Predicting Student Outcomes: Evaluating Regression Techniques in Educational Data

Manish Kumar Singla [1] *, **Faris H. Rizk**[2], **Mahmoud Elshabrawy Mohamed**[2], **Ahmed Mohamed Zaki**[2]
[1]Department of Interdisciplinary Courses in Engineering, Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India.
[2]Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA
Emails: manish.singla@chitkara.edu.in, faris.rizk@jcsis.org, mshabrawy@jcsis.org, Azaki@jcsis.org

Abstract

Student performance prediction is essential so that institutions can assist in identifying weak performers and initiate corrective measures. This research assesses different regression models by applying data from Kaggle, which involves data cleaning like managing missing values and scaling of the data, hence feature extraction, then model imposition and authenticity. The models followed are Linear Regression, SVR, MLPRegressor, Gradient Boosting, Catboost, Xgboost, Random Forest, Extratrees, Decision Tree and K-neighbors. The analysis shows that Linear Regression produced the best result as it has the lowest MSE score of 0. 000521 and high accuracy regarding other measures, including RMSE, MAE, and R^2. The results reveal that regression models can be used to predict students' performance and be helpful to the various stakeholders in the system. The findings of this study will help develop required models for decision-making to improve students' performance.

Keywords: Student performance prediction, regression models, educational data, data preprocessing, predictive analytics

1 introduction

Evaluation of students' performance is an essential component in obtaining education as it determines further approaches in curriculum and instruction and policy making. Thus, the regular assessment of students' results is crucial to identifying their needs and the institution's capacity to address them and to satisfy current educational requirements and social expectations. The frames raising in last years show the growing tendency to utilize outstanding data to analyze students' results and their improvement. Of these approaches, regression analysis has turned out to be one of the most useful statistical tools for the prediction and analysis of a host of factors affecting poor students' performance [1–3].

Regression analysis, a tool widely used in statistics, enables the researcher to uncover the connection between an independent and a dependent variable. Applied to the field of education, this technique will help find the essential indicators of students' performance, from socioeconomic background and truancy rates to the choice of instructional models and after-school participation. Socioeconomic status might affect the ability of the student to use resources like private tutoring or advanced books for learning. The same applies to class attendance, whereby students will likely be exposed more and perform well in instructional lessons and activities. Identifying these relations helps educators and policymakers get useful information regarding the efficacy of various interventions and strategies for enhancing educational outcomes [4–6].

https://doi.org/10.54216/JAIM.070205
Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

The following paper seeks to undertake a study on the role of regression analysis in predicting student performances. The analysis will be devoted to the determinants of academic achievement and how they affect the entire performance. For instance, variables like parental education level or school funding affect learners' achievement individually, but their combination might shed light on strengthened effects or diminution. Thus, the model's sensitivity and comprehensiveness by including a diverse range of demographic, academic, and behavioral variables for developing a strong dataset will be examined by this research study. While this model will show trends in how student performance depends on specific predictors, it will also reveal intertwined relationships of predictors [7–9].

The final direction is that this paper will discuss the implications of the findings for educational practice and policy. From the determinants of students' performance, it is possible to draw interventions, allocate resources and introduce changes that would make education fair and effective. For example, based on the findings that extracurricular activity enhances performance, policies may be developed to ensure schools focus on providing funds for after-school activities. Besides, the study will present the limitations of regression analysis utilization in educational research, such as high intercorrelation among predictor variables or possible cross-context transferability of the results. Giving proposals for further research in this field, the given work's goal is to contribute to developing effective perspectives in student performance as far as its different determinants are concerned [10–12].

In conclusion, it can be stated that the regression analysis method used in the analysis of student performance displayed considerable potential as the tool allowing the expansion of knowledge on the reasons for the effectiveness of students' learning. Through these studies, this research thus seeks to establish the predictors above by defining and quantifying them with the potential of improving future educational practices and results. The ultimate utilization is to equip educators, administrators, and policymakers with relevant information, which can contribute to promoting a proper climate for student's successful learning experiences.

