Predicting Math Performance in High School Students using Machine Learning Techniques

Yuan hui
Wuchang Institute of Technology, School of Information Engineering, China

***Abstract*—In the field of education, understanding and predicting student performance plays a crucial role in improving the quality of system management decisions. In this study, the power of various machine learning techniques to learn the complicated task of predicting students' performance in math courses using demographic data of 395 students was investigated. Predicting students' performance through demographic information makes it possible to predict their performance before the start of the course. Filtered and wrapper feature selection methods were used to find 10 important features in predicting students' final math grades. Then, all the features of the data set as well as the 10 selected features of each of the feature selection methods were used as input for the regression analysis with the Adaboost model. Finally, the prediction performance of each of these feature sets in predicting students' math grades was evaluated using criteria such as Pearson's correlation coefficient and mean squared error. The best result was obtained from feature selection by the LASSO method. After the LASSO method for feature selection, the Extra Tree and Gradient Boosting Machine methods respectively had the best prediction of the final math grade. The present study showed that the LASSO feature selection technique integrated with regression analysis with the Adaboost model is a suitable data mining framework for predicting students' mathematical performance.**

Keywords—Student performance; math grade prediction; feature selection; regression analysis; machine learning; data mining

I. Introduction

In the field of education, understanding and predicting student performance plays a crucial role in improving the quality of system management decisions. Through the utilization of machine learning methodologies, educators and administrators may effectively utilize data to detect pupils who may be prone to failure right from the outset of the course. By acting as an early warning system, these predictive models enable the implementation of focused support measures and intervention techniques to enhance student learning outcomes [1]. Machine learning algorithms and data mining techniques are commonly utilized in student performance prediction modeling [2]. These techniques analyze various attributes such as grades, educational background, psychological evaluation, and demographics to generate predictions about a student's future performance [3]. By utilizing machine learning techniques, educators can gain valuable insights into student behavior and patterns, allowing them to tailor their approach to meet individual students' needs. This not only improves student performance but also helps in identifying slow learners, predicting dropout rates, and enhancing overall educational outcomes. These predictive models help in improving the overall education system by identifying students who may require additional support and intervention. Additionally, machine learning techniques can help in improving student attendance and predicting learning behavior to warn students who are at risk [4]. Machine learning techniques have revolutionized the field of education by providing accurate and timely predictions about student performance [5].

The process of extracting valuable embedded information from data plays a crucial role in various scenarios, providing valuable insights to organizations, companies, and research analysts for addressing different challenges and making informed decisions [6], [7]. The present investigation looks at the challenge of evaluating math students' performance with respect to important variables that have a big influence on the possibility of repeating the course. This study examines the application of several machine-learning approaches for data mining, taking into account the diversity of factors impacting students' success or failure in a course [8]. By mining the available data, the aim is to determine the relative importance of different factors in students' academic achievements.

Several researchers have explored the field of data mining related to students' performance. Kostopoulos et al. [1] proposed a new semi-supervised regression algorithm to predict the final grade of students in an online course. They showed that their technique can improve the performance of student performance prediction models. Ranđelović et al. [2] proposed a multidisciplinary-applicable aggregated model based on analytic hierarchy process and ReliefF classifier to predict further students' education. Xu et al. [8] introduced a two-layer machine learning architecture consisting of multiple base predictors and a set of ensemble classifiers to predict student performance in degree programs. They proposed a data-driven approach based on probability matrix factorization and latent factor models to construct baseline predictors. Through extensive simulations on an undergraduate student dataset collected over three years at University of California, they showed that this technique may achieve superior performance over benchmark approaches. However, none of the mentioned studies managed to identify important factors in student performance. The decision tree's (DT) ID3 variation method was used in one study by Baradwaj and Pal [9] to forecast end-semester marks (ESM). Previous semester marks (PSM), seminar performance (SEP), assignment (ASS), class test grade (CTG), attendance (ATT), lab work (LW), and general proficiency (GP) were among the considerations. Through the implementation of the DT method, a set of IF-THEN rules were derived to predict students' ESM categorized as first, second, third, or fail. Kabakchieva [10] used a variety of algorithms in her study to estimate students' performance based on data obtained, including rule learners, K-nearest neighbors, DTs, and Bayesian classifiers. The results demonstrated that although these classifiers were suitable for the data mining task, none of the methods or classifiers achieved an accuracy of more than 75%, which is subpar considering how crucial it is to predict students' performance. The effectiveness of artificial neural networks and deep learning models (DTs) in simulating the academic standing of students of Nigeria's University of Ibadan, was examined in different research conducted by Osofisa et al [11]. The results showed that in terms of training and test data accuracy, the neural network model performed better than the DT model. 98.26% and 60.16%, respectively, for the training and test data were the classification accuracies of the multilayer perceptron model, which demonstrated the best performance. Roy et al. [12] investigated an adaptive dimensionality reduction algorithm for educational data mining. They showed that this algorithm can improve the performance of predictive models and provide useful insights into the important factors affecting student performance. However, the authors compared the proposed algorithm with some limited existing algorithms and were not able to introduce important factors affecting student performance.

In general, few studies have used a variety of data mining and machine learning methods to predict students' performance, and mostly limited artificial intelligence techniques have been investigated. Some existing studies have only reported the accuracy of classification using neural networks as a black box and have not investigated the important factors in predicting students' performance. Therefore, the current study aims to systematically examine and compare various filtered and wrapper data mining methods to determine important factors in predicting students' performance. For this purpose, a variety of filtered-based, L1- and L2-based, tree-based, and evolutionary-based methods are examined to predict students' performance. This study looked into the ability of several machine learning approaches to learn the challenging job of predicting students' success in math classes using 395 students' demographic data. Predicting students' performance through demographic information makes it possible to predict their performance before the start of the course. The article is arranged as follows: Section II presents the dataset used and the proposed framework. Section III presents the experimental results. Section IV provides a discussion of the findings and Section V provides a conclusion on the study.

II. Methods

A. Dataset

This information relates to the secondary school academic performance of two Portuguese schools [12]. The information was gathered through school reports and surveys, and its properties include student grades as well as demographic, social, and school-related information. A total of 395 students filled the questionnaires and the data set has no missing values. This dataset has 32 attributes which are shown in Table I. As shown, the variable G3, i.e. the final grade, is considered as the target variable, which is tried to be predicted by other variables.

TABLE I. 32 Attributes of the Students' Performance Dataset

Attributes	Type	Value	Description
School	Binary	GP/MS	Student's school
Sex	Binary	F/M	Student's sex
Age	Numeric	15-22 years	Student's age
Address	Binary	U/R	Student's home address type
Pstatus	Binary	T/A	Parent's cohabitation status
Famsize	Binary	LE3/GT3	Family size
Medu	Numeric	0-4	Mother's education
Fedu	Numeric	0-4	Father's education
Fjob	Nominal	Teacher, health services, at home, other	Father's job
Mjob	Nominal	Teacher, health services, at home, other	Mother's job
Guardian	Nominal	Mother, father, other	Student's guardian
Reason	Nominal	Home, reputation, course, other	Reason to choose this school
Traveltime	Numeric	1-4	Home to school time arrival
Studytime	Numeric	1-4	Weekly study time
Failures	Numeric	$1 \le n < 3$ else 4	Number of past class failures
Famsup	Binary	Yes/No	Family educational support
Schoolsup	Binary	Yes/No	Extra educational support
Nursery	Binary	Yes/No	Attended nursery school
Activities	Binary	Yes/No	Extra-curricular activities
Paid	Binary	Yes/No	Extra paid classes within the course subject
Internet	Binary	Yes/No	Internet access at home
Higher	Binary	Yes/No	Wants to take higher education
Romantic	Binary	Yes/No	With a romantic relationship
Freetime	Numeric	1-5	Free time after school
Famrel	Numeric	1-5	Quality of family relationships
Dalc	Numeric	1-5	Workday alcohol consumption
Goout	Numeric	1-5	Going out with friends
Walc	Numeric	1-5	Weekend alcohol consumption
Health	Numeric	1-5	Current health status
Absences	Numeric	0-93	Number of school absences
G1	Numeric	0-20	First-period grade
G2	Numeric	0-20	Second-period grade
G3	Numeric	0-20	Final grade (Target)

