IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Unlocking Success: A Smart System For Predicting Student Performance And Recommending Courses

[1]Aayushi Patel, [2]Sanika Pathak, [3]Nidhi Khadke, [4]Shreya Mukherjee, [5]Nayanshree Purbia, [6]Reshma Sonar

[1]Student, [2]Student, [3]Student, [4]Student, [5]Student, [6]Professor

[1]BTech CSE,

[1]MIT World Peace University, Pune, India

Abstract: This research paper describes a system for predicting student performance and recommending courses using predictive analytics. The system utilizes machine learning models such as Linear Regression, Support Vector Regression (SVR), and Random Forest to accurately forecast student performance based on historical academic data, study habits, and course preferences. Moreover, it generates personalized course recommendations specific to each student's profile, academic goals, and learning needs. Experimental evaluation demonstrates the effectiveness of the proposed approach in predicting performance and providing relevant course suggestions, which can significantly enhance academic outcomes and student satisfaction.

Keywords: student, performance, prediction, recommendation, machine learning, linear regression

***Index Terms* -** Index, Literature Review, System Architecture, Methodology, Proposed Model, Result Analysis, Conclusion, Acknowledgement, References

I. INTRODUCTION

In today's education sector, accurately predicting student performance and providing personalized course recommendations is highly valuable. Educational institutions and online learning platforms are constantly seeking innovative solutions to enhance student success rates and improve learning outcomes. Machine learning techniques, such as linear regression, support vector regression (SVR), and random forest, have emerged as powerful tools in this domain.

This research aims to develop a comprehensive Student Performance Prediction and Course Recommendation System using machine learning algorithms. The primary objective is to use historical student data to predict future performance accurately and recommend courses that cater to each student's needs and abilities. By integrating predictive analytics and personalized recommendations, this system aims to improve student engagement, retention, and academic success.

Our approach comprises data preprocessing, model training using linear regression, SVR, and random forest, performance evaluation, and course recommendation based on predicted outcomes. Through rigorous experimentation and analysis, we aim to demonstrate the efficacy and scalability of our predictive and recommendation system in real-world educational settings.

The rest of this paper is organized as follows: Section II provides a detailed review of related work in student performance prediction and course recommendation systems. Section IV describes the methodology, including data collection, preprocessing techniques, and the implementation of machine learning models. In Section VI, we present experimental results, including model performance metrics and course recommendation accuracy. Finally, Section VII concludes the paper with a summary of findings, contributions, and avenues for future research.

II. LITERATURE REVIEW

Singh, A. S. Sabitha and A. Bansal, "Student performance analysis using clustering algorithm," [1]. The paper showcases a proactive approach towards understanding and enhancing student performance within educational institutions by employing data mining techniques. This modern analytical method indicates a willingness to embrace technological advancements in improving educational outcomes, fostering an environment of continuous improvement and innovation.

S. Rajput and S. Ramesh, "Student Performance Analysis based on Machine Learning Algorithms," [2]. The paper introduces a rule-based recommender system designed to analyze and forecast student performance, utilizing demographic data and academic abilities to provide tailored recommendations. The emphasis on evaluating machine learning algorithms demonstrates a commitment to technological advancement and personalized student support.

B. Guo, R. Zhang, G. Xu, C. Shi and L. Yang, "Predicting Students Performance in Educational Data Mining," [3]. The research introduces a Deep Learning-based classification model for predicting student academic performance, demonstrating a commitment to utilizing advanced techniques for educational improvement. Experimental results suggest the effectiveness of the proposed method, indicating its potential for implementation in academic pre-warning mechanisms.

M. S. Ram, V. Srija, V. Bhargav, A. Madhavi and G. S. Kumar, "Machine Learning Based Student Academic Performance Prediction," [4]. By utilizing Linear Regression and Random Forest algorithms, the research demonstrates a proactive approach towards improving academic outcomes. The comparison of algorithm performance using metrics like accuracy, precision, recall, and F1 ranking indicates a thorough evaluation process to determine the most effective predictive model for academic achievement.

J. Xu, K. H. Moon and M. van der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," [5]. The research presents a novel machine learning method for predicting student performance in degree programs, addressing challenges such as diverse student backgrounds and varying course relevance. Through innovative features and a data-driven approach, the method achieves superior performance compared to benchmark approaches, as demonstrated by extensive simulations on real undergraduate student data from the University of California, Los Angeles.

