www.ijecs.in
International Journal of Engineering and Computer Science
Volume 14 Issue 03 March 2025, Page No. 26907-26919
ISSN: 2319-7242 DOI: 10.18535/ijecs/v14i03.4999

Enhancing Student Academic Performance Prediction Through Feature Selection Techniques and Machine Learning Algorithms

B. Nantheeswari PG Student [1], Dr. R. Porkodi Professo [2*]

[1] Department of Computer Science Bharathiar University Coimbatore-641046

[2] Department of Computer Science Bharathiar University Coimbatore-641046

Abstract

Student academic performance evaluates the level of achievement a student attains in their educational tasks. It is used to identify strengths, weaknesses, and trends in learning. Performance is often measured through grades, test scores, and overall participation in academic activities. High performance signifies strong comprehension and active engagement, whereas lower performance can highlight potential challenges or areas needing support. This evaluation plays a important role in designing personalized learning strategies. A dataset with 33 features has been utilized to analyse student performance. These features include demographic information (such as age, gender, and family background), academic results, and personal and social variables like family relationships, free time, health status, and alcohol consumption. To identify the most significant factors influencing performance, feature selection techniques such as Recursive Feature Elimination (RFE), Random Forest Regressor, and Chi-square were applied. Among these, the Random Forest Regressor demonstrated the highest predictive performance, achieving an accuracy of up to 93.67% in models such as logistic regression.

Keywords: Feature selection Techniques, Dataset, Machine Learning Algorithms, Recursive Feature Elimination, Random Forest Regressor, Chi-square

1. Introduction

Student academic performance reflects the level of success students achieve in their studies, typically measured through grades, test scores, and how well they understand the material. These measures help teachers assess how much students have learned and how they apply that knowledge. Academic success depends on many factors, such as intrinsic motivation, which encourages students to engage with their studies, and effective study habits that improve learning. Support from family and teachers is also essential for creating a positive learning environment.

External factors like mental health and socio-economic conditions can significantly impact a student's performance. Students facing mental health challenges or financial difficulties may struggle to focus, stay motivated, or access the necessary resources, which can hinder their progress. Regularly monitoring students' performance enables educators to identify these challenges early and provide the right support, such as counseling, financial assistance, or additional tutoring. This approach ensures that every student receives the help they need to succeed, leading to better learning outcomes and more equal opportunities for all.

A dataset capturing student academic performance typically includes 33 variables. These encompass demographic information alongside academic grades. Additionally, personal and social factors. Feature selection is important for optimizing algorithm performance. It focuses on identifying

the most relevant attributes from the dataset, which enhances model accuracy and interpretability. Methods such as Recursive Feature Elimination (RFE), Random Forest Regressor, and Chi-square tests are widely used to rank and select key features based on their predictive significance. Machine learning algorithms such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVMs) analyse patterns within the data to forecast student performance. By applying feature selection, can identify the most influential variables, enhancing model efficiency and improving predictive accuracy by eliminating irrelevant or redundant features.

2. Literature Review

J. Malini and Y. Kalpana [1] investigated how students' economic backgrounds influence course completion rates. Using educational data mining techniques, the study analysed various datasets to identify factors affecting academic performance. Machine learning models, including decision trees and neural networks, were employed, with bagging classifiers demonstrating the highest accuracy. Their findings highlight the critical role of economic characteristics in shaping learning habits and emphasize the need for customized teaching strategies.

Pranav Dabhade et al. [2] aims to predict performance and derive valuable insights from student data. The research methodology involves the use of surveys based on questionnaires and academic records to gather information on students' personal, educational, and behavioral characteristics. Through the application of various machine learning algorithms, such as support vector regression and multiple linear regression, academic achievement is predicted. The findings indicate a strong correlation between students' behavioral traits and their academic results, indicating the need for further investigation into machine learning algorithms and larger datasets.

Muhammad Imran et al. [3] increasing significance of predictive analytics in the field of education, specifically concentrating on predicting student outcomes through methods such as EDM. Stakeholders are in search of early warning systems to pinpoint students who are at risk and improve the learning experience. Different classification algorithms, such as K-nearest neighbor and Decision Trees, are utilized to create predictive models. The research underscores the importance of accurate performance prediction models in educational settings and addresses issues related to data quality.

Bindhia K. Francis [4] the crucial role of academic performance evaluation in educational institutions and the importance of early prediction to assist struggling students. Student performance is impacted by various factors, such as socioeconomic and environmental conditions.EDM techniques, including data mining, are utilized to extract patterns and information from educational databases for predictive analysis. decision trees are employed in research to forecast student outcomes. A hybrid technique that combines clustering and classification algorithms indicating its potential to enhance academic performance and reduce failure rates.

Alaa khalaf hamoud et al. [5] enhancing student achievement in education settings. The objective is to utilize decision tree algorithms and surveys to analyze collective student data, with the aim of predicting and categorizing student performance. By identifying the factors that influence success and failure rates, this research aims to support academic decision making and provide guidance to both lecturers and students. Through the implementation of various classification and prediction methods, the study strives to enhance the quality of education and identify underperforming students early in the semester and students for improving performance.