B. Proposed Framework

The proposed framework for math performance prediction using various machine-learning methods is shown in Fig. 1. As shown, at first, filtered and wrapper feature selection methods were used to find 10 important features in predicting students' final math grades. Then, all the features of the data set as well as the 10 selected features of each of the feature selection methods were used as input for the regression analysis with the Adaboost model. Finally, the prediction performance of each of these feature sets in predicting students' math grades was evaluated using criteria such as Pearson's correlation coefficient and mean squared error. In the following, each of the feature selection and regression analysis methods used will be briefly described.

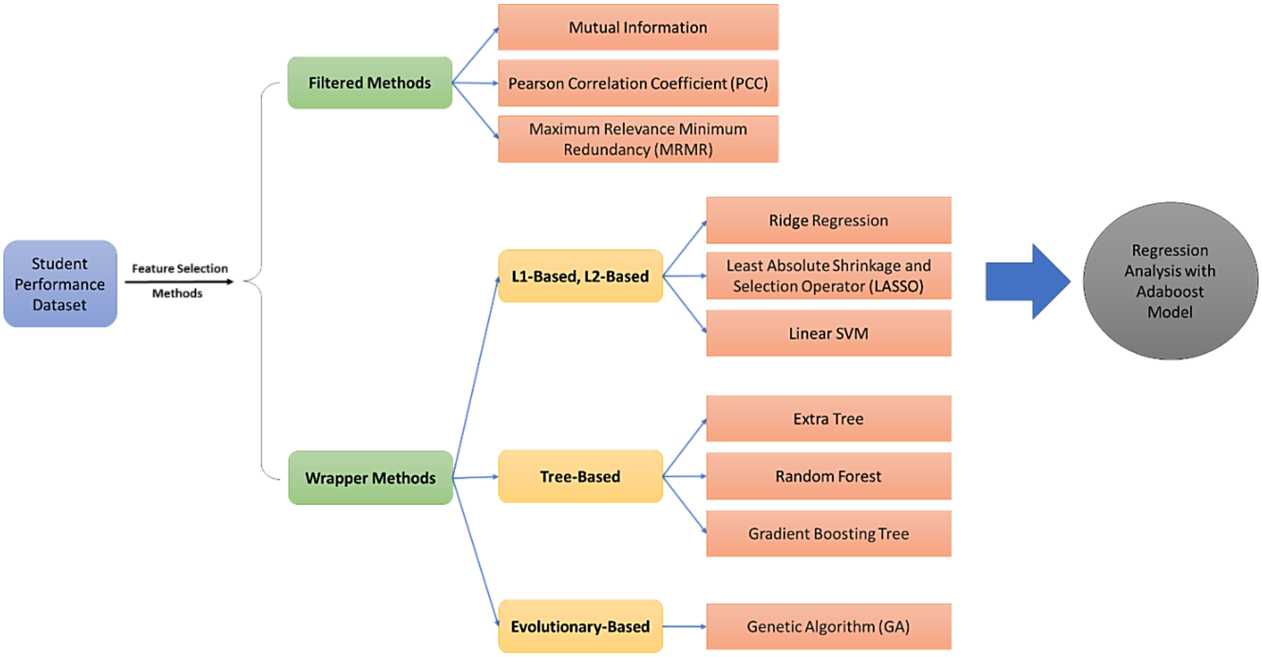
Student Performance Dataset
Feature Selection Methods
Filtered Methods
Mutual Information
Pearson Correlation Coefficient (PCC)
Maximum Relevance Minimum Redundancy (MRMR)
Wrapper Methods
L1-Based, L2-Based
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Linear SVM
Tree-Based
Extra Tree
Random Forest
Gradient Boosting Tree
Evolutionary-Based
Genetic Algorithm (GA)
Regression Analysis with Adaboost Model

Fig. 1. Proposed framework for math performance prediction using various feature selection methods.

C. Filtered Feature Selection Approaches

Feature selection methods, known as filtered methods, choose features based on their performance measure without considering the specific data modeling algorithm used. Once the optimal features are identified, they can be utilized by the modeling algorithms. Filtered approaches have the capability to assess individual features' rankings or evaluate entire subsets of features [13]. The information, consistency, distance, statistical metrics, and similarity that were established for feature filtering may all be generally classified into these categories [14]. In this research, three filtered feature selection methods were utilized: mutual information, Pearson correlation coefficient (PCC), and maximum relevance minimum redundancy (MRMR).

Mutual Information. The mutual information feature selection method is a technique used to evaluate the relevance between features and the target variable. It measures the amount of information that two variables share, indicating their dependency and the potential of a feature to contribute useful information for the prediction task. This method calculates the mutual information score for each feature by considering both its individual information content and its relationship with the target variable. Features with high mutual information scores are considered more informative and are selected for further analysis or model building. By focusing on the information shared between features and the target, the mutual information feature selection method aids in identifying the most relevant features and improving the overall performance of machine learning algorithms [15].

$$I(X,Y) = \sum\sum p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1}$$

where, I (X, Y) represents the mutual information between variables X and Y, p (x, y) denotes the joint probability distribution function of X and Y, p(x) and p(y) represent the marginal probability distribution functions of X and Y, respectively.

Pearson correlation coefficient (PCC). It is a widely used feature selection method in statistics and machine learning. It measures the linear relationship between two variables, typically a feature and a target variable. The PCC calculates the strength and direction of the linear association by calculating the covariance of the variables divided by the product of their standard deviations. A PCC value close to +1 indicates a strong positive correlation, while a value close to -1 suggests a strong negative correlation. Feature selection using PCC involves selecting the features with the highest absolute PCC values, as they are considered more informative for predicting the target variable. This method helps identify relevant features and can be particularly useful in applications where linear relationships between variables are expected [16].

$$PCC(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{2}$$

where, cov(X, Y) represents the covariance between X and Y, which measures their joint variability, and σ(X) and σ(Y) represent the standard deviations of X and Y, respectively.

Maximum Relevance Minimum Redundancy (MRMR). It is an approach used to select the most informative features from a given dataset while minimizing the redundancies among them. It evaluates the relevance of each feature with respect to the target variable and simultaneously considers the redundancy among the selected features. By incorporating both relevance and redundancy measures, MRMR aims to identify a subset of features that maximizes the discriminative power while minimizing the overlap or redundancy between them. This technique has proven useful in various applications such as pattern recognition, text mining, and bioinformatics, allowing researchers to extract a compact and informative feature subset for improved model performance and interpretability [17], [18].

$$MRMR(S) = argmax\left\{\sum\left[I(f_i; C) - \sum I(f_i; f_j)\right]\right\} \quad (3)$$

where, MRMR(S) denotes the feature subset S that maximizes the objective function, I(fi; C) represents the relevance or mutual information between feature fi and the target variable C, and I(fi; fj) represents the redundancy or mutual information between feature fi and feature fj.