T. Kour, R. Kumar and M. Gupta, "Analysis of student performance using Machine learning Algorithms," [6]. This project examines past student outcomes, as well as their individual characteristics such as family history, demographic distribution, age, study attitude, and put this information to the test using diverse machine learning (ML) algorithms in WEKA (Waikato Setting for Knowledge Analysis) tool. The performance of the various algorithms was assessed using the percentage split (80:20) as well as the test-case cross-validation(10-fold).

III. System Architecture

The proposed model was deployed using various steps. The figure below shows the architectural flow of the same.

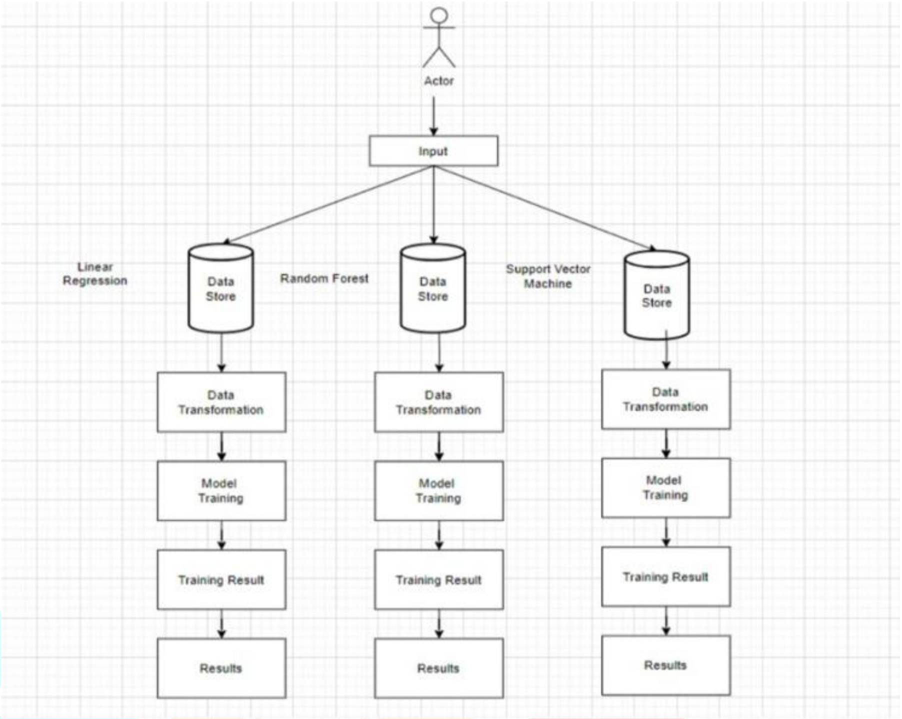


Fig.1. System Architecture

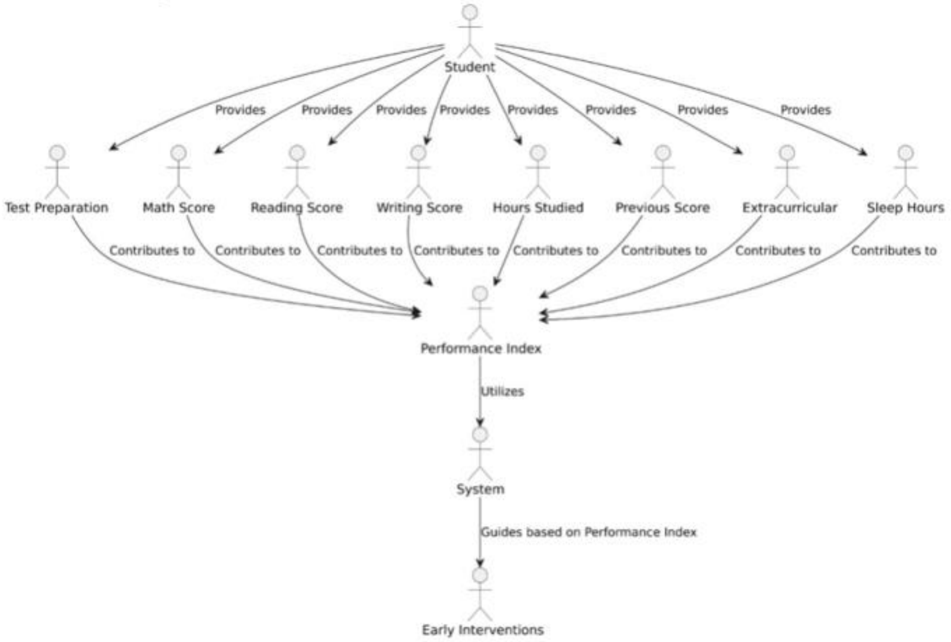


Fig.2. Use Case Diagram

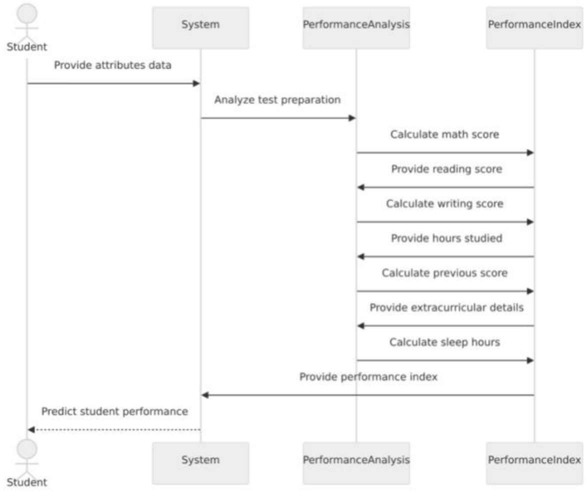


Fig.3. Sequence Diagram

IV. Methodology

A. Data Collection

We have collected data on student performance from datasets: "exams 1.xlsx - exams 1.csv" and "exams 2.csv" and a “real time” dataset and have merged them. The dataset has variables such as gender, race/ethnicity, parental level of education, lunch type, test preparation course completion, math score, reading score, writing score, performance index, and extracurricular activities.

B. Data Preprocessing

Data preprocessing is a crucial step for the preparation of data before model construction. It gives us space to remove duplicates, missing values, and deal with outliers. Load the datasets using pandas library. Check for missing values and handle them appropriately (e.g., imputation with mean/median, dropping rows/columns). Encode categorical variables using LabelEncoder from scikit-learn. Lastly concatenate the two datasets into a single dataframe.

C. Exploratory Data Analysis (EDA)

Visualizations, including histograms, scatter plots, box plots, and violin plots, were used to explore the distribution of the target variable (Performance Index) and relationships between the target variable and other features. Patterns, trends, and outliers in the data were identified.

D. Feature Engineering

In this we have selected relevant features for prediction, such as hours studied, previous scores, extracurricular activities, sleep hours, and sample question papers practiced and normalized or scale numerical features to ensure they have similar ranges.

