iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65404

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

Regression and Ensemble Models for Analyzing Students key Performance factors

Pranav Daware[1], Anurag Jagtap[2], Ashish Khenat[3], Gayatri Mohite[4], Dr. Manisha Mali[5]
Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India

Abstract: ***Predicting student performance can help identify when students need extra support and tailor interventions to improve outcomes. In this study, we use machine learning to explore the factors influencing student success and build the models that can predict. We apply several algorithms , including Linear Regression , "Random Forest Regression" , " Gradient Boosting Regression", " Support Vector Regression", " K-Nearest Neighbors Regression" , each chosen for their ability to capture different patterns in data , whether linear or complex . The models are trained on factors such as attendance, socioeconomic status, parental education , and academic history to understand how these impact student achievement and learning . Our findings show that ensemble techniques like Gradient Boosting and Random Forest tend to provide more accurate predictions than traditional regression models. By comparing these models, we aim to offer insights into how machine learning can help predict and support student performance, helping educators make more informed decisions about learning. This also helps personalized decision-making to support learners effectively.***

Keywords: Student performance, Regression models, Ensemble learning, Machine learning, Educational data mining (EDM), Predictive Analytics, Linear regression, " Random Forest", " Gradient Boosting", " Support Vector Regression", " K-Nearest Neighbors", Academic achievement, Socioeconomic status, Attendance, Parental education, Hyperparameter tuning, Predictive modeling, Student success, Early intervention

I. INTRODUCTION

Institutions are increasingly resorting to technology in the quickly changing field of education to address the difficulties associated with evaluating student performance. Accurate academic performance prediction is essential for identifying children who might require extra help and for developing successful intervention plans. Numerous studies have been conducted on the factors that affect student performance, including socioeconomic background, academic history, and demographic information.However, machine learning and educational data mining (EDM) have brought a new degree of analysis and prediction to the study of student achievement, coinciding with the growth of data-driven decision-making. "Educational data mining" (EDM), is a rapidly evolving area that leverages data mining techniques to analyze and extract insights from large educational datasets. In addition to predicting student achievement, educational institutions can employ the study of these datasets to comprehend the underlying factors that impact learning results. This process helps teachers make data-driven decisions, which enhances instruction and student engagement. With the development of technology, machine learning models—which offer a more dynamic and accurate way of forecasting results based on historical and real-time data—have become indispensable tools for predicting academic success. Our aim in this work is to investigate several elements of student performance by employing many popular machine learning models. These models were chosen because they demonstrated performance in prediction tests and had a variety of methodological approaches.

II. LITERATURE REVIEW

Predicting student achievement using educational data mining has seen a rise in the usage of machine learning (ML). These prediction models assist teachers in spotting potential problem pupils so they can intervene and provide support before it's too late. This review of the literature highlights three significant studies that predict student achievement using various supervised machine learning algorithms, emphasizing critical elements and the efficacy of each strategy.

A. Machine Learning Algorithms for Assessing Students' Performance (2020)[2]

Many applied machine learning research studies have been conducted. Several methods have been shown to be effective in predicting student performance. For example, the accuracies realized by the "Random Forest" and " Support Vector Machine" models were 79% and 75%, respectively. Deep Neural Networks achieved the highest accuracy of 84%.

These results illustrate how machine learning should use a combination of academic, behavioral, and demographic data to identify students in need of immediate intervention. Furthermore, decision trees are highly appreciated for being easy to read and useful for the user, they are very useful for teachers who want decisions to be fact based Using all these high-tech techniques in a learning environment that leads to learning outcomes, much- It can also be a supportive learning environment.

B. Using Supervised Learning Algorithms to Predict Student Success (2020)[3]

Hashim and his team [8] looked into how well different methods work to figure out if student's passing or failing . They checked out "Decision Trees", "Naive Bayes", "Logistic Regression", and "Support Vector Machine"(SVM). Logistic Regression came out on top with an 88.8% success rate when using data from various universities to predict failure. This research shows that mixing academic, behavioral, and demographic info can make predictions more accurate. It also helps schools provide support to students before they fail.

