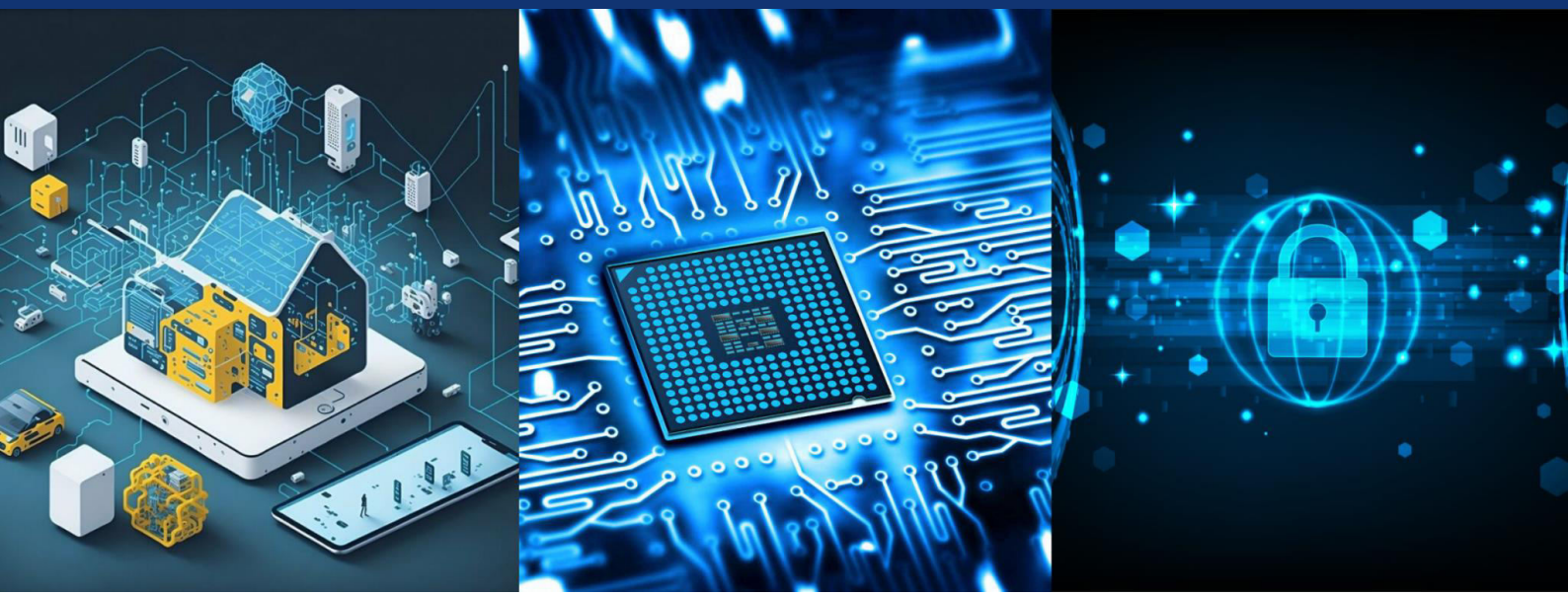
IJIRCCE

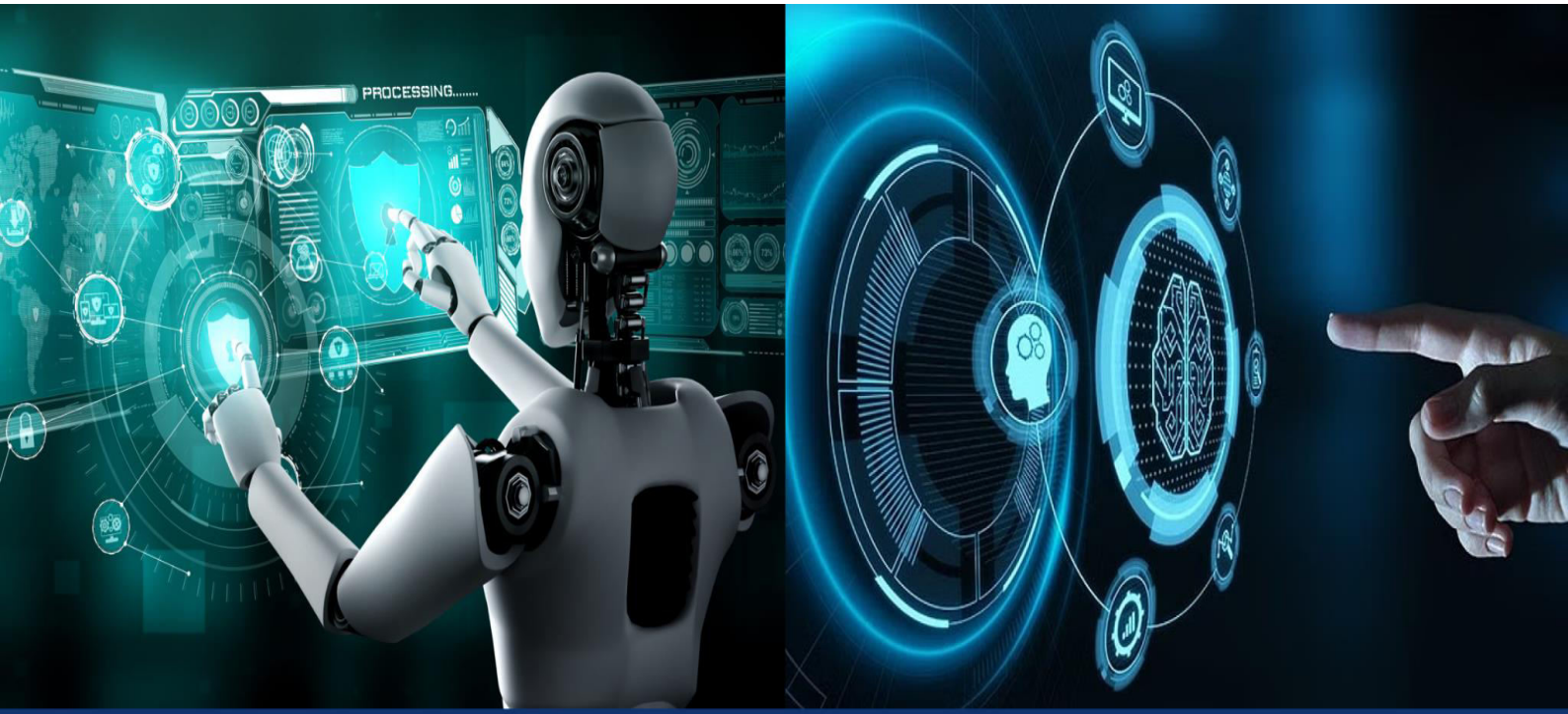
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025 |DOI:10.15680/IJIRCCE.2025.1304267

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Student Performance Prediction using Random Forest Machine Learning Algorithm

Durgeshwari Patil[1], Devendrasing Pawar [2], Prasanna Nile[3], Tejas Borase[4], Ashish T. Bhole[5]

UG Engineering Students, Department of Computer Engineering, SSBT's College of Engineering, KBC North Maharashtra University, Jalgaon, Maharashtra, India[1-4]

Associate Professor, Department of Computer Engineering, SSBT's College of Engineering, KBC North Maharashtra University, Jalgaon, Maharashtra, India[5]

ABSTRACT: Student performance prediction is a rapidly evolving focus within educational data mining and learning analytics, driven by the growing need for data-driven strategies in academic environments. The integration of Machine Learning (ML) techniques into education enables the exploration of complex relationships between diverse student-related factors such as academic history, attendance patterns, socio-demographic data, and level of education. The research emphasizes the development of predictive models capable of analyzing large-scale educational datasets to identify trends and forecast academic outcomes. By employing supervised learning algorithms and effective feature engineering, the study establishes a robust framework for understanding student trajectories. The approach not only facilitates proactive identification of learning challenges but also lays the groundwork for designing personalized educational pathways. Through the use of ML, educators and institutions gain actionable insights which enhance strategic planning, support mechanisms, and overall academic effectiveness. The findings suggest predictive modeling has the potential to transform traditional educational practices by fostering a more adaptive, equitable, and efficient learning environment. Moreover, research contributes to the foundation for future innovations in educational technology, intelligent systems, and learning support tools aimed at maximizing student potential. Approach opens up research opportunities and business applications in the field of learning analytics, offering data-driven solutions for enhancing student success.