M. Durairaj et al. [6] utilized data mining to uncover hidden patterns in student performance. The study applied K-Means clustering and Naive Bayes classifiers, demonstrating their effectiveness in predicting pass and fail rates. The WEKA tool facilitated efficient data mining

processes, highlighting the value of predictive analytics in education.

Harikumar Pallathadka et al. [7] emphasized the use of educational data mining to predict student outcomes and implement targeted interventions. Machine learning algorithms like SVM and Naive Bayes were applied, with SVM proving to be the most accurate. This research also explored the impact of teacher evaluations on student performance and the benefits of analyzing question papers for grading consistency.

Carina Granberg et al. [8] Investigated how formative assessment contributes to the development of self-regulated learning (SRL) skills. The study indicated that peer and self-assessment could enhance motivation and self-efficacy, despite mixed results. The research calls for experimental studies to explore the effects of daily formative assessment practices on SRL strategies in classroom settings.

Raza Hasan et al. [9] supervised classification model to predict academic success using data from SIS, Moodle, and edify. Random Forest achieved the highest accuracy of 88.3% among eight tested algorithms. Feature reduction techniques like PCA and genetic algorithms yielded inconclusive results, while multivariate analysis identified nine critical variables for prediction, outperforming the full feature set. The CN2 Rule Inducer algorithm closely followed with 87.4% accuracy. The study emphasizes the importance of blended learning and flipped classrooms, suggesting the implementation of a performance monitoring dashboard Jie Cai et al. [10] feature selection methods in machine learning and data mining, with a focus on high-dimensional data's impact on model performance. Key methods include filter, wrapper, and embedded approaches, each with unique evaluation criteria and applications. For example, filter methods often rely on statistical measures, while wrappers use model performance to assess feature subsets. Embedded methods integrate feature selection directly into the training process. Addresses supervised, unsupervised, and semi-supervised techniques, along with future challenges such as managing extreme data and improving stability in feature selection.

Drishty Sobnath et al. [11] focused on predicting employment outcomes for disabled graduates in the UK. Using machine learning models like decision trees and logistic regression, the study identified predictors such as age, institution, and disability type. The results highlighted the importance of tailored support systems to enhance employability.

Chitra Jalota et al. [12] explored the application of feature selection algorithms and machine learning models for predicting academic performance. The study utilized tools like WEKA and algorithms such as Ant Colony Optimization and Bayesian classifiers, emphasizing optimal feature-classifier combinations for improved predictions.

Maryam Zaffar et al. [13] compared different feature selection algorithms, including Chi-square, Gain Ratio, and Relief, to evaluate their effectiveness in predicting academic performance. The study demonstrated that the performance of feature selection techniques varies based on dataset characteristics, underscoring the importance of algorithm selection in improving prediction accuracy.

3. Methodology

The methodology outlined in fig 1 provides a structured approach for predicting student academic performance using feature selection techniques and machine Learning algorithm. This process follows several key steps: gathering relevant data, preparing it through preprocessing, selecting important features, training predictive models, evaluating their performance, and comparing results to determine the best approach. Each step builds on the previous one to refine data and improve the model's predictive accuracy, ultimately identifying the optimal model for forecasting student outcomes.

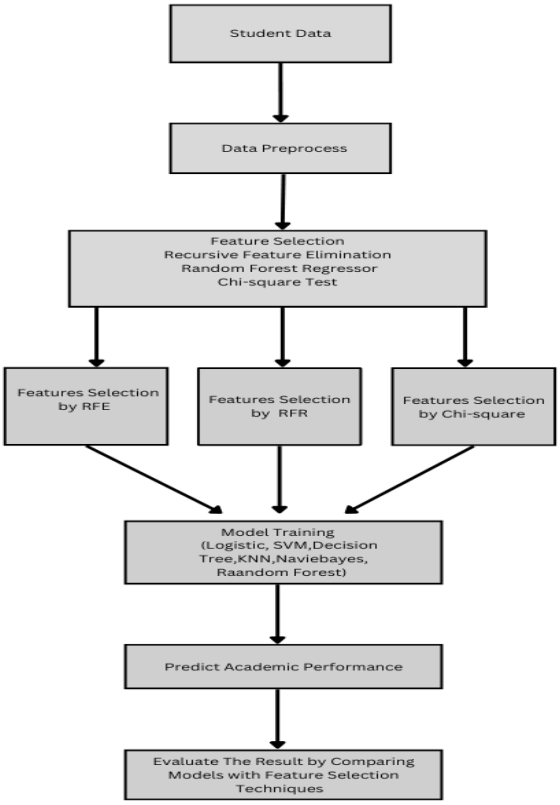

Fig 1: Flow diagram of the methodology

3.1 Dataset

The student performance dataset, designed to predict academic achievement in secondary education, originates from two Portuguese schools and is publicly available through the UCI Machine Learning Repository and Kaggle. The dataset provides insights into student outcomes for two subjects: Portuguese language and mathematics. It includes 33 variables that capture a wide range of factors influencing student performance, including social, demographic, academic, and school-related attributesThe dataset can be broadly divided into three groups of attributes:

Personal Attributes: Gender, age, type of residence (urban or rural), involvement in romantic relationships, and alcohol consumption patterns on weekdays and weekends. Financial and Family Background: Covers information about the family structure and financial background, such as parents' cohabitation status, parental education levels, family size, parents' occupations, reasons for school selection, primary guardian, availability of family support for education, and access to resources like internet and extra educational support.

