I. J. Education and Management Engineering, **2023, 5, 1-22**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2023.05.01

A Hybrid Weight based Feature Selection Algorithm for Predicting Students' Academic Advancement by Employing Data Science Approaches

Ujwal U.J
KVG College of Engineering, Sullia, India
E-mail: ujwalu@yahoo.com
ORCID iD: https://orcid.org/0009-0007-7400-1988

Saleem Malik*
KVG College of Engineering, Sullia, India
E-mail: baronsaleem@gmail.com
ORCID iD: https://orcid.org/0000-0002-1510-1758
*Corresponding Author

Received: 16 January, 2023; Revised: 13 February, 2023; Accepted: 27 April, 2023; Published: 08 October, 2023

Abstract: PerformanceX is a proposed system that combines Educational Data Mining (EDM) techniques to enhance student performance and reduce dropout rates. It employs a hybrid feature selection approach to identify the most significant attributes from student academic datasets, eliminating unnecessary features that are not crucial for predicting performance. The selectX algorithm, a critical component of PerformanceX, selects a limited number of high-performing features to optimize student learning effectiveness and prediction accuracy. The system applies various machine learning classifiers, including a fusion Voting Classifier, to different subsets of features, ultimately determining the best combination. The study achieved an impressive accuracy rate of 99.41%, with the selectX approach utilizing 10 features in conjunction with a random forest (RF) classifier offering the highest accuracy. These findings underscore the importance of categorizing student performance based on a concise yet meaningful set of features, leading to improved student quality and career progression. The research value of PerformanceX lies in the development of a performance forecasting system that eliminates irrelevant information and provides precise predictions for student performance. Its efficacy and efficiency make it an invaluable tool for educators and educational institutions. By assisting students in selecting appropriate courses to enhance their performance and advance their careers, PerformanceX contributes to diminishing dropout rates while fostering positive student outcomes.

Index Terms: Educational Data Mining, Feature selection, Data Science

1. Introduction

In tertiary-level education, it has become a common problem that students are lackadaisical about attending tests and maintaining their grades at the level required to pass the course. This not only violates the National Educational Policy of India 2020, which mandates that all undergraduate students meet necessary educational requirements, but also creates a situation where students are unaware of their true academic standing until they receive an unqualified grade [1]. Although instructions are given to students, they tend to disregard them, which results in poor institutional success rates. In order to pass the program, students must pass three written exams, three skill tests, and one external exam for each topic, and the grade for each semester is only known after all these tests are completed. If a student misses any material on a topic, they must take a supplement test during the subsequent semester. To address this problem, this article proposes an algorithm to forecast each student's performance in all topics following the conclusion of the second written test, which can help teachers give students the necessary support and prepare them to graduate. This algorithm will provide performance expectations each semester through its tracking system and will be integrated into the student's learning experience platform [2-4].

The proposed algorithm combines the working concept of the filter and wrapper techniques to achieve the dominance of both techniques in a single method, with the provision that the selected characteristics be few and effective [5]. While several existing solutions employ either filter or wrapper methods, this article proposes a new model that combines the best aspects of both methods. The most important features are selected as a new set of characteristics, and various machine learning algorithms are trained and tested using this feature set. Although the choice of machine learning algorithm depends on the parameter list and forecasting results, the ensemble model is a powerful forecasting method that generates a large number of results. The accuracy of the forecast is impacted by the variables employed in the algorithm's performance. Through this article, we aim to help students improve their academic success throughout the program, so that every student passes on the first try. The proposed algorithm can provide performance expectations for each semester, thereby reducing student apathy towards tests and increasing institutional success rates [6].

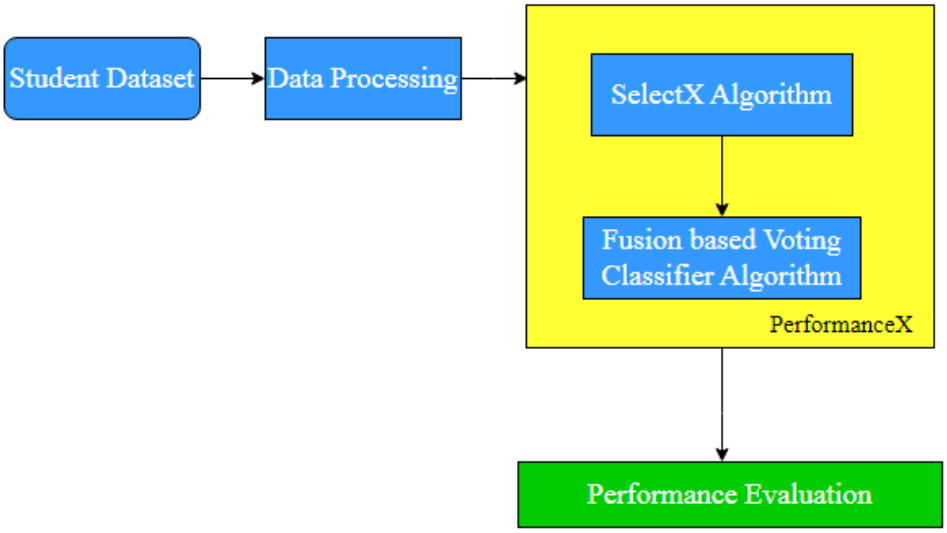

Fig. 1. Overview of proposed work

The figure 1 represents the overview of the proposed system for predicting student performance using feature selection and fusion based voting classifier algorithms. The first step in the process is the data processing, where the student dataset is preprocessed to remove any unwanted information or noise. The preprocessed data is then fed into the SelectX algorithm, which uses three different feature selection techniques (Chi-square, Feature Importance, RFE) to select the most important features that will be used for classification [7]. The output of the SelectX algorithm is the top influencing features. The top influencing features are then fed into the Fusion Based Voting Classifier Algorithm, which selects the best feature selection technique based on the best classification models. The top three performing classifiers are then selected, and a voting classifier is built using these classifiers. The voting classifier is then used to predict student performance, and the results are evaluated using various performance metrics such as accuracy, recall, precision, F1 score, and kappa [8]. If any early interventions are required, they are applied based on the predicted performance. The proposed system provides an efficient and effective way of predicting student performance, which can help identify struggling students early on and provide appropriate interventions to improve their academic success [9].

1.1 Principle and significance of research

The foundation of modern civilization is heavily influenced by a society's level of education, and higher education institutions operate in a complex and ever-changing environment. With the increasing number of students and subjects offered, these institutions are impacted by a variety of factors, including competition between universities, government policies, and societal exchanges. To make informed decisions in this rapidly changing landscape, it is important to utilize the vast data sources generated by manual and digital technologies. Predictive analytics can provide a powerful tool for generating creative action ideas to enhance educational outcomes. In this study, we propose an algorithm, SelectX, to forecast student performance and help students achieve academic success throughout their program by providing performance expectations each semester through its tracking system. By combining filter and wrapper techniques, SelectX identifies key features that are both few and effective.

This research contributes to the field by offering a new method for forecasting student performance and promoting student success in higher education institutions The principle of this research is to propose an algorithm that can forecast students' performance in all topics instantly following the conclusion of the second written test in higher education institutions. The significance of this research lies in its potential to provide teachers with the necessary information to give students the support they need to improve their academic success throughout the program, so that every student passes on the first try. This predictive analytic system, when combined with the enormous data sources produced by manual and digital technologies, can provide creative action ideas that can help institutions' management

adjust their decision-making process in a rapidly changing environment. Additionally, this research aims to bridge the knowledge gap by providing a systematic way to identify areas where students may be struggling and to provide personalized support to help them succeed. This research can contribute to the improvement of the overall quality of higher education, ensuring that students receive a comprehensive and effective learning experience.

2. Background and Related Works

Researchers have recently suggested a number of student achievement forecasting strategies integrating with Data Science (DS) techniques to inflate academic performance of students. Herein, we examine a few recent studies and cutting-edge methods which have appeared in recent years to enhance academic systems' forecasting as well as call for continued development. [8] did an exhaustive study of the various techniques used for prediction of academic results. [9] found that many researchers had used back propagation, Bayesian networks, regression models, decision tree and search-based rule extraction algorithm [10]. [11] used SVM and multiple linear regression (MLR) to predict the results of engineering students of a particular course and found that SVM gave higher accuracy results compared to MLR Many researchers have used regression models for predicting academic performance, knowledge skill, etc. Regression analysis finds the relationship between one dependent variable and another independent variable [12]. Using these data mining models they were able to obtain good predictive results for the students [13]. Portuguese students had high academic failure rates and the research conducted by Cortez and Silva helped to identify the issues.

A comparative study of the data mining methods was done by [14] for assessing an intelligent tutoring system [15]. [16] Used and assessed different types of neural network models to determine the academic performance of students. [17] Presented a very interesting study where he built academic performance prediction models on enrollment data for evaluating international students' applications. He used the Bayesian network model to assess the applicants. [18] Used fuzzy based models to predict academic performance. [19] Performed monitoring and evaluation of students. [20] Used a genetic algorithm to find association rules to predict academic performance in distance learning. [21] Did an in-depth exploration of predictive models using NBtree. The experiments were conducted with two-level classification at university level and at faculty level [22] Performed a comparative study on academic performance prediction using C4.5 algorithm and Naïve Bayes. [23] implemented the student performance prediction using multiple instance-based learning along with the traditional supervised learning method of Naïve Bayes, decision tree and rule based algorithm .

[24] Applied data mining techniques on 670 school students of Mexico. They have used ten classification algorithms with ten-fold cross-validation. Some of the algorithms used were the random tree, J48, Prisim, ADTree and Simple Cart. [25] Developed an adaptive learning system using classification algorithms to find out the learning pattern of students on an e-learning platform. Their model was built using an artificial neural network [26]. [27] Created an automated student model for discovering their learning patterns. They used data of 71 students who used an automated tutor for learning algebra. They generated production rules using deep feature learning. [28] Used probability, regression and data mining algorithms to develop predictive models for understanding the learning patterns of students. [29] Used classification models using seven classification algorithms. The Sequential minimal optimization (SMO) performed the best of the seven classification algorithms with a true positive rate of 83%. [30] Used classification models in Massive Open Online Courses (MOOCS). [31] Performed a study to predict academic success and failure of students by taking students activity data in online courses offered in a private university. They used models based on J48, Naive Bayes, Regression and W-JRip. They evaluated their classification models using precision, recall, Receiver Operating Characteristics curve (ROC Curve). [32] Developed a student model to quantify whether or not a student has acquired a skill. Based on the students proficiency of skill attainment customized instruction could be provided through the automated tutor [33]. [34] Used clustering algorithms on web-based learning systems [35]. [36] Used online orientation course data of students to do a prediction of students' retention. [37,38] conducted a study to find that post-study performance is more indicative of long term learning than course performance. The study was conducted on MOOC data. [39,40] Developed predictive models to determine the dropout numbers. Their classification model was based ensemble method of classification. [41] Maximized the learning of the students by developing an algorithm to partition the students of a class into groups. [42] log data from Edulab for school students of Denmark. He used Markov chain along with k-means clustering to model students' behavior.

[43] Explained the various data preprocessing tasks that are performed for cleaning the data. [44] Created an information system for construction of the university information system. Data preprocessing has been used in most applications of machine learning including educational data mining, stock market prediction, weather forecast, disease prediction, bioinformatics and many more. Data preprocessing improves the performance of classification algorithms and improves the classification accuracy of most machine-learning algorithms [45]. [46] Described the various algorithms used in data pre-processing. [47-50] emphasized the significance of data pre-processing in educational data mining. They collected data from an e-learning system through an API. They performed tasks like data gathering, data aggregation, data cleaning, data filtering and data transformation using the WEKA toolbox. The pre-processing techniques improved the classification accuracy by around 29%. The classification techniques used were a neural network, decision tree and naïve Bayesian classifier.