KEYWORDS: Machine Learning, Future Academic Progress, Prediction Model System, Student Performance Prediction, Random Forest Regression.

I. INTRODUCTION

Student performance prediction has emerged as a pivotal area in educational research, offering a data-driven approach to enhancing academic success and institutional efficiency. Accurately forecasting student outcomes enables early identification of individuals at risk of underperforming, allowing educators and administrators to implement timely interventions which can significantly improve academic achievement.

With the increasing availability of educational data, machine learning (ML) techniques have gained prominence due to the ability to model complex, non-linear relationships and handle large, multidimensional datasets. The models are trained using various student-related features, such as attendance records, historical grades, and study behaviors. By analysing such data, ML algorithms can detect hidden patterns which may not be evident through traditional analytical methods.

The core objective of the study is to design and evaluate a machine learning-based model capable of predicting student academic performance with high accuracy. The predictive framework functions as an early warning system, identifying students who may require academic support before the performance declines significantly. The proactive strategy facilitates more effective allocation of educational resources and contributes to the development of personalized learning pathways.

The success of a student performance prediction model largely depends on the quality and relevance of the input data. Educational datasets typically include diverse attributes such as demographic details, previous academic performance,

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

learning habits, and course completion. Integrating the multifaceted data points enables machine learning algorithms to capture nuanced relationships which influence the academic outcomes. A careful preprocessing such as handling missing values, encoding categorical data, and normalizing features is essential to ensure data consistency and enhance the model's predictive accuracy.

Model selection is another crucial component of the study. Various machine learning algorithms, including Decision Trees, Random Forests, Support Vector Machines (SVM), Linear Regression, and Neural Networks, are evaluated to determine the most effective approach for student performance prediction. Each model is assessed based on performance metrics such as accuracy, precision, recall, F1-score, R^2 and Area Under the Curve (AUC). Through comparative analysis, the study identifies the most suitable model for early detection of at-risk students, balancing both predictive performance and computational efficiency.

In addition to prediction, the interpretability of model results plays a significant role in educational settings. It is vital educators and administrators not only receive accurate predictions but also understand the underlying factors contributing to those outcomes. Technique such as feature importance scores is employed to provide insights into the key determinants of student success or failure. These insights not only support informed decision-making but also help tailor targeted interventions which address individual learning needs, thereby fostering a more supportive and responsive educational environment.

II. LITERATURE SURVEY

Covering research conducted between 2009 and 2021, the paper [1] primarily investigates two key areas: the prediction of students at risk of academic failure and the prediction of student dropouts. The authors highlight how machine learning models have been applied to identify students who may underperform or leave the studies early, enabling early interventions to support them more effectively.

The review discusses a variety of data sources used in these studies, including university records and e-learning platforms, and outlines the most commonly applied machine learning algorithms, such as Decision Trees, Logistic Regression, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines. It also emphasizes the importance of various predictive features, such as demographic details, academic history, and digital learning behaviors. A key takeaway from the paper is the identification of existing gaps in the literature, including the limited availability of large and diverse datasets and a lack of dynamic modeling approaches which reflect changes in student behavior over time. The authors suggest the future work should explore more advanced techniques, such as ensemble methods and time-aware models, to improve prediction accuracy and responsiveness in educational settings. [2] present a comprehensive review of machine learning and data mining techniques used for Student Performance Prediction (SPP). The study categorizes the SPP process into five key stages: data collection, problem formalization, modeling, prediction, and application. Experiments were conducted using datasets comprising 1,325 students and 832 courses, sourced from both institutional records and a public dataset. The results underscore existing challenges in model generalization, feature selection, and real-world deployment. The study emphasizes the importance of integrating contextual and behavioral features to enhance prediction accuracy and support adaptive learning systems. It also recommends exploring hybrid and ensemble models to overcome limitations of individual techniques and improve interpretability in educational settings.