Educational and Academic Background: Contains details about students' academic habits and experiences, including study time, number of past academic failures, participation in paid and extracurricular activities, nursery attendance, aspirations for higher education, social life (e.g., time spent with friends), and school attendance. To identify the most influential features affecting student performance, feature selection methods such as Recursive Feature Elimination (RFE), Random Forest Regressor, and Chi-square techniques are applied. These techniques assess the predictive power of each feature, systematically eliminate less significant ones, and rank the top 15 features that most strongly correlate with academic outcomes.

3.2 Data preprocessing

Data preprocessing is an essential phase in readying data for analysis, ensuring its quality and enhancing the performance of predictive models. missing data is handled by filling gaps or discarding incomplete records. data normalization or standardization scales numerical values to a common range, which helps models interpret the features effectively. Categorical features are then converted into numerical formats, often using techniques like one-hot encoding. Any outliers, which are extreme values that could distort predictions, are identified and managed. Additionally, irrelevant or redundant features are removed, simplifying the dataset and reducing computation time.

4. Feature Selection Algorithms

Feature selection is a important process in machine learning that involves identifying and retaining the most relevant features from a dataset while discarding those that are less significant or redundant. This approach enhances model performance by reducing complexity, improving accuracy, and minimizing the risk of overfitting. It helps streamline the computational process, making models faster and more efficient [13].

Recursiv e Feature Elimination (RFE) is a popular example, where the algorithm iteratively removes the least significant features. Embedded Methods perform feature selection as part of the model training process. For instance, techniques like Lasso regression incorporate feature selection by penalizing less important features, effectively shrinking their coefficients to zero.

By applying these techniques, models can focus on the features that most significantly impact the target variable, which leads to better interpretability and often higher accuracy. For instance, using RFE, Random Forest Regressor, and Chi-square tests, the top 15 features can be identified from an initial dataset of 33 attributes. RFE systematically eliminates weaker features, Random Forest ranks features based on their contribution to prediction accuracy, and Chi-square tests evaluate the statistical relationship between features and the target variable.

A. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a widely used feature selection technique in machine learning. Its primary goal is to reduce the dimensionality of a dataset by identifying and retaining the most relevant features, which can enhance the efficiency and performance of a predictive model. RFE operates by recursively fitting a model to the data, evaluating the importance of each feature, and progressively removing the least significant ones. The process begins by training a model on the complete set of features. Features are then ranked based on their significance, which is typically determined using model-specific metrics. [11] At each iteration, the feature with the lowest importance is discarded, and the model is retrained using the reduced feature set. This iterative process continues until the desired number of features is reached or only the most influential features remain.X= {x_1, x_2 ....x_n} represents the set of features, and y is the target variable. RFE evaluates the model X, identifies the feature with the smallest contribution to f, and recursively removes it. This process can be expressed as:

$$RFE(f, X, y) = X - \{ x_j \} \quad \text{where} \quad j = \arg FeatureImportance\ (\ x_j\ ) \quad (1)$$

Table 1 Ranked Features Based on RFE

Feature	Ranking
G2	1
Fjob_services	2
Fjob_health	3
Famrel	4
guardian mother	5
Higher yes	6
Activities yes	7
Schoolsup_yes	8
Reason other	9
Reason reputation	10
Nursery yes	11
Pstatus_t	12
Study time	13
Romantic yes	14
G1	15

This process is repeated until a specified number of features remains. The objective is to retain only the most significant features, reducing model complexity, enhancing interpretability, and mitigating the risk of overfitting, particularly when many features are redundant or irrelevant. RFE systematically identifies and ranks features based on their importance, ultimately yielding a reduced subset of features that are most predictive of the target variable. This approach was applied to predict student performance, with the most significant features ranked and presented in Table 1. These selected features were determined to have the greatest impact on the prediction outcomes.

Table 2 Ranked Features Based on RFR

Feature	Ranking
G2	0.7737
Absences	0.1181
Age	0.0149
Health	0.0083
Famrel	0.0066
Reason home	0.0066
Studytime	0.0060
G1	0.0051
Guardian_mother	0.0048
Activities yes	0.0042
Free time	0.0039
Feud	0.0036
Gout	0.0033
Walk	0.0030
School_ms	0.0030

B. Random Frest Regressor

Random Forest Regressor is a powerful feature selection technique in machine learning. This model works by constructing multiple decision trees during training and averaging their predictions for regression tasks. One common feature selection technique involves calculating feature importance scores based on the impurity reduction achieved by each feature across all trees. Features contributing more to reducing error are assigned higher importance scores. Another method uses permutation importance, where feature values are shuffled to assess the impact on model performance.

