*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
***Article Received:*** *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

# Dimensionality Reduction using Feature Selection Techniques on EDM for Student Academic Performance Prediction

**Kajal Mahawar[1], Punam Rattan[2]**
[1]School of Computer Applications, Lovely Professional University, Phagwara (Punjab), India
kajalmahawar01@gmail.com
[2]School of Computer Applications, Lovely Professional University, Phagwara (Punjab), India
punamrattan@gmail.com

**Abstract**— The aim of this study is to perform dimensionality reduction on student datasets using different feature selection strategies in order to better forecast students' academic progress and minimize high-dimensionality concerns. The effectiveness of models built utilizing classification algorithms and pertinent features chosen using different feature selection strategies has been experimentally evaluated in this study. Nine feature selection methods, including Chi-Square, Pearson Correlation, Variance Threshold, Feature Importance, Recursive Feature Elimination (RFE), Lasso, Ridge, Random Forest, and XGBoost, as well as one classification method, logistic regression, were applied to the student dataset to determine the results of the analysis. In this experiment, dimensionality reduction techniques were used to reduce the number of features from 70 to the optimal 21 feature subsets for analysis of student academic performance prediction. Five measures were used to assess the effectiveness of the techniques: $r^2$, precision, recall, f-score, and accuracy. With the logistic regression classifier, the feature subset chosen by the chi-square had the highest accuracy of 72.4%, precision of 71.4%, recall of 72.4%, f-score of 70%, and r-square of -0.221%.

**Keywords**- Dimensionality reduction, Students' academic performance, Feature selection, Logistic regression, Measures

## I. INTRODUCTION

Student academic performance prediction is a study area having applications of deep learning, machine learning, artificial intelligence, and statistics for the education sector. Machine learning and data mining approaches employ to shift through data from educational settings in search of patterns and predictions that describe students' actions and performance. It enhances students' brains and personalities. A crucial component of the future perspective in education is the ability to predict academic accomplishment in advance. Since students are the main participants in educational institutions, these institutions can adapt to the shifting needs of the society by examining student data and drawing various conclusions from data. Additionally, the outcomes of projections might be useful for developing plans to raise educational standards. Instructors, students, and surroundings are important sorts of participants in an academic institution. These participants' interactions provide a large amount of data that can be methodically grouped to harvest priceless knowledge. Predicting a student's success is a major concern for school administrators and the government. It could provide suitable warnings to students who are in danger by anticipating their grades and assisting them in avoiding problems and overcoming any study challenges. As a result, machine learning techniques are increasingly being used to investigate student data and identify hidden relevant patterns for predicting students' grades. To predict student performance, a range of machine learning-related algorithms have been developed [1]. The user can handle and evaluate the student's data acquired from many sources thanks to a developing study field called "mining of student's data with machine learning." An author used a meta-analysis to assess the overall efficacy of interactive white board-based training by finding empirical studies that appear on students' cognitive learning results [2]. Another author [3] experimented questioned college students in China who used blended learning for the first time in the semester of 2020 and 2021 to discover the significant features that influence beginners' intention. The associations between important technological variables and connected classroom climate in cloud classrooms are empirically supported according to the authors [4]. According to the [5], student academic performance prediction from their previous performance record is a challenging problem in the education sector because of student's microarray data.

There are certain distinctive features in the microarray data. In other words, there are few samples and a high degree of dimension. Dimensionality reduction is a strategy for transforming a dataset with multiple dimensions into a dataset with fewer dimensions while maintaining the same amount of information. It is a vital preprocessing step in the classification process since microarray data is huge data and has numerous dimensions [6]. Three major issues arise when the classification task is implemented due to the presence of numerous dimensions. These issues include the slow learning process,

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
***Article Received:*** *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

rising processing costs, and declining classification accuracy. Techniques for reducing the dimensions fall under one of two categories: feature extraction or feature selection [7]. The goal of the feature extraction method is to build the features into a new, lower-dimensional feature set. By removing the duplicate and irrelevant characteristics, the feature selection strategy, on the other hand, leverages the dataset that was originally collected to choose an ideal subset of important features.

The objective of this study is to ascertain how various feature selection methods, categorized as filter, wrapper, and embedding strategies, affect the accuracy of academic achievement prediction for students. The ramification of various feature selection techniques is detected using the primary student performance dataset. To determine the best feature set for predicting student academic achievement, the current study compares nine supervised feature selection strategies with logistic regression machine learning techniques. Chi-square, Pearson correlation, variance threshold, feature importance, recursive feature elimination, lasso, ridge, random forest, and XGBoost are feature selection techniques that have been employed. In other words, in order to determine the most effective feature selection techniques that could be used to forecast students' academic performance and benefit the education system, this paper examines techniques and assesses them against a variety of performance measures, including $r^2$, precision, recall, f-score, and accuracy. A summary of our technical contribution is presented below:

- In this study, 70 features were taken into account.
- To analyse the effectiveness of dimensionality reduction feature selection methods, chi-square, Pearson, Variance threshold, Feature importance, RFE, Lasso, Ridge, Random forest, and XGBoost approaches were utilized.
- In this study, a logistic regression classifier is utilised to evaluate the effectiveness of machine learning classification models.
- To assess the effects of dimensionality reduction strategies on model performance, prediction model was tested on the real and reduced student dataset.

The remainder of the article is structured as follows. In Section 2, a state-of-the-art literature is addressed; Sections 3 and 4 present the methodology and experiment. Results and discussion are presented in Section 5. Finally, Section 6 draws a conclusion based on the entire study and discusses the prospect of further research.