The research paper [3] provides a detailed comparison of machine learning techniques used to predict student academic outcomes. The study is designed around three distinct and task-oriented datasets, which vary in terms of educational context and data attributes, to thoroughly evaluate the performance and generalizability of seven machine learning algorithms. The algorithms include Random Forest, Decision Tree, Support Vector Machine, Artificial Neural Networks, and others, all of which were parameter-optimized to ensure fair comparison.

The researchers conducted experiments involving both binary and multi-class classification tasks to simulate different educational prediction scenarios such as pass/fail outcomes or performance tier classifications. Among the algorithms tested, Random Forest emerged as the most reliable and consistently accurate method across all datasets, demonstrating strong predictive power and robustness. Decision Trees and Artificial Neural Networks also performed well, making them suitable candidates for real-world educational prediction systems.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025 |DOI:10.15680/IJIRCCE.2025.1304267

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The study not only highlights the strengths and limitations of each algorithm but also emphasizes the importance of choosing the right model depending on the nature of the dataset and prediction goals. The authors suggest which future research should explore hybrid and ensemble models, as well as incorporate additional contextual data to further improve prediction accuracy and practical applicability in educational settings.

III. PROBLEM STATEMENT

In the evolving landscape of education, predicting student performance has become essential for timely interventions and personalized learning. However, existing methods often struggle to address the complexity of factors influencing academic success, such as student behavior, attendance, and historical performance. As educational institutions increasingly rely on data-driven solutions, the challenge remains to develop accurate machine learning models which can handle diverse, high-dimensional datasets and provide reliable predictions. The inability to predict student outcomes effectively leads to missed opportunities for early support, affecting both student success and resource allocation. Thus, a robust and adaptive machine learning framework is required to accurately forecast student performance and enable proactive intervention strategies.

IV.SYSTEM ARCHITECTURE

System architecture is the blueprint of a system outlines its components, their interactions, and the overall structure. It's a strategic plan which design and development of complex systems, ensuring they function efficiently, reliably, and securely. It encompasses both hardware and software elements, defining how they interact to achieve specific objectives. System architecture focuses on the high-level design and structure of a system, rather than the specific details of its implementation. The System Architecture is mentioned as follow in the Figure 1.

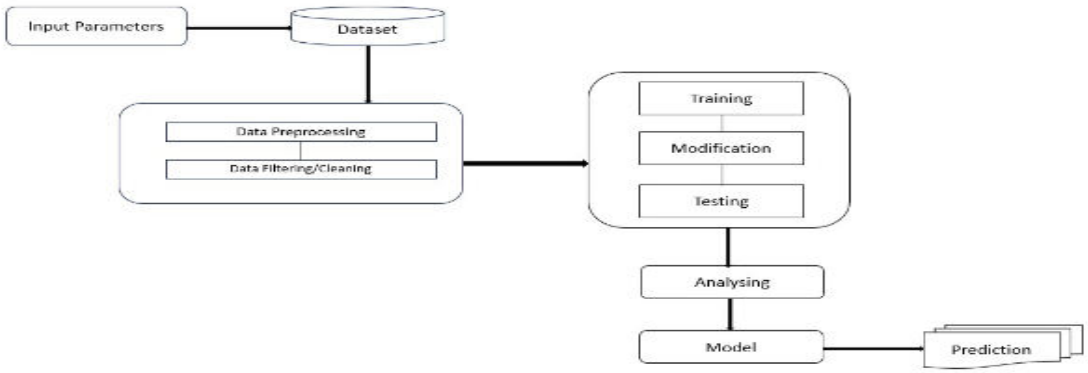


Figure 1- System Architecture

A. Input Parameters:
The block represents the initial set of variables which influence student performance.

B. Dataset Formation:
The input parameters are used to form a structured dataset. The dataset acts as the foundational input to the machine learning pipeline.