Random Forest Regressor calculates feature importance by evaluating how much each feature contributes to reducing prediction error. [7] Features that lead to greater error reduction are assigned higher importance scores, allowing the model to rank them based on their predictive value. By pinpointing the most influential features, this method helps streamline the model, improving its performance and interpretability while also reducing the complexity of the dataset by focusing on the key drivers of the target variable.

$$I(f_j) = \frac{1}{N} \sum_{t=1}^{N} \sum_{split\ on\ f_j} \Delta MSE_t \quad (2)$$

Where:

- $I(f_j)$ is the importance score of features f_j,
- N is the total number of trees in the forest,
- ΔMSE_t is the reduction in Mean Squared Error caused by splits on feature f_j in tree t.

The Random Forest algorithm inherently performs feature selection by assigning higher importance to features that play a important role in improving predictive accuracy. This used in identifying the most relevant features while filtering out those that are redundant or less impactful. As a result, the model becomes simpler without a significant loss in performance. In the Random Forest Regressor, multiple decision trees are built during the training process. Each tree is trained on a random subset of the data and features. The importance of a feature is determined by how much it contributes to reducing error or improving the model's decision-making across these trees. This approach was applied to predict student performance, with the most significant features ranked and presented in Table 2. These selected features were determined to have the greatest impact on the prediction outcomes.

A. Chi-square

A statistical technique for assessing the correlation between categorical characteristics and the target variable in a dataset is the chi-square feature selection algorithm. It is particularly useful in classification tasks to identify which features have the most significant influence on predicting the target class.[14] The algorithm relies on the Chi-square (χ^2) test of independence, which measures the difference between the observed and expected distributions of data.The Chi-square test works by calculating a statistic that quantifies how much the observed frequencies of a feature's values differ from the frequencies expected under the assumption that there is no association between the feature and the target. Features with higher Chi-square scores are considered more relevant, as they demonstrate a stronger dependence on the target variable.

The procedure involves comparing the observed frequencies of a feature's categories across different classes with the expected frequencies, assuming independence. A large Chi-square value indicates a significant discrepancy between the observed and expected distributions, suggesting a stronger relationship between the feature and the target class.

Table 3 Rank Based on Chi-square

Feature	Ranking
Absences	648.166848
G2	451.680952
G1	319.716427
Failures	140.934898
Walk	28.444167
Schoolsup_yes	28.319018
Fjob_teacher	23.133348
mjob_health	22.416110
Mjob_services	21.588835
Romantic yes	20.089526
Reason reputation	17.224393
Dalc	17.082868
Paid yes	16.129900
Reason other	16.029481
Fjob_services	15.848654

The algorithm ranks features based on their Chi-square scores, helping to prioritize those that contribute the most to classification performance. The Chi-square formula for each feature is given by:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- x^2 is the Chi-square statistic,
- O_i represents the observed frequency (the actual count of samples for each class and feature category), E_i represents the expected frequency (the count of samples that would be expected if the feature and class were independent).

The Chi-square algorithm evaluates the relationship between each feature and the target variable by calculating the Chi-square statistic. Features with higher Chi-

square scores are considered more relevant as they exhibit a stronger association with the target class, whereas those with lower scores may be removed due to their weak or negligible relationship. This method is especially useful for categorical data and aids in reducing dimensionality by selecting the most influential features for classification tasks. This approach was applied to predict student performance, with the most significant features ranked and presented in Table 3. These selected features were determined to have the greatest impact on the prediction outcomes. Common Features: Among Three feature selection algorithms such us Recursive Feature Elimination (RFE), Random Forest Regressor, and Chi-square are compared to identify the common attributes show in Table 4.

5. Prediction Techniques for Student Academic Performance Evaluation

A. Logistic Regression

Logistic regression is a statistical technique designed for binary classification tasks, where the objective is to estimate the probability of one of two possible outcomes (e.g., 0 or 1, yes or no) depending on a number of predictor variables. Unlike linear regression, which predicts continuous values, logistic regression employs a sigmoid function to constrain the predicted probabilities within the range of 0 and 1[15].

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Table 4 Common Features in Three Feature Selection Techniques

Common feature	Techniques
Fjob_services	RFE & chi-square
Schoolsup_yes	RFE & chi-square
Reason other	RFE & chi-square
Reason reputation	RFE & chi-square
Romantic yes	RFE & chi-square
Absences	RFE & chi-square
Activities yes	RFE & RFR
Study time	RFE & RFR
Famrel	RFE & RFR
guardian mother	RFE & RFR
G1	RFE, RFR, chi-square
G2	RFE, RFR, chi-square

where P(Y=1 | X) is the probability that the dependent variable Y is 1 given the predictors X; β0 is the intercept term, and β1, β2, βnare the coefficients of the predictor variables X1, X2, Xn. These coefficients are estimated using maximum

likelihood estimation to determine the best-fitting model for predicting the probability of an outcome. This probability can be compared against a threshold to classify observations into one of two categories. Logistic regression is widely valued for its simplicity, interpretability, and ability to effectively handle binary classification problems.