2 Related Works

Prediction of student performance has been one of the essential themes in educational data mining, emphasizing possible antecedents of students' success and failure and working towards creating prediction models. Different approaches and methods have been used to solve this problem, and they all offer different ways and methods of solving the problem and developing the technology.

Some researchers have applied initial multiple regression models to estimate performance outcomes. Linear regression models are prominent among them, and because of their simplicity and ease of interpretation, they are often applied. The potential of linear regression in using compendiously data, including attendance, previous grades, and extracurricular activities, to predict the final exam scores have been expounded in research [13]. Nevertheless, linear models do not always work effectively with educational data because of the curvilinear relations between the variables involved; thus, there is a need to look further into more complex equations.

In dealing with the shortcomings of linear regression, scholars have now started to adopt even higher-order machine learning models, including decision trees and random forests, widely in predicting the rates. The structure of decision trees, which is developed as a tree with branches, makes it possible to provide a common interpretation of the results and determine the conditional indicators [14]. For example, a study employed decision trees to define critical parameters affecting the learners' performance, including study habits and time management. Random continues from decision trees as an ensemble method for mitigating overfitting and enhancing stability using several trees [15]. This technique has demonstrated applicability in coping with nonlinearity in most educational data sets and can offer more stable prediction clues.

Techniques such as Support Vector Machines (SVM) have also been used to predict student performance, especially in cases where there may not be a direct correlation between input features and output [16]. SVMs operate well in high dimensionality of features; they do not suffer from overfitting and hence can always be used in analyzing educational datasets with many features. For instance, an SVM can work with data about student demographics, academic profiles, and psychological traits to determine the likelihood of success, which shows the model's versatility.

https://doi.org/10.54216/JAIM.070205

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

Over the last decade, deep learning models have become popular because they can achieve feature extraction and learning from the raw data. ANNs, mainly deep neural networks, have been used in predicting student performance with high levels of precision [17]. These models can fit more complex associations and dependencies in data that the conventional approaches might not. For example, a deep learning model may use interactions between students and learning management systems to forecast students' performance. This can then be employed as a complex student behavior analysis approach.

The integration of multiple models to achieve better results has also been considered within ensemble learning. Other methods, for instance, boosting and bagging, have been used to take advantage of the fact that each model is proficient in certain aspects and lacks proficiency in others. Hence, the weaknesses of one model can be compensated by another [18]. It is evident from the above methods that applying multiple models to predict student outcomes has been proven to be more efficient than individual models. For instance, applying decision trees together with SVM as one approach and neural networks as another supplementary approach would complete the selection of the best algorithms for prediction.

Statistical methods, like factor analysis and clustering analysis, have been used to determine the factors affecting students' performance. The above methods assist in identifying the underlying factors that may cause a student to succeed in his or her academics and may be applied to the development of forecast models as well [19]. For instance, it will be possible to identify the clusters of students in terms of learning behavior and the degree of participation, which will be helpful for educational intermediation.

In the specific field of educational data mining, ensemble models have also emerged, where the results of different algorithms are combined to mobilize all their respective strengths. For instance, integrating regression models with machine learning algorithms has helped increase the accuracy of predictions and given a wider outlook [20]. Linear regression might be applied to discover general trends, while machine learning algorithms improve these findings to produce more detailed and precise patterns concerning students' performance.

Besides, different works pointed out the significance of feature selection when considering students' performance. The choice of significant features from vast educational data is essential to enhance the models' efficiency and transparency. Other approaches used in determining the features highly correlated with students' success include, for instance, PCA and RFE [21]. For instance, while using PCA, one may work on decreasing the dimensionality to identify the main reasons sufficiently. Meanwhile, while using RFE, one has to drop the least important characteristics to increase the model's speed.