D. Wrapper Feature Selection Approaches

Wrapper approaches utilize a modeling algorithm as an opaque evaluator and use its performance to evaluate feature subsets. In classification tasks, these evaluators, like Naïve Bayes or SVM, evaluate subsets based on their classification performance, while in clustering tasks, they utilize clustering algorithms such as K-means to assess subsets. Similar to filters, wrappers employ a search strategy to generate subsets, repeating the evaluation process for each subset. However, wrappers are slower than filters as they depend on the computational demands of the modeling algorithm. Furthermore, the selected feature subsets can be biased towards the specific modeling algorithm used for evaluation, even with the employment of cross-validation. Therefore, for an accurate estimation of generalization error, it is crucial to validate the final subset with an independent sample and a different modeling algorithm [19]. On a positive note, empirical evidence suggests that wrappers outperform filters in obtaining subsets with higher performance due to the utilization of real modeling algorithms. While wrappers can be used in combination with various search strategies and modeling algorithms, they are most suitable for greedy search strategies and fast algorithms like linear SVM, and Naïve Bayes [20]. In this research, three main categories of wrapper feature selection methods were utilized: L1-based and L2-based (ridge regression, least absolute shrinkage and selection operator (LASSO), and linear SVM), Tree-based (extra tree, random forest, gradient boosting tree), and evolutionary-based (genetic algorithm).

Ridge Regression. Ridge regression, also known as Tikhonov regularization, is a feature selection method that introduces a regularization term to the linear regression model. It addresses the issue of multicollinearity among the predictor variables by shrinking the coefficients towards zero. In ridge regression, the objective is to find the subset of features that effectively contribute to the prediction while minimizing the impact of correlated or redundant variables. By controlling the regularization parameter, ridge regression allows for a balance between model simplicity and predictive accuracy. This method is particularly useful when dealing with high-dimensional datasets and helps prevent overfitting by reducing the variance of the model [21].

$$minimize: RSS + \alpha \sum \beta_i^2 \quad subject\ to \quad \sum \beta_i^2 \leq t \quad (4)$$

RSS represents the residual sum of squares, which measures the error between the predicted and actual values, βi refers to the regression coefficients for each predictor variable, α is the regularization parameter that controls the amount of shrinkage applied to the coefficients, and t is a threshold that determines the budget for the sum of squared coefficients.

Least Absolute Shrinkage and Selection Operator (LASSO). It is a feature selection method utilized in regression analysis to efficiently select relevant predictor variables. Unlike traditional regression models, LASSO incorporates a regularization term into the equation that penalizes the sum of the absolute values of the regression coefficients. This penalty encourages sparsity by driving some coefficients to exactly zero, effectively conducting feature selection. The LASSO method is beneficial in situations where there are many predictors, as it can help identify the most influential variables and disregard the less relevant ones, leading to a more interpretable and efficient model. By striking a balance between minimizing the residual sum of squares and reducing the magnitude of the coefficients, LASSO allows for automatic variable selection and works well in scenarios where there is a high degree of multicollinearity or when the number of predictors exceeds the number of observations [22].

$$minimize: \left(\frac{1}{2N}\right) \|Y - X * \beta\|^2 + \lambda \|\beta\|_1 \quad (5)$$

where, Y is the vector of target values, X is the design matrix containing the predictor variables, β is the coefficient vector, N is the number of samples, λ is the regularization parameter that controls the strength of the penalty term, and ||β||1 is the L1-norm (sum of absolute values) of the coefficient vector, which enforces sparsity.

Linear SVM. It works by optimizing a linear SVM model to find the hyperplane that best separates the classes of data points. In this process, SVM assigns weights or coefficients to each feature based on its importance in determining the class boundary. These weights can be used as a measure of feature relevance. By selecting features with large coefficients, which contribute significantly to class separation, the linear SVM feature selection method helps to identify the most informative features for classification tasks. This approach is effective in reducing dimensionality, improving model performance, and enhancing interpretability [23].

$$minimize: 0.5 \times \|w\|^2 + C \sum \xi \quad subject\ to \quad y_i(w^T x_i) \geq 1 - \xi_i\,, i = 1,2, \dots, N \quad (6)$$

where, ξi stands for the slack variables that permit some misclassifications, xi is the feature vector of the i-th data point, yi is the corresponding class label (+1 or -1), and ||w||2 is the L2 norm of the weight vector w. C is a regularization parameter that balances the trade-off between maximizing the margin and minimizing misclassifications.

Extra Tree. It is a variant of the Random Forest algorithm that further increases the randomness of the DTs. Extra Trees randomly selects subsets of features and thresholds to build a large number of DTs. The feature importance is calculated by measuring the average impurity decrease in overall features in the ensemble of trees. Features with high-importance scores are considered more relevant for prediction while low-scoring features can be discarded. The main advantage of Extra Trees is its ability to handle high-dimensional data and capture complex interactions among features. It can effectively reduce overfitting and improve model performance by selecting the most informative subset of features 24].

Random Forest. It involves constructing an ensemble of DTs, where each tree is trained on a random subset of features and the predictions are aggregated through voting or averaging. The importance of each feature is then determined by measuring how much the performance of the model decreases when that feature is randomly permuted. Features that lead to a significant drop in performance are considered more important, while those with minimal impact are deemed less relevant. By evaluating the importance scores across multiple trees, Random Forest feature selection provides a robust and efficient approach to highlighting the most influential features in a dataset. The feature importance score in this method is computed based on how much each feature contributes to the overall accuracy of the Random Forest model [25].

Gradient Boosting Tree. Unlike other methods, it doesn't rely on a specific mathematical equation but follows a sequential process. The algorithm starts by building weak DTs and then iteratively improves them by adding new trees that correct the errors made by previous trees. When choosing features for the Gradient Boosting Tree, each feature's contribution to lowering the model's total loss is taken into consideration. During the boosting process, features with higher significance ratings are prioritized since they are deemed more significant. By iteratively selecting and refining features, the Gradient Boosting Tree effectively identifies which features are most influential in predicting the target variable, leading to more accurate and efficient models [26].

Genetic Algorithm (GA). GA is a popular feature selection method inspired by the concept of natural selection and genetic evolution. It is a search algorithm that mimics the process of natural selection to find the best subset of features for a given problem. GA starts by representing each potential subset of features as a binary string, called a chromosome. These chromosomes then undergo reproduction, mutation, and crossover operations to create a new population of chromosomes in each generation. Fitness functions are defined to evaluate how well each subset performs. The subsets with the highest fitness values are given a higher probability of being selected for the next generation. This iterative process continues until a stopping criterion is met. By using genetic operators such as mutation and crossover, GA explores the solution space effectively and finds optimal or near-optimal feature subsets that can improve the performance of machine learning models [27].

E. Regression Analysis

Regression Analysis with Adaboost is a powerful machine learning technique that combines the principles of regression analysis and the Adaboost algorithm. Regression analysis is used to predict a continuous target variable based on one or more predictor variables. Adaboost, on the other hand, is an ensemble learning algorithm that combines the strengths of multiple weak classifiers to build a strong predictive model. In the context of regression, Adaboost works by iteratively training a series of weak regression models on different subsets of the training data. In each iteration, Adaboost assigns higher weights to the training instances that were poorly predicted by the previous models, thereby focusing on the most challenging cases. The weak models are then combined through a weighted average, where the weights are determined by their performance on the training data. By repeatedly refining the model based on the misclassified instances, Adaboost can ultimately create a robust and accurate regression model. This approach is helpful in handling complex regression problems with non-linear relationships between predictors and the target variable, as it effectively captures the underlying patterns and produces accurate predictions [28].