E. Models

We used two other models are used along with Linear Regression.
a. *Random Forest* is an ensemble learning method used for both classification and regression tasks. It works by constructing multiple decision trees during the training phase and outputs the mean prediction of the individual trees for regression tasks. Random Forest mitigates overfitting by averaging the predictions of multiple decision trees, resulting in a more robust model. In the code, RandomForestRegressor from the sklearn.ensemble module is utilized for regression. It's instantiated and trained alongside the Linear Regression model and Support Vector Machine to compare their performance.
b. *Support Vector Machine (SVM)* is a supervised learning algorithm commonly used for classification and regression tasks. In regression tasks, SVM aims to find the hyperplane that best fits the data points, with the goal of maximizing the margin between the hyperplane and the nearest data points (support vectors). SVM can handle both linear and non-linear relationships between features and the target variable through the use of different kernels (e.g., linear, polynomial, radial basis function). In the code, SVR (Support Vector Regressor) from the sklearn.svm module is employed for regression. It's instantiated and trained alongside the Linear Regression model and Random Forest to compare their predictive performance.

F. Model Selection

We Choose Linear Regression was chosen as the proposed model due to its simplicity and interpretability. It is expected to provide reasonable predictions of student performance based on multiple input features. is expected to provide accurate predictions of student performance based on multiple input features.

G. Model Training

For model training the Linear Regression model was trained using the selected features and the target variable (Performance Index).

H. Model Evaluation

Evaluate the performance of the trained model using metrics such as Mean Squared Error (MSE) and R-squared (R2) score on the testing set then compare the model's performance against other regression models like Random Forest Regression and Support Vector Machine. Cross-validation techniques may also be employed to assess model generalization and robustness.