C. A Comparative Analysis of Machine Learning Algorithms for Student Performance Prediction (2021)[5]

To estimate how students might perform using data from an online learning platform, El Guabassi et al. [7] compared seven machine learning techniques, including "Logistic Regression", "Support Vector Regression" (SVR), and "Random Forest Regression" (RFR). Log-linear regression yielded the best predictions for behavioral indicators such as frequency of participation in class activities or of use of learning materials. This work underscores the need to identify ahead students at risk of underperforming 28,29 and indicates that behavior is a particularly relevant type of early predictor.

D. Elements Influencing College Students' Forecast and Outcomes (2022)[6]

Wang et al. [9] studied the main factors that affect student performance and developed prediction models using "Naive Bayes", "Random Forest", "Support Vector Classifier" (SVC) and "Logistic Regression". Their results indicated that SVC classification among all the other classifiers produced the highest overall accuracy with 80.96%. They found that both academic and environmental circumstances along with students' study habits and attitudes towards learning affect their performance. Teachers can use these beneficial insights in order to design more effective intervention systems for those students who are at risk of failing.

E. Machine Learning Algorithms for Predicting Student Performance (2022)[7]

Dervenis et al. (2018) have used various machine learning methods to predict student performance. They noted that the inclusion of socioeconomic features together with previous academic data would enhance the prediction accuracy. Through a comparative study using several algorithms like "Decision Trees", "Random Forests" and "Deep Neural Networks" they have shown how machine learning can help identify students who are at risk by providing appropriate intervention in a timely manner.

F. Machine Learning Algorithms used to Predict Student's Performance (2023)[8]

The authors conducted a systematic analysis of machine learning methods applied and utilized to the performance prediction of students. They put to test several algorithms, such as Artificial Neural Networks, Decision Trees and Naïve Bayes, and stressed how important it is to use both cognitive and non-cognitive elements to increase the precision. Additionally, they argued that using behavioral data models allows achieving a considerable improvement of results – in some cases over 90% of accuracy. The study demonstrated the fact that with the use of machine learning we can facilitate prompt interventions, thus leading indirectly to better academic performance.

G. Machine Learning Algorithms for Predicting Student Performance (2024)[9]

To predict the outcomes of students, Dervenis et al. study several machine learning models. Furthermore, several socioeconomic features are used together with previous educational data to increase prediction rates. Models presented herein are able to predict low performance early enough for timely intervention using a set Decision Trees, Random Forests and Deep Neural Networks. The study exemplifies how machine learning can transform teaching and student performance.

H. Recurring Subjects in the Research

A couple of the basic ideas are common to all of these studies:

1) *Top Algorithms:* Logistic regression gave good results frequently, especially when predicting if students will pass or fail. SVM and Decision Trees were as good even though at times they required more processing power.

2) *Key Elements:* All studies find the most accurate predictions are obtained with models that incorporate behavioral, academic, and demographic data. Some of the most important features are how often students participate in class and utilize available resources.
3) *Early Intervention:* With these models teachers can identify troublesome children early in their development, which is necessary to intervene with timely help, and indeed this leads schools to improve performance and reduce dropout rates by doing so.

These examples are evident how powerful machine learning can be at predicting student's success. Thus, by using the data on student demographics performance and behavior, teachers and students can both anticipate the problems early and provide the necessary support to the students. This will in turn increase the academic performance and will help the students succeed.

III. MATERIALS AND METHODOLOGY

In recent times, "machine learning"(ML) technique has emerged as a crucial technology for student performance analysis in educational settings. This is because institutions of learning can better understand the factors which lead to better students' performance by employing advanced methods. Here, "Random Forest", "Gradient Boosting", "K-Nearest Neighbors "and "Linear Regression" machine learning models are utilized. These models were selected due to their capacity and strength in the detection of diverse data characteristic patterns. As such, we are optimistic that the significant determinants of academic achievement can be determined using these methods. With this study, our research aims to provide additional knowledge to teachers that can help improve the students' effectiveness allowing more targeted support.