## II. LITERATURE REVIEW

In this study, component of the suggested model has been extensively researched in recent years. Various techniques are available in machine learning to analyze students' academic performance. The two main approaches in machine learning to predict students' academic success are the supervised method and the unsupervised method. Many researchers prefer classification and regression techniques to develop performance prediction models. The researchers have applied a variety of machine learning classification and regression techniques, according to the most recent literature. The techniques are employed for their low complexity and high accuracy rate in comparison to others. In supervised and unsupervised techniques, some are linear regression, support vector machine, logistic regression (LR), association rule mining, k mean clustering and decision tree. SVM technique falls under the category of classification. This technology is used in numerous research studies and fields such as healthcare, 3-D object recognition, text classification, education, etc. [8]. It has been noted that a number of researchers preferred the logistic regression technique above other cutting-edge studies to forecast student achievement. Linear regression is another prediction method that predicts data. Various characteristics such as marks, age, and height can be predicted using this technique. According to the author [9], the majority of researchers employed logistic regression followed by support vector machine. In EDM, association rules are used to discover surprising and powerful rules from education sector-based datasets using "support" and "confidence" as pre-built measures. In association rule mining, the apriori algorithm is a well-known technique [10]. The group of items is classified using the K-mean technique on the k-number of attributes known as the k-mean clustering algorithm [11]. The steps in this algorithm are, first to find the centroid, second, determine the object's spacing at the center point, then cluster the objects based on the shortest distance between them. Another supervised technique is the decision tree. This technique is used in regression and classification problems. The decision tree has the structure of a flowchart where the leaf (inner) node represents the class label. This algorithm has the simplest techniques for solving real-life problems.

Feature selection is one of the important and challenging phases before model testing. Feature selection strategies such as Chi-Square Based Feature Evaluation (CSBFE), Correlation Based FS (CBFS), Information Gain Attribute Evaluation (IGATE), Info Gain, Chi-squared, Gain Ratio, Filter Feature Selection (FFS), Pearson Correlation (PC), Principal Components (PCo), Relief Weighted Based Algorithm (RWBA), Mutual Information (MI), Weight Based Feature Selection (WBFS), and Hybrid Feature Selection Framework (HFSF) were employed in cutting-edge research investigations. The Fast Correlation-Based Filter (FCBF) is the technique used by the many researchers. According to the researchers, the article was a first step in figuring out what influences students' academic achievement [8]. A set of data of 309 students' was defined in an experiment [12]. The collection of data was used to forecast the students' marks for the final exams of the term.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
***Article Received:*** *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

To help students forecast their performance in getting accepted into an engineering program, the researchers used classifier method for the research [13]. Before employing classification techniques, feature selection strategies were suggested to pinpoint the most important and essential characteristics [14]. The performance of filter feature selection techniques and classifiers on two independent student data records was examined [15]. According to the researchers, in the future, researchers will also be aided in their search for the ideal combination of filter feature selection algorithms and classifiers by the results produced from various feature selection algorithms and classifiers on two student data records with various numbers of features. To find the important characteristics that are related to the class label and forecast academic achievement, a hybrid feature selection framework (HFSF) using feature fusion was developed [16]. A number of practical feature selection techniques have been suggested to analyse student academic performance for choosing effective features in order to solve the dimensionality problem. According to the researchers [17], the authors identify relevant and irrelevant features that improve or decrease student academic performance by using Pearson correlation feature selection method using different machine learning techniques including logistic regression. The authors [18], studied the elements affecting students' academic success. Based on features such as course-category, attendance, grade, gender, school-type, GPA, and delivery mode, a fuzzy-neural (FNN) technique is used to develop a model that predicts and explains variances in grades across students. In one study [19], the author studied the relationship between the variables. The data analysis was done using the SPSS 22 programme. The hierarchical regression analysis and Pearson Product-Moment (PPM) correlation were employed to ascertain the associations between variables in addition to descriptive statistical methods.

Additionally, the author examined how the family support variable affected the math skills of third-grade students at a public elementary school in Rizal, the Philippines. For this, the results were evaluated using various statistical techniques (STATS), including frequency-counts, percentages, averages, weighted averages, and T-tests [20]. In order to enhance students' learning experiences and reduce the dropped-out ratio [21], the authors tried to find at-risk students early on by suggesting a prediction framework based on data mining. Various data mining techniques such as K-Neural Network, Decision Tree with gain ratio, Naïve Bayes, Random Forest with gain ratio with logistic regression were employed. Table I depicts the feature selection techniques used by state-of-the-art research studies.

TABLE I. STATE-OF-THE-ART FEATURE SELECTION TECHNIQUES

| References | FS Techniques | Highest Accurate Technique |
|---|---|---|
| [8] | FCBF | FCBF |
| [12] | CBFS, CSBFE, IGATE, WBFS | CBFS |
| [13] | Chi-square, Info-Gain, Gain-Ratio, FCBF | FCBF |
| [14] | Chi-square, PC, and MI | MI |
| [15] | CBFS, CSBFE, FFS, Gain-Ratio, PCo, RWBA | CBFS |
| [16] | HFSF using feature fusion | HFSF |
| [17] | Pearson correlation | Pearson correlation |
| [18] | FNN | FNN |
| [19] | PPM correlation | PPM |

## III. MATERIALS AND METHODS

The effects of dimensionality reduction techniques on the prognosis of the student dataset were investigated in this study. In feature selection system framework (Figure 1), generally, feature selection methods divided into three groups: filter, wrapper, and embedded methods. The reduced feature subset was produced after the implementation of feature selection algorithms. The following is a detailed explanation of various techniques.