C. Data Preprocessing and Cleaning:
The dataset undergoes preprocessing which includes handling missing values, normalization, encoding of categorical variables, and transformation of features. Additionally, data filtering and cleaning processes are applied to remove noise, outliers, and inconsistencies to enhance data quality and reliability [2].

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025 |DOI:10.15680/IJIRCCE.2025.1304267

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Model Development Process:
The stage involves three key operations:
Training: The cleaned dataset is used to train the machine learning model using algorithm such as the Random Forest.
Modification: Model parameters and features are tuned to optimize performance, often through hyperparameter tuning or feature selection techniques.
Testing: The trained model is evaluated using unseen test data to measure its accuracy, precision, recall, and other performance metrics.

E. Model Analysis:
The performance of the trained model is analyzed to assess its predictive capability. Evaluation metrics help in determining the effectiveness of the model and guide further refinement.

F. Final Model and Prediction:
After analysis and validation, the best-performing model is selected. The model is then used to predict student performance outcomes based on input data.

V.METHODOLOGY

The modules in project are describe as follows:

A. Machine Learning Model:
The core of the student performance prediction system is the Machine Learning Model module, which utilizes the Random Forest Regression algorithm to predict academic outcomes [3]. Random Forest, an ensemble learning method, is chosen for its robustness, accuracy, and ability to handle high-dimensional data without overfitting. The model is trained on a dataset comprising key academic features such as attendance, internal assessment scores, previous grades, and other behavioral indicators. Data preprocessing steps include handling missing values, feature scaling, and encoding categorical variables to ensure the dataset is clean and consistent. The Random Forest Regressor is then trained using the processed data, and its performance is evaluated using regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). The trained model is capable of providing continuous performance scores, which can be used to assess the risk level of each student.

B. User Interface of the Application:
The User Interface (UI) is designed to offer an interactive and user-friendly experience for faculty and administrators. It allows users to upload student data, view predictions, and access performance analysis. The interface includes forms for data entry and a marks evaluation page. Which enables non-technical users to interpret results easily and make informed decisions regarding academic interventions. The frontend is developed using standard web technologies and is designed for responsiveness and accessibility. Visual elements such as graphs, charts, and color-coded indicators enhance the interpretability of model outputs and allow educators to identify at-risk students effectively.

C. Backend and Integration Services:
The backend of the application is developed using Flask, a lightweight Python-based web framework. Flask facilitates the integration between the frontend interface and the machine learning model. It acts as an intermediary, receiving input data from the UI, passing it through the trained Random Forest Regressor, and returning prediction results to be displayed on the interface. Additionally, Flask handles HTTP requests, model loading, and real-time response generation, ensuring low-latency interactions and a seamless user experience. The backend also manages data storage and retrieval operations, enabling the application to maintain historical records for continuous improvement and future analysis.

VI.IMPLEMENTATION

The implementation of the student performance prediction system involves using machine learning algorithms to forecast academic results based on features such as attendance, test scores, course completion and level of education. The data undergoes preprocessing, including cleaning and normalization, to ensure accuracy and efficiency. The model Random Forest is trained and tested. The Random Forest offering higher accuracy and also avoid the problem of

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025 |DOI:10.15680/IJIRCCE.2025.1304267
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

overfitting [3]. The system is developed using Python and libraries like Scikit-learn, highlighting the effectiveness of Machine Learning in analyzing and predicting student performance.
The Algorithm for Student Performance Prediction is as follow:

Input:
Dataset D containing student records with features such as course completion, level of education, previous scores, etc.
Output:
Predicted student performance and model evaluation metrics (e.g., MAE, MSE, R^2 score).

Step 1: Input the Dataset
1.1 Load dataset D into memory using pandas.read_csv().

Step 2: Import Required Libraries
2.1 Import following Python libraries to perform various operations on the dataset such as:
-pandas, numpy which are the mainly used for data manipulation and to perform various mathematical operations.
-matplotlib.pyplot, seaborn for data visualization.
-sklearn modules for machine learning.