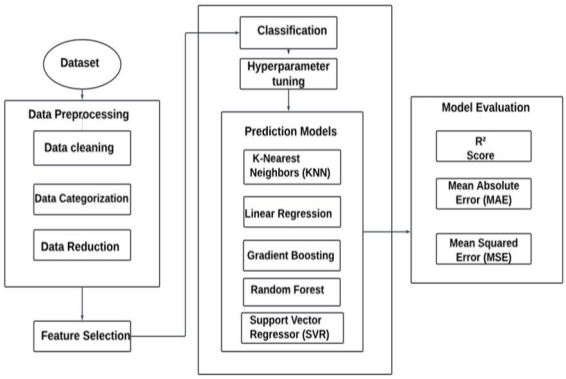


Figure 1 : Workflow diagram of methodology

A. *Methodology*

The proposal made in this work considers four constituent parts that are a recommender system, data preparation, hyperparameter tuning, and model evaluation. Each of these primary parts includes additional characteristics that enhance the performance of the various models. In Figure 1 model architectural representation is visually demonstrated in relation to the integration of these components and the analysis as well as the enhancement of the student performance. This systematic reasoning enhances the understanding as well as the application of the gained knowledge while reducing the time taken in assisting the educators in making decisions that are intended to positively impact their students.

B. Dataset

The dataset used for this study was gathered from Kaggle as well, and it consists of various performance-related features of students and contains variables like:

1) Study Hours: The quantity of time pupils spend studying.
2) Sleep Hours: How much sleep did students get?
3) Physical Activity: The quantity of hours spent working out.
4) Attendance: The proportion of a student's classes that they attend.
5) Prior Results: The results of earlier academic tests.
6) Exam scores: Shows how well students performed on the exam.

C. Data Preprocessing

Preprocessing was crucial to ensuring the data was ready for machine learning models.

The following measures were implemented:

1) *Handling Missing Values:* When there are missing data points, machine learning models may not function as well. The missing values in the dataset were appropriately handled. For numerical qualities, the mean or median values were utilized to add in the missing values based on the data distribution. For categorical features, the mode was used for imputation.
2) *Feature Encoding:* With the help of label encoding Categorical variables such as gender or study programs were transformed into numerical data . This allowed the machine learning algorithms to process categorical data efficiently.
3) *Feature Scaling:* Scaling applied to ensure that numerical variables were within the same range. Techniques like StandardScaler were used to standardize features like study hours, sleep hours, and physical activity, while MinMaxScaler was applied to normalize features such as attendance and previous scores. This ensured that no feature dominated the others due to its range.
4) *Outlier Detection and Removal:* Using statistical methods like IQR (Interquartile Range) outliers were detected in continuous variables method. These outliers were either transformed or removed based on their potential impact on model performance.

D. Model Selection

Several machine learning regression models were applied to predict students' exam scores.

The models include:

1) *Support Vector Regressor (SVR):* SVR, powerful algorithm for regression problems, especially when the relationship between the target variable and features is non-linear. We used a radial basis function (RBF) kernel to capture non-linearity in the data.
2) *Linear Regression:* This model establishes a linear relationship between the independent variables and the target variable. It served as a baseline model to compare the performance of more complex algorithms.
3) *Random Forest Regressor:* To improve prediction accuracy, the Random Forest ensemble learning technique makes use of a number of decision trees. It can capture complex associations between variables by averaging the results of several trees.
4) *Gradient Boosting Regressor:* Models are built sequentially using an ensemble technique, with each new model seeking to address the shortcomings of the prior one. It functions well with datasets of a respectable size and provides good accuracy.
5) *The K-Nearest Neighbors (KNN):* Regressor is a non-parametric model that uses the nearest neighbors in the dataset to predict values. It is particularly useful when there is a strong localized connection between the variables.

E. Model Training and Hyperparameter Tuning

After selection, each model was trained on the pre-processed dataset.

The following models' hyperparameters were changed using GridSearchCV to improve performance:

1) *SVR, or Support Vector Regressor:* The following hyperparameters have been tuned: C, kernel, and epsilon.
2) *Random Forest Regressor: The* Random Forest Regressor's adjusted hyperparameters are min_samples_split (the least number of samples needed to split a node), max_depth (the maximum depth of trees), and n_estimators (the number of trees).
3) *Gradient Boosting Regressor:* Tuned hyperparameters: n_estimators, learning_rate , max_depth.
4) *KNN Regressor:* Tuned hyperparameters: n_neighbors (number of neighbors), and weights.

