput.*, vol. 24, no. 24, pp. 18477–18487, 2020, doi: 10.1007/s00500-020-05075-4.
[15] M. Laxmi, I. Raju, G. Latha Pavani, and K. Vijaya Kumar, "ANALYSIS OF STUDENT ACADEMIC PERFORMANCE LEADING TECHNOLOGY USING CLASSIFICATION ALGORITHMS," *J. Crit. Rev.*, vol. 7, no. 19, pp. 7251–7259, 2020.
[16] M. Asiah, K. Nik Zulkarnaen, D. Safaai, M. Y. Nik Nurul Hafzan, M. Mohd Saberi, and S. Siti Syuhaida, "A Review on Predictive Modeling Technique for Student Academic Performance Monitoring," in *Engineering Applications of Artificial Intelligence Conference (EAAIC 2018)*, 2019, vol. 255, doi: 10.1051/matecconf/201925503004.
[17] N. Alangari and R. Alturki, "Predicting students final GPA using 15 classification algorithms," *Rom. J. Inf. Sci. Technol.*, vol. 23, no. 3, pp. 238–249, 2020.
[18] R. Tabassum and N. Akhter, "Effect of Demographic Factors on Academic Performance of University Students," *J. Res. Reflections Educ.*, vol. 14, no. 1, pp. 64–80, 2020.
[19] S. Viswanathan and S. Vengatesh Kumar, "Study Of Students ' Performance Prediction Models Using Machine Learning," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 2, pp. 3085–3091, 2021.
[20] A. Alshanqiti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203827–203844, 2020, doi: 10.1109/ACCESS.2020.3036572.
[21] M. Kumar, N. K. Rath, A. Swain, and S. K. Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," in *International Multi-Conference on Information Processing (IMCIP)*, 2015, vol. 54, pp. 301–310, doi: 10.1016/j.procs.2015.06.035.
[22] M. Bejani, D. Gharavian, and N. Charkari, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks," *Neural Comput. Appl.*, vol. 24, no. 2, 2014, doi: 10.1007/s00521-012-1228-3.
[23] S. Abdulsalam *et al.*, "Performance Evaluation of ANOVA and RFE Algorithms for Classifying Microarray Dataset Using SVM," *Eur. Mediterr. Middle East. Conf. Inf. Syst.*, vol. 402, pp. 480–492, 2020, doi: 10.1007/978-3-030-63396-7.
[24] J. Shana and T. Venkatachalam, "Identifying Key Performance Indicators and Predicting the Result from Student Data," *Int. J. Comput. Appl.*, vol. 25, no. 9, pp. 45–48, 2011, doi: 10.5120/3057-4169.
[25] M. Zaffar, "A Study of Feature Selection Algorithms for Predicting Students Academic Performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 5, 2018, doi: 10.14569/IJACSA.2018.090569.
[26] M. Zaffar and M. A. Hashmani, "Comparing the Performance of FCBF , Chi- Square and Relief-F Filter Feature Selection Algorithms in Educational Data Mining," in *International Conference of Reliable Information and Communication Technology (IRICT)*, 2018, vol. 843, pp. 151–160, doi: 10.1007/978-3-319-99007-1.
[27] N. Rachburee, "A Comparison of Feature Selection Approach Between Greedy , IG-ratio , Chi-square , and mRMR in Educational Mining," in *International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2015, pp. 420–424.
[28] D. Das, A. K. Shakir, and M. Rahman, "A Comparative Analysis of Four Classification Algorithms for University Students Performance Detection," *Int. Conf. Electr. Control Comput. Eng.*, 2019, doi: 10.1007/978-981-15-2317-5.
[29] U. Ashfaq, P. M. Booma, and R. Mafas, "Managing Student Performance : A Predictive Analytics using Imbalanced Data," *Int. J. Recent Technol. Eng.*, vol. 8, no. 6, 2020, doi: 10.35940/ijrte.E7008.038620.
[30] F. Catarina, L. I. Tec, and R. Sobral, "Predicting students ' performance using survey data," 2020.
[31] R. Alshabandar, L. J. Moores, L. J. Moores, R. Keight, L. J. Moores, and L. J. Moores, "Students Performance Prediction in Online Courses Using Machine Learning Algorithms," 2020.
[32] Y. Cui, F. Chen, and A. Shiri, "Scale up predictive models for early detection of at-risk students : a feasibility study," *Inf. Learn. Sci.*, vol. 121, no. 3/4, pp. 97–116, 2020, doi: 10.1108/ILS-05-2019-0041.
