Student performance classification: a comparison of feature selection methods based on online learning activities

Muhammad Aqif Hadi Alias, Mohd Azri Abdul Aziz, Najidah Hambali, Mohd Nasir Taib
School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), Selangor, Malaysia

Article Info

Article history:

Received Oct 29, 2023
Revised Mar 3, 2024
Accepted May 18, 2024

Keywords:

Artificial neural network
Classification
Feature selection
Multi-layer perceptron
Student performance

ABSTRACT

The classification of student performance involves categorizing students' performance using input data such as demographic information and examination results. However, our study introduces a novel approach by emphasizing students' online learning activities as a rich data source. To avoid misinterpretation during the classification, we therefore presented a study comparing several feature selection (FS) methods combined with artificial neural network (ANN), for classifying students' performance based on their online learning activities. At first, we focused on tackling the issue of missing values by implementing data cleaning using variance threshold. feature selection techniques were implemented which encompass both filter-based (information gain, chi-square, Pearson correlation) and wrapper-based, sequential selection (forward and backward) techniques. In the classification stage, multi-layer perceptron (MLP) was used with the default hyperparameters and 5-fold cross-validation along with synthetic minority oversampling technique (SMOTE) were also applied to each method. We evaluated each feature selection method's performance using key metrics: accuracy, precision, recall, and F1-score. The outcomes highlighted information gain and sequential selection (forward and backward) as the top-performing methods, all achieving 100% accuracy. This research underscores the potential of leveraging online learning activities for robust student performance classification within the specified constraints.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Mohd Azri Abdul Aziz
School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM)
Selangor, Malaysia
Email: azriaziz@uitm.edu.my

1. INTRODUCTION

In recent years, the performance of students in educational institutions has garnered increasing attention. Undoubtedly, a substantial number of institutions have recognized this as a pivotal determinant in enhancing both the overall quality of the institutions and the educational outcomes of their students [1]–[4]. Identifying students at risk early in the course allows for the implementation of timely interventions and initiatives aimed at improving academic performance [5]–[11]. Consequently, in the pursuit of a deeper comprehension of the learning process and the environmental factors influencing it, the field of educational data mining has gained notable momentum. This discipline assumes a critical role in the classification of students' academic achievements [12]–[17]. The term "educational data mining" pertains to the utilization of data mining techniques to enhance the educational quality by identifying areas for improvement, identifying students in need of additional support, and uncovering the various factors that impact student academic success [18]. It is worth mentioning that, despite the widespread use of data mining in the commercial sector, its integration into education is relatively recent [19].

The foremost factor that significantly influences student performance is the quality of the acquired data, which is pivotal for the efficient development of predictive models. Educational data mining is primarily concerned with formulating algorithms capable of unveiling latent patterns within educational data, a field that encompasses a multitude of student-related features requiring comprehensive analysis [20], [21]. However, a substantial portion of the gathered data is inherently intricate, encompassing unwanted features. Without adequate data preprocessing, these unwanted elements may lead to model misinterpretations, consequently undermining the accuracy of student performance predictions [22]. Hence, further research in this domain promises to provide valuable insights for the enhancement of student performance. Such endeavors entail a fresh examination of the features contributing to student performance and the development of methodologies for effectively classifying student performance.

As acknowledged by researchers, the process of data preprocessing is deemed as a pivotal role in enhancing data quality and bolstering the reliability of data mining algorithms [23]. Failure to undertake effective data preprocessing may lead to erroneous conclusions, as raw data often contains unwanted features and noise [22]. The research in [21] underscored the influence of gathered data and attributes on the quality of data mining. One viable strategy within the data preprocessing phase is feature selection, which entails the identification of the most relevant attributes while discarding undesired ones, thereby reducing data dimensionality. To tackle the difficulty of an unbalanced multi-classification dataset, a data-level approach based on oversampling and two feature selection methods, wrapper and filter, were employed as benchmarks [24]. Notably, some researchers have advocated the integration of feature selection techniques along certain classification algorithms to improve the predictive models [12], [22], [23], [25]–[27].

A filter-based technique is applied as a preprocessing step, utilizing statistical tests to assess the correlation with the dependent variable. Its primary purpose is to identify and eliminate irrelevant features, resulting in a dataset containing the most valuable feature columns based on their respective scores. One notable advantage of this approach is its speed and minimal computational complexity since it does not necessitate model training. Notably, researchers have employed various filter-based feature selection methods, including information gain [21], [22], [26], [27] Correlation [28]–[30], and chi-square [22], [31]. The researchers [26], [32]–[35] adopted a filter-based approach known as information gain (IG) or mutual information (MI), which employs statistical tests to identify the most significant features. Sixhaxa *et al.* [26] specifically applied the MI feature selection technique to obtain the optimal feature set. This method estimates entropy reduction by comparing the information gain of individual features with the information gain of the dependent feature and selects the feature with the highest information gain. Studies in [33], [34] discovered that IG performed better in signaling the relevant elements in each research after adopting feature selection strategies. In [33], [34], IG emerged as a robust feature selection strategy that performed well with certain classifiers. Notably, IG showed extraordinary efficacy, especially when combined with classifiers such as artificial neural network (ANN) and decision tree (DT), as shown in [34]. Furthermore, when combined with the random forest (RF) classifier, MI outperformed other approaches for detecting Internet cheaters among students [33]. The study in question used MI to pick the top five features, exhibiting improved compatibility and performance synergy with the RF classifier. The study also demonstrated the adaptability of feature selection strategies by combining MI and analysis of variance (ANOVA), with each method significantly contributing to the selection of 5 features from a pool of 13. The study in [24] found that the unbalanced multi-class issue and the overfitting problem were major concerns when developing multi-class prediction models for students' grade predictions. To address these concerns, an oversampling methodology called as synthetic minority oversampling technique (SMOTE) was created, along with two feature selection techniques, namely *WrapperSubsetEval*, *ClassifierSubsetEval*, and IG, which will be tested with a variety of classification algorithms. The results of the suggested technique, coupled with six classification algorithms, including DT (J48), naïve Bayes (NB), K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and RF, revealed that all feature selection (FS) performed similarly across all classifiers, with more than 90% of each measure. Based on the data, it appears that Wrapper approaches are unnecessary since they require more computational complexity and processing capability.