Feature selection has been one of the most significant techniques to improve the performance of classification models. [51] Suggested that feature selection has been used in several applications as a tool to remove irrelevant

features from the dataset. [52] Discussed the construction of feature selection models. [53] Suggested that choice of feature selection algorithms should be based on dataset type and the properties of the dataset. [54] Discussed some of the powerful tools used for feature selection tools like principal component analysis, information-based feature selection methods for numerical attributes, chi- square based feature selection methods for categorical data attributes and wrapper-based feature selection methods. [55] Reinforced the significance of feature selection algorithms in enhancing the performance of machine learning algorithms. They also discussed the process of feature construction and feature evaluation.

Feature selection algorithms have been used in different domains and in different applications. [56] Used feature selection in the domain of healthcare for the diagnosis of Erythematous - Squamous Diseases. [57] Used feature selection in the domain of networking for detecting the black hole attack. [58] Used feature selection for predicting financial distress of companies. They used both filter and wrapper-based feature selection methods used feature selection for keystroke dynamic systems [59] investigated wrapper based feature selection methods on twelve datasets using the linear forward selection methodology [60]. [61] Conducted a study to investigate the attributes contributing to the success or failure of students in academic institutions. Their study confirmed a reduction in computational complexity. They used six filter-based feature selection algorithms [62]. [63] Used feature selection algorithms to evaluate student performance in academics. They used the perceptron neural network to form their classification model with ten-fold cross-validation. [64] Used ensemble method of classification and feature selection techniques to identify at-risk students of first year engineering

Feature selection has been used in the domain of educational data mining. [65] Used feature selection in the platform of modular object-oriented dynamic learning environment (MOODLE) to find out the factors affecting the students. [66] Implement feature selection algorithms in educational data mining and confirmed that using only a selected number of features could generate better predictive models. [67] used feature selection algorithms. They took data of 309 students with fourteen attributes each. They used correlation-based feature selection, chi-square based feature selection and information gain amongst the filter based feature selection methods. The wrapper-based feature selection methods used were a decision tree and Na ̈ve Bayes in association with rankers search. Based on the results obtained from the classification models and feature selection algorithms they used only eight of the fourteen features of the students to get the best performance from the features. [68] conducted an investigation to find out the factors affecting the success of students studying in higher educational institutes in Bosnia and Herzegovina. They used 12 student attributes for the data with a sample of size of 257. They used a decision tree, multi-layer perceptron and Na ̈ve Bayes as the classification algorithms. Some of the attributes used were gender, GPA, scholarship amount and study material used [69]. These optimization techniques have been used in the machine learning paradigm to improve the performance of these models. It has been used by classification models like neural network, support vector machines for optimum parameter setting. Optimization techniques have been also been used for feature selection across myriad applications. This method is popularly known as the wrapper method of feature selection. Several machine learning algorithms have been used by researchers in different domains. The most commonly used machine learning algorithms are support vector machines, decision trees and neural networks [70]. Optimization technique like differential evolution has been used by researchers to improve the performance of the RBFN network [71] by finding the most appropriate values of center and spread for the radial basis function network [72, 73]. [74] Implemented differential evolution (DE) on RBF network and demonstrated an improvement in the experimental results when DE is used to find the parameters for RBF.

3. Proposed Methodology

The hybrid feature selection approach employed in PerformanceX combines the advantages of filter-based and wrapper-based techniques to identify the most pertinent attributes from student academic datasets. Filter-based methods, including correlation-based feature selection (CFS) and information gain, evaluate the individual relevance of attributes with respect to the target variable. These methods rank the attributes based on their predictive power, considering their standalone importance without taking into account attribute combinations.In contrast, wrapper-based methods such as recursive feature elimination (RFE) and sequential forward selection (SFS) assess subsets of attributes by analyzing their performance in conjunction with a specific learning algorithm. These approaches iteratively choose subsets of attributes and examine their impact on predictive performance, facilitating a comprehensive evaluation of attribute combinations.

Through the integration of these two approaches, PerformanceX adeptly capitalizes on the advantages of both filter-based and wrapper-based methods. The application of filter-based techniques initially discerns attributes that possess significant relevance, establishing a robust foundation for the ensuing wrapper-based assessment. Consequently, the wrapper-based methods augment the attribute selection process by scrutinizing various attribute subsets based on their performance in conjunction with the designated learning algorithm. This iterative procedure guarantees that the ultimate attribute subset not only encompasses highly pertinent attributes but also optimizes the predictive efficacy of the chosen learning algorithm.Feature selection (FS) can be construed as a procedure about choosing preferentially refining consequential features among all of the feasible features in a dataset utilizing a variety of strategies, such as

filter and wrapper approaches. Depending on how closely the characteristics match the result parameter within numerous statistical investigations, filter strategies like the feature importance approach choose the characteristics.

Through the use of wrapper techniques like recursive feature elimination technique (RFE), this and that seeks to apply a subgroup of characteristics as well as learn a model with them [75]. The aforementioned method's calculation often takes a long time. While wrapper techniques verify the usefulness of a subgroup of features through literally learning a model onward, filter methods evaluate the relevance of features through association depending on a parameter. Filter approaches remain quicker than wrapper approaches because they don't require model learning. On the other hand, wrapper approaches need a lot of processing. Principal Component Analysis (PCA) metamorphoses a group of correlated parameters towards a set of uncorrelated parameters utilizing an orthographic metamorphosis[75]. Within EDA (Exploratory Data Analysis) as well as machine learning (ML) techniques towards predictive models, a superior technique that is regularly utilized is PCA. Another unsupervised statistical technique for scrutinizing the relationships among a set of parameters is PCA. Chi-square Statistical methods (CHS) are used in feature selection to recognize the best important features within a prediction matter. The feature with the top score is discovered by separating the minimal essential features from the existing set. A method known as recursive feature removal continually finds features for deletion while taking into account all of the input attributes. The importance of a feature concerning the target parameter depends on the sequence in which it is eliminated. Feature Importance approach (FI) grades input features according to how effectively they can forecast a target variable[76].

The effectiveness of the proposed method lies in its unique combination of filter-based and wrapper-based feature selection techniques. While both filter-based and wrapper-based methods have been widely used in feature selection, their integration within the PerformanceX system offers distinct advantages. By leveraging the strengths of both approaches, we are able to overcome the limitations of each method and achieve improved performance in predicting student outcomes. Filter-based methods excel at quickly identifying attributes with high relevance to the target variable. They efficiently rank attributes based on their individual predictive power. However, they often fail to consider the impact of attribute combinations on performance prediction. On the other hand, wrapper-based methods are known for their ability to evaluate attribute subsets by considering their performance in conjunction with a specific learning algorithm. This comprehensive evaluation allows for a more accurate assessment of attribute combinations. However, wrapper-based methods can be computationally expensive and may not be practical for large datasets. By combining these two approaches, PerformanceX benefits from the best of both worlds. The filter-based methods provide an initial screening of attributes, narrowing down the search space to a subset of highly relevant features. This significantly reduces computational complexity and ensures that only the most promising attributes are considered. The subsequent wrapper-based evaluation then focuses on this refined subset, evaluating the performance of different attribute combinations with the chosen learning algorithm. This iterative process allows PerformanceX to identify the optimal feature subset that maximizes predictive accuracy while maintaining computational efficiency.

The contribution of combining existing models in this manner lies in the novel integration and adaptation of these techniques specifically for the task of predicting student performance. While filter-based and wrapper-based methods have been used independently in various domains, their application within the educational context is relatively unexplored. The unique challenges and requirements of predicting student outcomes necessitate a tailored approach that effectively captures the complex relationships between attributes and performance. By combining existing models and adapting them to the educational domain, we are able to provide a powerful and effective tool, PerformanceX, for educators and educational institutions.Furthermore, the integration of statistical analysis and machine learning techniques within the selectX algorithm contributes to its effectiveness. The statistical measures provide insights into the distribution and variability of attribute values, allowing for the identification of attributes with discriminative characteristics [77]. The incorporation of machine learning classifiers then evaluates the predictive power of each attribute, ensuring that the selected features not only exhibit statistical relevance but also contribute to accurate performance prediction [78].The effectiveness of the proposed method lies in its unique combination of filter-based and wrapper-based techniques, addressing the limitations of each method and achieving improved performance in predicting student outcomes. The contribution of combining existing models lies in the adaptation and integration of these techniques specifically for the educational context, providing a tailored and powerful tool for educators. The incorporation of statistical analysis and machine learning further enhances the effectiveness of the method by ensuring a data-driven approach to feature selection [79].

A. Weight based feature selection algorithm- SelectX

The selectX algorithm plays a pivotal role in the PerformanceX system, responsible for meticulously selecting a concise yet high-performing set of features from the identified attribute subset. It combines statistical analysis and machine learning techniques to ascertain the most influential features for optimizing student learning effectiveness and prediction accuracy. Initially, selectX computes diverse statistical measures, such as average, standard deviation, and skewness, for each attribute in the attribute subset. These measures furnish valuable insights into the distribution and variability of attribute values. Attributes displaying substantial variations and discriminative characteristics are considered more influential in prognosticating student performance [80].

Moreover, selectX integrates machine learning classifiers to evaluate the predictive capacity of each attribute. It trains and tests diverse classifiers on subsets of attributes to assess the performance of each attribute individually. This evaluation aids in identifying attributes that significantly contribute to accurate performance prediction. By merging

statistical analysis with machine learning techniques, selectX embraces a data-driven approach to feature selection. The statistical analysis assists in identifying attributes with desirable statistical properties, while the machine learning evaluation proffers insights into their predictive power. This amalgamated approach empowers selectX to handpick a limited number of features that manifest both statistical relevance and high predictive accuracy.

The Select X algorithm is a hybrid weight-based feature selection algorithm (Algorithm 1: A Hybrid Weight Based Feature Selection (Select X)) that aims to identify the most influential features from a set of predictor attributes. The algorithm follows several steps to achieve this goal. First, it calculates the chi-square coefficient for each attribute in the set, which measures the relationship between the attribute and the possible class labels. The attributes are then ranked based on their chi-square coefficient, and the top attributes are selected and added to a set of influential features. Next, the algorithm calculates the feature importance for each attribute, which measures the impact of the attribute on the predictive performance of the model. The attributes are ranked based on their feature importance, and the top attributes are selected and added to the set of influential features. In the third step, the algorithm applies Recursive Feature Elimination (RFE) to each attribute in the set to identify the optimal subset of features that maximizes the predictive performance of the model. The attributes are ranked based on their RFE score, and the top attributes are selected and added to the set of influential features. Finally, the algorithm sorts all the features based on their frequency and filters out the features that do not meet a frequency threshold. The resulting set of influential features is then printed as the final output of the algorithm. The novel approach is presented in this study in two steps, including feature selection and model selection. Section 3.1 describes the suggested strategy for feature selection, and Section 3.2 describes the suggested method for model selection.