Using grid search and 5-fold cross-validation, the best set of hyperparameters for every model were discovered. Cross-validation helps ensure that the model is not overfitting to the training data and generalizes well to unseen data.

F. Model Evaluation

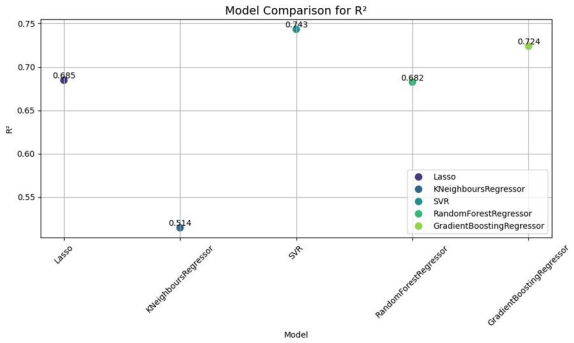
For the performance evaluation following metrics were used :

1) *"R^2 Score (Coefficient of Determination)":* This metric indicates how well the independent variables explain the variability in the target variable. A higher R^2 value signifies a better fit of the model.
2) *"Mean Absolute Error (MAE)":* MAE measures the average magnitude of the prediction errors. It is the average over the test data of the absolute differences between predicted and actual values. Lower MAE values indicate better accuracy.
3) *"Mean Squared Error (MSE)":* MSE computes the average of the squares of the prediction errors, giving more weight to larger errors. It is useful when you want to penalize large errors more heavily.

IV. RESULTS

Model	Test R^2	MAE	MSE
Linear Regression	0.684107	1.056981	4.465165
Random Forest Regressor	0.650749	1.1459	4.936688
Gradient Boosting Regressor	0.723698	0.859856	3.905554
SVR	0.735527	0.743988	3.73835
K Neighbors Regressor	0.497162	1.679274	7.107655

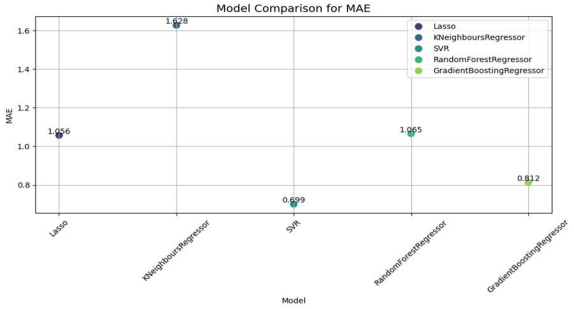
Table 5.1 Results Before Tuning



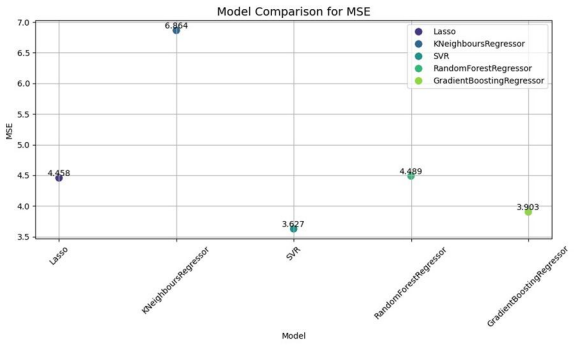
Graph 5.1 Model Comparison for R^2

Model	Best Hyperparameters	R^2	MAE	MSE
Linear Regression	{'alpha': 0.01}	0.6846	1.0564	4.4581
K-Neighbours Regressor	{'n_neighbors': 7, 'weights': 'distance'}	0.5143	1.6275	6.8642
SVR	{'C': 10.0, 'kernel': 'rbf'}	0.7433	0.6991	3.6273
Random Forest Regressor	{'max_depth': None, 'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 200}	0.6824	1.0648	4.4890
Gradient Boosting Regressor	{'learning_rate': 0.2, 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 100}	0.7238	0.8117	3.9032

Table 5.2 Results After Tuning



Graph 5.2 Model Comparison for MAE



Graph 5.3 Model Comparison for MSE

A. Linear Regression

- Test R^2: 0.6846 — It explains about 68.41% of the variance in the target variable.
- MAE: 1.0564, MSE: 4.4581 — Moderate errors, indicating a decent fit but not the most accurate model.