### *A. Filter methods*

Filter methods are commonly used in pre-processing step. In computation, this method is very fast, inexpensive, and effective for removing duplicate and redundant features from the dataset. Two sub-categories in this method are, statistical (stats) and feature importance methods. Each source variable's relationship to the target feature is evaluated using statistical analysis, and the source variable with the strongest correlation to the target feature is selected. The chi-squared test, pearson correlation, variance threshold, information gain, fisher score, correlation-based selection, mutual information, and reliefF are a few examples.

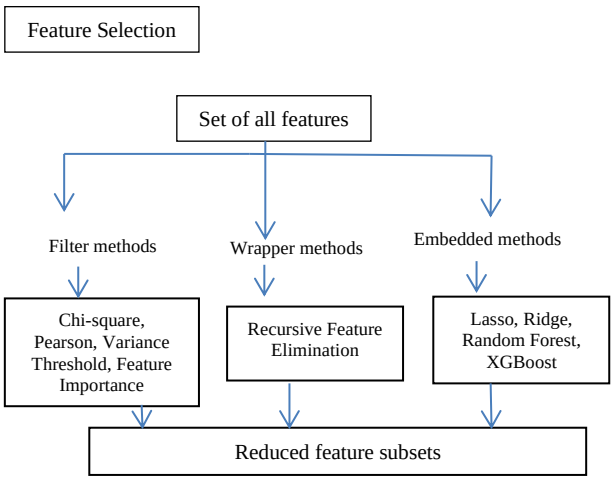


Figure 1. Feature selection system framework.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
***Article Received:*** *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

- *Chi-square:* To ascertain if a mismatch between observed and predicted data is the product of chance or a connection between the variables under inquiry, the chi-square test contrasts observed and expected findings.
- *Pearson Correlation:* To determine how closely two variables are related to one another, correlation coefficients are employed. Pearson's correlation, which is frequently used for linear regression, is one of the most well-known formulas for calculating correlation coefficients.
- *Variance Threshold:* A straightforward fundamental method for feature selection is known as variance threshold. All features that have variances below a particular threshold are removed using this technique.
- *Feature Importance:* In the feature importance method, the list of features that the system deems important is considered. Every attribute is given a significant value that indicates how important that attribute is for the results.

### *B. Wrapper methods*

Another method in supervised feature selection is the wrapper method. The wrapper method is defined as greedy algorithms that train the model by using a subset of parameters in the repetitive form then the addition and deletion of parameters process take place. The RFE (Recursive Feature Elimination) algorithm is one that is employed in this technique.

- *RFE:* This algorithm eliminates the weak attribute (or attributes) from a system until the required set of attributes is obtained.

### *C. Embedded methods*

The embedded method has its built-in parameter selection methods. This method is very popular among researchers because it eliminates the previous two (filter and wrapper) supervised feature selection methods' drawbacks and has its advantages. This methodology has two subcategories: basic and trees approach. Regularisation, random forest, gradient boosting, L1 (LASSO), ridge, and XGBoost are a few examples. By including additional data and lowering the size of the variables, regularisation is a strategy used to reduce the over-fitting or under-fitting situation. Gradient boosting is another method used in the embedded feature selection methodology where each model tries to fix the flaws in the prior process. This technique, also referred to as a gradient boosting machine (GBM), grows progressively.

- *Lasso:* The coefficients' absolute sum is represented by lasso phrase. This phrase penalises the model, causing it to reduce the value of the coefficients in order to lessen loss, when the coefficient value increases from 0 to 1.
- *Ridge:* In Ridge regression, authors include a penalty term that has the same value as the square root of the coefficient. The magnitude of the coefficients squared is what determines the L2 term. To regulate that penalty term, researchers additionally include the coefficient lambda.
- *Random Forest:* Another technique in the embedded method is random forest. Up to 1200 decision trees can be used in the random forest technique, and each tree step is assembled in an arbitrary way to extract the findings from the microarray data and extract the feature sets. Each tree in a random forest is a series of yes-and-no issues based on a single or several feature sets.
- *XGBoost:* XGBoost is a distributed gradient boosting library that has been optimized for quick and scalable machine learning model training. One of XGBoost's important strengths is its effective handling of missing values, which enables it to handle actual data with missing values without the need for a lot of pre-processing.

Intuitionally, the most recently taken arena for feature selection study is a hybrid approach. To identify the optimal feature set, several researchers investigated numerous combinations of feature selection techniques. In hybrid strategies, the advantages of the filter and wrapper approaches are merged. In hybrid methods, after employing one or more filter-based approaches initially, the best feature subset is chosen using the wrapper strategy. The output of hybrid methods can occasionally outperform that of independent ones.

The performance was evaluated using accuracy, $r^2$, precision, recall, and F1 score.

- *Accuracy:* Accuracy is described as the proportion of all forecasts made to the number of accurate predictions made. With the aid of the following formula, researchers can calculate the accuracy using the confusion matrix.

$$\frac{\textbf{True Positive + True Negative}}{\textbf{True Positive + False Positive + True Negative + False Negative}}$$

- *$R^2$:* A set of predicted output values and actual output values are typically compared using the R Squared metric, which shows how well the predicted and actual values fit together.

$$\mathbf{R^2 = 1 - (SS_{res} / SS_{tot})}$$

$SS_{res}$= residual sum of square

$SS_{tot}$ = total sum of square

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
***Article Received:*** *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

- *Precision:* The quantity of accurate data retrieved by the model is known as precision, which is employed in data retrieval. With the aid of the following formula, researchers can calculate the precision using the confusion matrix.

$$\frac{\textbf{True Positive}}{\textbf{True Positive + False Positive}}$$

- *Recall:* The number of positive results that the model returns is known as recall. Researchers can calculate the recall using the confusion matrix with the help of the following formula. Another name for recall is sensitivity.

$$\frac{\textbf{True Positive}}{\textbf{True Positive + False Negative}}$$

- *F1 score:* Researchers can calculate the harmonic mean of recall and precision by using the F1 score. The precision and recall are weighted and averaged to create the F1 score. Using the following formula, researchers can determine the F1 score.