The correlation-based feature selection technique is employed as a method for feature selection without reliance on the final classification model. It assesses the strength of the linear relationship between two variables, assigning values ranging from -1, indicating a strong negative correlation, to 1, indicating a strong positive correlation, with 0 signifying no correlation. In pursuit of improved prediction accuracy, Nidhi *et al.* [28] introduced a hybrid approach that incorporates correlation attribute evaluation (CAE), ensemble learning, and seven distinct machine learning algorithms. In this context, the CAE was utilized to execute the feature selection process, ultimately selecting only the top 10 features to assess the accuracy of the classification algorithms. As per the experimental findings, classification algorithms constructed using a heterogeneous combination of ensemble learning and the CAE demonstrated superior performance when compared to methods employing ensemble learning without CAE. This performance enhancement can be attributed to the integration of CAE in the feature selection process. In the study of college students' learning behavior and its learning impact [29], the researchers employed a threshold value based on the dependence value given by Pearson correlation. In this scenario, a 0.50 value was used to analyze the relationship between learning behavior and its learning impact. Similarly, in research aimed at finding the bare minimum of characteristics required for effective analysis [30], a threshold value of 0.7 was intentionally used. This threshold functioned as a criterion for determining the connection between distinct aspects, making it easier to identify and then exclude strongly associated features. By putting the criterion at 0.7, the study intended to compress the feature set, eliminate redundancy, and improve the efficiency of following studies.

The chi-square method represents a prominent feature selection technique. This statistical test evaluates the degree of disparity between observed values and expected outcomes, providing insights into the predictor variable [36]. In their research, Hashemi *et al.* [22] observed that both the chi-square and IG algorithms exhibited superior performance, as determined through an analysis involving the Kappa statistic and F-measure. In a different study, Trivedi *et al.* [37] conducted a statistical examination of the features, specifically using the chi-square test to calculate p-values, thus establishing the significance levels of these features. Furthermore, in a separate investigation [30], the chi-square test was applied alongside Pearson correlation to assess feature significance. In this context, a reference p-value of 0.05 was employed to gauge the features' significance, with values exceeding this threshold resulting in their exclusion from consideration.

Wrapper methods for feature selection rely on specific machine learning algorithms that are tailored to a given dataset. These methods employ a greedy search strategy, systematically evaluating all potential feature combinations based on predefined evaluation criteria. In a study by [12], a genetic algorithm (GA) was employed. It was characterized by a binary representation of individual solutions, straightforward crossover and mutation operators, and a proportional selection mechanism, all aimed at identifying the most optimal feature combinations. Li *et al.* [38] utilized GA as a dimension reduction method to streamline calculations and eliminate uncorrelated features. Their findings indicated that GA contributed to enhancing the fitness of gene sequences, reducing data dimensionality from 7,070 to 3,579 and identifying 3,491 features as uncorrelated. Trivedi *et al.* [37] focused on predicting the intent of using social media in online blended learning, a dataset initially containing 61 attributes underwent a reduction to 24 and subsequently 5 attributes. This feature reduction was achieved through the application of both a greedy technique and a wrapper method, respectively. In [39], a research was conducted to assess the effectiveness of various feature selection techniques on several classification algorithms using educational datasets. Three wrapper-based feature selection approaches were implemented: sequential forward selection (SFS), sequential backward selection (SBS), and differential evolution. Based on the prediction accuracy mean values, these three approaches outperformed the other filter-based methods utilized in the study, with DE scoring the highest. In [40], the greedy forward selection algorithm picked the fewest features from 15 features, but the other three approaches, mRMR, chi-square, and IG-ratio, selected 9,10, and 10 features, respectively. Using an ANN classifier improved the performance of the Greedy forward selection method.

In summary, several feature selection techniques will be proposed in this paper by considering previous studies which used several types of feature selection. The rest of this paper is organized as follows: section 2 is for the methods used; section 3 discusses the results obtained from the experiment and section 4 concludes the whole process of this study.

2. METHOD

In this study, a comparison of several feature selection methods involving two types which are filter-based and wrapper-based methods. Generally, both methods will be evaluated by using a machine learning algorithm which is known as multi-layer perceptron (MLP). The data included 102 samples of programming students who enrolled under the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), Malaysia, where the input data was extracted from an online learning platform which indicates the students' online learning activities. Thus, in this study, we will observe the relationship between online learning activities and students' academic performance within the semester. The whole process of this study is shown as in Figure 1.