[33] Y. Liang, S. Zhang, H. Qiao, and Y. Yao, "iPromoter-ET : Identifying promoters and their strength by extremely randomized trees-based feature selection," *Anal. Biochem.*, vol. 630, p. 114335, 2021, doi: 10.1016/j.ab.2021.114335.
[34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely Randomized Trees," *Mach. Learn.*, vol. 63, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.
[35] R. Marée, P. Geurts, and L. Wehenkel, "Random subwindows and extremely randomized trees for image classification in cell biology," *BMC Mol. Cell Biol.*, vol. 8, pp. 1–12, 2007, doi: 10.1186/1471-2121-8-S1-S2.
[36] S. Wiyono and T. Abidin, "Comparative Study of Machine Learning Knn, Svm, and Decision Tree Algorithm To Predict Student'S Performance," *Int. J. Res. -GRANTHAALAYAH*, vol. 7, no. 1, pp. 190–196, 2019, doi: 10.29121/granthaalayah.v7.i1.2019.1048.
[37] E. N. Ogor and C. Islands, "Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques Department of Natural Sciences Turks & Caicos Islands Community College Visualization and Articulation," in *Electronics, Robotics and Automotive Mechanics Conference (CERMA)*, 2007, pp. 0–5, doi: 10.1109/CERMA.2007.78.
[38] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Educ. Inf. Technol.*, vol. 26, pp. 205–240, 2021, doi: 10.1007/s10639-020-10230-3.
[39] M. Tiwari, R. Singh, and N. Vimal, "An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education," *Int. J. Comput. Sci. Mob. Comput.*, vol. 2, no. 2, pp. 53 – 57, 2013.
[40] R. O. Salem, N. Al-Mously, N. M. Nabil, A. H. Al-Zalabani, A. F. Al-Dhawi, and N. Al-Hamdan, "Academic and socio-demographic factors influencing students' performance in a new Saudi medical school," *Med. Teach.*, vol. 35, no. SUPPL. 1, pp. 83–89, 2013, doi: 10.3109/0142159X.2013.765551.
[41] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: An application of data mining methods with an educational web-based system," *33rd ASEE/IEEE Front. Educ. Conf. 1*, 2003.
[42] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/ijiet.2016.v6.745.
[43] A. M. Remali, M. A. Ghazali, M. K. Kamaruddin, and Y. K. Tan, "Understanding academic performance based on demographic factors, motivation factors and learning styles," *Int. J. Asian Soc. Sci.*, vol. 3, no. 9, pp. 1938–1951, 2013.
[44] M. Kumar and Y. K. Salal, "Systematic review of predicting student's performance in academics," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 3, pp. 54–61, 2019, doi: 10.13140/RG.2.2.26667.69923.
[45] K. Prasada, M. V. P. Chandra, and B. Ramesh, "Predicting Learning Behavior of Students using Classification Techniques," *Int. J. Comput. Appl.*, vol. 139, no. 7, pp. 15–19, 2016, doi: 10.5120/ijca2016909188.
[46] E. E. Ebenuwa-Okoh, "Influence of Age, Financial Status, and Gender on Academic Performance among Undergraduates," *J. Psychol.*, vol. 1, no. 2, pp. 99–103, 2010, doi: 10.1080/09764224.2010.11885451.
[47] M. Osmanbegovic, Edin; Suljic and Article, "Data Mining Approach for Predicting Student Performance," *Econ. Rev. J. Econ. Bus.*, vol. 10, no. 1, pp. 3–12, 2012.
[48] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 2, pp. 447–459, 2018, doi: 10.11591/ijeecs.v9.i2.pp447-459.
[49] S. Urkude and K. Gupta, "Student intervention system using machine learning techniques," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6 Special Issue 3, pp. 2061–2065, 2019, doi: 10.35940/ijeat.F1392.0986S319.
[50] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, 2013, doi: 10.2478/cait-2013-0006.
[51] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," *Int. J. Comput. Sci. Manag. Res.*, vol. 1, no. 4, pp. 686–690, 2012.