Finally, it is proposed that demographic and socio-economic variables be incorporated into the context of external factors affecting students' performance within such models. It has been established that variables like family income, parents' education and educational resources influence academic performance, and their incorporation into models increases the predictive capabilities [22]. For example, if the model incorporated socio-economic indicators, it would give a clearer picture of the issues affecting the students' experience, and advancement or mitigation measures would be much more efficient and just.

3 Data Preprocessing

Data pre-processing is vital in model building because different data pre-processing techniques may yield different results. This helps ensure that the dataset employed in training and testing the models is clean, well-formatted, and coherent with the intended model. In this paper, we go through steps such as handling missing values, normalizing and scaling the data, categorical data encoding, dividing the data into training and testing sets, and feature selection and feature creation.

3.1 Handling Missing Values

Such a problem impacts bias and reduces the prediction capacity of the model in case of data deficiency. We employed various techniques to handle missing values in the dataset:

https://doi.org/10.54216/JAIM.070205

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

- **Imputation**: Numerical features were imputed either by taking the variables' mean or median value depending on the data's distribution. Specifically, for categorical features, mean imputation was used, where the mode was taken as the variable's mean value.
- **Removal**: In other cases where the percentage of the missing values in a particular feature was very large, it was excluded to lessen bias.

3.2 Data Normalization and Scaling

To ensure that the features contribute equally to the model, we normalized and scaled the data:

- **Normalization**: This process ensured that all the features were put on the same scale while preserving the ratio of the ranges of the values for different features. Before feature scaling, we used min-max normalization to bring the features to the range [0, 1].
- **Standardization**: For any model that considers the scale of data, like Support Vector Regression and Multi-Layer Perceptron Regressor, we normalized the features by calculating $\frac{X-\text{mean}(X)}{\text{sd}(X)}$.

3.3 Encoding Categorical Variables

Many methods used in machine learning processes need numerical input data. Therefore, categorical variables were encoded using appropriate techniques:

- **One-Hot Encoding**: For the nominal categorical features that do not bear an ordinal relation to each other, we used the one-hot encoding method that created a new binary column for each of the many available categories.
- **Label Encoding**: Regarding the ordinal categorical features, label encoding was used where the categories are ordered in a manner that they can be ranked.

3.4 Dividing the Data into Training and Test Data

To evaluate the performance of our models, we split the dataset into training and testing sets:

- **Training Set**: We employed an 80/20 data split where 80 percent of the data was used for training the models.
- **Testing Set**: Twenty percent of the data was saved for assessing the model's efficacy. This split also guarantees that the models eventually have to calculate scores based on unseen observations, thus better estimating how accurate the models are.

3.5 Feature Selection and Engineering

Feature selection and engineering are critical for improving model performance and reducing overfitting:

- **Feature Selection**: Feature selection methods include correlation analysis and feature importance scores obtained using tree-based models. This included removing features with low relevance and high correlation with other features.

https://doi.org/10.54216/JAIM.070205

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

- **Feature Engineering**: Some analysis on new features was done based on domain knowledge and exploratory data analysis. For instance, interaction terms and polynomial features were created to capture non-linear relations in the data.

In Figure 1, a dependency heatmap of all the dataset features is shown, giving an outlook of how different variables are related. This heatmap is useful when one is concerned with knowing the nature and intensity of the linear relationship of a feature to other features. Combining values and properties with high values makes it possible to understand the selection and engineering of features. Addressing multicollinearity is important since it affects the accuracy of estimating the regression parameters and the exclusion of irrelevant predictors to increase the performance of the predictive models.