**F1= 2*(precision * recall) / (precision + recall)**

## IV. SYSTEM FRAMEWORK

In this experiment, dimensionality reduction techniques were used to reduce the number of features from 70 to the optimal 21 feature subsets for analysis of student performance prediction. The proposed study aims to improve classification accuracy by reducing the number of features in a student dataset. In Figure 2, the study's framework is depicted. Data collection, data preprocessing, feature selection, data splitting, classifier-based model training, and model evaluation are the framework's main constituents. The fundamental components of the system framework are explained in the ensuing sections.

### *A. Acquisition Of Dataset*

Through the use of a Google Form, we created a questionnaire that we distributed to students at state colleges. The acquired student dataset (Table II) has 142 records and 70 features.

### *B. Data preprocessing phase*

The data preprocessing stage is crucial for cleaning the data and preparing it for use in creating the ML model, which will improve the model's accuracy and effectiveness. The four steps listed below contribute to the data preprocessing phase.

- *Data Cleaning:* Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Since every question in the specially constructed questionnaire has the necessary validation, the collected student dataset in this experiment has been manually cleansed and contains no missing values.

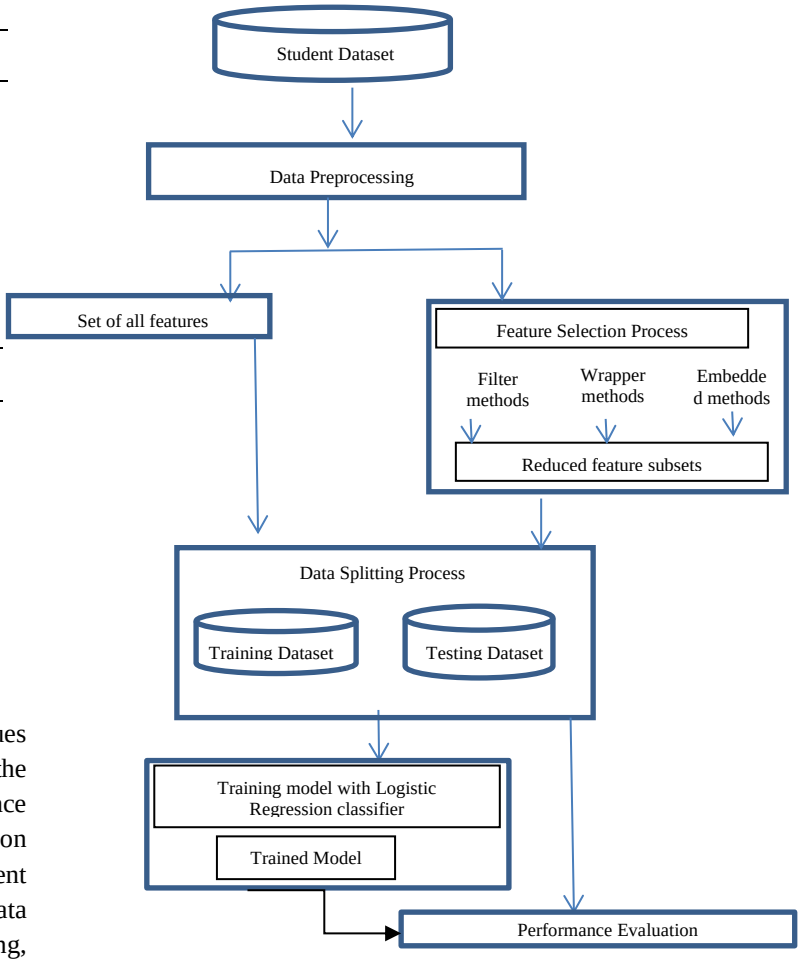


Figure 2. System framework.

TABLE II. ATTRIBUTE DETAILS OF STUDENT DATASET

| Figure | Features Name | Features Description |
|---|---|---|
| F1 | G | Gender |
| F2 | A | Age |
| F3 | SPC | Single Parent Child |
| F4 | CNF | Confident |
| F5 | EMT | Emotional |
| F6 | AT | Admission Type |
| F7 | PT | Part-Time Job |
| F8 | CAC | Co-Curricular Activities In College |
| F9 | ATTD | Attendance |
| F10 | GIQ | Gender Inequality |
| F11 | FAQ | Father Academic Qualification |
| F12 | MAQ | Mother Academic Qualification |