B. Naive Bayes

Naive Bayes is a classification technique grounded in Bayes' Theorem, primarily used for probabilistic machine learning tasks. [9] It operates on the assumption of conditional independence, meaning it presumes that each feature in a dataset contributes independently to the probability of a given outcome, even if this isn't strictly true in practice. This independence assumption allows for a simplified computation of conditional probabilities, making Naive Bayes computationally efficient, especially when handling large datasets. The model selects the class with the highest probability after calculating the posterior probability of a class given a collection of features.

C. Decision Tree

A decision tree is a model that is organized as a hierarchical tree, with each internal node testing a particular condition or characteristic, each branch representing the test's result, and each leaf node offering the final classification or prediction. Starting at the root node, the dataset is split up into smaller groups according to the property that best divides the data. This attribute is selected using a criterion such as Information Gain. The tree-building process continues recursively, with each subsequent node further refining the separation of data. Information Gain is calculated using entropy, which quantifies the level of disorder or uncertainty within a dataset. By maximizing Information Gain, the model minimizes entropy, resulting in more distinct and meaningful splits.

$$\text{Information Gain}(S, A) = \text{Entropy} - \sum_{v-Values(A)} \frac{|S_v|}{|S|} \text{Entropy}\,(S_v)$$

The splitting process continues recursively until stopping criteria like maximum depth or minimum samples per node are met, ensuring a balance between complexity and overfitting.

D. Random Forest

A Random Forest is an ensemble learning method commonly used for both classification and regression tasks. It constructs multiple decision trees during training, with the aim of improving accuracy and stability in predictions [10]. Each tree is trained on a different random subset of the data, selected through bootstrapping, where samples are drawn with replacement. Moreover, during the construction of each tree, only a random subset of features is considered for each split, which helps reduce correlations between the individual trees. This added randomness helps prevent overfitting and improves the model's ability to generalize. For classification, the prediction is based on the majority vote from all the trees, while for regression, the final output is obtained by averaging the predictions from all the trees.

E. K-Nearest Neighbors (K_Nn)

A data point is classified using the k-Nearest Neighbors (k-NN) method by looking at the categories of its closest neighbors. [8] It assumes that points that are close to each other in the feature space are likely to have similar outcomes. k-NN finds the k closest points from the training data when a new data point is added, then utilizes their labels to ascertain the new point's value or categorization. The algorithm calculates the distances between the new point and all the existing points to find the closest neighbors. The distance formula for two points x = (x1, x2, ..., xn) and y = (y1, y2, ..., yn) in an n-dimensional space.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}$$

After identifying the k nearest neighbors, the k-Nearest Neighbors (k-NN) algorithm assigns the new data point to the class that is most frequent among these neighbors. The value of k plays a crucial role in the model's performance: a small k makes the algorithm more sensitive to noise in the data, whereas a larger k can smooth the decision boundaries, but it may also lead to misclassification due to excessive generalization.

F. Support Vector Machine

A Support Vector Machine (SVM) is a supervised learning algorithm primarily used for classification tasks. It works by finding the optimal hyperplane that best separates data points of different classes in a given feature space. The core concept of SVM is to identify a boundary, called the hyperplane, that maximizes the margin between the two classes. The data points closest to this hyperplane are referred to as support vectors, and they play a key role in determining the position and orientation of the hyperplane. The optimization process allows the SVM to create a robust decision boundary that minimizes classification errors. When the data is not linearly separable, SVM uses kernel functions, such as the polynomial or radial basis function (RBF) kernel, to transform the data into higher dimensional spaces where separation becomes more feasible. The kernel trick enables SVM to operate effectively in complex environments without needing to directly compute the transformation to higher dimensions.

Table 5 Machine Learning Models in Recursive Feature Elimination

	Precision	Recall	F1-Score	Accuracy
Logistic regression	0.636364	0.954545	0.763636	0.672269
Decision tree	0.638889	0.696970	0.666667	0.613445
Random forest	0.639175	0.939394	0.760736	0.672269
K-NN	0.643678	0.848485	0.732026	0.655462
SVM	0.611111	1.000000	0.758621	0.647059
Naive bayes	0.640000	0.969697	0.771084	0.680672

6. Result and Discussion

Performance measures for feature selection techniques in predicting student academic performance are important to ensure that the selected features significantly contribute to accurate predictions of academic outcomes. Accuracy values indicate the proportion of correct predictions made by the model using the selected features, ensuring that the most relevant predictors are identified for improving the model's predictive power. These evaluation techniques use to validate that the feature selection process enhances the overall performance and reliability of the prediction model.

Accuracy:

$$\text{Accuracy} = \frac{tp+tn}{tp+fn+tp+tn}$$

Accuracy is a metric used to assess the overall correctness of a model's predictions. It is determined by dividing the number of correct predictions (including both true positives and true negatives) by the total number of predictions made.