III. Results

Table II shows the characteristics of the dataset used, which includes the 32 features shown in Table I. In addition, Fig. 2 shows the histogram of some variables of this dataset to display the status of students. Fig. 3 also represents the outcome of the correlation evaluation between all the variables of this dataset. As it is clear, the G1 and G2 variables have a correlation greater than 0.8 with the target variable (G3).

TABLE II. Characteristics of the Dataset

Attributes	**Type/Value**
Dataset	Student performance (Math course)
Number of samples	395
Number of features	32
Number of target feature	1
Missing values	0

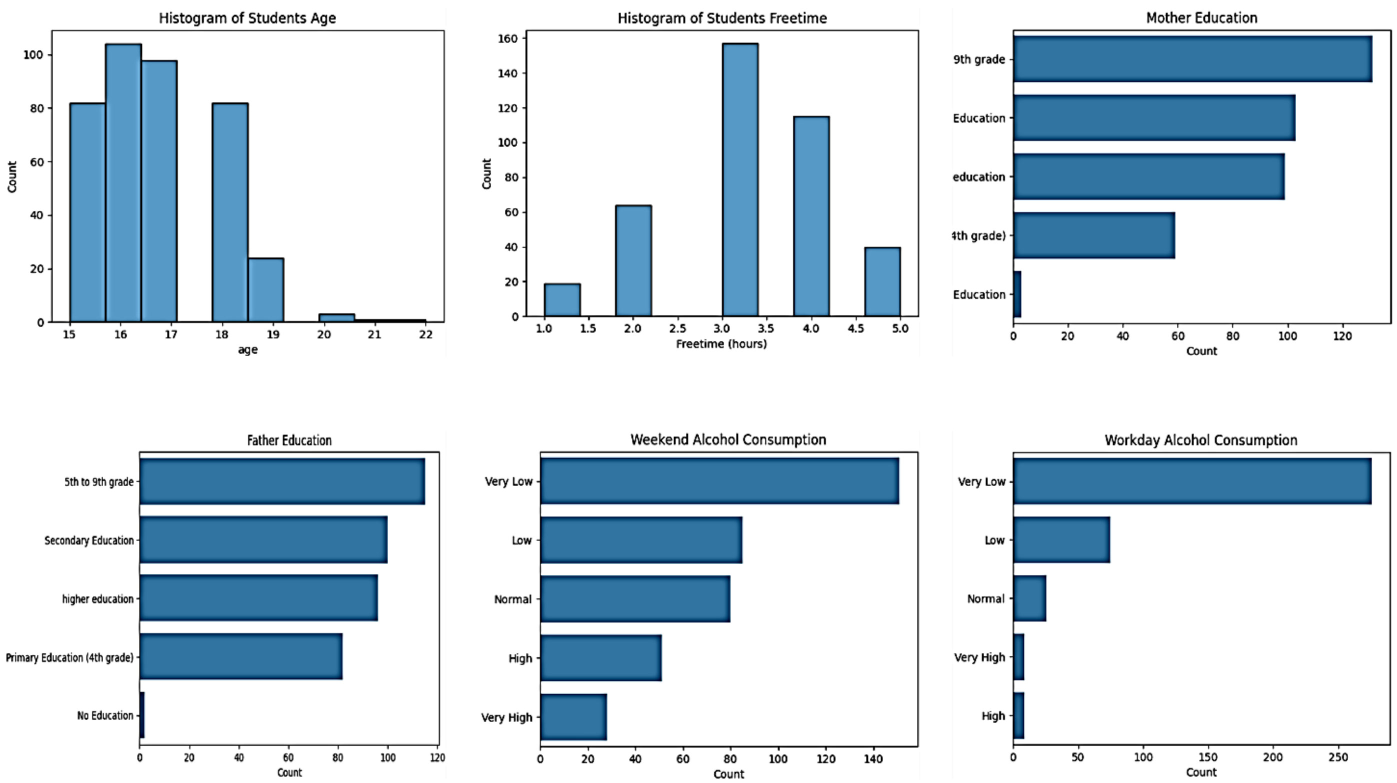

Fig. 2. Histogram of different variables.

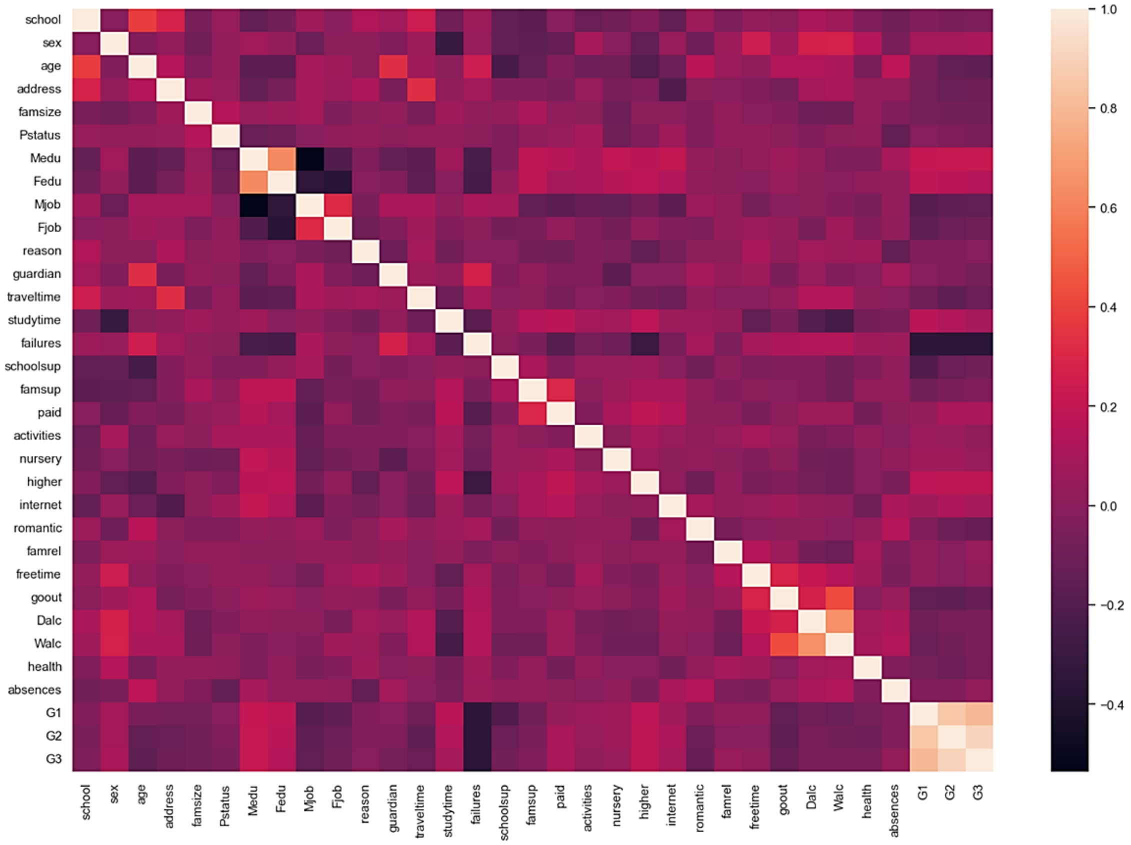

Fig. 3. FCorrelation between 32 dataset variables.

Fig. 4 to Fig. 6 show the results of feature ranking by different feature selection methods to predict final math grades. Also, Table III shows the Top 10 features selected by MRMR and GA feature selection techniques. As shown, G1 and G2 scores had the highest correlation with the final grade (G3), and in most of the feature selection methods, they were among the best features.