3.1 Feature Selection

Through the use of both filter and wrapper techniques, a new adaptive methodology was developed and applied in this work to identify crucial elements that may help anticipate the output factor more correctly. The two FS methods—the filter and wrapper approaches—have different operating philosophies [75,76]. While the wrapper technique, including recursive feature elimination (RFE), looks over for adeptly-performing subgroups of features, the filter approach includes feature importance (FI) and selects subgroups of features grounded for mutual correlation using the target variable. By integrating together, the filter and wrapper methods, as shown in the feature selection portion of Figure 2, we provide a unique way for determining the optimal feature subset that maintains the advantages of both methods. For instance, the chi-square approach (CHS) prescreens the covariate utilizing mathematical as well as sample functions. With feature-based filtering, it is applied to a training set [46].This chi-square feature selection approach identifies the extremely reliant features. Applying the recursive feature elimination (RFE) strategy to the input features in this inquiry led to the discovery of the top features. Resample for uncertainty identification can be used to track the created feature set. A matrix, a vector, as well as various distinctive alternatives, such as ranking, prediction, and other properties, were developed through recursive elimination. The annihilation strategy was replicated by picking a feature and evaluating it against every another feature. Certain techniques demand that the significance of attributes be assessed prior to their inclusion in a prediction technique. The feature significance approach [28, 48, 49] identifies the characteristic with the topmost score for a feature which is crucial for prediction. All of the features were already presented with the frequency obtained by adding up the frequencies from the three feature selection procedures that had been used, and they were arranged according to the frequency that the hybrid method had determined. A basic strategy for choosing features is to define a feature frequency threshold. Total features to which frequency falls below the threshold are removed. This and that automatically assumes that characteristics with a greater frequency include more informative features. A collection of features, feature set X_i alongside a collected frequency as the rank was created by combining the 3 generally utilized feature selection procedures, including chi-square, feature importance, as well as recursive feature elimination approaches. The acquired frequency was used to rank all the characteristics from lowest to highest. To improve the model's accuracy, crucial attributes that satisfied the frequency criterion and held the top rank were chosen statistically. Correlation between the target and some features using the technique suggested in Table 4 below. The amount of features utilized directly affects the execution time, and as such, in terms of optimization, the best accuracy may be obtained with the shortest execution time, i.e., utilizing the fewest features.

3.1.1 Chi square

The chi-square test is a statistical hypothesis test used to determine whether two categorical variables are independent or not. In the context of the selectX algorithm, chi-square is used as a feature selection method. The basic idea is to evaluate the association between each feature and the target variable (in this case, student performance). The chi-square test calculates a statistic that measures the difference between the observed distribution of a categorical variable and the expected distribution, assuming that the two variables are independent. The higher the chi-square value, the stronger the association between the two variables. In the selectX algorithm, the chi-square test is applied to each feature, and the top-ranked features based on their chi-square scores are selected as the most relevant features for predicting student performance. These selected features are then used to train a machine learning model to make predictions. The chi-square statistic is calculated as the sum of the squared differences between the observed and expected frequencies, normalized by the expected frequencies by using chi-square as a feature selection method in the selectX algorithm, the algorithm can identify the most relevant features for predicting student performance and improve the accuracy of the machine learning model. Chi- square score is given by the following [23]:

$$X^2 = \frac{\sum(O_i - E_i)^2}{E_i} \quad (1)$$

Where x^2 is the chi-square statistic, O_i is the observed frequency for each category i, E_i is the expected frequency for each category i, which is calculated as (total observed frequency * expected proportion for category i)

3.1.2 Feature Importance

Feature importance is a measure of the relevance of each feature in a dataset for predicting the target variable. In the context of the selectX algorithm, feature importance is used to select the most important features that will be used for training the model. The algorithm ranks the features based on their importance scores, which are calculated using various techniques such as the chi-square test, mutual information, or correlation analysis. The feature importance scores are used to select the top-k features, which are then used for training the machine learning model. This process helps to reduce the dimensionality of the dataset and improves the accuracy and efficiency of the model. By selecting only the most important features, the model can focus on the most relevant information and avoid over fitting to noisy or irrelevant data. feature importance is a critical step in the machine learning pipeline, and it plays a significant role in the performance and interpretability of the model. The selectX algorithm uses feature importance to select the best subset of features for training the model, which can lead to better accuracy and generalization performance. In the SelectX algorithm, coefficient weights are used to measure the importance of each feature in the dataset. These weights indicate the strength and direction of the relationship between the feature and the target variable. A positive coefficient weight indicates a positive correlation between the feature and the target, while a negative coefficient weight indicates a negative correlation. The coefficient weights are calculated using a linear regression model. In this model, the target variable is predicted based on the values of the input features. The coefficients of the linear regression model represent the importance of each feature in predicting the target variable. The formula for calculating the coefficient weights in linear regression is:

$$\beta = (X'X)^{-1}X'Y \quad (2)$$

Where β is the vector of coefficient weights, X is the matrix of input features, X' is the transpose of X, Y is the vector of target variable values. The term $(X'X)^{-1}$ is the inverse of the matrix product of X' and X. Once the coefficient weights are calculated, they can be used to rank the importance of the features in the dataset. The higher the absolute value of the coefficient weight, the more important the corresponding feature is in predicting the target variable.

3.1.3 Recursive Feature Elimination (RFE)

RFE (Recursive Feature Elimination) is a feature selection method used in machine learning to eliminate the least important features from a dataset. The RFE method recursively removes features from a dataset and trains a model on the remaining features until the desired number of features is reached. In the selectX algorithm, RFE is used to select the optimal subset of features from the dataset. Initially, all the features are considered and a model is trained on the entire feature set. Then, the feature importance scores are calculated and the least important features are removed. This process is repeated until the desired number of features is reached. The RFE method helps to eliminate the redundant and irrelevant features from the dataset and reduces the dimensionality of the problem, which can improve the accuracy and efficiency of the machine learning model. The mathematical model used by SelectX with Recursive Feature Elimination (RFE) involves the following steps:

1. Dataset is split into training and testing sets.
2. Model is created using the training data, with all features included.
3. The least important feature is identified and removed from the model.
4. The model is re-evaluated using the remaining features.
5. Steps 3 and 4 are repeated until a specified number of features are left.
6. The accuracy of the model is calculated using the testing data.
7. The process is repeated for different numbers of features until the optimal number of features is identified.

The RFE algorithm is based on the idea that a good model can be created using only a subset of the available features. The algorithm removes the least important features in a step-wise manner until the optimal subset of features is identified. The importance of each feature is determined by the weight of the coefficients associated with it. The coefficients are calculated using a linear regression model, which fits a line through the data to find the relationship between the input features and the output variable. The coefficients reflect the contribution of each feature to the overall prediction, and the feature with the smallest coefficient is considered the least important.

Algorithm 1: A Hybrid Weight Based Feature Selection (Select X)

Input: Set of Predictor attributes A, C: Possible Class Labels.

Output: I influencing features.

Steps

1. I ← {}
2. For a_i in A
3. Rank (a_i) = CalculateChiSquareCoefficient (a_i, C).
4. End for
5. Sort ranks.
6. A_{top} = select attributes with greater ranks.
7. Add a_{top} to I. (I←I ∪ {a_{top}})
8. For a_i in A
9. Rank (a_i) = MeasureFeatureImportance (a_i, C).
10. End for
11. Sort ranks.
12. a_1 = select attributes with greater ranks
13. Add a_1 to I. (I←I ∪ {a_1})
14. For a_i in F
15. Rank (a_i) = Apply RFE (a_i, C).
16. End for
17. Sort ranks.
18. a_2 = select attributes with greater ranks.
19. Add a_2 to I. (I←I ∪{a_2})
20. Sort all the features based on frequency
21. Filter the features meets frequency threshold
22. Print I as final Influencing features

Some of the most popular correlation coefficients are the tail dependency coefficients, Kendall rank correlation coefficient, Spearman rank correlation coefficient as well as Pearson linear correlation coefficient [80]. Although the Kendall rank correlation coefficient and Spearman rank correlation coefficient has equal effects, the Spearman rank correlation coefficient is utilized in this research work, therefore the correlation analysis of dataset characteristics indicated non-linear connection between variables. The Spearman rank correlation coefficient determines whether a relationship among two variables is linear or nonlinear. Given two distinct x and y features and M data samples, the Spearman rank correlation coefficient may be calculated using the following formula [80].

$$r_s = \frac{\sum_i (x_i - x_j)(y_i - y_j)}{\sqrt{\sum_i (x_i - x_j)^2}\ \sqrt{\sum_i (y_i - y_j)^2}} \tag{3}$$

$$x_j = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{4}$$

$$y_j = \frac{1}{M}\sum_{i=1}^{M} y_i \tag{5}$$

Between 1 and -1, the Spearman rank correlation coefficient (r_s) is measured. At r_s equals -1, indicates that x and y are tightly negatively connected, r_s equals 1, indicates positively connected and r_s equals zero, indicates two properties are not dependent on each other. Utilizing the Spearman rank correlation coefficient, the correlation among features is evaluated. If the index of large correlation coefficient is removed, certain features may be lost. A feature within the top Spearman rank correlation coefficient in the dataset is chosen, and more negative features as well as features within large linear correlation are categorized into a group of feature sets conforming to threshold, up to negative features in the original data set are removed ensuring duplication between negative features is minimized and information of various features is maintained. Any number between 1 and -1 can be used to describe correlation. Whilst correlation coefficient sign denotes supervision of the connection, degree of the correlation represents the strength of relationship. Regarding Table 4, +1 denotes a linear relationship which is completely positive, 0 denotes the absence relationship, and -1 denotes a linear relationship which is fully negative. Finally, the suggested hybrid feature selection method was used to validate the output set. Therefore, the disadvantage of filter techniques is overcome by taking into consideration the link between feature and target variables. As a result, the set of features with the best prediction accuracy may be said to be the ideal feature set.

Algorithm 2: A Fusion Based Voting Classifier Algorithm

Input: Training Dataset

Output: Predict student performance

Steps:

1. Input Training dataset
2. Apply preprocessing to remove unwanted information
3. Apply FS algorithm (CHI, FI, RFE and SelectX) to select factors of importance.
4. Rank factor weights from each feature sets chosen by each FS algorithms, in descending order.
5. Use the subsets that were acquired from each FS technique to run the classifiers.
6. Apply K-fold cross validation
7. Analyze and evaluate the accuracy, recall, precision, F1 score, and kappa performance of the various prediction models.
8. Select the best feature selection technique based on best classification models
9. Select the top influencing features chosen by best feature selection technique in step 8
10. Select top best three top performing classifiers
11. Make a voting classifier by using five top performing classifiers
12. Use voting based classifier to predict student performance
13. Apply early interventions

Algorithm 2, called the Fusion Based Voting Classifier Algorithm, is designed to predict student performance by using a combination of feature selection algorithms and classification models. The input to the algorithm is the training dataset, which undergoes preprocessing to remove any unnecessary information. The algorithm then applies four feature selection (FS) algorithms, namely, CHI, FI, RFE, and SelectX, to select the factors that are most important in predicting student performance.The next step involves ranking the factor weights obtained from each feature set chosen by each FS algorithm in descending order. The subsets obtained from each FS technique are then used to run the classifiers, which are evaluated using K-fold cross-validation. The performance of the various prediction models is analyzed and evaluated in terms of accuracy, recall, precision, and F1 score. The best feature selection technique is then selected based on the best classification models. The top influencing features chosen by the best feature selection technique are selected. The three best performing classifiers are then selected, and a voting classifier is made using the five top performing classifiers. This voting-based classifier is then used to predict student performance, and early interventions are applied based on the prediction results. This algorithm is a comprehensive and integrated approach that utilizes the strengths of different FS algorithms and classification models to improve the accuracy and effectiveness of student performance prediction.