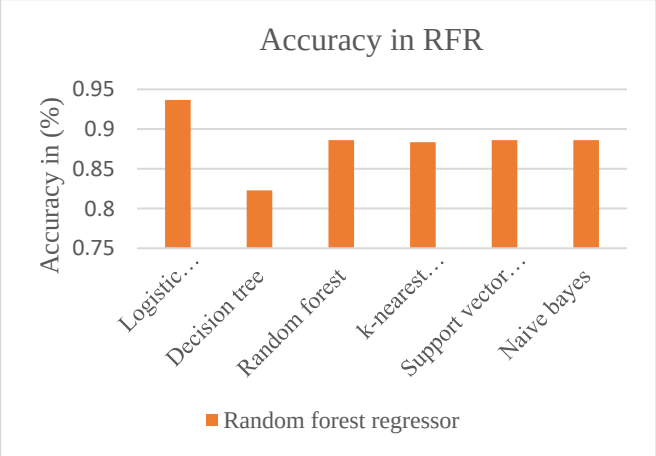

Fig 2 Accuracy of Machine Learning Models in Recursive Feature Elimination

Its strong precision and recall values highlight its reliability in predicting student performance, demonstrating its ability to effectively utilize the selected features. Performing feature selection using the Random Forest Regressor (RFR) Table 6 show the performance of various machine learning models was evaluated. The Logistic Regression model achieved the highest accuracy of 93.67%, showcasing its robust predictive

capabilities when utilizing the features identified by the Random Forest Regressor. This highlights the model's effectiveness in leveraging the most relevant features for predicting student performance.

Table 6 Machine Learning Models in Random Forest Regressor

	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.956522	0.936170	0.946237	0.936709
SVM	0.895833	0.914894	0.905263	0.886076
KNN	0.895833	0.914894	0.905263	0.886076
Random Forest	0.931818	0.872340	0.901099	0.886076
Naive Bayes	0.872340	0.872340	0.872340	0.848101
Naive bayes	0.883721	0.808511	0.844444	0.822785

The performance of various machine learning models was evaluated using features selected by the Chi-square algorithm. Among the tested models, the Random Forest achieved the highest performance, with an accuracy of 25.21% and an F1-score of 24.17%. These results demonstrate its ability to effectively utilize the features identified by the Chi-square method, despite the overall modest predictive accuracy show in Table 7.

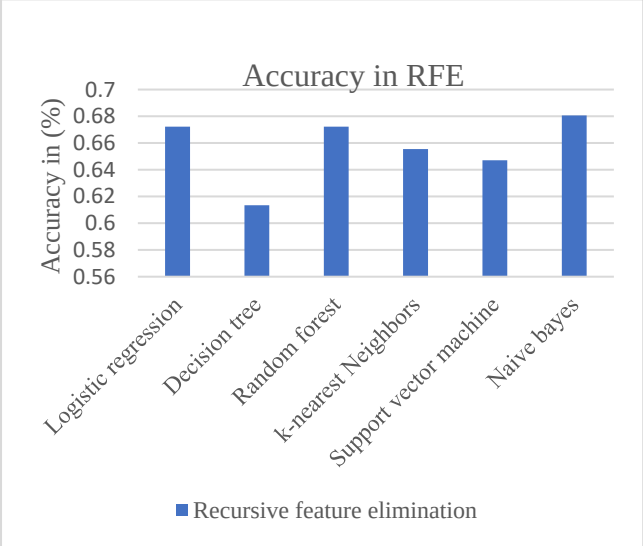

Fig 3 Accuracy of Machine Learning Models in Random Forest Regressor

The performance of various machine learning models was evaluated using features selected by the Chi-square algorithm. Among the tested models, the Random Forest achieved the highest performance, with an accuracy of 25.21% and an F1-score of 24.17%. These results demonstrate its ability to effectively utilize the features identified by the Chi-square method, despite the overall modest predictive accuracy show in Table 7.

Three feature selection methods Recursive Feature Elimination (RFE), Random Forest Regressor, and Chi-square are compared to assess their effectiveness in selecting the most relevant features for predicting student academic performance. These techniques are applied to a dataset containing various student performance attributes. The performance of each method is evaluated by measuring the accuracy of machine learning models trained with the selected features as shown in Fig 2, Fig 3, Fig 4.