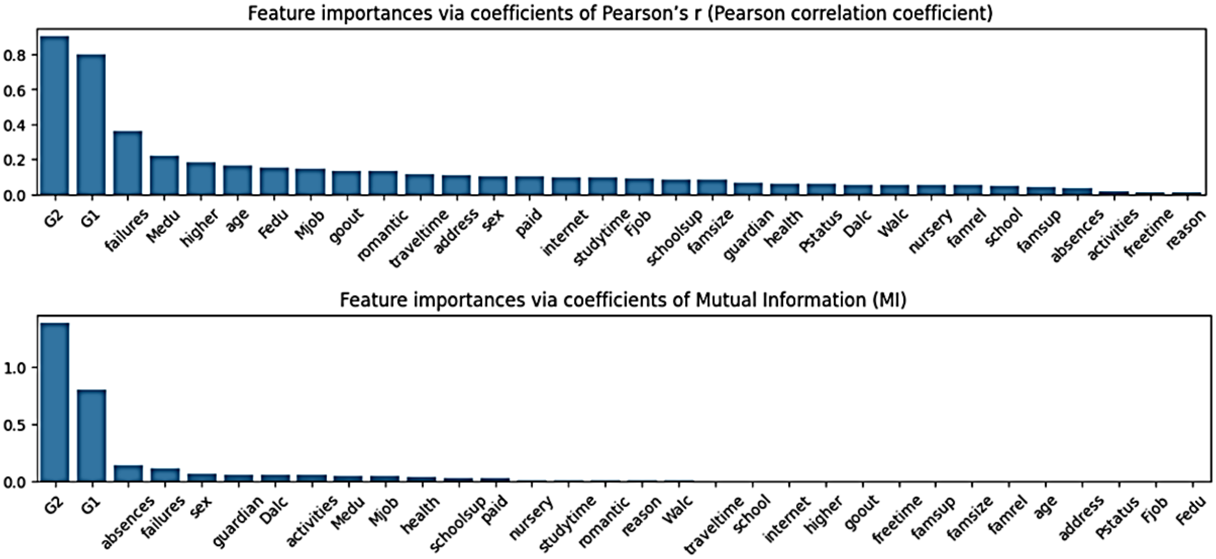

Fig. 4. Dataset feature ranking using two filtered feature selection techniques (PCC and MI) to predict final math grades.

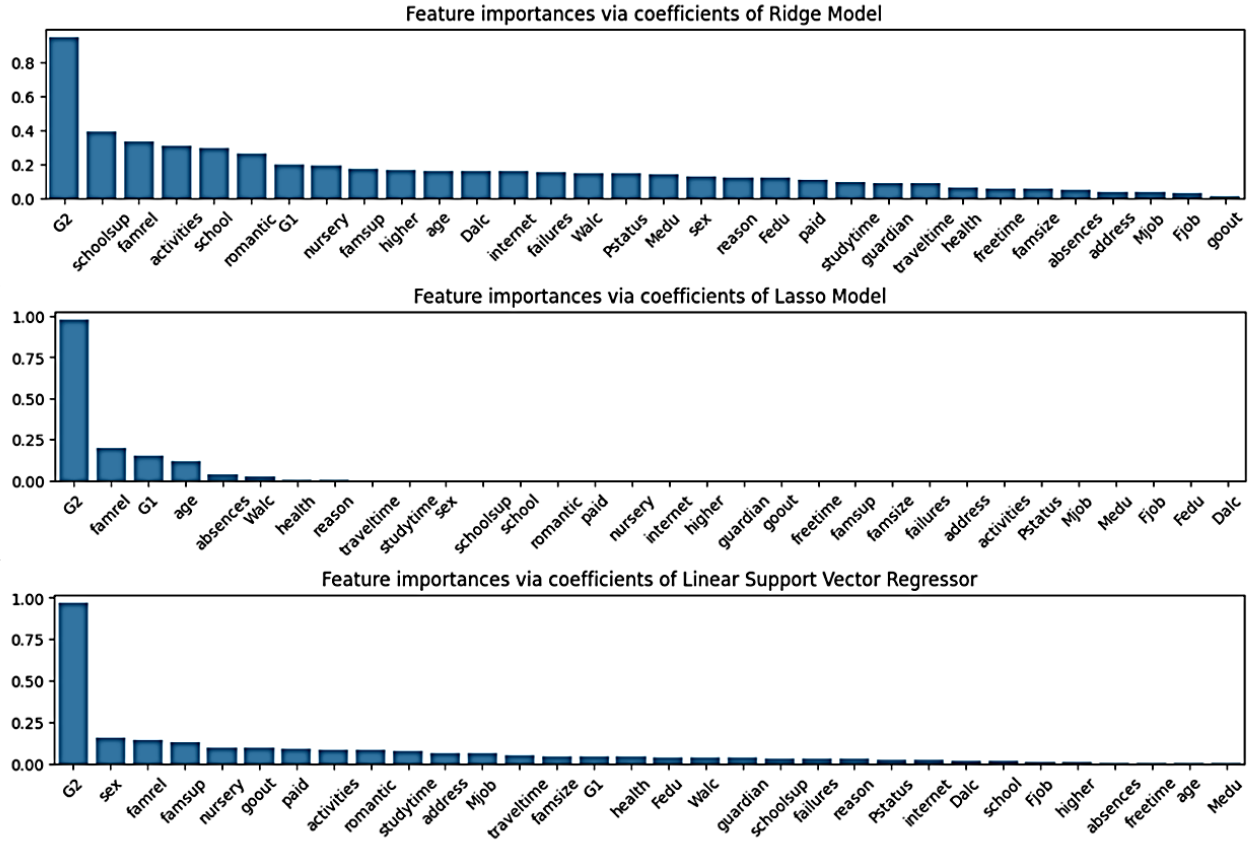

Fig. 5. Dataset feature ranking using three wrapper feature selection techniques (Ridge, LASSO, and SVM) to predict final math grades.