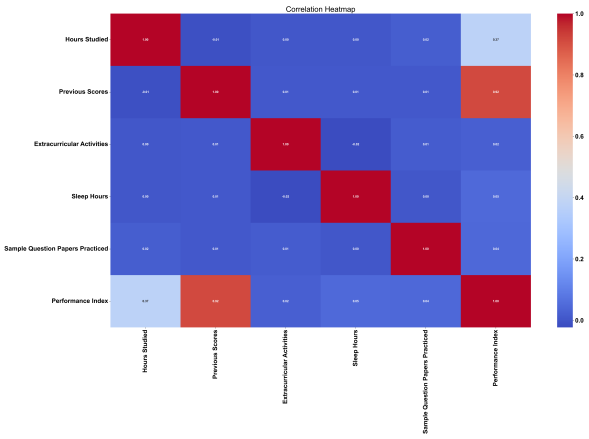

Figure 1: Correlation Heatmap of the Dataset Features

Figure 2, being the box plot of the dataset features, offers information about the distribution of the characteristics of variables in the dataset. Thus, the boxplots are an appropriate tool for determining the location, spread, and, potentially, outliers within the obtained set of values. This helps explain the pattern of the data distribution and whether any outliers or kurtosis might hamper model performances. Therefore, from these boxplots, it is possible to decide how to proceed with the data normalization or scaling as well as how to deal with outliers concerning the model training.

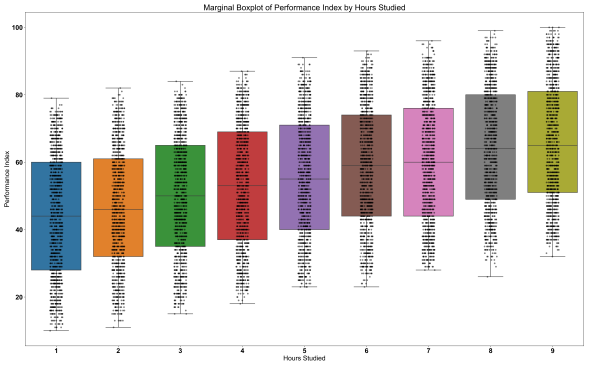

Figure 2: Boxplot of the Dataset Features

Therefore, through these preprocessing steps, we were able to prepare the dataset for model training and evaluation, enhancing the models' accuracy.

https://doi.org/10.54216/JAIM.070205

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

4 Methodology

4.1 Overview of Regression Models Used in the Study

Regarding student performance prediction, we compared several regression models in this study. The set of models chosen for the evaluation includes basic linear regression and sophisticated machine learning techniques to understand the latter's limitations. Below is an overview of each model used:

- **Linear Regression**: Probably the simplest but one of the most effective models, which uses a linear relationship between the input features and the variable to be predicted. It is easy to apply and understand. However, it may not detect multiple characteristics of the dataset.
- **Support Vector Regression (SVR)**: A method derived from support vector machines but developed with an emphasis on regression problems. SVR is constructed to work in high-dimensional spaces, which makes it suitable when the number of dimensions is larger than the number of samples. It seeks a function that differs from the observed values by a value less than a certain acceptable level.
- **Multi-Layer Perceptron Regressor (MLPRegressor)**: A kind of ANN that allows for modeling non-linear associations in the information via several hidden layers of nodes. Each layer includes some neurons in which a weighted sum of inputs is taken and the results are put through a non-linear activation function that allows the network to learn complex patterns.
- **Gradient Boosting Regressor**: A technique used for making successive models for solving a particular problem. The new model, known as the slave model, attempts to correct errors made by the master model. This approach can be used to create a good model by summing up weak models because they combine their strengths and overcome their weaknesses.
- **CatBoost**: An algorithm based on gradient boosting for categorical features without preliminary data transformation. It is characterized by the best performance, simple operations, and the capacity to predict with less changing of parameters precisely.
- **XGBoost**: A fast and efficient algorithm for gradient boosting that has proven significantly successful in many machine learning challenges. XGBoost includes enhancements like weight regularization, which helps avoid overfitting, and parallel processing, which increases computation speed.
- **Random Forest Regressor**: Combines multiple decision trees in parallel, where each tree uses a different subset of the data and features. Random Forest reduces high rates of overfitting by averaging numerous trees, thus increasing the level of predictability.
- **Extra Trees Regressor**: Similar to Random Forest but introduces more randomness when choosing splits, which results in even further reduction of variance and better generalization.
- **Decision Tree Regressor**: A model that is easy to understand and distills the data into regions based on the input features. Decision Trees provide high accuracy for training data but typically suffer from overfitting.
- **K-Neighbors Regressor**: Does not assume any specific distribution and makes predictions by averaging the outcomes of the k-nearest neighbors in the training dataset. It is easy to implement and understand but depends on the factor of k and the distance measure used.