2.1. Data preparation

Our proposed study was supposed to classify the data which contains three aspects of online learning activities that need to be observed which are notes, exercises, and tutorials. The target variable, grade, comprises of pass or fail for all students depending on their course grade point average (GPA). Our dataset was generated over the course of one semester from electronic and electrical engineering students enrolled in the programming course known as ECE431 at Universiti Teknologi MARA (UiTM), Malaysia. Table 1 shows the information of students' online learning activities where Feature1-Feature5 indicates students' access to online learning platform notes, Feature6-Feature9 represents students' attempts to complete online learning platform exercises, and Feature10-Feature12 reflects the patterns of students' tutorial answers in online learning platform. The collected data contains about 102 samples which comprised of electronic/electrical engineering students of the early semester for programming course. Thus, their online learning activities were assessed based on the three categories implying to their efforts in achieving better academic performance.

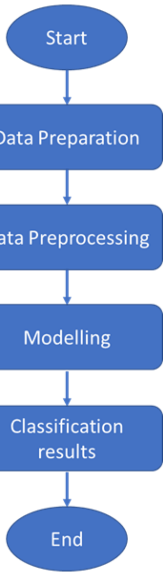

Figure 1. Flowchart of the process

Table 1. Details of the online learning activities

Category	Description	Label
Notes	Students' access to the notes before the class for the upcoming lesson begins	Feature1
	Students' access to the notes after the class lesson has started	Feature2
	Students' access to the notes after the first class has ended	Feature3
	Students' access to the notes after all classes have ended	Feature4
	Length of notes left by the students	Feature5
Exercises	Students do the exercise before the class for the upcoming lesson begins	Feature6
	Students do the exercise after the class lesson has started	Feature7
	Students do the exercise after the first class has ended	Feature8
	Students do the exercise after all classes have ended	Feature9
Tutorials	Students get 3 questions and above correct	Feature10
	Students answer all tutorial questions	Feature11
	Students get wrong answer for the questions before the questions of correct answer	Feature12

2.2. Data preprocessing

This section is the main part of this study where feature selection techniques will be explored with the intention to compare several methods in selecting the most significant features from the dataset. Two methods known as filter-based (information gain, chi-square and correlation based) and wrapper-based (Forward selection and Backward selection) feature selection methods were used where the performance of both types are compared based on several evaluation metrics by using machine learning algorithm which is MLP. Once relevant features are obtained from each method, a distinct subset of the data will be generated for each method based on their selected features. This subset was then divided into training and testing sets with a ratio of 80:20 respectively.

2.2.1. Information gain

It is a term used in decision tree algorithms, specifically in feature selection. It aids in determining a feature's relevance or importance in categorizing or predicting a target variable. The entropy concept underpins the IG formula where it calculates entropy reduction by comparing each independent feature's information gain to the information gain of the dependent feature and selecting the feature with the highest information gain [26]:

– Entropy ($H(S)$): Entropy quantifies the amount of impurity or disorder in a set of data. It is calculated in the context of a dataset with many classes using (1).

$$H(S) = -\sum_{i=1}^{n} p_i log_2(p_i) \quad (1)$$

where n is the number of classes and p_i is the proportion of samples in a certain class i

– Entropy of a feature ($H(A)$): The entropy of a feature A with regard to a target variable S is determined as a weighted sum of the entropies of subsets formed by partitioning the data depending on feature A values (2):

$$H(A) = \sum_{v \in values(A)} \frac{|S_v|}{|S|} \times H(S_v) \quad (2)$$

where $values(A)$ represents the set of all potential values for feature A, $|S_v|$ denotes the number of samples in which feature A has a value of v. $|S|$ describes the total number of samples, and $H(S_v)$ denotes the entropy of the subset corresponding to value v of feature A.

– The reduction in entropy produced by dividing data based on a certain attribute is measured as information gain (IG) as calculated in (3).

$$IG(A) = H(S) - H(A) \quad (3)$$

where $H(S)$ signifies the entropy of the initial dataset and $H(A)$ represents the entropy of the dataset following the split based on feature A.

2.2.2. Correlation-based

Pearson correlation feature selection is a method for determining the degree and direction of a linear connection in a dataset between a feature and a target variable. It measures how much the target variable changes when the feature changes. The Pearson correlation coefficient, indicated as r, has a value between -1 and 1, with -1 indicating a strong negative linear relationship, 0 indicating no linear relationship, and 1 indicating a perfect positive linear relationship. The Pearson correlation coefficient between a feature X and a target variable Y is calculated as (4):

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (4)$$

where individual values of the feature and target variable are denoted by X_i and Y_i, respectively. $\bar{X}$ represents the average value of the feature variable, while $\bar{Y}$ represents the average value of the target variable. n represents the quantity of data points.

In employing Pearson correlation for feature selection, features exhibiting greater absolute values of r (approaching 1 in either direction) are regarded as being more pertinent or significant for predicting the target variable. Conversely, features with low correlation (near 0) are typically seen as less significant for the given task. Both [29]-[30] used a threshold value based on the dependency value produced by Pearson correlation which are 0.5 and 0.7 respectively. In this study, we will consider a value in between which is 0.6 as a new distinct value.