Step 3: Visualize and Analyze the Data
3.1 Use python libraries seaborn and matplotlib to generate plots (e.g., heatmaps, histograms, scatter plots).
3.2 Identify data patterns, correlations, and detect missing or outlier values.

Step 4: Preprocess the Data
4.1 [2] Handle missing values within the given dataset using various techniques such as mean/mode.
4.2 Encode categorical features within the given dataset using label encoding or one-hot encoding.
4.3 Normalize or scale numerical features using StandardScaler.

Step 5: Split the Dataset
5.1 Use train_test_split from sklearn.model_selection:
-Training
-Testing

Step 6: Apply Random Forest Regression
6.1 Import RandomForestRegressor.
6.2 Instantiate the model with suitable parameters.

Step 7: Train the Model
7.1 Fit the Random Forest model on D_train using .fit() method.

Step 8: Make Predictions
8.1 Use .predict() on D_test features to generate predictions.

Step 9: Evaluate the Model
9.1 Calculate performance metrics such as Mean Square Error (MSE), Mean Absolute Error (MAE) and R^2 score.
9.2 Output the evaluation results.

VII. RESULTS

The Student Performance Prediction system takes input data related to student features, such as previous grades, attendance, study habits, and socio-economic factors, and predicts their future performance using machine learning models. Instead of traditional Linear Regression, Random Forest Regression is employed to capture complex, non-linear relationships within the data, resulting in improved prediction accuracy and robustness. The main output of the system includes:

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025 |DOI:10.15680/IJIRCCE.2025.1304267

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. Predicted Performance

The system predicts the performance of students, providing estimated grades or scores based on historical data and the selected features. Predictions can be shown for individual student.

B. Data Visualizations

Interactive graphs and charts are presented to show predicted vs. actual results. Users can analyze patterns such as the impact of study habits or attendance on performance.

Table 1 illustrates the prediction results generated by the proposed machine learning model for student performance evaluation using the Random Forest Regression algorithm. The model accepts three academic input parameters: Math Score, Reading Score, and Writing Score. Based on these input features, the model predicts the average score for each student, which is used to classify the performance into qualitative categories. The performance categories are defined as Needs Improvement, Good, Very Good, and Excellent, based on the predicted average score.

Table 1- Prediction Table

Input			Output	Prediction
Math Score	**Reading Score**	**Writing Score**	**Average Score**	**Marks Evaluation**
72	72	74	72.66	Good
69	90	88	82.33	Very Good
90	95	93	92.66	Excellent
47	57	44	49.33	Needs Improvement
76	78	75	76.33	Good

Table 2 provides a comparative analysis of Random Forest and Linear Regression models used for predicting student aca demic performance. The ”Input” column denotes the average predicted scores, while ”Execution Time” reflects the time taken by each model to produce results. The Random Forest model consistently achieves high accuracy, whereas Linear Regression shows moderate accuracy. It indicates the Random Forest is more suitable for handling complex educational datasets. Linear Regression, being a simpler model, offers faster computation but lower accuracy. The results demon strate, the choice of model impacts both prediction accuracy and processing efficiency. Random Forest is ideal for scenarios where accuracy is a priority. These findings highlight the importance of selecting appropriate machine learning algo rithms for educational data analysis. Overall, machine learning shows strong potential for improving academic performance prediction.

Table 2- Results

		Prediction Score Accuracy	
Input (Average Score)	**Execution Time**	**Random Forest**	**Linear Regression**
72.66	42.75	98.42	87.18
82.33	72.05	99.71	87.43
92.66	80.30	99.01	88.01
49.33	86.12	99.31	88.34
76.33	93.42	99.62	88.57

VIII. CONCLUSION AND FUTURE WORK

The project successfully demonstrates the practical application of machine learning techniques in the educational domain. By processing and analyzing various student-related datasets, the model able to make meaningful predictions regarding academic performance with a high degree of accuracy. It not only validates the effectiveness of ML algorithms in handling educational data but also highlights the potential for such systems to be integrated into real-

world academic settings. The insights derived from the model can assist educators and administrators in developing targeted interventions, enhancing student engagement, and fostering better academic outcomes.