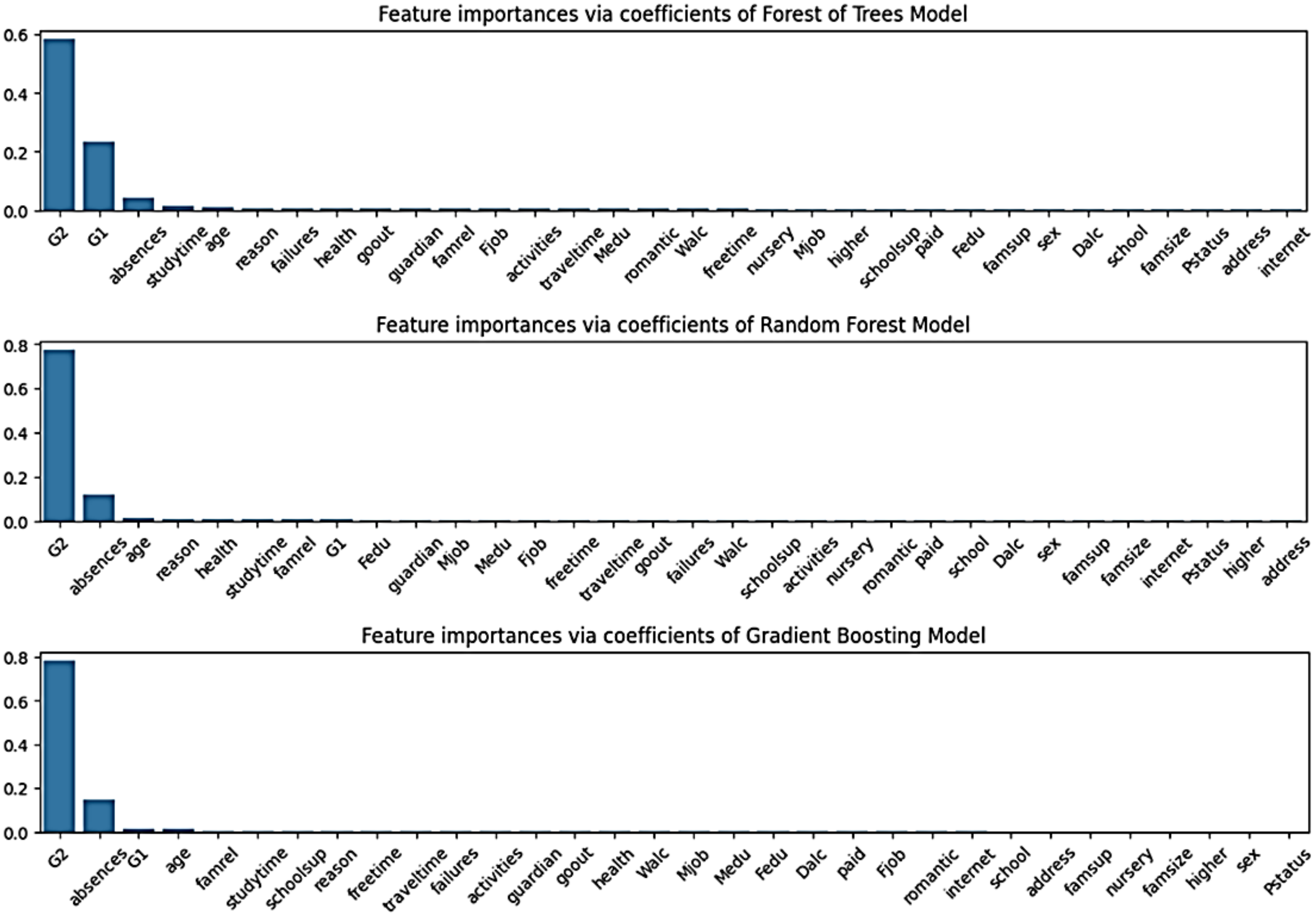

Fig. 6. Dataset feature ranking using three wrapper tree-based feature selection techniques (extra tree, random forest, and gradient boosting model) to predict final math grade.

TABLE III. TOP 10 FEATURES SELECTED BY MRMR AND GA FEATURE SELECTION TECHNIQUES

Technique	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
MRMR	G1	Failures	Medu	Romantic	Higher	Goout	Famsize	Age	Traveltime	Mjob
GA	Medu	Studytime	Romantic	G2	Freetime	Goout	Dalc	Walc	Health	G1

After selecting the feature, 80% of the data was used as a training sample and the remaining 20% was used as a test sample. Table IV shows the result of regression analysis using the Adaboost model and features selected by different machine learning techniques to predict the final math grade. Pearson correlation coefficient (PCC), mean absolute error (MAE), and mean squared error (MSE) were used to evaluate the results of the regression analysis. As shown, the best result was obtained from feature selection by the LASSO method with PCC = 94.26%, MAE = 1.12, and MSE = 2.53. After the LASSO method for feature selection, the Extra Tree (PCC = 94.00%, MAE = 1.13, MSE = 2.64) and Gradient Boosting Machine (PCC = 93.55, MAE = 1.15, MSE = 2.73) methods respectively had the best prediction of the final math grade. Fig. 7 shows the scatter plots of the top 10 features selected by the LASSO technique. The bolder the data is, the higher the final math score. However, the rest of the feature selection techniques, except the random forest (PCC = 93.35%, MAE = 1.13, MSE = 2.76), achieved a lower precision than the original dataset for predicting the final math grade.

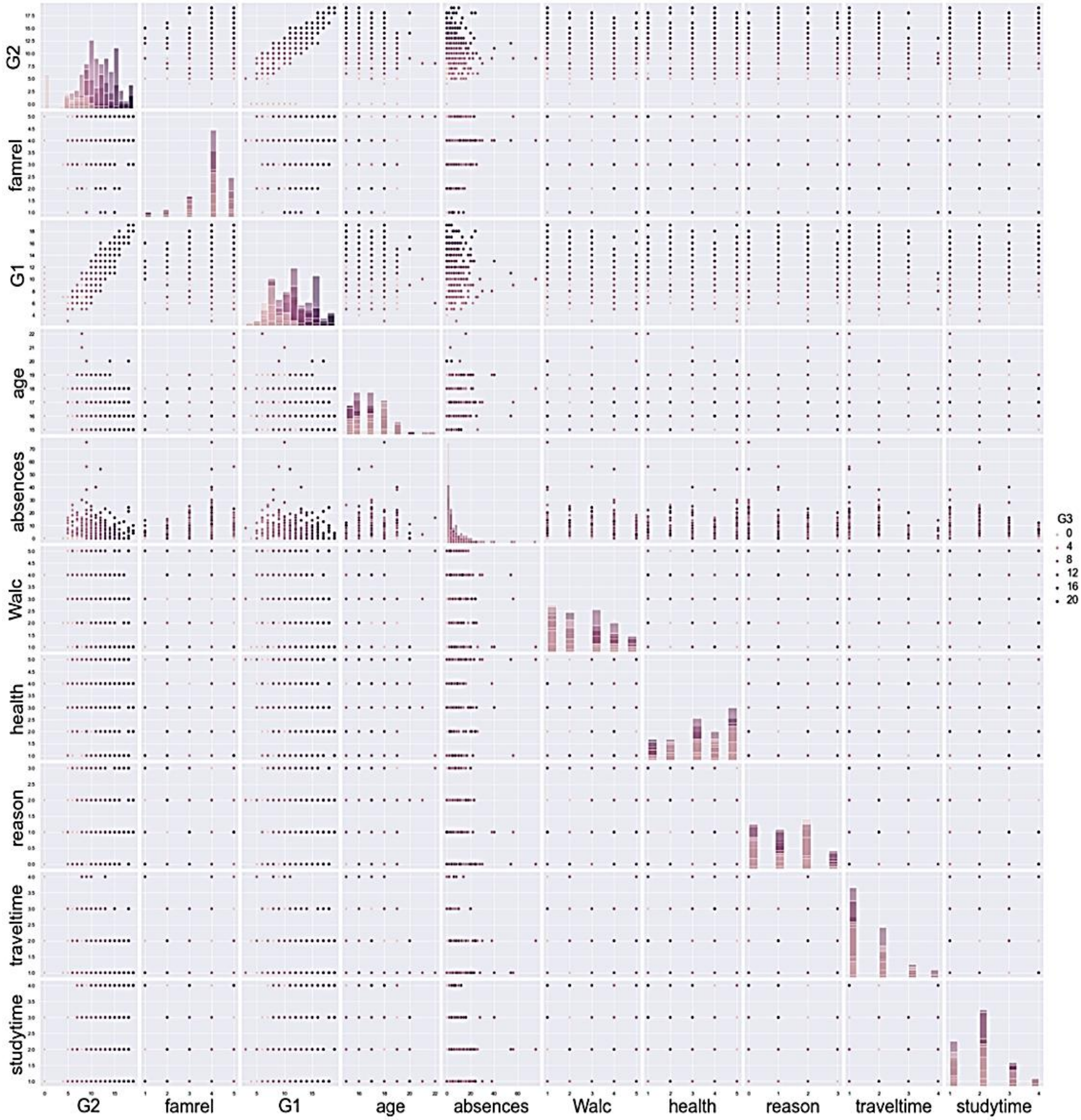

Fig. 7. Scatter plots of the top 10 features selected by the LASSO technique. The bolder the data is, the higher the final math score.