2.2.3. Chi-square

Chi-square (x^2) feature selection is a statistical method for identifying the most important features in a dataset for classification tasks. In a categorical dataset, it assesses the independence between a feature and the target variable. The Chi-square test determines how much the observed data distribution differs from the predicted distribution, given that the features and the target variable are independent [36]. The chi-square is calculated as in (5):

$$x^2(X,Y) = \sum \frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \quad (5)$$

where O_{ij} denotes the observed frequency of occurrence for a certain combination of a feature category X (category i) and a target variable category Y (category j) while E_{ij} is the expected frequency assuming independence calculated as (6):

$$\frac{(total\ count\ of\ samples\ in\ category\ i\ for\ feature\ X)\times(total\ count\ of\ samples\ in\ category\ j\ for\ target\ variable\ Y)}{total\ count\ of\ samples} \quad (6)$$

Larger x^2 values suggest that the feature is strongly linked to the target variable, making it very useful for predicting or categorizing.

2.2.4. Sequential selection

Sequential feature selection (SFS) is a technique used to choose a subset of features from a larger feature set based on their relevance to a target variable. It entails analyzing several feature combinations and choosing the optimum subset that optimizes a stated criterion, such as accuracy, in a machine learning model. The goal is to utilize a search strategy (forward or backward) to efficiently explore the feature space, adding or deleting features repeatedly and assessing the impact on model performance until the desired stopping condition is reached.

a. Forward selection:
   - Start with an empty set: Begin with no features selected.
   - Choose a feature: Evaluate each feature separately and select the one that performs best based on a certain criterion (for example, accuracy).
   - Add the feature: Add the selected feature to the feature set.
   - Iterate: Repeat steps 2 and 3, considering the current set of selected features and evaluating the addition of one more feature at a time until the desired number of features is reached.
   - Stop criterion: Stop when you have the appropriate number of features or when the performance improvement falls below a certain threshold.

b. Backward selection:
   - Start with all features: Begin with all features in the set.
   - Choose a feature to remove: Evaluate the model's performance with each feature and choose the one that has the least influence on the specified criterion (e.g., accuracy) when eliminated.
   - Remove the feature: Discard the selected feature from the set.
   - Iterate: Repeat steps 2 and 3 while reviewing the smaller set of features and evaluating the removal of one more feature at a time until the required number of features is obtained.
   - Stop criterion: Stop when you have the appropriate number of features or when the performance improvement falls below a certain threshold.

2.3. Modelling

In this part, MLP is solely being used as the model for assessing the performance of both types of feature selection methods. MLP, the most common feed-forward neural network, contains a minimum of three layers or more, which comprises of input layer, hidden layer, and output layer as shown in Figure 2. It works by sending data from the input layer to the neurons in the output layer. For the execution of the model, we implemented the MLP using a Python-based software, namely Jupyter Notebook. Following dataset preprocessing, which encompassed the selection of pertinent features through the utilization of five distinct feature selection methods outlined in the preceding step, the chosen features from each method underwent training with the MLP model which took into account both the prudent application of cross-validation, in this case 5-fold was used, and the employment of synthetic over-sampling technique (SMOTE). This integration aimed to fortify the model's robustness, demonstrating its efficacy in addressing the challenge of class imbalance based on the target variable where SMOTE was adeptly employed to generate new synthetic instances for the minority class, matching the number of instances in the majority class, across each fold.

The assessment of each feature selection combination, in conjunction with MLP, will be conducted by evaluating metrics derived from the confusion matrix to ascertain their respective performances. Below are several measurements used in evaluating the model:

$$Accuracy = \frac{True\ positive+True\ negative}{Total\ samples} \quad (7)$$

$$Precision = \frac{True\ positive}{True\ positive+False\ positive} \quad (8)$$

$$Recall = \frac{True\ positive}{True\ positive+False\ negative} \quad (9)$$

$$F1-score = 2\ \times\ \frac{Precision\times Recall}{Precision+Recall} \quad (10)$$

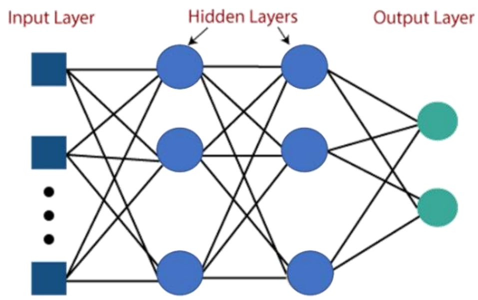

Figure 2. Structure of MLP

3. RESULTS AND DISCUSSION

The experiment was conducted on online learning activities dataset obtained from the online learning platform for ECE431 subject code at Universiti Teknologi MARA UiTM, Malaysia. The dataset contains 102 samples of Electronic/Electrical Engineering students for the programming course of the early semester. In order to avoid any misinformation and noise, data cleaning was done at the early phase of classification whereby one feature (Feature5) was removed since it included no values at all. In the next phase, several feature selection methods were employed as can be seen in Tables 2 and 3, which present the tables of features' importance for IG and Chi-square respectively, while Figure 3 displays the correlation matrix indicating the correlation of each feature towards each other. The figure was used to discard those features that show high correlations between the features. In the modelling part, we used MLP with the default hyperparameters such as (*hidden_layer_sizes* =100, activation = 'relu') to classify the dataset with the inclusion of 5 folds cross validation along with SMOTE which applied for each fold in the training set. A higher value of folds used will consume more processing capacity and time complexity.