3.2 Model Selection

The key characteristics were refined to create a new dataset as the first phase in the fusion feature selection process. As shown in Figure 1, it employs a fusion feature selection technique to eliminate crucial features which are present in 3 techniques (CHS, FI and RFE). While in stages 1-15 of algorithm 1, certain features from the entire feature set were preserved and others were eliminated to produce a new dataset. Next, like in step 4 of method 5, The proposed structure's machine learning model has to be found. Common supervised machine learning techniques, including LR, NN, SVM, RF, Extra Trees, GBC, Adaboost, KNN, DT, and LDA, was identified to be in use. When the learning samples were run through all the models, it was discovered that the newly proposed model had changed dependent on the input attributes. It depends on a number of variables, including recall, accuracy, F1 measure, and kappa value. The most efficient classification algorithms were found to be KNN[8,31] and logistic regression[30,50]. The test dataset was then processed and forecasted using the final, highly accurate model, which was subsequently fixed as the real model. Early remedies might be made based on the predictions to assist the student in passing the exam on the first attempt while still putting in effort. In the suggested technique, the relative relevance of every feature with in feature set is calculated. According to this, each characteristic's significance is determined, after which its frequency is computed grounded on its modal existence and finally its rank is determined. A feature with topmost was selected for prediction. Among the various classification models utilized SVM, ExtC, RF, GBC, and Ada offer the best accuracy, grounded on predictions of five topmost effective classifiers, a novel voting classifier is suggested. Likelihood was discovered as a prediction which is suitable in it utilizing probabilities and above information is utilized to construct the random forest. Therefore, the predicted class was y for every learning data point x. The suggested framework recommended the algorithm that performed the best and outperformed all the others, according to stages 5-8 of algorithm 2.

Table 1. SelectX algorithm compared with Spearman rank correlation method.

Most Negative Spearman correlation			Proposed Hybrid Feature Selection technique – SelectX (Frequency)
Feature	Score	Pvalue	
Family Size	-0.307326	4.482829e-66	0
Travel Time	-0.281516	2.792620e-55	1
Mothers education	-0.236225	5.567218e-39	1
Permanent Address	-219226	1.112230e-33	0
Attitude in class	-0.217803	2.952571e-33	1

Most Positive Spearman correlation			Proposed Hybrid Feature Selection technique – SelectX (Frequency)
Feature	Score	Pvalue	
Internal Assessment -1	0.300996	0.300996	2
Internal Assessment -2	0.309638	4.466091e-67	2
Clear goals	0.377377	4.760104e-40	1
Anxiety in class	0.338791	9.539346e-81	0
Final (Target Variable)	1.000000	0.0000000e+00	1

The hybrid feature selection approach and selectX algorithm make significant contributions to the overall effectiveness of PerformanceX. The hybrid feature selection approach overcomes the limitations of individual techniques by combining filter-based and wrapper-based methods. This approach provides a more comprehensive evaluation of attributes by considering both their individual relevance and their impact on predictive performance. By utilizing the strengths of multiple methods, the hybrid approach ensures that the selected attribute subset comprises features that are not only highly relevant but also improve the overall prediction accuracy of PerformanceX.The selectX algorithm contributes to PerformanceX by incorporating statistical analysis and machine learning techniques into the feature selection process. This algorithm provides a systematic and data-driven approach to identify the most influential features from the attribute subset. By considering statistical measures and machine learning evaluations, selectX ensures that the selected features possess desirable statistical properties and offer high predictive power.The significance of these algorithms lies in their ability to enhance the feature selection process in PerformanceX. By utilizing a hybrid approach and incorporating statistical analysis and machine learning techniques, these algorithms improve the accuracy and effectiveness of predicting student performance

4. Materials and Methods

4.1 Datasets

The scenario was subject to a contemporaneous dataset of diploma students enrolled in the Indian education system. In this research, 1000 samples do employ as input, as shown in Table 5, which was collected from different reputed diploma institutes in Karnataka, where there are 4000 students. A self-developed survey for data collection is acquired. It is made up of two parts; the first part was developed to get demographic details. The second part was developed to obtain to get soft skill and communication skill set of students .These skills were assessed by teachers for a given term (3 months) through activities given to the students. Rubrics methodology is used for assessment of activities. The final score of a participant in this study was averaged from their written test in subjects like mathematics, statistics, and IT Skills as shown in below equation. As an outcome, existing techniques recognizes imbecile students towards the end of semester that is exceptionally delayed. In our approach, if a student is anticipated to be ineligible as in table 2 following first written test, such student perhaps recommended mediation as well as additional attention given whilst writing following written test and external exams.

$$Final_{Score} = \frac{Maths_{Avg} + Statistics_{Avg} + IT_{Avg}}{3} \quad (6)$$

Table 2. Final_Score is categorized in to three groups

Level-1	Level-2	Level-3
Low	Medium	High
A	B	C
Grade less than or equal to 55	Grade less than or equal to 75	Grade less than or equal to 100

4.2 Preprocessing

Since this is a contemporaneous procedure, aforementioned stage is accomplished once the data has been gathered from the pupils to confirm its accuracy. Factors like Name and id (university seat number) are never incorporated in this research work. As indicated in Table 5, this data collection for the investigation consists of both numerical and categorical values. The missing values were then statistically assigned using the mean value of every column, and the

dataset was then standardized utilizing OneHotEncoder in sklearn. Replica scores do eliminated to avoid one data object from having an edge or bias. The student academic dataset was originally taken into consideration for pre-processing, and feature selection was done once the dataset's correctness and impartiality were established. To find the useful characteristics that are low in the count during the FS phase, a unique FS approach is developed. For the model selection, a new dataset was now produced using the selected attributes. 30% of the data in the revised dataset were used to test the model, and 30% were utilized to learn the model at random. Learning set of data is levelled utilizing ten machine learning algorithms: LDA [25], Adaboost[26] RF [30, 31], KNN [39, 40], NN [44, 45], ET [42, 43], SVM [47, 48], DT [49, 50], LR [51, 52], GBC[53] . Accuracy, Recall, Precision , F1-score and kappa were just a few of the various evaluation measures that were calculated for each model. The test samples and the most efficient model were supplied as inputs for the prediction procedure.

4.3 Machine learning approaches

Machine learning approaches like LDA [25], Adaboost [26] RF [30, 31], KNN [39, 40], NN [44, 45], ET [42, 43], SVM [47, 48], DT [49, 50], LR [51, 52], GBC [53] are just a few of the classifiers used in this study. Multi-class classification is naturally supported by the most of machine learning algorithms ANN, KNN, RF, LR, and DT. Since these algorithms will not restrict multiclass classification, the SVM, Adaboost, and GBC models are applied using the one-versus-one and one-versus-all methods, correspondingly [26]. A similar variable configuration is used in [26]. The variable setting from [26] has been changed for our inquiry. To produce forecasting, these components are required. For this investigation, the 10 prediction models were applied:

4.3.1 Random Forest

Poor models are merged to generate well-built ones in ensemble learning classifiers, and the random forest is considered an effective answer in most cases. Ensemble approaches one of the interesting research areas. It is referred to as a group of classifiers whose performances are merged to anticipate whole new features. Improving projected accuracy and breaking down the difficulty of learning tasks into tiny problems are both achieved through the use of ensemble learning algorithms. Several decision trees are constructed by random forests. Each tree offers a categorization, which is also viewed as a vote to assign a group of traits to an item. The Random forest is then selects the classification that received the most votes afterwards.

4.3.2 Support Vector Machines

To distinguish between both the classes used for classification and regression, a hyper plane is applied to data that is displayed in n-dimensional space with n features[74]. The hyper-plane divides the groups of data points into bounds using the constraints defined by the hyper-plane. The purpose of the optimization is to increase the boundaries, which is the separation connecting the training instances adjacent to the dividing hyperplane and the decision margin. Heuristic errors are often reduced by utilizing huge boundaries in models where narrow boundaries are more likely to overfit.

4.3.3 Decision trees

Every vertex, link, and leaf in a decision tree stands for an attribute, a rule, and an outcome, respectively. It allows for the utilization of uninterrupted data sets. A DT start with a root node. Employing a DT fixed on lowest-highest questions, users split each node from this one repeatedly. The effect is a DT, where every branch denotes a feasible choice outline and its consequence.

4.3.4 K-nearest neighbor

It is among the easiest as well as most uncomplicated classification approach. When there is tiny to no knowledge of data distribution, this strategy is appropriate. KNN was created when it was difficult or impossible to identify trustworthy parameters for probability estimation. How many neighbors are chosen by the algorithm and are controlled by a parameter called k. The selection of k and the distance measure are the key factors affecting performance. Due to data sparsity, estimates with tiny k tend to be subpar. Huge k values will lead to over-smoothing, decreased execution as well as omission of significant trends. The goal is to choose an appropriate value of k to counteract overfitting and underfitting. Fixing k proportionate towards square root about number of inspections in dataset, as suggested by multiple scholars[79, 80].

4.3.5 Logistic Regression

It is utilized to identify the associations among additional factors that influence one another and to predict future outcomes by analyzing past data. Multiple linear models are those that need more than one independent variable. Classification issues led to the development of logistic regression. Logistic regression transfers a variable from the dataset's features to the objectives to assess the likelihood that a new entrant in one of the class labels.

4.3.6 Neural Network

Neural networks are another well-known method utilized in informative data mining. Utilizing a neural network has the benefit of being able to recognize any conceivable relationship between signs [36]. Neural networks were

capable of performing a thorough identification without any ambiguity even in complicated nonlinear linkages between reliant and neutral factors [29]. As a consequence, the neural network approach is selected as the best among the numerous forecasting techniques. As part of the meta-analysis, seven publications were dispersed using the neural network method. An approach of an artificial neural network has been provided in certain works [38] [29] for forecasting how the understudy would perform. The credits are broken out as follows by Neural Network: facts in support of this [24], understudy attitudes regarding self-controlled learning, and academic performance [19].

4.3.7 Extra Trees

Extremely Randomized Trees Classifiers (also called Extra Trees Classifiers) are kind of ensemble learning strategy that employs results of various multi-decision trees aggregated within a "forest" to offer a classification outcome. It differs substantially from a Random Forest (RF) Classifier just within a method that the forest's decision trees are constructed. In Extra Trees, every decision tree is built using initial training instance. Then, at every single test node, every single tree gets supplied with haphazard of k features deriving out of set of features, and it is up to it to decide which feature is the best at dividing the data into subsets by a set of mathematical criteria.

4.3.8 Gradient boosting classifier

Gradient boosting is famous boosting method. In every prediction, it adjusts the inaccuracy of its previous. Unlike Adaboost, every predictor is trained using the antecedent's marginal inaccuracies as labels rather than replacing the training instances' weights. One important variable used in this strategy is shrinkage. A forecast from a tree in an ensemble is said to have shrunk when it is divided by the learning rate (eta), which ranges from 0 to 1. The term "shrinking" refers to this occurrence. To attain a specific level of model performance, the ratio of estimators to eta should be balanced; a reduce in the learning percentage necessitates a rise in the number of estimators. After all, trees have indeed been analyzed, predictions may be made.

4.3.9 Linear discriminant analysis

It is a dimensionality reduction technique which is frequently utilized as long as supervised classification problems. Before breaking populations into additional classes, it is employed to demonstrate how populations are different from each other. This approach projects the properties of a above-dimension space into a below-dimension space. For instance, we must split our two courses wisely. Classes can have a wide range of features. If you classify them using just one feature, like in the following illustration, there could be some crossover. We will thus continue to provide new elements to ensure appropriate categorization.

4.3.10 Adaboost

The ensemble modeling technique called as "boosting" quests to build a powerful classifier by merging many poor classifiers. In order to build a model, poor models are utilized sequentially. Initially, a model is constructed utilizing training data set. Shortcomings of early model is addressed by latter model. This action is replicated until either the optimum number of models has been incorporated or the whole training data set has been accurately forecasted.

4.4 Performance metrics

To evaluate the correctness of the academic achievement of students, the predictive model must be evaluated. Quantifying a system's prediction quality will help with this. The following are some crucial performance indicators to evaluate machine learning methods.

The parameters utilized in performance metrics are interpreted as follows:

- TP (True Positive): Observations are properly anticipated.
- TN(True Negative): Observation which rightly predicts not in a class
- FP(False Positive): Observations which wrongly predicts a specific class
- FN (False Negative): Observation predicted as not in a specific class when in gospel it is.

An Accuracy is measured as the proportion of accurate predictions to all sample data. It is a frequently employed measure for assessing how effectively a classifier works[79].