4.2 Description of the Validation Techniques Employed

To ensure the robustness and generalizability of our models, we employed rigorous cross-validation techniques and evaluated performance using a variety of metrics:

https://doi.org/10.54216/JAIM.070205

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

- **Cross-Validation**: K-fold cross-validation was employed as the validation technique and is regarded as one of the most commonly utilized approaches. In k-fold cross-validation, the dataset is split into k-equal subsets. k-1 subsets are used to train the model, while one subset is used to validate the model. This process is repeated k times (e.g., k = 5, 10, 15), so each subset becomes a validation set once. The performance estimates from each iteration are averaged to give a final estimate. This method reduces the overfitting problem and generally provides a better estimation of how well the model performs on unseen data.

4.3 Performance Metrics Used for Evaluation

We utilized several performance metrics to comprehensively evaluate the models, ensuring a thorough assessment of their predictive power:

- **Mean Squared Error (MSE)**: Measures the variability of the errors in the data by averaging the squared differences between the predicted and actual values. Models with lower MSE values perform better on the given dataset.
- **Root Mean Squared Error (RMSE)**: The square root of the MSE, measuring the average size of the prediction errors. RMSE is in the same units as the target variable, making it easy for analysts to understand the comparison to the target variable.
- **Mean Absolute Error (MAE)**: The average absolute value of the differences between the predicted and actual values. It is simple to compute and provides an immediate indicator of forecast accuracy. MAE is less affected by outliers than MSE.
- **Mean Bias Error (MBE)**: Calculates the mean of the absolute differences between target values and predictions, clarifying the direction of the predictions' bias. A value close to zero indicates negligible bias.
- **Correlation Coefficient (r)**: Describes how well the model fits a straight line through the data points by summarizing the positive or negative association between the predicted and actual values. Values closer to 1 or -1 indicate a stronger relationship.
- **Coefficient of Determination (R^2)**: Measures the extent to which the independent variables explain the variation in the dependent variable. Higher values indicate a better model fit.
- **Nash-Sutcliffe Efficiency (NSE)**: Measures the accuracy of the models in terms of prediction, with higher values indicating more accurate models. NSE tests the model's predictive skills by comparing it to the average of the observed data.
- **Willmott's Index (WI)**: Calculates the accuracy by squaring the errors between observed and predicted values and summing them. A figure closer to 1 indicates better performance. WI takes both the size and the sign of the errors into account.

These multiple measures were used to comprehensively assess the outcomes of different models and identify the best-performing models for predicting student performance.

5 Experimental Results

5.1 Presentation of the Results Derived from Each of the Models

To test the performance of the regression models developed to predict student performance, the following measures were conducted on each model. The following table summarizes the results for each model:

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

Table 1: Performance Metrics for Various Regression Models

Models	**MSE**	**RMSE**	**MAE**	**MBE**	**r**	**R^2**	**NSE**	**WI**	**Fitted Time (s)**
Linear Regression	0.00052	0.0228	0.0181	0.0007	0.9943	0.9887	0.9887	0.9499	0.0258
Pipeline	0.00052	0.0228	0.0181	0.0006	0.9943	0.9887	0.9887	0.9499	0.2771
SVR	0.00052	0.0229	0.0181	0.0008	0.9943	0.9887	0.9887	0.9499	27.6520
MLP	0.00053	0.0231	0.0184	0.0003	0.9942	0.9884	0.9884	0.9492	21.1404
Gradient Boosting	0.00055	0.0235	0.0186	0.0007	0.9940	0.9880	0.9880	0.9485	0.0056
CatBoost	0.00055	0.0235	0.0187	0.0008	0.9940	0.9880	0.9880	0.9483	24.1672
XGBoost	0.00056	0.0237	0.0188	0.0008	0.9939	0.9879	0.9878	0.9480	50.6825
Random Forest	0.00062	0.0249	0.0198	0.0012	0.9933	0.9866	0.9865	0.9451	3.4385
Extra Trees	0.00062	0.0251	0.0198	0.0010	0.9932	0.9864	0.9863	0.9451	0.0060
Decision Tree	0.0008	0.0284	0.0226	0.0004	0.9912	0.9825	0.9825	0.9374	0.3027
K-Neighbors	0.00119	0.0346	0.0276	0.0008	0.9886	0.9774	0.9740	0.9237	0.0847

Analyzing the results depicted in Table 1, Linear Regression was seen to have the lowest MSE score of 0.0005216, which suggests that it performed better than the other models used in the study to predict student performance. Additionally, Linear Regression showcased satisfactory results regarding other accuracy measures, including RMSE, MAE, and R^2, implying that the model sufficiently captured the underlying relationship between forecasting variables.

5.2 Comparison of the Models Based on Performance Metrics

- **Linear Regression**: Achieved the best results with the lowest MSE, RMSE, and MAE values and a high R^2, indicating that the increase in input features had a direct impact on the target variable.
- **Pipeline and SVR**: Both models had metrics close to the Linear Regression model, though the MSE and RMSE were slightly higher. SVR demonstrated an ability to handle non-linear relationships effectively.
- **MLP Regressor**: Displayed slightly more errors than Linear Regression but still performed reasonably well, showcasing its ability to analyze diverse patterns due to its neural network structure.
- **Ensemble Models (Gradient Boosting, CatBoost, XGBoost, Random Forest, Extra Trees)**: These models also worked well but were comparatively slower in computation. Gradient Boosting and CatBoost provided satisfactory accuracy but were surpassed by less complex models in terms of error metrics.
- **Decision Tree and K-Neighbors Regressor**: These models yielded much higher error rates, indicating their limitations in capturing the rich features within the data. Decision Trees tend to overfit the training data, while the K-Neighbors algorithm is sensitive to the choice of k and distance metrics.

5.3 Analysis of the Strengths and Weaknesses of Each Model

- **Linear Regression**: Strengths include ease of implementation and interpretation, and good performance on this dataset. However, it may fail to manage complex relationships in other datasets.
- **SVR and MLP Regressor**: Both can find non-linear relationships, with SVR being more efficient in high-dimensional spaces. MLP Regressor can capture interactions between predictors but requires proper hyperparameter selection.
- **Ensemble Models**: Robust and efficient in solving problems with non-linear relations. However, they are computationally intensive and require significant parameter tuning.
- **Random Forest and Extra Trees**: Reduce overfitting by averaging predictions from multiple models but are less interpretable than simpler models.

https://doi.org/10.54216/JAIM.070205

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

- **Decision Tree and K-Neighbors**: Conceptually simple but computationally complex, prone to overfitting, and sensitive to hyperparameters, resulting in higher errors.

Figure 3 presents the correlation matrix that depicts the findings gotten from the assorted regression analysis models. This heatmap shows the correlation of performance of different models, and thus, it shows which models have similar or dissimilar tendencies for different questions. Determining the relation of models can be examined by comparing their high correlation value, which shows that the models give out similar results and quantifies their reliability. On the other hand, low correlation values show that the models are dissimilar, pointing to their specific strengths and/or weaknesses. It is useful when choosing simpler models that can be used in the ensemble technique to enhance the model's predictive accuracy.