[52] Y. K. Saheed, T. O. Oladele, A. O. Akanni, and W. M. Ibrahim, "Student performance prediction based on data mining classification techniques," *Niger. J. Technol.*, vol. 37, no. 4, pp. 1087–1091, 2018, doi: 10.4314/njt.v37i4.31.

[53] V. Ramesh, "Predicting Student Performance : A Statistical and Data Mining Approach," *Int. J. Comput. Appl. (0975 – 8887)*, vol. 63, no. 8, pp. 35–39, 2013.

[54] Ş. Aydoğdu, "Predicting student final performance using artificial neural networks in online learning environments," *Educ. Inf. Technol.*, vol. 25, no. 3, pp. 1913–1927, 2020, doi: 10.1007/s10639-019-10053-x.

[55] B. K. Bhardwaj, "Data Mining : A prediction for performance improvement using classification," *Int. J. Comput. Sci. Inf. Secur.*, vol. 9, no. 4, 2011.

[56] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of Student Academic Performance By an Application of Data Mining Techniques," *Manag. Artif. Intell.*, vol. 6, no. January, pp. 110–114, 2011.

[57] J. Martínez-Líbano, M. M. Yeomans, and J. C. Oyanedel, "Psychometric Properties of the Emotional Exhaustion Scale (ECE) in Chilean Higher Education Students," *Eur. J. Investig. Heal. Psychol. Educ.*, vol. 12, no. 1, pp. 50–60, 2022, doi: 10.3390/ejihpe12010005.

[58] P. A. Lowe, "The Test Anxiety Measure for College Students-Short Form: Development and Examination of Its Psychometric Properties," *J. Psychoeduc. Assess.*, vol. 39, no. 2, pp. 139–152, 2021, doi: 10.1177/0734282920962947.

[59] S. Barkat Ali, N. un Nisa khan, S. Ejaz, and S. Shamsy, "Psychometric Features of Motivated Strategies for Learning Questionnaire (MSLQ) Among the Students of Higher Education Sector in Karachi-Pakistan," *Rev. Manag. Sci.*, vol. III, no. 2, pp. 85–100, 2021, doi: 10.53909/rms.03.02.090.

[60] A. J. Martin, H. W. Marsh, D. M. Mclnerney, J. Green, and M. Dowson, "Getting Along with Teachers and Parents: The Yields of Good Relationships for Students' Achievement Motivation and Self-Esteem," *Aust. J. Guid. Couns.*, vol. 17, no. 2, pp. 109–125, 2007, doi: 10.1375/ajgc.17.2.109.

[61] D. Magdalene Delighta Angeline, "Association Rule Generation for Student Performance Analysis using Apriori Algorithm," *SIJ Trans. Comput. Sci. Eng. its Appl.*, vol. 1, no. 1, pp. 12–16, 2013.

[62] S. Rajesh Suyal and M. Mukund Mohod, "Quality IMPROVISATION OF STUDENT PERFORMANCE USING DATA MINING TECHNIQUES," *Int. J. Sci. Res. Publ.*, vol. 4, no. 1, pp. 2250–3153, 2014, [Online]. Available: www.ijsrp.org.

[63] A. Dutt and M. A. Ismail, "Can We Predict Student Learning Performance from LMS Data? A Classification Approach," in *International Conference on Current Issues in Education (ICCIE 2018)*, 2018, vol. 326, pp. 24–29, doi: 10.2991/iccie-18.2019.5.

[64] K. Yadav and R. Thareja, "Comparing the Performance of Naive Bayes And Decision Tree Classification Using R," *Int. J. Intell. Syst. Appl.*, vol. 11, no. 12, pp. 11–19, 2019, doi: 10.5815/ijisa.2019.12.02.

[65] M. Asiah, K. Nik Zulkarnaen, D. Safaai, M. Y. Nik Nurul Hafzan, M. Mohd Saberi, and S. Siti Syuhaida, "A Review on Predictive Modeling Technique for Student Academic Performance Monitoring," *MATEC Web Conf.*, vol. 255, p. 03004, 2019, doi: 10.1051/matecconf/201925503004.

[66] J. -P. Vandamme, N. Meskens, and J. -F. Superby, "Predicting Academic Performance by Data Mining Methods," *Educ. Econ.*, vol. 15, no. 4, pp. 405–419, 2007, doi: 10.1080/09645290701409939.

[67] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Educ. Sci.*, vol. 11, no. 9, 2021, doi: 10.3390/educsci11090552.

[68] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks," *IEEE Access*, vol. 9, pp. 140731–140746, 2021, doi: 10.1109/ACCESS.2021.3119596.

[69] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," *Mater. Today Proc.*, vol. 47, no. xxxx, pp. 5260–5267, 2021, doi: 10.1016/j.matpr.2021.05.646.

[70] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00192-z.