Table 7 Machine Learning Models in Random Forest Regressor

	Accuracy	Precision	Recall	F1 Score
Logistic regression	0.226891	0.203145	0.226891	0.205056
Decision tree	0.184874	0.180691	0.184874	0.171992
Random forest	0.252101	0.269404	0.252101	0.241695
K-NN	0.235294	0.308024	0.235294	0.240900
SVM	0.142857	0.098727	0.142857	0.094267
Naive bayes	0.159664	0.207837	0.159664	0.090334

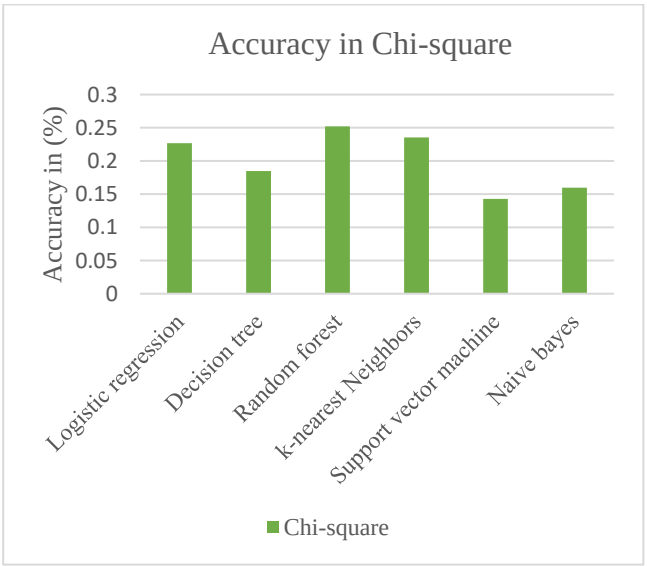

Fig 4 Accuracy of Machine Learning Models in Chi-Square

All three feature selection algorithms provided valuable insights into feature importance, but the Random Forest Regressor emerged as the most effective, delivering the highest accuracy and overall better performance across various models.

As shown in Table 8 and Fig. 5, Recursive Feature Elimination (RFE) demonstrated moderate accuracy, with logistic regression achieving 0.63 and Naive Bayes reaching 0.68. While this method is useful for understanding the relative importance of features, it did not perform as well in terms of predictive accuracy. In contrast, the Random Forest Regressor achieved superior accuracy, with logistic regression reaching up to 0.93. This method's strength lies in its ability to handle large datasets and capture non-linear relationships between features. The Chi-square algorithm, on the other hand, yielded lower accuracy, with logistic regression achieving only 0.22 and Naive Bayes at 0.15. These results highlight the need for selecting feature selection techniques that align with the characteristics of the dataset and the specific requirements of the predictive model. The Random Forest Regressor's embedded feature importance scoring is particularly valuable for identifying key variables in large datasets, making it an essential tool in educational data mining, where numerous interrelated features influence academic outcomes.

Table 8 Accuracy of Feature Selection Algorithms in Predicting Student Performance

	Accuracy in		
	RFE	RFR	Chi-square
Logistic regression	0.672269	0.936709	0.226813
Decision tree	0.613445	0.822785	0.184874
Random forest	0.672269	0.886076	0.252101
K-NN	0.655462	0.883418	0.235294
SVM	0.647059	0.886076	0.142857
Naive bayes	0.680672	0.848101	0.159664

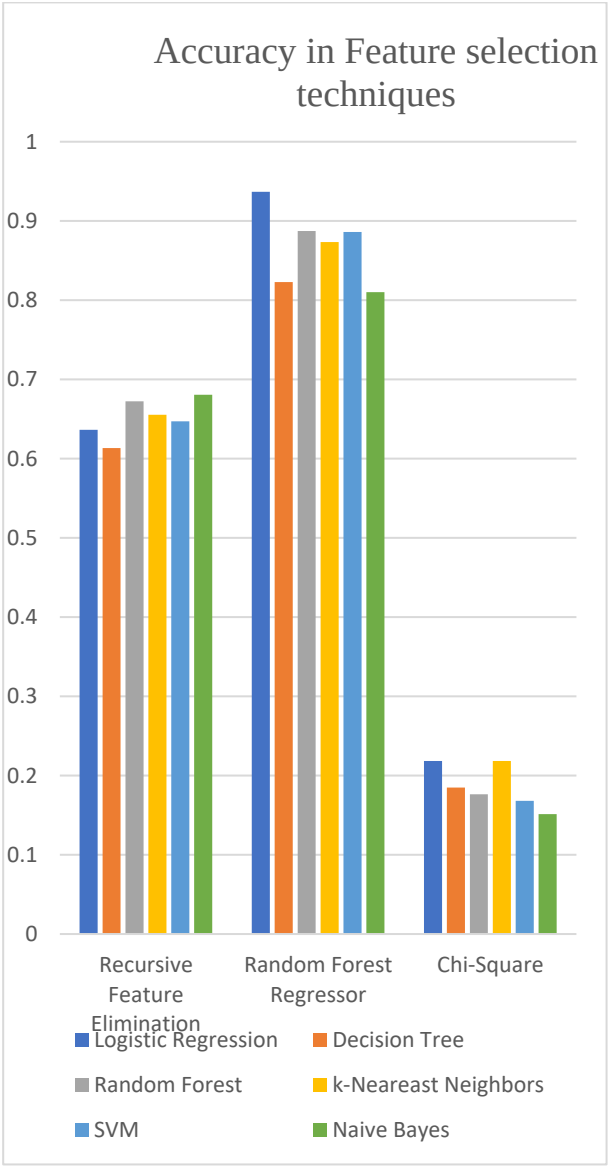

Fig 5. Performance of machine learning model with Feature selection algorithm

The Random Forest Regressor (RFR) stands out as the most effective feature selection algorithm for predicting student performance due to its ability to handle large, complex datasets and accurately capture non-linear relationships among features. In RFR feature importance scores are calculated based on how much each feature reduces prediction error across multiple decision trees, allowing the model to identify and prioritize the most influential variables. This selection process helps streamline the model by removing irrelevant or redundant features, reducing computational complexity, and minimizing the risk of overfitting. In the student performance dataset, RFR highlighted key predictors like prior grades (G2), family relationships, and attendance

(Absences), which significantly impacted outcomes. By focusing on these variables, the model achieved higher predictive accuracy 93.67% in logistic regression model demonstrating RFR's effectiveness in isolating impactful features.