| Figure | Features Name | Features Description |
|---|---|---|
| F13 | FOP | Father Occupation |
| F14 | TI | Total Income |
| F15 | CHD | Number Of Children's |
| F16 | HIC | Hrs. In Coaching |
| F17 | HIS | Hrs. In Social Sites |
| F18 | HIG | Hrs. In Online Video Games |
| F19 | HICF | Hrs. In Café |
| F20 | HISS | Hrs. In Self-Study |
| F21 | HIL | Hrs. In Lab, Library |
| F22 | FCHD | First Child |
| F23 | FSUP | Family Support |
| F24 | FCONF | Family Conflict |
| F25 | FFRD | Family Freedom |
| F26 | FMOV | Family Moral Values |
| F27 | TRPB | Transport Problem |
| F28 | SCH | Scholarship |
| F29 | HOS | Hostler |
| F30 | LMP | Learning Mode Preference |
| F31 | REWC | Relationship With Classmates |
| F32 | RET | Relationship With Teachers |
| F33 | QUAI | Quality Of Institute |
| F34 | CARIMP | Career Importance Awareness |
| F35 | AREL | Area Student Live |
| F36 | ENGMT | Engagement in Class |
| F37 | PGMC | Program Choice Satisfaction |
| F38 | COV_19 | Effect of COVID-19 Pandemic |
| F39 | ILLACT | Involve in Illegal Activities |
| F40 | FRIATTD | Friend Attendance |
| F41 | SIB | Number Of Siblings |
| F42 | FAHT | Family Help, Trust |
| F43 | MOV | Motivated |
| F44 | READ | Relationship With Administrative Staff |
| F45 | QUAPS | Quality Of Pass Out School |
| F46 | MOP | Mother Occupation |
| F47 | COUTP | Course Type |
| F48 | CAT | Category |
| F49 | MARS | Marital Status |
| F50 | DISC | Distance From College |
| F51 | FRIBH | Friend Bad Habits |
| F52 | TRAU | Trauma |
| F53 | LANGB | Language Barrier |
| F54 | CLMP | Current Learning Mode |
| F55 | FRIENG | Friend Engagement in Class |
| F56 | REL | Religion |
| F57 | DIS | Disability |
| F58 | SGCHD | Single Girl Child |
| F59 | FAMTP | Family Type |
| F60 | CAH | Co-Curricular Activities In Home |
| F61 | PROINT | Profession Interest |
| F62 | PALACT | Past Life Activities |
| F63 | EDUIMP | Education Importance Awareness |
| F64 | BHBT | Bad Habits |
| F65 | HLISS | Health Issues |
| F66 | SURR | Surroundings |
| F67 | PLPBLM | Personal Life Problems |
| F68 | LNGBRR | Language Barrier |
| F69 | FRIMOT | Friend Motivation |
| F70 | LAGD | Last year grade (Target variable) |

- *Data Integration:* The merging of data from multiple data stores is called data integration. Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. It might be better understood as a method of combining data from different unrelated sources. A single excel datasheet is created for this study.
- *Data Transformation:* Data transformation is the technological process of translating data collected from a single structure or format to a different one, without altering the dataset's content. This is often done to make the data more usable by researchers and to enhance the accuracy of the data. The student dataset for this study has been translated from an excel file to a .csv file.
- *Data Discretization:* Data discretization is the procedure of grouping together continuous data to create distinct containers. Label encoding in the Python environment is used to convert category columns into numerical ones for machine learning models that only accept numerical data.

### *C. Feature Selection phase*

To evaluate the impact of feature selection techniques, tests were carried out both with and without feature selection at this point. The main goal of feature selection is to extract optimal features subset according to their importance. The selection of features is an extremely important phase since the addition of duplicate and significant features has a considerable detrimental impact on the model's performance. The learning model is enhanced in several ways by choosing pertinent features from the original dataset, including (i) enhancing efficiency, and (ii) eliminating noisy and over-fitting data. In addition, working with more informative features contributes to preventing student failure risk. In this experiment, the effect of several feature selection techniques under the three main categories is evaluated. Three categories of feature selection

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
***Article Received:*** *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

strategies have been used separately in the initial dataset, as shown in figure 1. The feature selection procedure that each of the three categories used to determine the best feature subsets.

### *D. Spilitting of dataset*

To split a dataset into two halves with a ratio of 70:30, this study employed Python's SKlearntrain_test_split function. Furthermore, the classifier's training phase was validated using a 10-fold cross-validation technique.

### *E. Classification*

The data classification technique is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data). Based on the recent literature on student academic performance prediction, the present study implements a logistic regression technique.

## V. RESULTS AND DISCUSSION

The implementation of this experiment took place using the Jupyter Notebook with Python 3.8.14 and was run on a laptop with specifications (Windows 7, Intel Core I7-5500U, and 32 GB of RAM). The best feature subsets for classification were chosen using nine feature selection algorithms. The student dataset was primarily collected from the state colleges of Madhya Pradesh through Google Form. The dataset contains 70 features as shown in Table II. This section is divided into two sub-sections i.e., results before applying feature selection techniques and results after applying nine feature selection techniques.

### *A. Results before experiment*

Prior to feature selection, the student dataset is examined using the most popular logistic regression method. The experiment's results were shown in Table III. As per the analysis's findings, the original dataset of 70 features was able to attain 65% accuracy, 65.5% precision, 65.5% recall, 66% f-score, and -0.526% r square using the logistic regression classifier. Figure 3 represents the performance metrics before feature selection.

TABLE III. MACHINE LEARNING TECHNIQUE PERFORMANCE BEFORE FEATURE SELECTION

| Technique | $R^2$ | Precision | Recall | F-Score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | -0.526 | 0.655 | 0.655 | 0.66 | 0.65 |

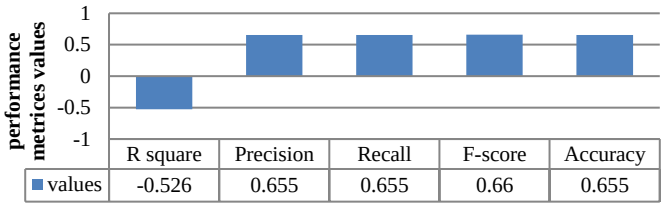


Figure 3. Performance metrics before feature selection.

### *B. Results after experiment*

After analysing the dataset result before applying any feature selection techniques, nine feature selection strategies from three fundamental feature selection categories are used in this experiment. These approaches were selected from the feature selection techniques' subcategories, which include statistics, feature importance, recursive elimination, basic techniques, and tree methods.

The chi-square, ridge, and random forest techniques all showed a considerable improvement in accuracy after applying all feature selection methods. The optimal accuracy performance of each feature selection method is compiled in Table IV. With the logistic regression classifier, the feature subset chosen by the chi-square feature selection technique has the best classification accuracy of 72.4%, precision of 71.4%, recall of 72.4%, f-score of 70%, and r square of -0.221%.