I. Recommendation System Integration

A recommendation system is incorporated with the model to suggest relevant study materials based on the predicted performance index of students. Leveraging the predictions generated by the Linear Regression model, the recommendation system identifies the subject area for which the student may require additional support. Depending on the forecasted performance index and the selected subject, the system suggests appropriate study resources such as textbooks, online courses, or reference materials tailored to the student's academic needs. Specifically, if the predicted performance index falls below 50%, the system recommends study materials deemed beneficial for students aiming to improve their academic performance. Conversely, if the predicted performance index exceeds 50%, the system suggests advanced study resources suitable for students performing well academically. This integration enhances the model's utility by providing personalized guidance to students, thereby fostering their academic success.

J. Interpretation of Results

We analyzed the Coefficients provided by the Linear Regression model that were analyzed to understand the factors influencing student performance. Insights were drawn from the analysis, considering the limitations and assumptions of the model and data. After that draw insights and conclusions from the analysis, considering the limitations and assumptions of the model and data.

V. PROPOSED MODEL

Linear Regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. In the context of predicting student performance, Linear Regression serves as a valuable tool for understanding how various factors such as hours studied, previous scores, extracurricular activities, sleep hours, and sample question papers practiced impact overall academic achievement. By fitting a linear equation to the observed data points, Linear Regression enables us to make predictions about the dependent variable based on the values of the independent variables. This approach not only provides insights into the relative importance of different predictors but also allows for the interpretation of the magnitude and direction of their effects on student performance. With its simplicity and interpretability, Linear Regression offers a solid foundation for analyzing and predicting student outcomes in educational settings.

A. Model Selection

Linear Regression was chosen for its simplicity, interpretability, and ability to provide reasonable predictions of student performance based on multiple input features.

B. Features

Features used for prediction include hours studied, previous scores, extracurricular activities, sleep hours, and sample question papers practiced. These features capture various aspects of a student's academic preparation, lifestyle, and study habits, which may influence their overall performance.

C. Training and Evaluation

The Linear Regression model is trained using the selected features and the target variable (Performance Index). Model performance is evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R2) score on the testing set.

D. Interpretation of Results

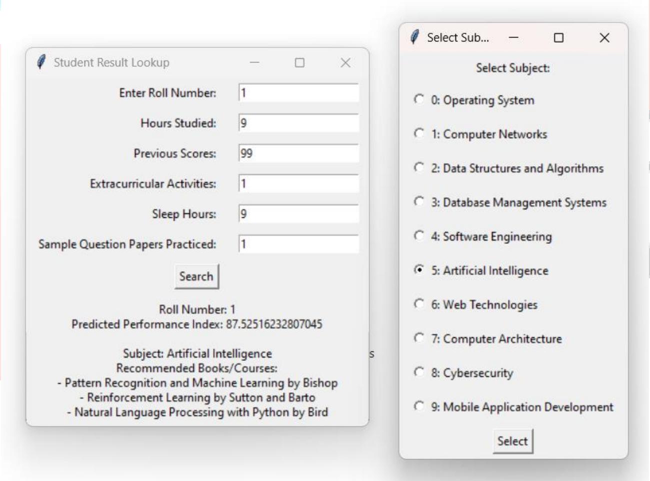
Coefficients provided by the Linear Regression model are analyzed to understand the relative importance of each feature in predicting student performance. Insights are drawn from the analysis, highlighting the key factors contributing to student performance and potential areas for improvement.

VI. Result Analysis

Critical insights into the efficacy of machine learning algorithms in forecasting academic results are revealed through the analysis of student data and performance evaluation. Mean Squared Error (MSE) and R-squared Score (R2 Score), two essential measures, are used to carefully evaluate the models' performance and provide insightful recommendations for educational decision-making. Three algorithms are examined: Support Vector Machine (SVM), Random Forest, and Linear Regression. Out of these, 3 stand out as having different advantages and disadvantages when it comes to capturing the subtleties of student performance. With the lowest MSE and greatest R2 Score, Linear Regression stands out as the leader in terms of predictive accuracy. This thorough research opens the door for more sophisticated educational interventions and improved student outcomes by supporting Linear Regression as the method of .choice for analyzing student data.

ML MODEL	MSE	R2 Score
Linear Regression	4.082	0.989
Random Forest	5.142	0.986
Support Vector Machine	5.386	0.985

Table.1. Comparion table of ML models



Enter Roll Number: 1
Hours Studied: 9
Previous Scores: 99
Extracurricular Activities: 1
Sleep Hours: 9
Sample Question Papers Practiced: 1

Roll Number: 1
Predicted Performance Index: 87.52516232807045

Subject: Artificial Intelligence
Recommended Books/Courses:
- Pattern Recognition and Machine Learning by Bishop
- Reinforcement Learning by Sutton and Barto
- Natural Language Processing with Python by Bird

Select Subject:
○ 0: Operating System
○ 1: Computer Networks
○ 2: Data Structures and Algorithms
○ 3: Database Management Systems
○ 4: Software Engineering
● 5: Artificial Intelligence
○ 6: Web Technologies
○ 7: Computer Architecture
○ 8: Cybersecurity
○ 9: Mobile Application Development

Fig.4. Output

A. Performance Evaluation

The models' performance is evaluated using two metrics: Mean Squared Error (MSE) and R-squared Score (R2 Score). These metrics give information about the models' performance in terms of goodness of fit and prediction accuracy.

B. Comparison of Algorithms:

1. Out of the three algorithms, Linear Regression had the lowest mean square error (MSSE) of 4.0826, meaning it was the most effective in reducing prediction mistakes.
2. The MSE values of Random Forest and Support Vector Machine (SVM) were marginally higher than those of Linear Regression, suggesting a somewhat inferior prediction accuracy.
3. Nonetheless, the R2 Scores of all three algorithms were high; Linear Regression had the highest score of 0.989, closely followed by Random Forest (0.986) and SVM (0.985). These high R2 Scores imply that a significant portion of the variance in the student results can be explained by the models.

C. Recommendation for Model Selection:

Linear Regression seems to be the best method for analyzing student data, according to the evaluation results because of its higher R2 Score and lower MSE when compared to Random Forest and SVM.

VII. Conclusion

It is imperative to recognize the constraints of the research, including the particular dataset employed, the feature selection procedure, and possible biases. More features, different techniques, or larger datasets may be investigated in future studies to enhance the performance and generalizability of the model. Furthermore, examining the models' interpretability and transparency may provide light on the variables affecting students' performance and enable educators to make well-informed decisions.

Acknowledgment

We would like to express our sincere gratitude to Dr. Reshma Sonar, our guidance teacher, for their invaluable guidance and support throughout the preparation of this paper. Their expertise, encouragement, and constructive feedback have been instrumental in shaping the content and structure of this document.

We would also like to acknowledge the support provided by MIT World Peace University and express our thanks to all those who have contributed to ourlearning experience.

References

[1] I. Singh, A. S. Sabitha and A. Bansal, "Student performance analysis using clustering algorithm," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, India, 2016, pp. 294-299, doi: 10.1109/CONFLUENCE.2016.7508131

[2] S. Rajput and S. Ramesh, "Student Performance Analysis based on Machine Learning Algorithms," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-4, doi: 10.1109/CONIT59222.2023.10205602

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

[4] M. S. Ram, V. Srija, V. Bhargav, A. Madhavi and G. S. Kumar, "Machine Learning Based Student Academic Performance Prediction," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 683-688, doi: 10.1109/ICIRCA51532.2021.9544538

[5] J. Xu, K. H. Moon and M. van der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 5, pp. 742-753, Aug. 2017, doi: 10.1109/JSTSP.2017.2692560

[6] S. Kour, R. Kumar and M. Gupta, "Analysis of student performance using Machine learning Algorithms," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 1395-1403, doi: 10.1109/ICIRCA51532.2021.954493

[7] Shen, Xiaoying, and Chao Yuan. "A college student behavior analysis and management method based on machine learning technology." Wireless Communications and Mobile Computing 2021 (2021): 1-10.

[8] Acharya, Anal, and Devadatta Sinha. "Early prediction of students performance using machine learning techniques." International Journal of Computer Applications 107.1 (2014): 37-43.

[9] Thangavel, S. K., Bkaratki, P. D., & Sankar, A. (2017, January). “Student placement analyzer: A recommendation system using machine learning”. In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1-5). IEEE.

[10] Yadav, N. R., & Deshmukh, S. S. (2023, May). “Prediction of Student Performance Using Machine Learning Techniques”: A Review. In International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022) (pp. 735-741). Atlantis Press.

[11] Agrawal, H., & Mavani, H. (2015). “Student performance prediction using machine learning”. International Journal of Engineering Research and Technology, 4(03), 111-113.

[12] Priya, S., Ankit, T., & Divyansh, D. (2021). “Student performance prediction using machine learning”. In Advances in parallel computing technologies and applications (pp. 167-174). IOS Press.