Table 2. Features' importance for information gain

Features	Importance values
Feature12	0.253611
Feature11	0.229787
Feature2	0.225398
Feature10	0.078053
Feature6	0.023170
Feature4	0.001085
Feature1	0
Feature3	0
Feature7	0
Feature8	0
Feature9	0

Table 3. Features' importance for chi-square

Features	Importance values
Feature2	0.000077
Feature4	0.000082
Feature12	0.000374
Feature3	0.003844
Feature9	0.062828
Feature1	0.113873
Feature11	0.292479
Feature6	0.470298
Feature10	0.492987
Feature7	0.781754
Feature8	0.935113

3.1. Filter-based feature selection

The values displayed in Table 2 provide valuable insights into the extent of each feature's influence on the target variable. The values, ranging from 0 to 0.253611, signify the degree of dependence of a feature in predicting the target variable. A lower value indicates a reduced level of influence or relevance of that particular feature in determining the outcome. Certain attributes have modest relevance values, suggesting that they have little influence on the target variable. Features 1, 3, 7, 8, and 9, which concern students' access to notes and involvement in activities, all have significance levels of zero. This suggests that these qualities are fundamentally independent of the objective variable, therefore including them in the analysis may not add significantly to forecasting the outcome. This focused selection of relevant features is crucial in optimizing model performance and enhancing the efficiency of the analysis.

Applying the chi-square method introduced a statistical aspect into our analysis. By considering p-values, a key determinant of statistical significance, we could evaluate the dependency of each feature on the target variable. We had set a standard p-value threshold of 0.05, a widely recognized level of significance, to categorize features accordingly. Upon reviewing Table 3, it became evident that seven features (Feature 1, 6, 7, 8, 9, 10, and 11) crossed this critical p-value threshold which indicates less influence. Features exceeding the threshold were discarded from the original dataset so that only relevant features will be used by creating a new subset of data. Hence, by having a new subset of data, including only the relevant features, it will somehow aid in enhancing the model's performance.

The Pearson correlation technique involved a meticulous review of the correlation matrix illustrated in Figure 3. Here, any values surpassing the established threshold of 0.6 were regarded as exhibiting significant correlation with other features. Consequently, four features (specifically, Feature 6, 7, 8, and 9) were removed due to their pronounced correlations with other features, highlighting redundancy in their inclusion. As a result, the remaining features indicating the most independent features will be trained in the next stage as another subset of data.

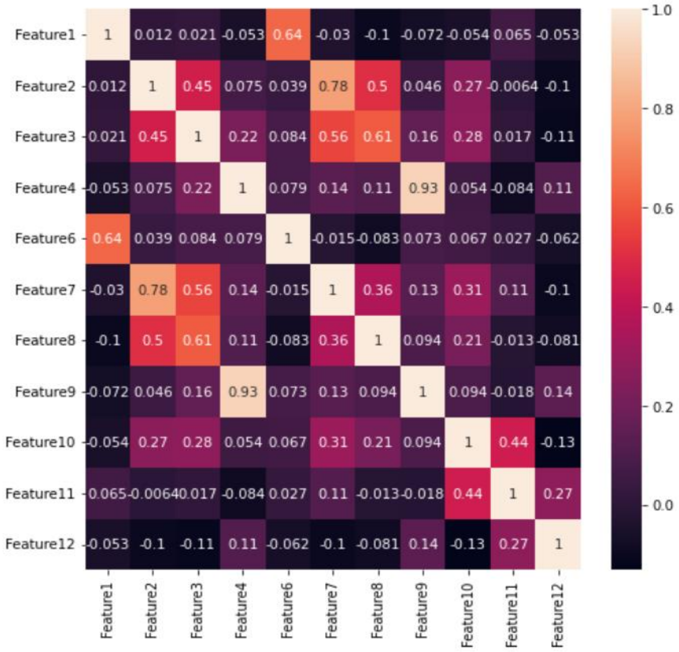

Figure 3. Correlation matrix for Pearson correlation method

3.2. Wrapper-based feature selection

Figures 4 and 5 illustrate the outcomes of forward and backward feature selection, showcasing the accuracy associated with varying numbers of features. In Figure 4, the model's accuracy remained consistently at 100% when utilizing 1 to 6 features. On the other hand, Figure 5 demonstrated the results of backward feature selection, revealing 100% accuracy with the first 5 features, which slightly diminished to 99.9% accuracy with the inclusion of 6 features used. A noticeable decline pattern in accuracy was depicted in both tables for more features used. However, backward feature selection showcased a more favorable overall performance, with the lowest accuracy stabilizing at 95% for more than 8 features utilized.