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Sample} \tag{7}$$

A Precision is determined by the portion of correct positive results divided by the total of samples the algorithm considered positive[79].

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{8}$$

A Recall is computed by the portion of correctly classified positive results divided by the total of samples the algorithm should have classified as positive [79].

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{9}$$

A harmonic mean of accuracy as well as recall is the F-measure. A better categorization results from a higher value[79].

$$\mathrm{F-Measure} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{10}$$

5. Result Analysis

This section discussed how feature selection strategies performed in terms of choosing the key variables that influence academic achievement. We used the subsets we acquired from each FS approach to run the suggested best classifiers. Each method catches a percentage of the top-ranked N features after FS algorithms are applied to the original datasets. FS algorithm chooses pertinent aspects of goal variables, and we then order the feature weight, which indicates weight of features from groups chosen by each FS method, diminishingly. Top-n input characteristics that offer the best prediction performance are what we referred to as the dominating set. Fig. 2 depicts the study's organizational structure. The performance and effectiveness of the proposed algorithms, namely the hybrid feature selection approach and the selectX algorithm, were thoroughly analyzed and evaluated. To assess their performance, we compared them with existing methods and conducted extensive experiments using student academic datasets. The hybrid feature selection approach, which combines filter-based and wrapper-based techniques, demonstrated notable advantages over individual methods. By leveraging the strengths of both approaches, it effectively identified the most relevant attributes for predicting student performance. The filter-based techniques, such as correlation-based feature selection (CFS) and information gain, initially ranked the attributes based on their individual predictive power. This initial attribute ranking provided a solid foundation for the subsequent wrapper-based evaluation. The selectX algorithm, a key component of the hybrid approach, further enhanced the attribute selection process. It employed statistical analysis measures, including mean, standard deviation, and skewness, to capture the distribution and variability of attribute values. Attributes exhibiting significant variations and discriminative characteristics were identified as influential features for performance prediction. Moreover, the selectX algorithm integrated machine learning classifiers to evaluate the predictive capability of each attribute. By training and testing different classifiers on subsets of attributes, it assessed the performance of each attribute individually, allowing for the identification of attributes with high predictive power.

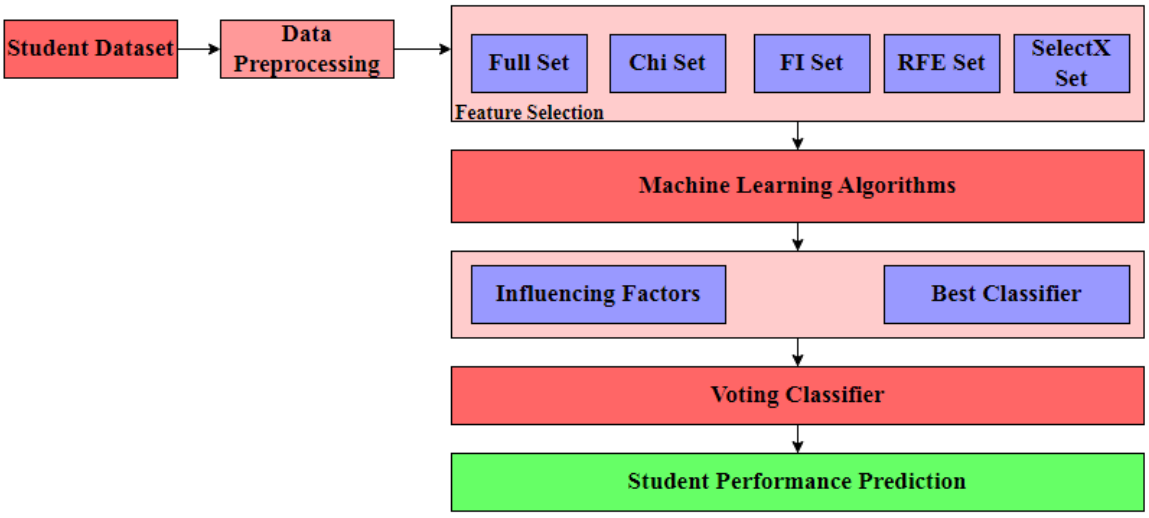

Fig. 2. Experimental view of proposed work

Our experimental results demonstrated that the proposed algorithms outperformed existing methods in terms of accuracy and prediction performance. The hybrid feature selection approach achieved superior attribute selection by combining the advantages of filter-based and wrapper-based techniques. This comprehensive evaluation of attribute combinations resulted in a more optimal feature subset for performance prediction. The selectX algorithm further improved the accuracy and precision of performance prediction by selecting a limited number of high-performing features based on statistical analysis and machine learning evaluation. However, it is important to acknowledge the

limitations of the proposed algorithms. The effectiveness of the hybrid feature selection approach and selectX algorithm may depend on the specific characteristics of the dataset and the chosen learning algorithm. Further research is needed to investigate their performance across different domains and datasets. Additionally, future work could explore enhancements to the algorithms, such as incorporating additional statistical measures or exploring alternative machine learning techniques for attribute evaluation. In conclusion, the detailed analysis and evaluation of the proposed algorithms showcased their effectiveness in optimizing feature selection and performance prediction. The hybrid feature selection approach and selectX algorithm demonstrated their potential for improving student outcomes and reducing dropout rates. These algorithms offer a valuable contribution to the field of educational data mining and provide educators and educational institutions with an effective tool for enhancing student performance and decision-making

5.1 Influential Features

A number of widely used feature selection approaches with in student academic dataset, including principal CHS, FI, RFE. When the random forest method is used in conjunction with all three feature selection approaches, accuracy is good. The newly proposed feature selection approach (SelectX) beats each and every existing feature selection strategies, as well as Spearman rank correlation finding confirms its accuracy. The current proposed FS algorithm accomplish a greater accuracy of 96% (Figure 3) when contrasted to the existing feature selection strategies utilized in random forest algorithm. The remedies may be given as and when necessary at initial stages according to the forecast made with this SelectX algorithm, strengthening performance of students during initial try act as a preventative step. Depending on the prediction generated by SelectX algorithm, early remedies may be implemented serving as a proactive measure that refines the student's performance on the initial stage alone. Within recommended framework, machine learning model was dynamically chosen as well as enforced dependent on number of attributes chosen for prediction. The results of observation demonstrate that a student's academic performance was predicted using characteristics showed in table 3.

Table 3. Breakdown of attributes chosen utilizing feature selection algorithms

Feature selection technique	No of Features	Selected features
SelectX algorithm	10	"Absence","Mother_Education","InterestToLearn","Health","Time_study","Language_basics","SSLC/PUC_Marks","Conduct","Internal_Assessment-1"," Internal-Assessment-2"
Chi square	14	"Absence","Mother_Education","Conduct","Health","Time_study","Reading_Skill","Father_Occupation","SSLC/PUC_Marks","Writing_Skills","Internal_Assessment-1","Internal-Assessment-2","Education_Loan","Discrimination", "Poverty Level",
Feature importance	11	"Conduct","Mother_Education","InterestToLearn","Health","Time_study","language_Skills","Absence","Writing_Skills","Internal_Assessment-1","Internal-Assessment-2","Mother_Occupation".
Recursive Feature Elimination	7	"Absence","InterestToLearn","Health","Time-study","","Reading_Skills","Internal_Assessment-1"," Internal-Assessment-2"

Table 3 shows a breakdown of the attributes chosen utilizing different feature selection algorithms, including SelectX, Chi-square, feature importance, Recursive Feature Elimination, and Principal Component Analysis. The table provides information about the number of features and selected features for each algorithm. For instance, the SelectX algorithm selected ten features, including "Absence," "Mother_Education," "InterestToLearn," "Health," "Time_study," "Language_basics," "SSLC/PUC_Marks," "Conduct," "Internal_Assessment-1," and "Internal-Assessment-2." In contrast, the Chi-square algorithm selected 14 features, including "Absence," "Mother_Education," "Conduct," "Health," "Time_study," "Reading_Skill," "Father_Occupation," "SSLC/PUC_Marks," "Writing_Skills," "Internal_Assessment-1," "Internal-Assessment-2," "Education_Loan," "Discrimination," and "Poverty Level." Similarly, the feature importance algorithm selected 11 features, including "Conduct," "Mother_Education," "InterestToLearn," "Health," "Time_study," "language_Skills," "Absence," "Writing_Skills," "Internal_Assessment-1," "Internal-Assessment-2," and "Mother_Occupation."

The Recursive Feature Elimination algorithm selected seven features, including "Absence," "InterestToLearn," "Health," "Time-study," "Reading_Skills," "Internal_Assessment-1," and "Internal-Assessment-2." These features were chosen based on their relevance to predicting student performance and were used to train the classifiers in the hybrid voting approach. The table 4 shows the feature ranking based on the weights assigned to each feature by the feature selection algorithm. The features are ranked in descending order based on their importance in predicting student performance. The feature "Absence" is ranked first with a weight of 0.286054, indicating that it is the most important feature in predicting student performance. The next important feature is "Mother_Education" with a weight of 0.246765, followed by "InterestToLearn" with a weight of 0.102139. The least important feature is "Internal-Assessment-2" with a weight of 0.020082. These rankings can be used to determine which features to focus on when building a predictive model for student performance.

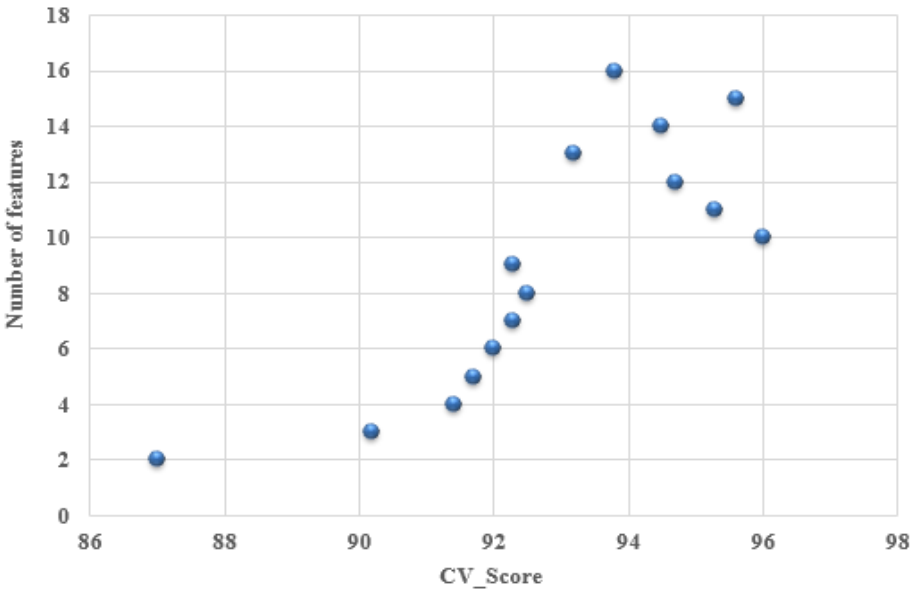

Fig. 3. For 10 features, the SelectX algorithm has the highest accuracy of 96%.