7. Conclusion

Feature selection techniques like Recursive Feature Elimination (RFE), Random Forest, and Chi-square are applied to datasets containing information on student demographics, academic history, and lifestyle factors. These methods help identify the most influential features for academic performance, such as grades from previous periods (e.g., G2 and G1) and family support. By selecting the most relevant features, these techniques optimize model performance models such as Support Vector Machine (SVM), Lasso Regression, and Random Forest Regression are utilized to identify the most significant predictors of student performance. Among these, SVM and Lasso Regression achieved the highest accuracy, indicating the strong influence of various social, demographic, and academic factors on student outcomes. These models highlight how understanding these factors can aid in developing effective intervention strategies to support students. Feature selection is crucial in enhancing the accuracy and efficiency of models used to predict student academic success. Among the three techniques, Random Forest Regressor achieved the highest predictive accuracy 93.67% in Logistic Regression. Random Forest Regressor highlights its effectiveness in strengthening the prediction of academic outcomes. These integrating feature selection methods with predictive models can improve the identification of important predictors, reduce model complexity, and support the design of targeted academic interventions.

References

1. J. Malini and Y. Kalpana, "Investigation of factors affecting student performance evaluation using education materials data mining technique," Mater. Today Proc., vol. 47, pp.6105–6110,2021, doi: 10.1016/j.matpr.2021.05.026.
2. Dabhade P., Agarwal R., Alameen K.P., Fathima A.T., Sridharan R., Gopakumar G"Educational data mining for predicting students' academic performance using machine learnialgorithms"(2021) Materials Today: Proceedings, 47 , pp. 5260-5267.
3. M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student Academic Performance Prediction using Supervised Learning Techniques,"vol. 14, no. 14, pp. 92-104, 2019, doi: 10.3991/ijet. v14i14.10310.
4. B. K. Francis and S. S. Babu, "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach,"part of Springer Nature, 2019, pp. 1-12, doi:10.1007/s00500-019-04052-9.
5. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," Basra, Iraq, Nov. 3, 2017. Accepted Jan. 22, 2018, published Feb. 9, 2018.
6. M. Durairaj and C. Vijitha, M. "Educational Data Mining for Prediction of Student Performance Using Clustering Algorithms," International Journal of Computer Science and Information Technologies, vol. 5, no. 4, pp. 5987–5991, 2014.
7. H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," Materials Today: Proceedings, vol. 80, no. 3, pp. 3782-3785, 2023.
8. C. Granberg, T. Palm, and B. Palmberg, "A case study of a formative assessment practice and the effects on students' self-regulated learning," Studies in Educational Evaluation, vol. 68, Mar. 2021, Art. no. 100955.

9. R. Hasan, S. Palaniappan, A. R. Abdul Raziff, S. Mahmood, and K. U. Sarker, "Student Academic Performance Prediction by Using Decision Tree Algorithm," in Proc. 4th Int. Conf. Computer and Information Sciences, Kuala Lumpur, Malaysia, Aug. 2018, pp. 1-6, doi: 10.1109/ICCOINS.2018.8479234.
10. J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” Neurocomputing, vol. 300, pp. 70–79, 2018, doi: 10.1016/j.neucom.2017.11.077.
11. D. Sobnath, T. Kaduk, I. U. Rehman, and O. Isiaq, "Feature Selection for UK Disabled Students' Engagement Post Higher Education: A Machine Learning Approach for a Predictive Employment Model," IEEE Access, vol. 8, pp. 159530-159541, Sep. 2020
12. D. Sobnath, T. Kaduk, I. U. Rehman, and O. Isiaq, "Feature Selection for UK Disabled Students' Engagement Post Higher Education: A Machine Learning Approach for a Predictive Employment Model," IEEE Access, vol. 8, pp. 159530-159541, Sep. 2020
13. M. Zaffar, M. A. Hashmani, K. S. Savita, and S. K. Sajjad, "A Study of Feature Selection Algorithms for Predicting Students Academic Performance," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 5, pp. 541–549, 2018
14. S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, "Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection," Informatica, vol. 47, no. 1, pp. 11–20, 2023
15. S. Hussain and M. Q. Khan, "Student Per formulator: Predicting Students’ Academic Performance at Secondary and Intermediate Level Using Machine Learning," Springer-Verlag GmbH, pp. 1-12, Apr. 2021.