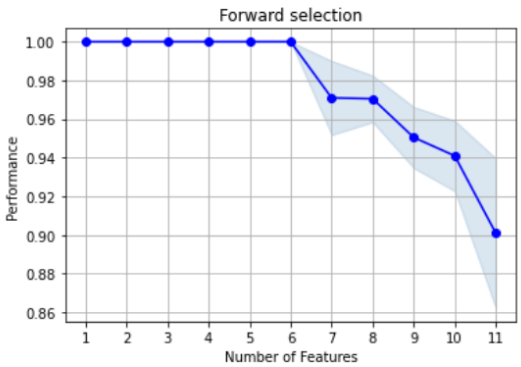

Figure 4. Graph for forward feature selection performance

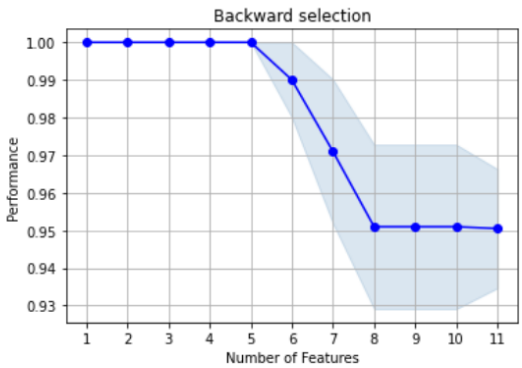

Figure 5. Graph for backward feature selection performance

3.3. Selected features and model evaluation

In Table 4, each feature selection method comes along with the selected features whereby it was generated during data preprocessing stage. In terms of the number of selected features, Chi-square had the least number of features which was 4 features while the highest was 7 features selected by the Pearson correlation method. In general, a discernment of the most pivotal features can be gleaned from prioritizing the methods exhibiting the highest performance. Notably, Feature 1, 2, 4, 6, 11, and 12 consistently emerged as the prominently selected features. Among these, Feature 1 to 4 primarily pertain to students' accessibility to course materials, Feature 6 highlights students' proactive engagement with exercises preceding forthcoming lessons, and Feature 11 to 12 shed light on students' answers pattern on tutorial questions.

Based on Table 5, the model with both wrapper methods, which are forward selection and backward selection had accurately predicted the testing dataset with 100% accuracy. As for the filter-based methods, IG significantly topped the chart with 100% accuracy, followed by Chi-square and Pearson correlation with the same accuracy of 95.24%.

Table 4. Selected features based on several feature selection methods

Methods	Selected features
Information gain	2,4,6,10,11,12
Chi-square	2,3,4,12
Pearson correlation	1,2,3,4,10,11,12
SFS	1,6,8,10,11,12
SBS	1,6,8,9,11,12

Table 5. Model's performance for five feature selection methods

	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Information gain	100	100	100	100
Chi-square	95.24	95.24	100	97.56
Pearson correlation	95.24	95.24	100	97.56
SFS (Forward)	100	100	100	100
SBS (Backward)	100	100	100	100

4. CONCLUSION

This paper undertook a comparative analysis of various feature selection techniques to unveil their effectiveness in classifying students' performance. It has been demonstrated that feature selection is critical in determining the most important features while also minimizing computing time complexity in classification. The results indicated that information gain, along with both forward and backward selection, wielded substantial influence in the classification of student performance where the three methods' performance recorded 100% for all evaluation metrics including Accuracy, Precision, Recall and F1-score. Substantially, information gain took less computation time whereas both wrapper methods took much longer in computing the tasks given. The study also highlighted the relevance of utilizing online learning activities as a dataset, as the selected features encompassed categories like notes, exercises, and tutorials. Moving forward, we posit that this research could be further explored from different perspectives, including the implementation of these techniques with diverse sets of algorithms.

ACKNOWLEDGEMENT

The authors would like to express gratitude to College of Engineering, Universiti Teknologi MARA (UiTM) Shah Alam for the funding of this study.

REFERENCES

[1] 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," in *Materials Today: Proceedings*, 2021, vol. 47, pp. 5260–5267, doi: 10.1016/j.matpr.2021.05.646.

[2] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: automated machine learning approach," *Computers and Electrical Engineering*, vol. 89, Jan. 2021, doi: 10.1016/j.compeleceng.2020.106903.

[3] X. Tao *et al.*, "Data analytics on online student engagement data for academic performance modeling," *IEEE Access*, vol. 10, no. September, pp. 103176–103186, 2022, doi: 10.1109/ACCESS.2022.3208953.

[4] S. D. A. Bujang *et al.*, "Imbalanced classification methods for student grade prediction: a systematic literature review," *IEEE Access*, vol. 11, no. October 2022, pp. 1970–1989, 2023, doi: 10.1109/ACCESS.2022.3225404.

[5] M. Riestra-González, M. del P. Paule-Ruíz, and F. Ortin, "Massive LMS log data analysis for the early prediction of course-agnostic student performance," *Computers and Education*, vol. 163, Apr. 2021, doi: 10.1016/j.compedu.2020.104108.

[6] V. Christou *et al.*, "Performance and early drop prediction for higher education students using machine learning," *Expert Systems with Applications*, vol. 225, no. April, 2023, doi: 10.1016/j.eswa.2023.120079.