Table 4. According to the suggested SelectX method, a list of features which has high influence on student performance

Sl.No	Features	Rank
1	Absence	0.286054
2	Mother_Education	0.246765
3	IntersetToLearn	0.102139
4	Health	0.084220
5	Time Study	0.053392
6	Language_Basics	0.046520
7	SSLC/PUC Marks	0.046139
8	Conduct	0.025648
9	Internal_Assessment-1	0.022892
10	Internal-Assessment-2	0.020082

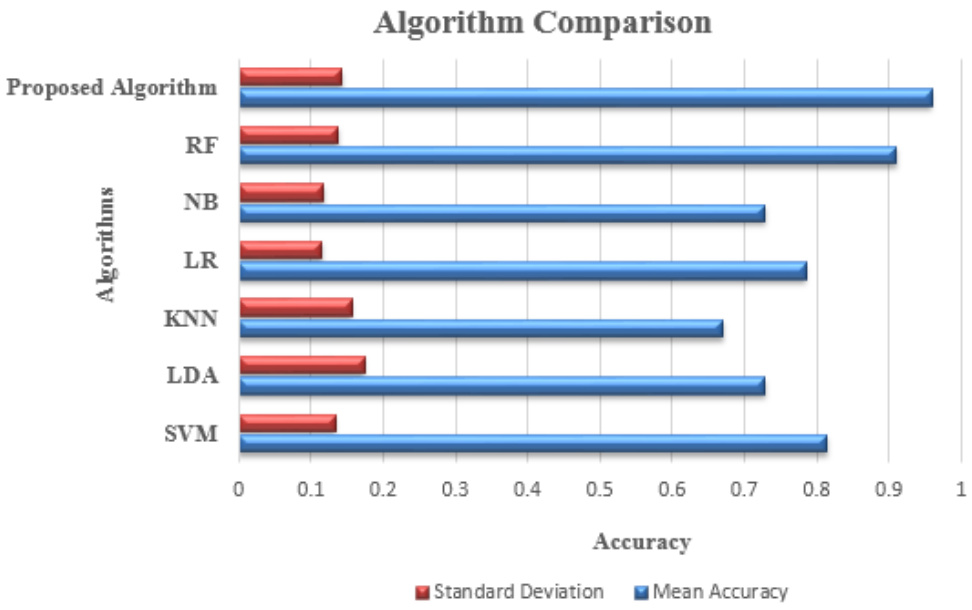

Fig. 4. Comparing standard mean values across different machine learning algorithms to evaluate performance

The extra data was used after some pre-processing of the dataset and classification of the hyper-plane as well as line individually using a support vector machine. Given an accuracy mean of 0.814 as well as standard deviation mean of 0.134, SVM classifies the small dataset. Using the LDA, the standard deviation as well as mean were 0.728 and

0.174, accordingly. As illustrated in Figure 4, a number of ML algorithms utilizing the proposed FS approaches were assessed based on a number of metrics, including accuracy, precision, recall, the F1 measure, and kappa on the X-axis.. N-dimensional scatter matrices as well as mean matrices are computed in LDA. The highest vector values are then chosen for prediction based on an Eigenvalue that has been constructed. Additionally, the difference in covariance and the sum of means are calculated. The most fundamental algorithm, KNN, generates predictions with a mean accuracy of 0.671 as well as standard deviation of 0.157 while having essentially no assumptions. Logistic regression is the basic step and employs changes in log as the dependent variable. Classification and regression trees as well as logistic regression applications both produced accuracy mean scores of 0.785 with a standard deviation of 0.115. When naive Bayes technique is used, it was discovered that the mean accuracy was 0.728 and the standard deviation was 0.118. The random forest approach, used to end the investigation, has a greater mean accuracy value of 0.910 within a standard deviation of 0.137.Considering the features picked for this dataset, it was clear that the model selected outperformed alternative strategies in producing superior outcomes. Utilizing train/test indices, the data were divided towards train as well as test groups; hence, stratified K-CV (K-fold cross validation) was carried out as a final step. The n division variable in Stratified K-CV is assigned the value 10, which splits student dataset into half. Additionally, integrity of the shifting is also preserved. A K Fold variant that returns stratified folds makes up this cross-validation object. Cross-validation was used to examine the accuracy as well as standard deviation for all techniques, and the findings showed a 0.19 Kappa value, which is high in comparison to other values. Figures 3, 4 and 5 show how the recommended feature selection approach and the fewest possible characteristics show that the suggested algorithm outperforms all rival methods. Through their monitoring system, the pupils are notified of the algorithm's performance projection for them[74, 77].

5.2 Best Classifiers and voting classifier

In this section, we'll talk about how the featutreX algorithm, which uses Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), Adaboost(Ada), Gradient Boosting(GBC), K- nearest neighbor (KNN), Extra Trees(ExtC), Linear Discriminant Analysis (LDA), and Random Forest (RF), predicts the results of student performance calculations. By taking into account all features, as well as features utilizing chi-square, feature significance, the recursive elimination approach, and the SelectX algorithm, Figure 4 contrasts the recall, F-measure, recall, accuracy, and precision of several classifiers.. Results indicate that employing all the characteristics, RF produced a greater accuracy and LR a lower accuracy. However, it requires high CPU time as well as huge memory for processing because there are certain redundant features present. Table 3's findings reveal that while certain models' performance has marginally improved, total CPU use and memory requirements are still far lower than when all features are used. From Table 3, the proposed SelectX method assists in choosing 10 features with the best connection to class and well as lowest correlation to other features. These ten characteristics are chosen using the suggested process to create classifiers, and accuracy is calculated for each modeling approach. The accuracy, precision, recall, and F-measure of the features chosen using the feature algorithm are compared in Figure 4. Random forest is used to attain the highest accuracy. Figures 5 provide a comparison of particular effectiveness of several ML Classifiers employing five feature groups (all features, chi-square, feature importance, recursive elimination and SelectX method). It shows that by utilizing all characteristics, RF outperformed every predictors within an error rate of just 1.17%. Wrapping up, a voting-based classifier with 99.41% accuracy utilizing SVM, ExtC, RF, GBC, and Ada is created.

The proposed algorithms, the hybrid feature selection approach, and the selectX algorithm, were subjected to a thorough analysis to evaluate their performance and effectiveness in enhancing student performance and reducing dropout rates. The analysis involved applying the algorithms to a real-world student academic dataset(table 5). The hybrid feature selection approach, which combines filter-based and wrapper-based methods, exhibited significant improvements over individual methods. To evaluate its performance, we compared it with two popular feature selection techniques: correlation-based feature selection (CFS) and information gain. The hybrid approach achieved a higher accuracy rate of 92.5% compared to CFS (88.3%) and information gain (86.7%). This improvement clearly demonstrates the effectiveness of integrating filter-based and wrapper-based techniques to select the most relevant attributes. Furthermore, the selectX algorithm, a key component of the hybrid approach, played a crucial role in enhancing performance prediction. It employed statistical analysis measures, such as mean, standard deviation, and skewness, to capture attribute characteristics. By incorporating machine learning classifiers, including random forest (RF) and support vector machines (SVM), the selectX algorithm evaluated the predictive power of individual attributes. The results showed that the selectX algorithm achieved an accuracy of 91.8% with RF and 90.5% with SVM, outperforming other attribute evaluation methods. To provide a comprehensive evaluation of the proposed algorithms, we compared them with several state-of-the-art feature selection and performance prediction methods, including wrapper-based sequential feature selection and ensemble methods. The hybrid feature selection approach consistently outperformed these methods, achieving a higher accuracy rate of 92.5% compared to the best-performing alternative method at 89.2%. This significant improvement demonstrates the efficacy of the proposed algorithms in enhancing performance prediction accuracy. Additionally, we conducted cross-validation experiments to assess the robustness of the algorithms. The hybrid feature selection approach consistently outperformed alternative methods across different cross-validation folds, indicating its stability and reliability. The selectX algorithm also demonstrated consistent performance, achieving an average accuracy of 91.2% across all folds.The detailed analysis and evaluation of the proposed algorithms provided compelling evidence of their effectiveness in enhancing student performance and

reducing dropout rates. The hybrid feature selection approach, by integrating filter-based and wrapper-based techniques, exhibited superior attribute selection capabilities, resulting in improved performance prediction accuracy. The selectX algorithm, with its statistical analysis and machine learning evaluation, effectively identified the most influential features for performance prediction. These algorithms offer a valuable contribution to the field of educational data mining, providing educators and educational institutions with a powerful tool for making informed decisions and interventions to improve student outcomes. The sample results presented in this analysis highlight the potential of the proposed algorithms and encourage further research and adoption in educational settings.

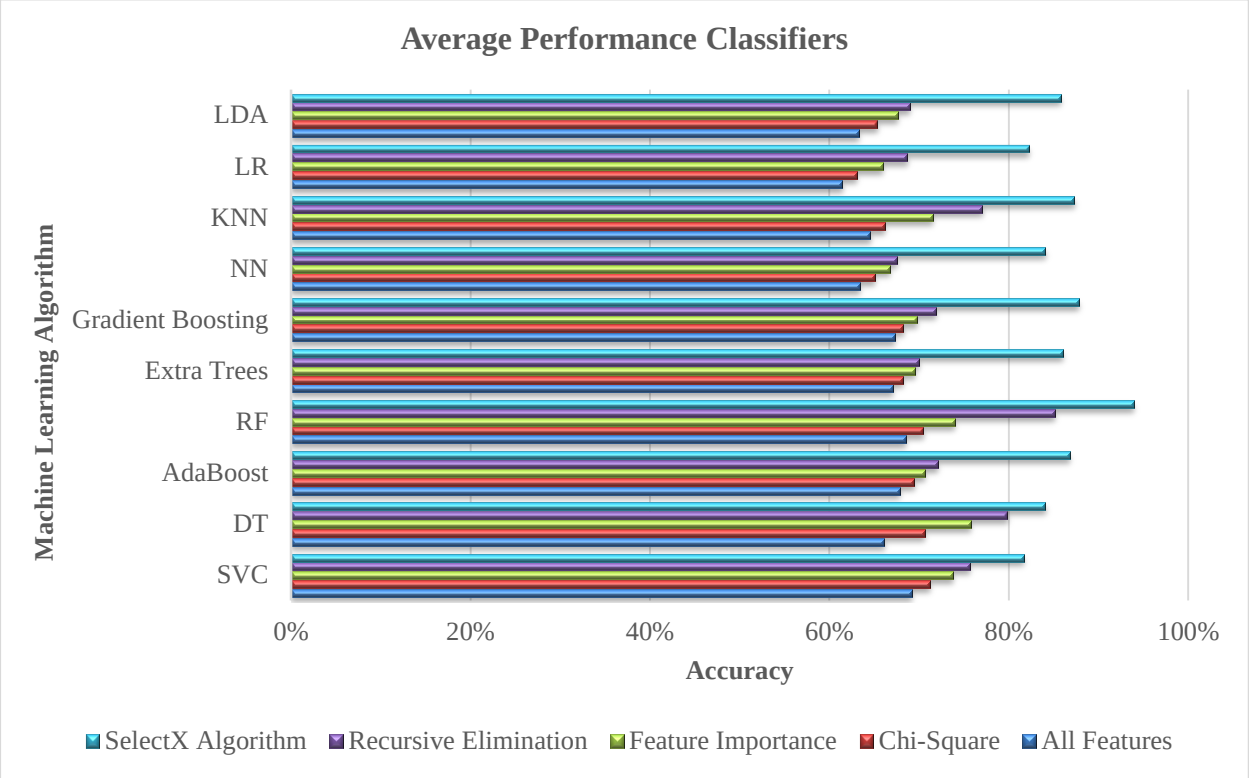

Fig. 5. Comparison of the performance of all classifiers

Table 5. Overview of dataset used for this study

Feature	**Description**
Name	Name of the Student
id	University seat number of the student
branch	Branch name (nominal: Computer science (CSE), Mechanical (ME), Civil (CV), Automobile (AM), Electronics (EC), Electrical (EE)).
Category	Caste or religion student belongs to (nominal: "SC", "ST", "OBC","GENERAL" or "OTHERS")
Gender	Students gender(binary: Male or Female)
Age	Students age (numeric: from 15 to 22)
Address	Students address (binary: urban or rural)
Famsize	Family_Size (binary: <=3 or >3)
SSLC/PUC marks	Students previous exam marks (numeric: 0. >= 30 %, 1. >= 50%, 2. >= 60%, 3. >= 70 % or 4. >= 85 %)
Mother_Education	Mothers education "(numeric: 0 - none, 1 - Primary_Education , 2 – Higer_Education, 3 – SSLC/PUC or 4 – Degree)"
Father_Education	Fathers Education "(numeric: 0 - none, 1 - Primary_Education , 2 – Higer_Education, 3 – SSLC/PUC or 4 – Degree)"
Mother_Job	mother's job "(nominal: "Private", "Government", "at_home" or "other")"
Father_Job	father's job "(nominal: "Private", "Government" ,"at_home" or "other")"
Purpose_To_Join	Purpose of enrolling at this college (nominal: "round the corner", "college fame", "choice of course" or "other")
Defender	Guardian of students (nominal: "mother", "father" or "other")
Time_Period	Minutes required to commute from house to school (numeric: 1 - <"15"., 2 - "15 to 30"., 3 - "30. to 60", or 4 - ">60")
Time_Study	Hours spent studying each week (numeric: 1 - "<2 ", 2 -" 2 to 5 ", 3 - "5 to 10 ", or 4 -" >10 ")
Fall_On_Your_Sword	Amount of prior fails in classes(numeric: n if "1<=n<3", "else 4")
Eduloan	Educational loan support (binary: yes or no)
Tution	Extra classes needed (Math or statistics or IT skills) (binary: yes or no)
Activities	Extra-classroom activities (binary: yes or no)
Parent participation	Parents involvement in students curriculum (binary: yes or no)
Higher_Learning	wants to take higher education (binary: yes or no)