B. Random Forest Regressor

- Test R^2: 0.6835 — Explains slightly less variance than Linear Regression.
- MAE: 1.0655, MSE: 4.4727 — Higher error values than Linear Regression, indicating larger prediction errors.

C. Gradient Boosting Regressor

- Test R^2: 0.7236 — Performs better than Linear Regression and Random Forest.
- MAE: 0.8129, MSE: 3.9065 — Lower errors, showing better prediction accuracy.

D. Support Vector Regressor (SVR)

- Test R^2: 0.7433 — Best R^2 values, explaining the most variance.
- MAE: 0.6991, MSE: 3.6273 — Lowest error metrics, making it the most accurate model in this comparison.

E. K-Neighbors Regressor

- Test R^2: 0.5143 — The lowest R^2 values, explaining the least variance.
- MAE: 1.6275, MSE: 6.8642 — Highest errors, showing the least accurate predictions.

Best Model: Support Vector Regressor (SVR)
Highest Test R^2 (0.7433) — Explains the most variance in the target variable.
Lowest MAE (0.6991) — Smallest absolute prediction errors.
Lowest MSE (3.6273) — Smallest squared errors, indicating better generalization and lower error magnitude.
SVR is the best model due to its superior performance across R^2, MAE, and MSE metrics.

V. CONCLUSION

In this research , we explored various machine learning models to predict student performance by examining key factors such as attendance , socioeconomic status , and academic history . We applied models like “Linear Regression” , “Random Forest Regressor” , “Gradient Boosting Regressor” , “Support Vector Regressor” , and “K-Nearest Neighbors Regressor” , with the goal of understanding which approaches provide the most accurate predictions .

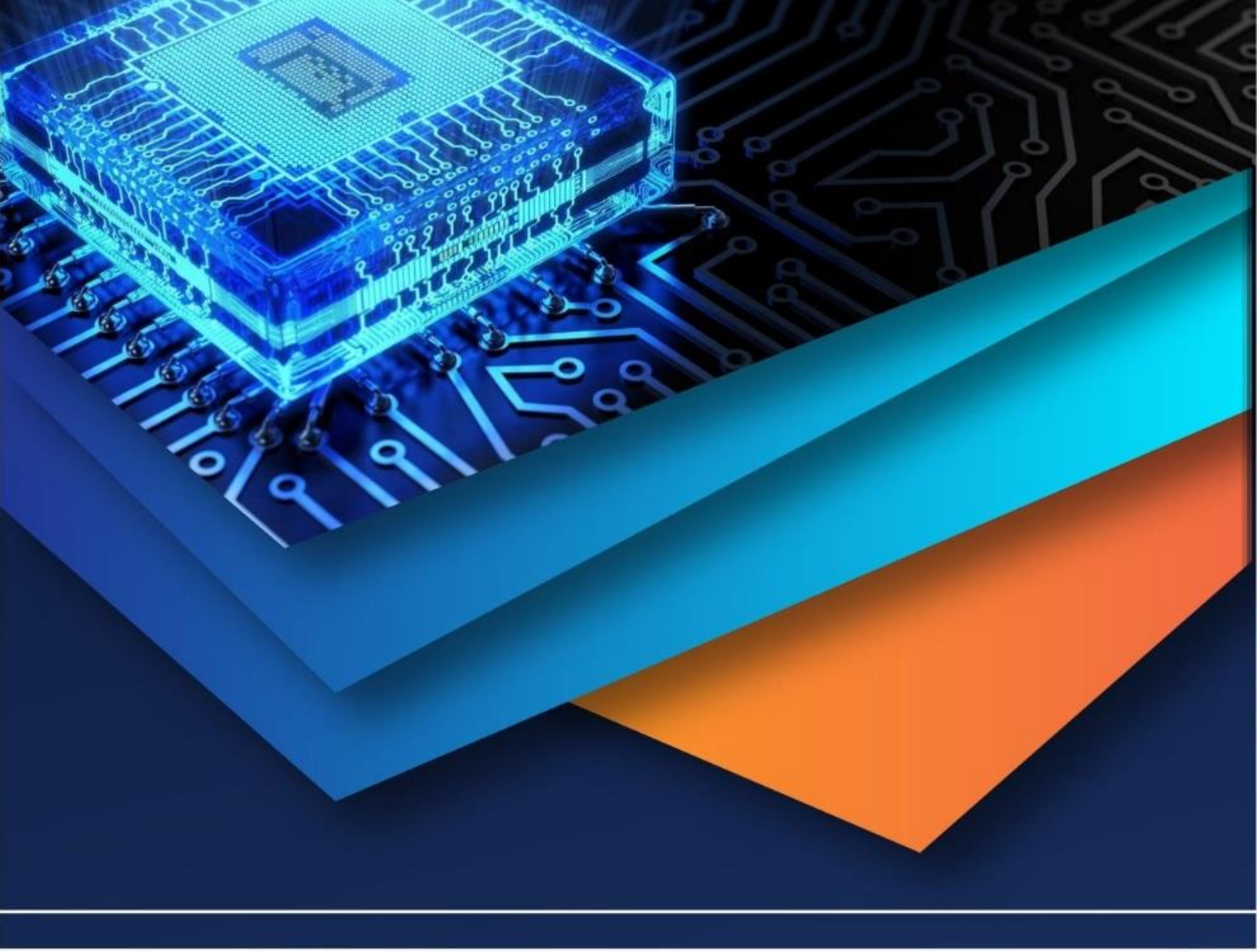
Based on our results, the best performance level allows for these models; especially, the “Gradient Boosting Regressor” and “Support Vector Regressor” (SVR) – performed better than other implemented models . In particular, the SVR model showed the best result as it had the least prediction errors and the highest R2. This indicates that learners are able to outperform when a model is able to explain the nonlinear relationships which exist among the various variables, in this case the student outcomes.