Figure 4-5-6-7-8 depicts the model evaluation of each feature selection techniques. Table V displays the features that were chosen based on these nine strategies.

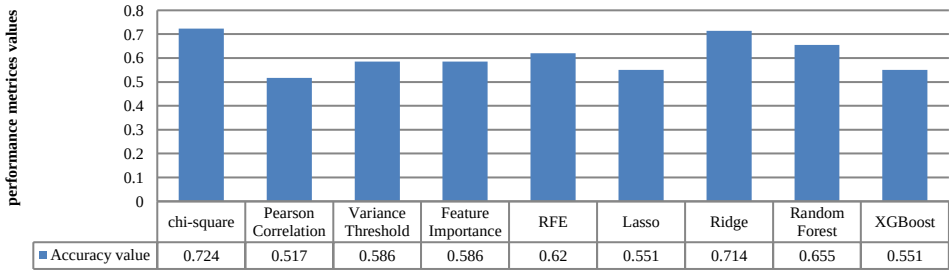


Figure 4. Accuracy values.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
***Article Received:*** *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

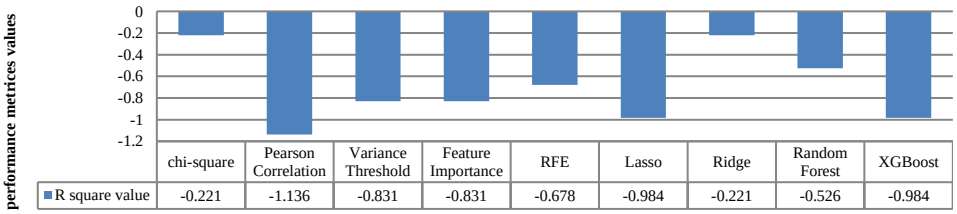


Figure 5. R Square values.

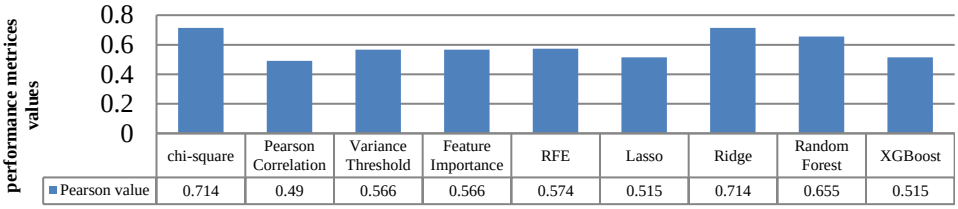


Figure 6. Precision values.

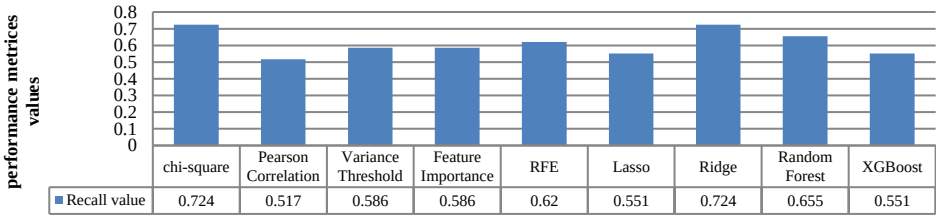


Figure 7. Recall values.

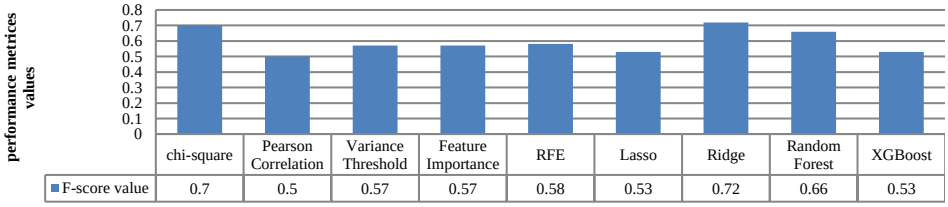


Figure 8. F - score values.

TABLE IV. COMPARATIVE ANALYSIS OF FEATURE SELECTION TECHNIQUES

| **Categories** | **Sub-Categories** | **Techniques** | **$R^2$** | **Precision** | **Recall** | **F-Score** | **Accuracy** |
|---|---|---|---|---|---|---|---|
| Filter | ***Statistics*** | *Chi-square* | -0.221 | 0.714 | 0.724 | 0.70 | 0.724 |
| | | *Pearson* | -1.136 | 0.49 | 0.517 | 0.50 | 0.517 |
| | | *Variance threshold* | -0.831 | 0.566 | 0.586 | 0.57 | 0.586 |
| | ***Feature importance*** | *Feature importance* | -0.831 | 0.566 | 0.586 | 0.57 | 0.586 |
| Wrapper | ***Recursive Feature Elimination*** | *RFE* | -0.678 | 0.574 | 0.620 | 0.58 | 0.620 |
| Embedded | ***Basic Techniques*** | *Lasso* | -0.984 | 0.515 | 0.551 | 0.53 | 0.551 |
| | | *Ridge* | -0.221 | 0.714 | 0.724 | 0.72 | 0.714 |
| | ***Trees*** | *Random forest* | -0.526 | 0.655 | 0.655 | 0.66 | 0.655 |
| | | *XGBoost* | -0.984 | 0.515 | 0.551 | 0.53 | 0.551 |

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 10*
***Article Received:*** *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

This study focuses on the top 21 ideal features out of 70 features that increase the accuracy of the model for student academic achievement. For each technique, the top 21 optimal features are listed in Table V. In this Table, "1" stands for the most optimal features, while "0" stands for the least optimal features, as determined by the procedure. These optimal features have the greatest impact on student academic performance prediction models. The graph of the relevance score is shown in Figure 9.

According to the final findings (Table VI), it is concluded that demographic factors such as gender, age, etc. have a positive relationship with academic performance, followed by social factors such as gender inequality, part-time job, etc., psychological factors such as self-confidence, emotional, etc., and economic factors like total income, father's occupation, etc. Furthermore, among the nine strategies, the chi-square method has the highest accuracy.

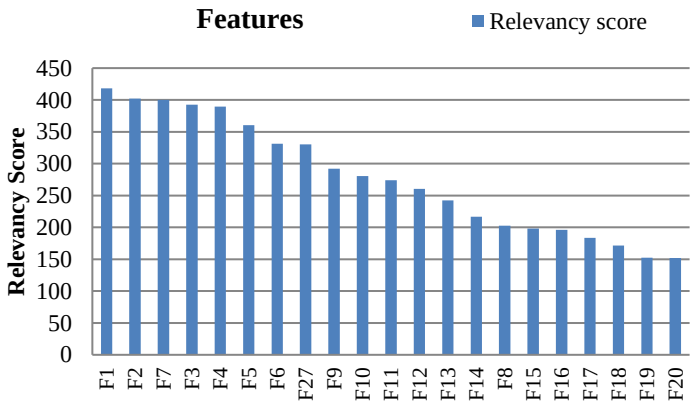


Figure 9. Feature relevancy score.

TABLE V. TOP OPTIMAL FEATURES SET ACCORDING TO EACH TECHNIQUE

| Features | Chi-square | Pearson | Feature importance | RFE | Lasso | Ridge | Random Forest | XGBoost | Variance Threshold |
|---|---|---|---|---|---|---|---|---|---|
| F1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| F2 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| F3 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| F4 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| F5 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| F6 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| F7 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| F8 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| F9 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| F10 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| F11 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| F12 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| F13 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| F14 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| F15 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| F16 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| F17 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| F18 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| F19 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| F20 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| F21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| F22 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| F23 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| F24 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| F25 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |
| F26 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| F27 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| F28 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| F29 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| F30 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| F31 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| F32 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| F33 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| F34 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| F35 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| F36 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| F37 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| F38 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| F39 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| F40 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| F41 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| F42 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| F43 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| F44 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| F45 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| F46 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| F47 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| F48 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| F49 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| F50 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| F51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F52 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| F53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F54 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

TABLE VI. FINAL SUBSET OF FEATURES THAT ACHIEVES HIGHEST ACCURACY

| Figures | Features Name | Relevancy score |
|---|---|---|
| F1 | G | 418.452 |
| F2 | A | 402.221 |
| F7 | PT | 400.199 |
| F3 | SPC | 392.780 |
| F4 | CNF | 389.455 |
| F5 | EMT | 360.221 |
| F6 | AT | 331.489 |
| F27 | TRPB | 330.112 |
| F9 | ATTD | 291.890 |
| F10 | GIQ | 280.662 |
| F11 | FAQ | 273.869 |
| F12 | MAQ | 260.324 |
| F13 | FOP | 242.310 |
| F14 | TI | 216.563 |
| F8 | CAC | 202.412 |
| F15 | CHD | 198.321 |
| F16 | HIC | 195.990 |
| F17 | HIS | 183.324 |
| F18 | HIG | 171.330 |
| F19 | HICF | 152.212 |
| F20 | HISS | 151.673 |

### *C. Other observations*

It was also noted in this experiment that 14 of the 70 features were not deemed significant by any of the feature selection strategies. Table VII provides a list of these features.

TABLE VII. LIST OF IRRELEVANT FEATURES

| Figures | Features Name |
|---|---|
| F57 | DIS |
| F58 | SGCHD |
| F59 | FAMTP |
| F60 | CAH |

| | |
|---|---|
| F38 | COV_19 |
| F61 | PROINT |
| F62 | PALACT |
| F63 | EDUIMP |
| F64 | BHBT |
| F65 | HLISS |
| F66 | SURR |
| F67 | PLPBLM |
| F68 | LNGBRR |
| F69 | FRIMOT |

## VI. CONCLUSION AND FUTURE WORK

This paper presents a comparative analysis-based feature selection for predicting students' academic grades using micro-array datasets. The number of features in this experiment was reduced from 70 to the ideal 21 feature subsets for the investigation of student performance prediction using dimensionality reduction techniques. To ascertain the impact of feature selection techniques, experiments were conducted both with and without feature selection. The feature selection techniques used were Chi-square, Pearson Correlation, Variance Threshold, Feature Importance, Recursive Feature Elimination (RFE), Lasso, Ridge, Random Forest, and XGBoost. Logistic regression was the only classification approach used for the analysis.

The result utilising the logistic regression classifier yield s 65% model accuracy without feature selection; this value was increased to 72.4% using chi-square feature selection and a logistic classifier. From the results, it is concluded that demographic factors are positively related to academic performance followed by social, psychological and economic factors. The chi-square method has also been demonstrated to be the most promising of all methods. In the future, this study approach can be modified and tested on larger, more feasible datasets for benchmarking purposes. In addition, other machine learning classifiers should be investigated. Additionally, the implementation of the proposed model should be investigated using tools such as R and Hadoop.

**Acknowledgment**

I immensely express my gratitude and appreciation to my supervisor, Dr. Punam Rattan, Faculty School of Computer Application, Lovely Professional University, Punjab, India, a great role model. Her encouragement and guidance allowed me to perform to my best potential.

**Declarations**

**Funding:** The author(s) received no financial support for the research, authorship, and/or publication of this article.

**Conflict of interest:** No relevant financial or non-financial competing interests to report.

**Ethics approval:** Not applicable.

**Consent to participate:** Not applicable.

**Consent for publication:** Not applicable.

**Availability of data and material:** Not applicable.

**Code Availability:** Not applicable.

**Authors' Contribution:** Kajal Mahawar performed the analysis of the research concerns and was a major contribution in writing the manuscript. Punam Rattan helped to find out the relevant machine learning technique to perform the analysis. All authors read and approved the final manuscript.

## REFERENCES

[1] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting Students Performance in Educational Data Mining," Proceedings - 2015 International Symposium on Educational Technology, ISET 2015, pp. 125–128, 2015, doi: 10.1109/ISET.2015.33.

[2] Y. Shi, J. Zhang, H. Yang, H. H. Yang, and Y. Shi, "Effects of Interactive Whiteboard-based Instruction on Students ' Cognitive Learning Outcomes : A Meta-Analysis," Interactive Learning Environments, vol. 0, no. 0, pp. 1–18, 2020, doi: 10.1080/10494820.2020.1769683.

[3] H. Yang, J. Cai, H. Hao, and X. Wang, "Examining key factors of beginner ' s continuance intention in blended learning in higher education," Journal of Computing in Higher Education, vol. 35, no. 1, pp. 126–143, 2023, doi: 10.1007/s12528-022-09322-5.

[4] J. Macleod, H. H. Yang, S. Zhu, and Y. Shi, "Technological Factors and Student-to-Student Connected Classroom Climate in Cloud Classrooms," Journal of Educatonal Computing Research, 2017, doi: 10.1177/0735633117733999.

[5] B. K. Yousafzai, M. Hayat, and S. Afzal, "Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student," Education and Information Technologies, 2020.

[6] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction : a review," Complex & Intelligent Systems, vol. 8, no. 3, pp. 2663–2693, 2022, doi: 10.1007/s40747-021-00637-x.

[7] E. Mahsereci, S. Ay, and İ. Turgay, "A comparative study on the effect of feature selection on classification accuracy," Procedia Technology, vol. 1, pp. 323–327, 2012, doi: 10.1016/j.protcy.2012.02.068.

[8] M. Zaffar, M. A. Hashmani, K. S. Savita, S. S. H. Rizvi, and M. Rehman, "Role of FCBF Feature Selection in Educational Data Mining," Mehran University Research Journal of Engineering and Technology, vol. 39, no. 4, pp. 772–778, 2020, doi: 10.22581/muet1982.2004.09.

[9] Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/4151487.

[10] L. Khanna, S. N. Singh, and M. Alam, "Educational data mining and its role in determining factors affecting students academic performance: A systematic review," India International Conference on Information Processing, IICIP 2016 - Proceedings, 2016, doi: 10.1109/IICIP.2016.7975354.

[11] R. Saxena, "Educational Data Mining : Performance Evaluation of Decision Tree and Clustering Techniques Using WEKA Platform," International Journal of Computer Science and Business Informatics, vol. 15, no. 2, 2015.

[12] A. Acharya and D. Sinha, "Application of Feature Selection Methods in Educational Data Mining," International Journal of Computer Applications, vol. 103, no. 2, pp. 34–38, 2014, doi: 10.5120/18048-8951.

[13] M. Doshi and S. K. Chaturvedi, "Correlation Based Feature Selection (CFS) Technique to Predict Student Perfromance," International journal of Computer Networks & Communications, vol. 6, no. 3, pp. 197–206, 2014, doi: 10.5121/ijcnc.2014.6315.

[14] W. Punlumjeak, N. Rachburee, and J. Arunrerk, "Big data analytics: Student performance prediction using feature selection and machine learning on microsoft azure platform," Journal of Telecommunication, Electronic and Computer Engineering, vol. 9, no. 1–4, pp. 113–117, 2017.

[15] M. Zaffar, M. A. Hashmani, K. S. Savita, and S. S. H. Rizvi, "A study of feature selection algorithms for predicting students academic performance," International Journal of Advanced Computer Science and Applications, vol. 9, no. 5, pp. 541–549, 2018, doi: 10.14569/IJACSA.2018.090569.

[16] M. Zaffar et al., "A hybrid feature selection framework for predicting students performance," Computers, Materials and Continua, vol. 70, no. 1, pp. 1893–1920, 2021, doi: 10.32604/cmc.2022.018295.

[17] N. Robert et al., "Determining factors that affect student performance using various Determining factors that affect student performance using various machine learning methods machine learning methods," Procedia Computer Science, vol. 216, no. 2022, pp. 597–603, 2023, doi: 10.1016/j.procs.2022.12.174.

[18] M. A. Naaj, R. Mehdi, and E. A. Mohamed, "Analysis of the Factors Affecting Student Performance Using a Neuro-Fuzzy Approach," education sciences, 2023.

[19] M. E. Işikgöz, "Analysis of the Relationship between Online Learning Activities and Academic Achievement of Physical Education and Sports School Students," The Turkish Online Journal of Educational Technology, vol. 22, no. 1, 2023.

[20] I. D. Juguilon, "Impact of Family Support System in the Academic Performance of Grade 3 Pupils at a Public Elementary School in Rizal, Philippines," INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY: APPLIED BUSINESS AND EDUCATION RESEARCH, vol. 4, no. 1, pp. 174–187, 2023, doi: 10.11594/ijmaber.04.01.16.

[21] K. Mahboob, R. Asif, and N. Ghani, "Quality enhancement at higher education institutions by early identifying students at risk using data mining," Mehran University Research Journal of Engineering and Technology, vol. 42, no. 1, pp. 120–136, 2023.