[71] I. Issah, O. Appiah, P. Appiahene, and F. Inusah, "A systematic review of the literature on machine learning application of determining the attributes influencing academic performance," *Decis. Anal. J.*, vol. 7, no. October 2022, p. 100204, 2023, doi: 10.1016/j.dajour.2023.100204.

[72] B. K. Yousafzai *et al.*, "Student-performulator: Student academic performance using hybrid deep neural network," *Sustain.*, vol. 13, no. 17, pp. 1–21, 2021, doi: 10.3390/su13179775.

[73] M. Kokoç and A. Altun, "Effects of learner interaction with learning dashboards on academic performance in an e-learning environment," *Behav. Inf. Technol.*, vol. 40, no. 2, pp. 161–175, 2021, doi: 10.1080/0144929X.2019.1680731.

[74] S. M. Aslam, A. K. Jilani, J. Sultana, and L. Almutairi, "Feature Evaluation of Emerging E-Learning Systems Using Machine Learning: An Extensive Survey," *IEEE Access*, vol. 9, pp. 69573–69587, 2021, doi: 10.1109/ACCESS.2021.3077663.

[75] C. Sha, X. Qiu, and A. Zhou, "Feature Selection Based on a New Dependency Measure," *IEEE Comput. Soc.*, pp. 266–270, 2008, doi: 10.1109/FSKD.2008.515.

[76] J. Li, K. Cheng, S. Wang, and F. Morstatter, "Feature Selection : A Data Perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017.

[77] J. Miao and L. Niu, "A Survey on Feature Selection," in *Information Technology and Quantitative Management (ITQM 2016)*, 2016, vol. 91, pp. 919–926, doi: 10.1016/j.procs.2016.07.111.

[78] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification : A Review," *Data Classif. Algorithms Appl.*, pp. 37–64, 2014.

[79] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," *Knowl. Inf. Syst.*, vol. 53, 2017, doi: 10.1007/s10115-017-1059-8.

[80] Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," *Adv. Bioinformatics*, vol. 2015, pp. 1–13, 2015, doi: 10.1155/2015/198363.

[81] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, 2021, doi: 10.3389/fenrg.2021.652801.

[82] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[83] L. M. Connelly, "Introduction to Analysis of Variance (ANOVA)," *J. Adult Heal.*, vol. 30, no. 3, p. 218, 2021.

[84] M. L. Mouritsen, D. Ph, J. T. Davis, D. Ph, and S. C. Jones, "ANOVA Analysis of Student Daily Test Scores in Multi-Day Test Periods," *J. Learn. High. Educ.*, vol. 12, no. 2, 2016.

[85] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and H. Moore, "Relief-Based Feature Selection : Introduction and Review," *J. Biomed. Inform.*, 2018, doi: 10.1016/j.jbi.2018.07.014.

[86] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," *Chemom. Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, 2006, doi: 10.1016/j.chemolab.2006.01.007.

[87] L. Chen and M. Xia, "A context-aware recommendation approach based on feature selection," *Appl. Intell.*, vol. 51, pp. 1–11, 2021.

[88] R. Kaviyarasi and T. Balasubramanian, "Exploring the High Potential Factors that Affects Student s ' Academic Performance," *I.J Educ. Manag. Eng.*, vol. 6, no. 8, pp. 15–23, 2018, doi: 10.5815/ijeme.2018.06.02.

[89] S. Zulfiker, N. Kabir, A. A. Biswas, P. Chakraborty, and M. Rahman, "Predicting Students ' Performance of the Private Universities of Bangladesh using Machine Learning Approaches," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 3, pp. 672–679, 2020, doi: 10.14569/IJACSA.2020.0110383.

[90] W. Xing, B. Pei, S. Li, G. Chen, and C. Xie, "Using learning analytics to support students ' engineering design : the angle of prediction," *Interact. Learn. Environ.*, 2019, doi: 10.1080/10494820.2019.1680391.

[91] H. Abubaker, A. Ali, S. M. Shamsuddin, and S. Hassan, "Exploring permissions in android applications using ensemble-based extra tree feature selection," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 1, pp. 543–552, 2020, doi: 10.11591/ijeecs.v19.i1.pp543-552.

[92] K.-B. Duan and H. Wang, "Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data," *IEEE Trans. Nanobioscience*, vol. 4, no. 3, pp. 228–234, 2005.

[93] K. K. Tae, "Understanding one-way ANOVA using conceptual figures," *Korean J. Anesthesiol.*, vol. 70, no. 1, pp. 22–26, 2017, doi: https://doi.org/10.4097/kjae.2017.70.1.22.