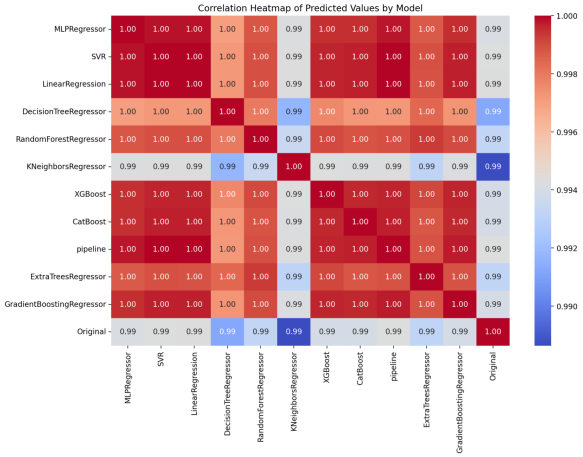

Figure 3: Correlation Heatmap of the Regression Models Results

In what follows, we present a story of results extracting a boxplot of the presented regression models' performance metrics of interest in Figure 4. From this boxplot, it is easy to compare models' accuracy, dispersion, and the existing outliers in their predictions. By analyzing the statistical dispersion of the performance measures, including Mean Square Error, Root Mean Square Error, Mean Absolute Error and the coefficient of determination, one can differentiate between models with stable and relatively low volatility and those with high volatility and possible outlier values. It aids in some critical choices about the deployment of models, as it classifies those the models most likely to be trusted and indicates how stable their forecast is.

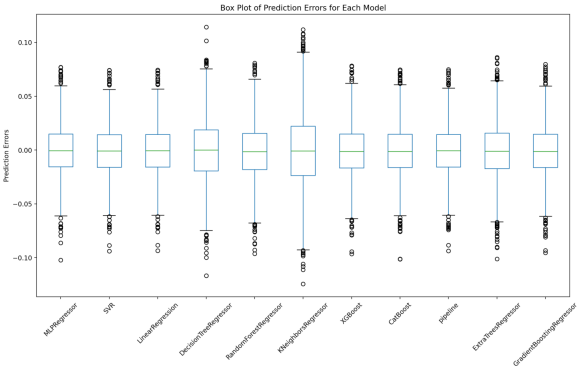

Figure 4: Boxplot of the Regression Models Results

5.4 Implications of the Study about Student Performance Prediction

From the results of this study, it can be concluded that regression models are useful in measuring student performance. Linear Regression was the most straightforward method, yet it offered the best performance, which aligns with the linear presented relationships. The performance of SVR and MLP Regressor confirms

that these models are appropriate for identifying more detailed patterns in the data. Ensemble models are powerful but require more computational resources and parameter tuning. Overall, the study demonstrates the applicability of various regression models in predicting student performance and highlights the trade-offs between simplicity and predictive power.

6 Conclusion

This research aimed to assess the efficiency of several regression models for estimating the learner's performance based on the educational records data set. Using the techniques of preprocessing data such as media, converting the data into float64, handling the missing values, and applying models including Linear Regression, SVR, MLPRegressor, Gradient Boosting, CatBoost, XGBoost, Random Forest, Extra Trees, Decision Tree, and K-Neighbors models the best model is determined to be Linear Regression with MSE of 0. 000521. Although models like SVR and MLPRegressor offered more accurate results because they captured higher-order relationships, they used more time in their computations. The study shows how regression models apply in forecasting student performance, and among them, linear regression is the simplest yet most effective method. As for the limitations and future research themes, it can be recommended to focus on the following: advanced data preprocessing and feature engineering to increase effectiveness; the utilization of other machine learning techniques, including deep learning and ensemble methods, to improve the results; active use of domain knowledge to improve the relevance of the results and the predictive performance of the models. Moreover, establishing the real-time prediction feature and further improvement of the model generalization and scaling in various fields of educational datasets are important for the increased application performance in educational environments.