TABLE IV. THE RESULT OF REGRESSION ANALYSIS USING THE ADABOOST MODEL AND FEATURES SELECTED BY DIFFERENT MACHINE LEARNING TECHNIQUES TO PREDICT THE FINAL MATH GRADE

Feature selection method	Pearson correlation coefficient (PCC)	Mean absolute error (MAE)	Mean squared error (MSE)	Number of features
Without feature selection	93.28±0.01	1.20±0.09	3.01±0.58	32
Ridge	87.87±0.03	1.63±0.11	5.30±1.22	10
LASSO	94.26±0.01	1.12±0.08	2.53±0.46	10
Linear SVM	86.16±0.03	1.71±0.08	6.23±1.51	10
Gradient boosting	93.55±0.01	1.15±0.09	2.73±0.46	10
Extra tree	94.00±0.01	1.13±0.11	2.64±0.58	10
Random forest	93.35±0.02	1.13±0.06	2.76±0.35	10
PCC	87.76±0.02	1.69±0.11	5.60±0.96	10
MI	91.04±0.01	1.25±0.05	3.57±0.56	10
MRMR	87.18±0.02	1.69±0.20	5.88±1.17	10
GA	74.35±0.04	2.50±0.25	10.39±1.14	10

IV. DISCUSSION

In this research, machine learning methods were used for data mining from a dataset of students' performance. For this purpose, a variety of filtered and wrapper feature selection methods were used to determine the important demographic factors involved in predicting students' math scores. Finally, the features selected by each method predicted the final math grade using regression analysis with the Adaboost model. The results showed that the wrapper LASSO feature selection technique selects the best subset of features to predict the final math grade. LASSO offers several advantages in the field of data analysis. Firstly, it provides a solution for handling high-dimensional datasets, where the number of predictors exceeds the number of samples. By imposing a penalty term on the regression coefficients, LASSO encourages sparsity by shrinking some coefficients to zero, effectively selecting the most relevant features. This can lead to improved model interpretability and the identification of key predictors driving the observed outcomes. Moreover, LASSO is robust against multicollinearity, a common issue when predictors are correlated, as it tends to select one representative variable among highly correlated features [29]. Additionally, LASSO aids in avoiding overfitting by preventing the model from becoming excessively complex, which can generalize well to unseen data. Therefore, the LASSO method provides a powerful and efficient approach to feature selection by effectively handling high-dimensional datasets, promoting interpretability, and robustness against multicollinearity, and preventing overfitting [30]. The important factors selected by LASSO involved in predicting the final math grade of students were first and second-period grades, quality of family relationships, age, number of school absences, weekend alcohol consumption, current health status, the reason for choosing the school, weekly study time, and home to school time arrival. Schools encounter a range of predominant difficulties, such as performance analysis, delivering exceptional education, devising effective methods to assess student progress, and planning for future initiatives[31]. To tackle the issues students may encounter while pursuing their studies, it becomes imperative for these institutions to establish student intervention programs. These intervention plans aim to address and resolve the challenges faced by students throughout their academic journey [32]. However, to have an effective intervention, important factors must be identified and this study was able to do this by using different data mining methods.

There are some previous attempts to survey the literature on academic performance [33]; however, most of them are general literature reviews and targeted towards the generic students' performance prediction. Table V compares the results obtained by the proposed framework in this study with previous techniques. As shown, the proposed framework outperforms other techniques in predicting student performance. As a result, this study could improve previous techniques in predicting student performance.

TABLE V. COMPARISON OF THE RESULTS OBTAINED BY THE PROPOSED FRAMEWORK WITH PREVIOUS TECHNIQUES

Reference	Machine learning technique	MAE	MSE
[1]	Semi-supervised regression algorithm	1.23	2.70
[34]	Model Tree (MT), NN, Linear Regression (LR), Locally Weighted Linear Regression, and Support Vector Machine (SVM)	1.21	-
[35]	Scoring algorithm called WATWIN and linear regression	-	6.91
[36]	Support Vector Machine (SVM), Random Forest, Logistic Regression, Adaboost, and Decision Tree	1.40	3.15
[37]	multilevel regression trees	1.33	2.97
[38]	Linear regression for supervised learning, linear regression with deep learning and neural network	3.26	7.19
[39]	Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek	4.11	10.76
Current study	LASSO and regression	1.12	2.53

Anticipating the academic performance of students assumes significance within educational settings like schools and universities. This enables the development of efficient mechanisms that enhance academic outcomes and deter dropout rates, among other benefits [40]. The automation of various tasks involved in students' regular activities, leveraging vast amounts of data obtained from technology-enhanced learning software tools, plays a pivotal role in achieving these advantages. Consequently, meticulous analysis and processing of this data can furnish valuable insights into students' aptitude and their correlation with academic assignments [41]. Such information serves as the foundation for propitious algorithms and methodologies capable of prognosticating students' performance. The present study showed that the proposed framework can be used for such work in educational environments. This framework can predict students' performance by analyzing large datasets and taking into account the past and current status of students. However, there are limitations in this study as in many other studies that should be mentioned. First, this model requires external validation using unseen datasets. Second, most of the variables in this data set were demographic factors, while there are certainly other important factors that should be investigated in future studies. Thirdly, although the obtained results were good and acceptable, future studies should seek to improve the current results by optimizing the model parameters.

V. Conclusion

In this research, a comparative study was conducted between different data mining techniques to predict the mathematical performance of students. For this purpose, various filtered and wrapper feature selection methods were compared to select the most useful factors in predicting math grades. The present study showed that the LASSO feature selection technique integrated with regression analysis with the Adaboost model is a suitable data mining framework for predicting students' mathematical performance. This framework was able to identify the most relevant factors related to math performance and predict student performance with low error rate. These methods that rely on data analysis can be employed alongside other educational techniques to assess students' progress effectively. They offer insightful feedback to academic advisors, enabling them to suggest appropriate follow-up courses and implement necessary pedagogical interventions. Moreover, this research will greatly influence the development of curricula within degree programs and contribute to the formulation of education policies at large. Future research should take advantage of optimization algorithms to adjust parameters to improve the structure of the proposed framework and achieve better results. In addition, it is necessary to examine the external validity of the proposed framework by applying it to other datasets.

References

[1] G. Kostopoulos, S. Kotsiantis, N. Fazakis, G. Koutsonikos, and C. Pierrakeas, “A semi-supervised regression algorithm for grade prediction of students in distance learning courses,” International Journal on Artificial Intelligence Tools, vol. 28, no. 04, p. 1940001, 2019.

[2] M. Ranđelović, A. Aleksić, R. Radovanović, V. Stojanović, M. Čabarkapa, and D. Ranđelović, “One Aggregated Approach in Multidisciplinary Based Modeling to Predict Further Students' Education,” Mathematics, vol. 10, no. 14, p. 2381, 2022.

[3] J. Cjuno, J. Palomino-Ccasa, R. G. Silva-Fernandez, M. Soncco-Aquino, O. Lumba-Bautista, and R. M. Hernández, “Academic procrastination, depressive symptoms and suicidal ideation in university students: a look during the pandemic,” Iran J Psychiatry, vol. 18, no. 1, p. 11, 2023.

[4] A. Krishna and A. Y. Orhun, “Gender (still) matters in business school,” Journal of Marketing Research, vol. 59, no. 1, pp. 191–210, 2022.

[5] M. R. Mohammadi, A. Khaleghi, K. Shahi, and H. Zarafshan, “attention deficit hyperactivity disorder: Behavioral or Neuro-developmental Disorder? Testing the HiTOP Framework Using Machine Learning Methods,” Journal of Iranian Medical Council, vol. 6, no. 4, pp. 652–657, 2023.

[6] A. Khaleghi, M. R. Mohammadi, G. P. Jahromi, and H. Zarafshan, “New ways to manage pandemics: using technologies in the era of COVID-19: a narrative review,” Iran J Psychiatry, vol. 15, no. 3, p. 236, 2020.

[7] A. Khaleghi, M. R. Mohammadi, K. Shahi, and A. M. Nasrabadi, “Computational neuroscience approach to psychiatry: a review on theory-driven approaches,” Clinical Psychopharmacology and Neuroscience, vol. 20, no. 1, p. 26, 2022.

[8] J. Xu, K. H. Moon, and M. van Der Schaar, “A machine learning approach for tracking and predicting student performance in degree programs,” IEEE J Sel Top Signal Process, vol. 11, no. 5, pp. 742–753, 2017.

[9] B. K. Baradwaj and S. Pal, “Mining educational data to analyze students' performance,” arXiv preprint arXiv:1201.3417, 2012.

[10] D. Kabakchieva, “Student performance prediction by using data mining classification algorithms,” International journal of computer science and management research, vol. 1, no. 4, pp. 686–690, 2012.

[11] A. O. Osofisan, O. O. Adeyemo, and S. T. Oluwasusi, “Empirical study of decision tree and artificial neural network algorithm for mining educational database,” Afr J Comput Ict, vol. 7, no. 2, pp. 187–196, 2014.

[12] K. Roy, H.-H. Nguyen, and D. M. Farid, “Impact of dimensionality reduction techniques on student performance prediction using machine learning,” CTU Journal of Innovation and Sustainable Development, vol. 15, no. Special issue: ISDS, pp. 93–101, 2023.

[13] X. Li, Y. Zhang, and R. Zhang, “Semisupervised feature selection via generalized uncorrelated constraint and manifold embedding,” IEEE Trans Neural Netw Learn Syst, vol. 33, no. 9, pp. 5070–5079, 2021.

[14] A. Jović, K. Brkić, and N. Bogunović, “A review of feature selection methods with applications,” in 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), Ieee, 2015, pp. 1200–1205.

[15] X. Song, Y. Zhang, D. Gong, and X. Sun, “Feature selection using bare-bones particle swarm optimization with mutual information,” Pattern Recognit, vol. 112, p. 107804, 2021.

[16] Y. Liu, Y. Mu, K. Chen, Y. Li, and J. Guo, “Daily activity feature selection in smart homes based on pearson correlation coefficient,” Neural Process Lett, vol. 51, pp. 1771–1787, 2020.

[17] A. Khaleghi et al., “EEG classification of adolescents with type I and type II of bipolar disorder,” Australas Phys Eng Sci Med, vol. 38, pp. 551–559, 2015.

[18] M. R. Mohammadi, A. Khaleghi, A. M. Nasrabadi, S. Rafieivand, M. Begol, and H. Zarafshan, “EEG classification of ADHD and normal children using non-linear features and neural network,” Biomed Eng Lett, vol. 6, pp. 66–73, 2016.

[19] C. Chen, Y. Tsai, F. Chang, and W. Lin, “Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results,” Expert Syst, vol. 37, no. 5, p. e12553, 2020.

[20] O. M. Alyasiri, Y.-N. Cheah, A. K. Abasi, and O. M. Al-Janabi, “Wrapper and hybrid feature selection methods using metaheuristic algorithms for English text classification: A systematic review,” IEEE Access, vol. 10, pp. 39833–39852, 2022.

[21] T. D. la Tour, M. Eickenberg, A. O. Nunez-Elizalde, and J. L. Gallant, “Feature-space selection with banded ridge regression,” Neuroimage, vol. 264, p. 119728, 2022.

[22] P. Ghosh et al., “Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques,” IEEE Access, vol. 9, pp. 19304–19326, 2021.

[23] D. Albashish, A. I. Hammouri, M. Braik, J. Atwan, and S. Sahran, “Binary biogeography-based optimization based SVM-RFE for feature selection,” Appl Soft Comput, vol. 101, p. 107026, 2021.

[24] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, “Ai meta-learners and extra-trees algorithm for the detection of phishing websites,” IEEE access, vol. 8, pp. 142532–142542, 2020.

[25] X. Li, W. Chen, Q. Zhang, and L. Wu, “Building auto-encoder intrusion detection system based on random forest feature selection,” Comput Secur, vol. 95, p. 101851, 2020.

[26] A. Alsahaf, N. Petkov, V. Shenoy, and G. Azzopardi, “A framework for feature selection through boosting,” Expert Syst Appl, vol. 187, p. 115895, 2022.

[27] F. Amini and G. Hu, “A two-layer feature selection method using genetic algorithm and elastic net,” Expert Syst Appl, vol. 166, p. 114072, 2021.

[28] G. Shanmugasundar, M. Vanitha, R. Čep, V. Kumar, K. Kalita, and M. Ramachandran, “A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining,” Processes, vol. 9, no. 11, p. 2015, 2021.

[29] S. Afrin et al., “Supervised machine learning based liver disease prediction approach with LASSO feature selection,” Bulletin of Electrical Engineering and Informatics, vol. 10, no. 6, pp. 3369–3376, 2021.

[30] L. Cui, L. Bai, Y. Wang, S. Y. Philip, and E. R. Hancock, “Fused lasso for feature selection using structural information,” Pattern Recognit, vol. 119, p. 108058, 2021.

[31] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student’performance prediction using machine learning techniques,” Educ Sci (Basel), vol. 11, no. 9, p. 552, 2021.

[32] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, “Analyzing and predicting students’ performance by means of machine learning: A review,” Applied sciences, vol. 10, no. 3, p. 1042, 2020.

[33] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," International Journal of Educational Technology in Higher Education, vol. 17, no. 1, p. 3, 2020.

[34] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: a decision support system for forecasting students’ grades," Artificial Intelligence Review, vol. 37, pp. 331-344, 2012.

[35] C. Watson, F. W. Li, and J. L. Godwin, "Predicting performance in an introductory programming course by logging and analyzing student programming behavior," in 2013 IEEE 13th international conference on advanced learning technologies, 2013: IEEE, pp. 319-323.

[36] H. Lakkaraju et al., "A machine learning framework to identify students at risk of adverse academic outcomes," in Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1909-1918.

[37] C. Masci, G. Johnes, and T. Agasisti, "Student and school performance across countries: A machine learning approach," European Journal of Operational Research, vol. 269, no. 3, pp. 1072-1085, 2018.

[38] A. O. Oyedeji, A. M. Salami, O. Folorunsho, and O. R. Abolade, "Analysis and prediction of student academic performance using machine learning," JITCE (Journal of Information Technology and Computer Engineering), vol. 4, no. 01, pp. 10-15, 2020.

[39] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students’ performance using machine learning techniques," IEEE Access, vol. 8, pp. 67899-67911, 2020.

[40] A. Asselman, M. Khaldi, and S. Aammou, “Enhancing the prediction of student performance based on the machine learning XGBoost algorithm,” Interactive Learning Environments, vol. 31, no. 6, pp. 3360–3379, 2023.

[41] S. Hussain and M. Q. Khan, “Student-performulator: Predicting students’ academic performance at secondary and intermediate level using machine learning,” Annals of data science, vol. 10, no. 3, pp. 637–655, 2023.