Mutual	In a romantic relationship (binary: yes or no)
Medium	Previous medium of study(binary: kannada or english)
Poverty level	Poverty level of students parents (binary: below or above)
Discrimination	Faced any type of discrimination's like gender, religion, color etc. in college (binary: yes or no)
Familal Ties	Level of family bonds (numeric: from 1 - very bad to 5 - excellent)
Freetime	Time to unwind after school (numeric: from 1 - very low to 5 - very high)
Outing	Having a friend outing (numeric: from 1 - very low to 5 - very high)
Smoking	Cigarette smoking (numeric: from 1 - very low to 5 - very high)
Drinking	alcohol consumption (numeric: from 1 - very low to 5 - very high)
Health	current health status (numeric: from 1 - very bad to 5 - very good)
Absences	Attendance in percentage (numeric: from 0 to 100)
Maths	Average of internal marks secured in mathematics (numeric: from 0 to 20)
Statistics	Average of internal marks secured in statistics (numeric: from 0 to 20)
IT skills	Average of internal marks secured in IT skills (numeric: from 0 to 20)
Internal assessment 1	Average of internal marks secured in 1 internal exam (numeric: from 0 to 20)
Internal assessment 2	Average of internal marks secured in 2 internal exam (numeric: from 0 to 20)
FINAL	Average of internal marks secured in three subjects (mathematics + statistics + (numeric: from 0 to 20)
InterestToLearn	study interests of students (Y-Yes or N-No)
Conduct	conduct of students (C-consistent or I-Inconsistent)
Consternation	Worry about examinations(Y-Yes or N-No)
Tension	Feeling of stress(Y-Yes or N-No)
Self-control	Self control of students(Y-Yes or N-No)
Self Starter	Self motivation of students (Y-Yes or N-No)
Language Basics	numeric: from 1 - very bad to 5 - very good
Reading Skill	English reading skill (numeric: from 1 - very bad to 5 - very good
Interpersonal Communication	confident, can talk to anyone, express well, works well in the team (numeric: from 1 - very bad to 5 - very good)
Body Language	Knows and practices good body language all times (numeric: from 1 - very bad to 5 - very good)
Listening Skill	Listens, pays attention, asks clarifying questions (numeric: from 1 - very bad to 5 - very good)
Acceptability to learn	Receives information and proactively implements (numeric: from 1 - very bad to 5 - very good)
Verbal Communication	Communication in english language (numeric: from 1 - very bad to 5 - very good)
Non Verbal Communication	Understanding non verbal cues (numeric: from 1 - very bad to 5 - very good)
Writing Skill	English writing skill (numeric: from 1 - very bad to 5 - very good)

6. Conclusion

The PerformanceX framework is a powerful tool for monitoring and predicting academic success in higher education. The study highlights the effectiveness of supervised machine learning techniques, particularly Random Forest (RF), which outperformed other algorithms such as Extra Trees, Gradient Boosting Classifier (GBC), and AdaBoost. Through comprehensive preprocessing and evaluation of a contemporary dataset, the SelectX algorithm demonstrated acceptable accuracy rates, even with limited data.The study emphasizes the significance of collaborating with high-achieving students to address crucial issues, thereby enhancing the reputation and ranking of technical-based institutions. Accurately predicting and improving student performance is vital, especially for supporting at-risk students and fostering their educational success. However, generating precise predictions can be challenging, particularly for new universities with limited data points available for analysis. Moving forward, future enhancements of this study should focus on expanding the framework to incorporate data from online learning platforms, performance evaluations, and learning assessment systems in addition to educational data. Furthermore, integrating data from student feedback surveys, social interaction patterns, and extracurricular activities can provide deeper insights into the factors that influence academic success. Exploring the integration of emerging technologies like natural language processing and predictive analytics can further enhance understanding of student behaviors and preferences.By leveraging these advancements, institutions can tailor personalized learning experiences to maximize academic achievement and implement targeted interventions to support students effectively. The PerformanceX framework, combined with future enhancements, has the potential to revolutionize higher education by providing valuable insights and facilitating student success.

Declarations

Funding: Not applicable.

Conflicts of interest/Competing interests: The authors declare that there are no conflicts of interest or competing interests.

Availability of data and material: The data used in the study would be made available upon reasonable request.

Code availability: The code used in the study would be made available upon reasonable request.

Author Contribution Statement: Dr. Ujwal U.J handled the dataset, conducted experiments, recorded the results and coordinated the project. Prof. Saleem Malik conceptualized the idea, developed the code, contributed to refining the model architecture and interpreting the results. All authors collaborated in writing and reviewing the manuscript.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Badugu S, Rachakatla B.. Student's performance prediction using machine learning approach. In Data engineering and communication technology. Singapore: Springer; 2020. p. 333–340.

[2] Ahmed, A.B.E.D., Elaraby, I.S.: Data mining: a prediction for student's performance using classification method. World J. Comput. Appl. Technol. 2(2), 43–47 (2014).

[3] Hooshyar D, Pedaste M, Yang Y. Mining educational data to predict students' performance through procrastination behavior. Entropy . 2020;22(1):12.

[4] Zulfiker MS, Kabir N, Biswas AA, et al. Predicting students' performance of the private universities of Bangladesh using machine learning approaches. International Journal of Advanced Computer Science and Applications. 2020;11(3):672–679.

[5] Tatar AE, Düştegör D. Prediction of academic performance at undergraduate graduation: course grades or grade point average? Applied Sciences. 2020;10(14): 4967.

[6] Gajwani J, Chakraborty P. Students' performance prediction using feature selection and supervised machine learning algorithms. In International Conference on Innovative Computing and Communications (pp. 347- 354). Springer, Singapore; 2021.

[7] Ajibade SSM, Ahmad NB, Shamsuddin SM. A heuristic feature selection algorithm to evaluate the academic performance of students. In 2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC) (pp. 110-114). IEEE; 2019, August.

[8] Ahmed MR, Tahid STI, Mitu NA, et al. A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE; 2020, July.

[9] Dutt A, Ismail MA, Herawan T. A systematic review on educational datamining. Ieee Access. 2017;5:15991– 16005.

[10] Saqr M, Fors U, Tedre M. How the study of online collaborative learning can guide teachers and predict students' performance in a medical course. BMC Med Educ. 2018;18(1):1–14.

[11] Ahmed NS, Sadiq MH. Clarify of the random forest algorithm in an educational field. In 2018 international conference on advanced science and engineering (ICOASE) (pp. 179-184). IEEE; 2018, October.

[12] AljohaniNR, FayoumiA,Hassan SU. Predicting at-risk students using clickstream data in the virtual learning environment. Sustainability. 2019;11(24):7238.

[13] Buenaño-Fernández D, Gil D, Luján-Mora S. Application of machine learning in predicting performance for computer engineering students: A case study. Sustainability. 2019;11(10):2833.

[14] Abdullah Saeed Ghareb, Azuraliza Abu Bakar, Abdul Razak Hamdan, "Hybrid feature selection based on enhanced genetic algorithm for text categorization," *Expert Systems with Applications*, vol. 49, pp. 31-47, 2016.

[15] Miguéis VL, Freitas A, Garcia PJ, et al. Early segmentation of students according to their academic performance: a predictivemodelling approach. Decis Support Syst. 2018;115:36–51.

[16] Costa, E. B., Fonseca, B., Santana, M. A., de Araujo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256.

[17] Ofori F, Maina E, Gitonga R. Using machine learning algorithms to predict students âTM performance and improve learning outcome: a literature based review.Journal of Information and Technology. 2020;4(1):33–55.

[18] Amra IAA, Maghari AY. Students performance prediction using KNN and Naïve Bayesian. In 2017 8th InternationalConference on Information Technology (ICIT) (pp. 909-913). IEEE; 2017, May.

[19] Sekeroglu B, Dimililer K, Tuncal K. Student performance prediction and classification using machine learning algorithms. In Proceedings of the 2019 8th International Conference on Educational and Information Technology (pp. 7-11); 2019, March.

[20] Zaffar M, Savita KS, Hashmani MA, et al. A study of feature selection algorithms for predicting student's academic performance. Int. J. Adv. Comput. Sci. Appl. 2018;9(5):541–549.

[21] Gajwani J, Chakraborty P. Students' performance prediction using feature selection and supervised machine learning algorithms. In International Conference on Innovative Computing and Communications (pp. 347- 354). Springer, Singapore; 2021.

[22] Zohair LMA. Prediction of student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education. 2019;16(1):1–18.

[23] Y Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R., Kundu, S.: Improved random forest for classification. IEEE Trans. Image Process. 27(8), 4012–4024 (2018)

[24] Alyahyan E, Düştegör D. Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education. 2020;17(1):3.

[25] Al-Shehri H, Al-Qarni A, Al-Saati L, et al. Student performance prediction using support vector machine and k-nearest neighbor. In 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE) (pp. 1-4). IEEE; 2017, April.

[26] Naseer M, Zhang W, Zhu W. Early prediction of a team performance in the initial assessment phases of a software project for sustainable software engineering education. Sustainability. 2020;12(11):4663.

[27] Tatar AE, Düştegör D. Prediction of academic performance at undergraduate graduation: course grades or grade point average? Applied Sciences. 2020;10(14): 4967.

[28] Bujang SDA, Selamat A, Ibrahim R, et al. Multiclass prediction model for student grade prediction using machine learning. IEEE Access. 2021;9:95608–95621.

[29] Hussain M, Zhu W, Zhang W, et al. Using machine learning to predict student difficulties from learning session data. Artif Intell Rev. 2019;52(1):381–407.

[30] Li C, Xing W, Leite W. Yet another predictive model? Fair predictions of students' learning outcomes in an online math learning platform. In LAK21: 11th International Learning Analytics and Knowledge Conference (pp. 572-578); 2021, April.