[7] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," *Trends in Neuroscience and Education*, vol. 33, 2023, doi: 10.1016/j.tine.2023.100214.
[8] L. Vives *et al.*, "Prediction of students' academic performance in the programming fundamentals course using long short-term memory neural networks," *IEEE Access*, vol. 12, pp. 5882–5898, 2024, doi: 10.1109/ACCESS.2024.3350169.
[9] W. Qiu, A. W. H. Khong, S. Supraja, and W. Tang, "A dual-mode grade prediction architecture for identifying at-risk students," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 803–814, 2024, doi: 10.1109/TLT.2023.3333029.
[10] M. Adnan *et al.*, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021, doi: 10.1109/ACCESS.2021.3049446.
[11] R. Z. Pek, S. T. Ozyer, T. Elhage, T. Ozyer, and R. Alhajj, "The role of machine learning in identifying students at-risk and minimizing failure," *IEEE Access*, vol. 11, no. December 2022, pp. 1224–1243, 2023, doi: 10.1109/ACCESS.2022.3232984.
[12] M. Wafi, U. Faruq, and A. A. Supianto, "Automatic feature selection for modified k-nearest neighbor to predict student's academic performance," in *Proceedings of 2019 4th International Conference on Sustainable Information Engineering and Technology, SIET 2019*, Sep. 2019, pp. 44–48, doi: 10.1109/SIET48054.2019.8986074.
[13] V. Mhetre and M. Nagar, "Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA," in *Proceedings of the International Conference on Computing Methodologies and Communication, ICCMC 2017*, 2018, vol. 2018-Janua, no. Iccmc, pp. 475–479, doi: 10.1109/ICCMC.2017.8282735.
[14] E. Alhazmi and A. Sheneamer, "Early predicting of students performance in higher education," *IEEE Access*, vol. 11, no. February, pp. 27579–27589, 2023, doi: 10.1109/ACCESS.2023.3250702.
[15] A. Khan, S. K. Ghosh, D. Ghosh, and S. Chattopadhyay, "Random wheel: an algorithm for early classification of student performance with confidence," *Engineering Applications of Artificial Intelligence*, vol. 102, Jun. 2021, doi: 10.1016/j.engappai.2021.104270.
[16] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, 2019, doi: 10.1016/j.heliyon.2019.e01250.
[17] D. Sun *et al.*, "A university student performance prediction model and experiment based on multi-feature fusion and attention mechanism," *IEEE Access*, vol. 11, no. August, pp. 112307–112319, 2023, doi: 10.1109/ACCESS.2023.3323365.
[18] S. Roy and A. Garg, "Predicting academic performance of student using classification techniques," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, UPCON 2017*, 2017, vol. 2018-Janua, pp. 568–572, doi: 10.1109/UPCON.2017.8251112.
[19] M. Sivasakthi, "Classification and prediction based data mining algorithms to predict students' introductory programming performance," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, Nov. 2017, pp. 346–350, doi: 10.1109/ICICI.2017.8365371.
[20] F. J. Kaunang and R. Rotikan, "Students' academic performance prediction using data mining," *Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018*, 2018, doi: 10.1109/IAC.2018.8780547.
[21] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2015*, 2015, doi: 10.1109/AEECT.2015.7360581.
[22] H. Z. Hashemi, P. Parvasideh, Z. H. Larijani, and F. Moradi, "Analyze students performance of a national exam using feature selection methods," in *2018 8th International Conference on Computer and Knowledge Engineering, ICCKE 2018*, Dec. 2018, pp. 7–11, doi: 10.1109/ICCKE.2018.8566671.
[23] W. Nuankaew and J. Thongkam, "Improving student academic performance prediction models using feature selection," in *17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2020*, 2020, pp. 392–395, doi: 10.1109/ECTI-CON49241.2020.9158286.
[24] S. D. A. Bujang *et al.*, "Multiclass prediction model for student grade prediction using machine learning," *IEEE Access*, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.
[25] M. R. Ahmed, S. T. I. Tahid, N. A. Mitu, P. Kundu, and S. Yeasmin, "A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms," *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, 2020, doi: 10.1109/ICCCNT49239.2020.9225341.
[26] K. Sixhaxa, A. Jadhav, and R. Ajoodha, "Predicting students performance in exams using machine learning techniques," in *Proceedings of the Confluence 2022 - 12th International Conference on Cloud Computing, Data Science and Engineering*, 2022, pp. 635–640, doi: 10.1109/Confluence52989.2022.9734218.
[27] W. Punlumjeak and N. Rachburee, "A comparative study of feature selection techniques for classify student performance," in *Proceedings - 2015 7th International Conference on Information Technology and Electrical Engineering: Envisioning the Trend of Computer, Information and Engineering, ICITEE 2015*, 2015, pp. 425–429, doi: 10.1109/ICITEED.2015.7408984.
[28] N. Nidhi, M. Kumar, and S. Agarwal, "Comparative analysis of heterogeneous ensemble learning using feature selection techniques for predicting academic performance of students," *Proceedings - 2021 2nd International Conference on Computational Methods in Science and Technology*, pp. 212–217, 2021, doi: 10.1109/ICCMST54943.2021.00052.
[29] Y. Shi, F. Sun, H. Zuo, and F. Peng, "Analysis of learning behavior characteristics and prediction of learning effect for improving college students' information literacy based on machine learning," *IEEE Access*, vol. 11, no. April, pp. 50447–50461, 2023, doi: 10.1109/ACCESS.2023.3278370.
[30] S. Sengupta, "Towards finding a minimal set of features for predicting students' performance using educational data mining," *International Journal of Modern Education and Computer Science*, vol. 15, no. 3, pp. 44–54, 2023, doi: 10.5815/ijmecs.2023.03.04.
[31] M. Zaffar, M. A. Hashmani, and K. S. Savita, "Performance analysis of feature selection algorithm for educational data mining," *2017 IEEE Conference on Big Data and Analytics*, vol. 2018-Janua, pp. 7–12, Feb. 2018, doi: 10.1109/ICBDAA.2017.8284099.
[32] K. Sabaneh and R. Jayousi, "Prediction of students' performance in e-learning courses," in *Proceedings - 2021 International Conference on Promising Electronic Technologies, ICPET 2021*, 2021, pp. 52–57, doi: 10.1109/ICPET53277.2021.00016.
[33] M. Garg and A. Goel, "Preserving integrity in online assessment using feature engineering and machine learning," *Expert Systems With Applications*, vol. 225, no. April, 2022, doi: 10.1016/j.eswa.2023.120111.
[34] Dafid and Ermatita, "Filter-based feature selection method for predicting students' academic performance," in *2022 International Conference on Data Science and Its Applications, ICoDSA 2022*, 2022, pp. 309–314, doi: 10.1109/ICoDSA55874.2022.9862883.
[35] B. Azhagusundari, "An integrated method for feature selection using fuzzy information measure," in *2017 International Conference on Information Communication and Embedded Systems (ICICES)*, Feb. 2017, pp. 1–6, doi: 10.1109/ICICES.2017.8070711.
[36] V. Shanmugarajeshwari and R. Lawrance, "Analysis of students' performance evaluation using classification techniques," *International Conf. on Computing Technologies and Intelligent Data Engineering*, 2016, doi: 10.1109/ICCTIDE.2016.7725375.

[37] S. K. Trivedi, A. Sharma, P. Patra, and S. Dey, "Prediction of intention to use social media in online blended learning using two step hybrid feature selection and improved SVM stacked model," *IEEE Transactions on Engineering Management*, pp. 1–16, 2022, doi: 10.1109/TEM.2022.3212901.
[38] X. Li, K. Jiang, H. Wang, X. Zhu, R. Shi, and H. Shi, "A novel K-means classification method with genetic algorithm," *Proceedings of 2017 International Conference on Progress in Informatics and Computing, PIC 2017*, pp. 40–44, 2017, doi: 10.1109/PIC.2017.8359511.
[39] S. S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "An heuristic feature selection algorithm to evaluate academic performance of students," in *ICSGRC 2019 - 2019 IEEE 10th Control and System Graduate Research Colloquium, Proceeding*, 2019, no. August, pp. 110–114, doi: 10.1109/ICSGRC.2019.8837067.
[40] N. Rachburee and W. Punlumjeak, "A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining," in *Proceedings - 2015 7th International Conference on Information Technology and Electrical Engineering: Envisioning the Trend of Computer, Information and Engineering, ICITEE 2015*, 2015, pp. 420–424, doi: 10.1109/ICITEED.2015.7408983.

BIOGRAPHIES OF AUTHORS

Muhamad Aqif Hadi Alias received the B.Eng. degree in electronic engineering from Universiti Teknologi MARA (UiTM), Malaysia, in 2022. Currently, he is a master's degree student at the School of Electrical Engineering, Universiti Teknologi MARA. His research interest includes the area of classification which specifically focuses on students' performance by obtaining their online learning activities. He can be contacted at email: aqifhadi7@gmail.com.

Mohd Azri Abdul Aziz is a senior lecturer at the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), Malaysia. He received the B.Eng. (Hons) degree in electronic engineering from University of Manchester Institute of Science and Technology (UMIST), UK. M.Sc. degree in electrical engineering and Ph.D. degree in electrical engineering, both from Universiti Teknologi MARA, Malaysia. He is also a fellow at Innovative Electromobility Research Lab (ITEM) UiTM, primarily researching on autonomous vehicles. His research area includes pattern classification, object detection, and IoT-based application. Currently, he is supervising and co-supervising 5 master's and 6 PhD students. He has authored and co-authored 8 journals and 16 conference proceedings with 5 H-index and 73 citations. He can be contacted at email: azriaziz@uitm.edu.my.

Najidah Hambali completed her PhD in advanced process control at the Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM). She received bachelor of engineering (B. Eng. Hons.) in electronics engineering from Universiti Sains Malaysia (USM) in 2004. In 2005, she joined Universiti Malaysia Pahang (UMP) as a tutor. After received her master's engineering science (M.Eng.Sc.) in systems and control from The University of New South Wales (UNSW), Sydney, Australia in 2006, she continued working in UMP as a lecturer until 2010. In January 2011, she joined UiTM and later in October 2015, she started her PhD study in UiTM. Currently, she is a senior lecturer in the Centre of System Engineering Studies (CSES), School of Electrical Engineering, College of Engineering UiTM, Shah Alam. Her current research interests are control system and nonlinear modelling for process control. She can be contacted at email: najidah@uitm.edu.my.

Mohd Nasir Taib is a professor of control and instrumentation, Universiti Teknologi MARA, Selangor, Malaysia. His area of expertise are control system and signal processing. He is a former director of Malaysia Institute of Transport, Malaysia. He is also Head of the Advanced Signal Processing Research Group (ASPRG), UiTM Shah Alam. He has published more than 500 research and academic journals around the world and won many gold awards in innovations that he participated in. He can be contacted at email: dr.nasir@uitm.edu.my.