References

[1] Raed Alamri and Basma Alharbi. Explainable student performance prediction models: A systematic review. *IEEE Access*, 9:33132–33143, 2021.

[2] Bashayer Albreiki, Nazar Zaki, and Hamdi Alashwal. A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, 11(9):Article 9, 2021.

[3] Abderrahmane Asselman, Mohamed Khaldi, and Salma Aammou. Enhancing the prediction of student performance based on the machine learning xgboost algorithm. *Interactive Learning Environments*, 31(6):3360–3379, 2023.

[4] Şenay Aydoğdu. Predicting student final performance using artificial neural networks in online learning environments. *Education and Information Technologies*, 25(3):1913–1927, 2020.

[5] Sobia Batool, Jibran Rashid, Muhammad Waqas Nisar, Jinyoung Kim, Hyung-Yoon Kwon, and Abid Hussain. Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1):905–971, 2023.

[6] Ghaith Ben Brahim. Predicting student performance from online engagement activities using novel statistical features. *Arabian Journal for Science and Engineering*, 47(8):10225–10243, 2022.

[7] Prashant Dabhade, Ruchir Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar. Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, 47:5260–5267, 2021.

[8] Caroline M. A. Gomes, Antonio Amantes, and Edmilson G. Jelihovschi. Applying the regression tree method to predict students' science achievement. *Trends in Psychology*, 28(1):99–117, 2020.

[9] Shabbir Hussain and Muhammad Qamar Khan. Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of Data Science*, 10(3):637–655, 2023.

https://doi.org/10.54216/JAIM.070205

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024

[10] Peng Jiao, Fang Ouyang, Qiang Zhang, and Amir H. Alavi. Artificial intelligence-enabled prediction model of student academic performance in online engineering education. *Artificial Intelligence Review*, 55(8):6321–6344, 2022.

[11] Ajmal Khan and Suman Kumar Ghosh. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, 26(1):205–240, 2021.

[12] Imtiaz Khan, A. R. Ahmad, Noura Jabeur, and Mohammed N. Mahdi. An artificial intelligence approach to monitor student performance and devise preventive measures. *Smart Learning Environments*, 8(1):17, 2021.

[13] Ahmed Nabil, Mina Seyam, and Alaa Abou-Elfetouh. Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access*, 9:140731–140746, 2021.

[14] Ahmad Namoun and Abdullah Alshanqiti. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1):Article 1, 2021.

[15] Salma Rebai, Faycal Ben Yahia, and Hatem Essid. A graphically based machine learning approach to predict secondary schools performance in tunisia. *Socio-Economic Planning Sciences*, 70:100724, 2020.

[16] Marta Riestra-González, Marta del P. Paule-Ruíz, and Fernando Ortin. Massive lms log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, 163:104108, 2021.

[17] Camilo F. Rodríguez-Hernández, Mariel Musso, Eva Kyndt, and Eduardo Cascallar. Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, 2:100018, 2021.

[18] Nikola Tomasevic, Nataša Gvozdenovic, and Sanja Vranes. An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, 143:103676, 2020.

[19] Huma Waheed, Sameer-Ul Hassan, Nasser R. Aljohani, Jan Hardman, Saleh Alelyani, and Rameez Nawaz. Predicting academic performance of students from vle big data using deep learning models. *Computers in Human Behavior*, 104:106189, 2020.

[20] Mustafa Yağcı. Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1):11, 2022.

[21] Bilal Khushal Yousafzai, Muhammad Hayat, and Sadaf Afzal. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. *Education and Information Technologies*, 25(6):4677–4697, 2020.

[22] Hussein Zeineddine, Udo Braendle, and Anton Farah. Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 89:106903, 2021.

https://doi.org/10.54216/JAIM.070205

Received: June 01, 2023 Revised: September 28, 2023 Accepted: February 25, 2024