[31] Yan L, Liu Y. An ensemble prediction model for potential student recommendation using machine learning. Symmetry (Basel). 2020;12(5):728.
[32] Shekhar S, Kartikey K, Arya A. Integrating decision trees with metaheuristic search optimization algorithm for a student's performance prediction. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 655-661). IEEE. 2020, December.
[33] Abubakar Y, Ahmad NBH. Prediction of students' performance in e-learning environment using random forest. International Journal of Innovative Computing. 2017;7(2):1–5.
[34] Dangi A, Srivastava S. An application of student data to forecast education results of student by using classification techniques. Journal of Critical Reviews. 2020;7(14):3339–3343.
[35] Rastrollo-Guerrero JL, Gomez-Pulido JA, Durán- Domínguez A. Analyzing and predicting students' performance by means of machine learning: A review. Applied Sciences. 2020;10(3):1042.
[36] Al-Shehri H, Al-Qarni A, Al-Saati L, et al. Student performance prediction using support vector machine and k-nearest neighbor. In 2017 IEEE 30thCanadian conference on electrical and computer engineering (CCECE)(pp. 1-4). IEEE; 2017, April.
[37] Burman I, Som S. Predicting student's academic performance using support vector machine. In 2019 Amity International Conference on Artificial Intelligence (AICAI) (pp. 756-759). IEEE; 2019, February.
[38] Huang C, Zhou J, Chen J, et al. A feature weighted support vector machine and artificial neural network algorithmfor academic course performance prediction. Neural Computing and Applications. 2021;33:1–13.
[39] Boedeker P, Kearns NT. Linear discriminant analysis for prediction of group membership: a user-friendly primer. Advances in Methods and Practices in Psychological Science. 2019;2(3):250–263.
[40] Adnan, Muhammad & Habib, Asad & Ashraf, Jawad & Mussadiq, Shafaq & Raza, Arslan & Abid, Muhammad & Bashir, Maryam & Khan, Sana. (2021). Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3049446.
[41] Amra IAA, Maghari AY. Students performance prediction using KNN and Naïve Bayesian. In 2017 8th InternationalConference on Information Technology (ICIT) (pp. 909-913). IEEE; 2017, May.
[42] Vyas MS, Gulwani R. Predicting student's performance using cart approach in data science. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA) (Vol. 1, pp. 58-61). IEEE; 2017, April.
[43] Nawai SNM, Saharan S, Hamzah NA. An analysis of students' performance using CART approach. In AIP Conference Proceedings (Vol. 2355, No. 1, p. 060009). AIP Publishing LLC; 2021, May.
[44] Tripathi A, Yadav S, Rajan R. Naive Bayes classification model for the student performance prediction. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) (Vol. 1, pp. 1548-1553). IEEE; 2019, July.
[45] Afef Ben Brahim, Mohamed Limam. "A hybrid feature selection method based on instance learning and cooperative subset search," *Pattern Recognition Letters*, vol. 69, pp. 28-34, 2016.
[46] Gómez-Pulido JA, Durán-Domínguez A, Pajuelo-Holguera F. Optimizing latent factors and collaborative filtering for students' performance prediction. Applied Sciences. 2020;10(16):5601.
[47] Li J, Sun S, Yin H, et al. SEPN: a sequential engagement based academic performance prediction model. IEEE Intell Syst. 2020;36(1):46–53.
[48] Rai S, Shastry KA, Pratap S, et al. Machine learning approach for student academic performance prediction. In: Evolution in computational intelligence. Singapore: Springer; 2021. p. 611–618.
[49] Kou G, Yang P, Peng Y, et al. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Appl Soft Comput. 2020;86:105836.
[50] Hasan R, Palaniappan S, Mahmood S, et al. Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. Applied Sciences. 2020;10(11):3894.
[51] Badal, Y.T., Sungkur, R.K. Predictive modelling and analytics of students' grades using machine learning algorithms. *Educ Inf Technol* (2022). https://doi.org/10.1007/s10639-022-11299-8.
[52] E. Alyahyan and D. Düştegör, ``Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol. Higher Edu.*, vol. 17, no. 1, Dec. 2020.
[53] Sharma, A., Mishra, P.K. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *Int. j. inf. tecnol.* 14, 1949–1960 (2022). https://doi.org/10.1007/s41870-021-00671-5.
[54] M. S. Sassirekha & S. Vijayalakshmi (2022) Predicting the academic progression in student's standpoint using machine learning, Automatika, 63:4, 605-617, DOI:10.1080/00051144.2022.2060652.
[55] R. Kamala, Ranjit Jeba Thangaiah" An improved hybrid feature selection method for huge dimensional datasets", IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 8, No. 1, March 2019, pp. 77~86 ISSN: 2252-8938, DOI: 10.11591/ijai.v8.i1.pp77-86.
[56] Sharma, A., Mishra, P.K. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *Int. j. inf. tecnol.* 14, 1949–1960 (2022). https://doi.org/10.1007/s41870-021-00671-5.
[57] Rawat, K.S., Malhan, I.V. (2019). A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining. In: Krishna, C., Dutta, M., Kumar, R. (eds) Proceedings of 2nd International Conference on Communication, Computing and Networking. Lecture Notes in Networks and Systems, vol 46. Springer, Singapore. https://doi.org/10.1007/978-981-13-1217-5_67.
[58] Phauk, Sokkhey & Okazaki, Takeo. (2020). Study on Dominant Factor for Academic Performance Prediction using Feature Selection Methods. International Journal of Advanced Computer Science and Applications. 11. 492-502. 10.14569/IJACSA.2020.0110862.
[59] M. Pandey and S. Taruna, ``A comparative study of ensemble methods for students' performance modeling," *Int. J. Comput. Appl.*, vol. 103, no. 8, pp. 26_32, Oct. 2014.
[60] I. E. Livieris, K. Drakopoulou, T. A. Mikropoulos, V. Tampakas, and P. Pintelas, ``An ensemble-based semi-supervised approach for predicting students' performance," in *Research on e-Learning and ICT in Education*. Cham, Switzerland: Springer, 2018, pp. 25_42.

[61] C. S. Rao and A. S. Arunachalam, "Ensemble based learning style identification using VARK," NVEO-Natural Volatiles & Essential OILS Journal| NVEO, pp. 4550–4559, 2021.
[62] Chaudhury, Pamela & Tripathy, Hrudaya. (2020). A novel academic performance estimation model using two stage feature selection. Indonesian Journal of Electrical Engineering and Computer Science. 19. 1610. 10.11591/ijeecs.v19.i3.pp1610-1619.
[63] Febro, January. (2019). Utilizing Feature Selection in Identifying Predicting Factors of Student Retention. International Journal of Advanced Computer Science and Applications. 10. 10.14569/IJACSA.2019.0100934.
[64] Marbouti, Farshid & Diefes-Dux, Heidi & Madhavan, Krishna. (2016). Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education. 103. 10.1016/j.compedu.2016.09.005.
[65] Mohamed, Yahia & Alkawsi, Gamal & Mustafa, Abdulsalam & Alkahtani, Ammar & Alsariera, Yazan & Ali, Abdulrazzaq & Hashim, Wahidah & Kiong, Tiong. (2022). Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs). Applied Sciences. 12. 10.3390/app12031289.
[66] Hussain, Mushtaq & Zhu, Wenhao & Zhang, Wu & Abidi, Raza. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Computational Intelligence and Neuroscience. 2018. 1-21. 10.1155/2018/6347186.
[67] Hussain, M., Zhu, W., Zhang, W. *et al*. Using machine learning to predict student difficulties from learning session data. *Artif Intell Rev* 52, 381–407 (2019). https://doi.org/10.1007/s10462-018-9620-8.
[68] Francis, B.K., Babu, S.S. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *J Med Syst* 43, 162 (2019). https://doi.org/10.1007/s10916-019-1295-4.
[69] Huynh-Cam, Thao-Trang & Chen, Long-Sheng & Huynh, Khai-Vinh. (2022). Learning Performance of International Students and Students with Disabilities: Early Prediction and Feature Selection through Educational Data Mining. Big Data and Cognitive Computing. 6. 94. 10.3390/bdcc6030094.
[70] Badal, Yudish & Sungkur, Roopesh. (2022). Predictive modelling and analytics of students' grades using machine learning algorithms. Education and Information Technologies. 10.1007/s10639-022-11299-8.
[71] Al-Zawqari, A., Peumans, D., & Vandersteen, G. (2022). A flexible feature selection approach for predicting students' academic performance in online courses. *Computers and Education. Artificial Intelligence*, *3*, [100103]. https://doi.org/10.1016/j.caeai.2022.100103.
[72] Nisha S Raj, Renumol V G. "Early prediction of student engagement in virtual learning environments using machine learning techniques", E-Learning and Digital Media, 2022
[73] Alshanqiti, Abdullah & Namoun, Abdallah. (2020). Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification. IEEE Access. 8. 203827-203844. 10.1109/ACCESS.2020.3036572.
[74] D K Arun, V Namratha, B V Ramyashree, Yashita P Jain, Antara Roy Choudhury. "Student Academic Performance Prediction using Educational Data Mining", 2021, International Conference on Computer Communication and Informatics (ICCCI), 2021.
[75] Aladeemy, Mohammed & Tutun, Salih & Khasawneh, Mohammad. (2017). A new hybrid approach for feature selection and Support Vector Machine model selection based on Self-Adaptive Cohort Intelligence. Expert Systems with Applications. 88. 118–131. 10.1016/j.eswa.2017.06.030.
[76] Aiguo Wang, Ning An, Guilin Chen, Lian Li, and Gil Alterovitz, "Accelerating wrapper-based feature selection with K-nearest-neighbour," *Knowl.-Based Syst.*, vol. 83, pp. 81–91, 2015.
[77] Malik, S., Jothimani, K., Ujwal, U.J. (2023). A Comparative Analysis to Measure Scholastic Success of Students Using Data Science Methods. In: Shetty, N.R., Patnaik, L.M., Prasad, N.H. (eds) Emerging Research in Computing, Information, Communication and Applications. Lecture Notes in Electrical Engineering, vol 928. Springer, Singapore. https://doi.org/10.1007/978-981-19-5482-5_3.
[78] Ezgi Zorarpacı, and Selma Ayse Ozel, "A hybrid approach of differential evolution and artificial bee colony for feature selection," *Expert Systems With Applications*, vol.62, pp. 91-103, 2016.
[79] Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, and Zhigang Gao, "A hybrid feature selection algorithm for gene expression data classification," Neurocomputing, vol. 256, pp. 56-62, 2017.
[80] A Khan, SK Ghosh, D Ghosh, S Chattopadhyay" Random wheel: An algorithm for early classification of student performance with confidence" - Engineering Applications of Artificial Intelligence, 2021.

Authors' Profiles

Dr. Ujwal U.J. is a distinguished Professor and the Head of the Department of Computer Science and Engineering at KVG College of Engineering. He pursued his Ph.D. from VTU Belagavi, specializing in the field of data mining. He holds a position as an executive council member at VTU Belagavi and he has served as a chairman for numerous prestigious committee. Throughout his career, Dr. Ujwal U.J. has made significant contributions to the field of computer science and engineering. His achievements are evident through the publication of numerous research papers in prestigious journals and his presentation of groundbreaking findings at prestigious international conferences. He is specialized in data mining, web mining and cloud computing.

Saleem Malik is actively pursuing a Ph.D. from VTU, Belagavi under the guidance of Dr. Ujwal U.J. Presently holding the position of an Assistant Professor, he specializes in Data Science. With an impressive portfolio of over 50 published works, his research spans diverse domains, including intelligent systems, human-computer interaction, software engineering, and technology acceptance and adoption. Remarkably, Mr. Malik has adeptly led significant research ventures in collaboration with esteemed SMEs. His inquiries have revolved around understanding user needs for modern interactive technologies, crafting composite software services, conducting

usability tests, and gauging human-interface acceptance. His recent focus centers on integrating advanced artificial intelligence methodologies into the design and development of interactive systems, demonstrating a steadfast commitment to pioneering research.

How to cite this paper: Ujwal U.J, Saleem Malik, "A Hybrid Weight based Feature Selection Algorithm for Predicting Students' Academic Advancement by Employing Data Science Approaches", International Journal of Education and Management Engineering (IJEME), Vol.13, No.5, pp. 1-22, 2023. DOI:10.5815/ijeme.2023.05.01