The model can be enhanced by incorporating real-time data and expanding feature sets to improve prediction accuracy and support dynamic student interventions.

REFERENCES

1. Albreiki, Balqis, Nazar Zaki, and Hany Alashwal. "A systematic literature review of student'performance prediction using machine learning techniques." *Education Sciences* 11, no. 9 (2021): 552.
2. Zhang, Yupei, Yue Yun, Rui An, Jiaqi Cui, Huan Dai, and Xuequn Shang. "Educational data mining techniques for student performance prediction: method review and comparison analysis." *Frontiers in psychology* 12 (2021): 698490.
3. Chen, Yawen, and Linbo Zhai. "A comparative study on student performance prediction using machine learning." *Education and Information Technologies* 28, no. 9 (2023): 12039-12057.
4. Yadav, Nitin Ramrao, and Sonal Sachin Deshmukh. "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, 2023.
5. Ahmed, Esmael. "Student performance prediction using machine learning algorithms." *Applied Computational Intelligence and Soft Computing* 2024, no. 1 (2024): 4067721.
6. Hashim, Ali Salah, Wid Akeel Awadh, and Alaa Khalaf Hamoud. "Student performance prediction model based on supervised machine learning algorithms." In *IOP conference series: materials science and engineering*, vol. 928, no. 3, p. 032019. IOP Publishing, 2020.
7. Dhilipan, J., N. Vijayalakshmi, S. Suriya, and Arockiya Christopher. "Prediction of students performance using machine learning." In *IOP conference series: Materials science and engineering*, vol. 1055, no. 1, p. 012122. IOP Publishing, 2021.
8. Alsalem, Gheed M., Noor Sarhan, Mustafa Hammad, and Bayan Zawaideh. "Predicting Students' Performance using Machine Learning Classifiers." In *2024 25th International Arab Conference on Information Technology (ACIT)*, pp. 1-5. IEEE, 2024.
9. Gupta, Shelly, and Jyoti Agarwal. "Machine Learning Approaches for Student Performance Prediction." In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pp. 1-6. IEEE, 2022.
10. Alhazmi, Essa, and Abdullah Sheneamer. "Early predicting of students performance in higher education." *Ieee Access* 11 (2023): 27579-27589.
11. Pallathadka, Harikumar, Alex Wenda, Edwin Ramirez-Asís, Maximiliano Asís-López, Judith Flores-Albornoz, and Khongdet Phasinam. "Classification and prediction of student performance data using various machine learning algorithms." *Materials today: proceedings* 80 (2023): 3782-3785.
12. Agrawal, Havan, and Harshil Mavani. "Student performance prediction using machine learning." *International Journal of Engineering Research and Technology* 4, no. 03 (2015): 111-113.
13. Sekeroglu, Boran, Kamil Dimililer, and Kubra Tuncal. "Student performance prediction and classification using machine learning algorithms." In *Proceedings of the 2019 8th international conference on educational and information technology*, pp. 7-11. 2019.
14. Priya, S., T. Ankit, and D. Divyansh. "Student performance prediction using machine learning." In *Advances in parallel computing technologies and applications*, pp. 167-174. IOS Press, 2021.
15. Ha, Dinh Thi, Pham Thi To Loan, Cu Nguyen Giap, and Nguyen Thi Lien Huong. "An empirical study for student academic performance prediction using machine learning techniques." *International Journal of Computer Science and Information Security (IJCSIS)* 18, no. 3 (2020): 75-82.
16. Alsariera, Yazan A., Yahia Baashar, Gamal Alkawsi, Abdulsalam Mustafa, Ammar Ahmed Alkahtani, and Nor'ashikin Ali. "Assessment and evaluation of different machine learning algorithms for predicting student performance." *Computational intelligence and neuroscience* 2022, no. 1 (2022): 4151487.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details