The insight obtained from this research could benefit the schools in terms of locating students who would need extra help. The teachers are able to follow these through the use of these machine learning models to predict and therefore target and implement interventions at an early stage and those interventions are specific to those individual students’ potentials Morrison, Jones, and Swanson 359. Here, the effectiveness of data-based working in schools is shown and complementary ways are opened up for the differentiation of the instruction and the enhancement of students’ achievement.

Further research may look at the methods of predicting models more accurately by focusing on additional variables or active data collection. It is reasonable to consider that the results of the research provide a solid foundation for the implementation of machine learning methods to understand and foster academic achievement in students' contexts.

REFERENCES

[1] J. Sultana, H. Farquad, and M. U. Rani, "Student's Performance Prediction using Deep Learning and Data Mining methods," ResearchGate, Article, Jun. 2019
[2] S. F. Aziz, "Students' Performance Evaluation Using Machine Learning Algorithms," University of AL-Hamdaniya, Mosul, Iraq, Jul. 2020.
[3] A. S. Hashim, R. Hamoud, and M. A. Obaid, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," IOP Conference Series: Materials Science and Engineering, vol. 928, p. 032019, 2020
[4] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," Universidad de Extremadura, Cáceres, Spain, Feb. 2020.
[5] I. El Guabassi, Z. Bousalem, R. Marah, and A. Qazdar, "Comparative Analysis of Supervised Machine Learning Algorithms to Build a Predictive Model for Evaluating Students' Performance," International Journal of Online and Biomedical Engineering, vol. 17, no. 2, pp. 20025, 2021.
[6] D. Wang, D. Lian, Y. Xing, S. Dong, X. Sun, and J. Yu, "Analysis and Prediction of Influencing Factors of College Student Achievement Based on Machine Learning," Hebei Agricultural University, China .2022
[7] S. F. Aziz, "Students' Performance Evaluation Using Machine Learning Algorithms", Researchgate.2022.
[8] S. O. Oppong, "Predicting Students' Performance Using Machine Learning Algorithms: A Review," Asian Journal of Research in Computer Science, vol. 16, no. 3, pp. 351, 2023.
[9] E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," College of Informatics, Wollo University, Dessie, Ethiopia, Apr. 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY