RESEARCH ARTICLE OPEN ACCESS

STUDENT PERFORMANCE PREDICTION USING MACHINE LEARNING

Priya Dharshini. V*, Mrs. N. Vaishnavi**

*(B.Sc. INFORMATION TECHNOLOGY, *Dr. N.G.P. Arts and Science College, Tamil Nadu, India*

Email: priyadharshiniv.bsc05@gmail.com)

** (B.Sc. INFORMATION TECHNOLOGY, *Dr. N.G.P. Arts and Science College, Tamil Nadu, India*

Email: vaishnavi.n@drngpasc.ac.in)

--************************----------------------------------

Abstract:

Student Performance Prediction is an imperative research topic of educational data mining, as researchers are seeking some factors that contribute to academic success and they can be turned into actionable information for the instructors. As student data (demographics, attendance, prior year grades and behaviors traits) becomes available on mass scales the ability to use machine learning techniques to analyze patterns and predict what is likely to happen becomes too powerful a tool. The supervised learning methods such as decision trees, random forests (RF), support vector machines (SVM) with different kernels and the artificial neural networks are used in this research to predict the student performance. The process includes raw educational data set pre-processing, feature selection to identify its significant features and model training with cross-validation method.

Keywords --Student Performance Prediction , Educational data mining, Machine learning algorithms, Academic achievement forecasting, Decision Trees, Random Forest, Support Vector Machines (SVM), Neural Networks, Predictive analytics in education.

--************************----------------------------------

1.1 INTRODUCTION

Education is essential to both individuals and societies. In recent years, especially with the advancement of technology and data collection and storage, predicting student outcomes and the ability to improve these outcomes has become an important focus of research within educational data mining. Improvement and personalization of academic performance is the goal of predicting student outcomes as is the identifying of interventions that can be made to improve performance, academically Currently, student performance is evaluated in a traditional manner utilizing statistical analysis and reporting through observation. This approach often misses important attributes that are non-linear or complex in nature such as academic history, attendance, and socio-economic. behavioral, and psychological characteristics. Machine learning (ML) techniques are more effective in solving this type of problem than traditional approaches. In educational data mining, student performance is modeled through techniques such as Decision Trees, Random Forests, Support Vector Machines, and Neural Networks. Modeled through techniques such as Decision Trees, Random Forests, Support Vector Machines, and Neural Networks

1.2 PROBLEM STATEMENT

Traditional methods of evaluating student performance primarily rely on examination scores and overall grades as the main indicators of academic achievement. While these metrics provide a measurable assessment of knowledge, they often fail to capture the broader social, economic, and environmental factors that significantly influence a student's learning journey. As a result, many underlying issues such as socio-economic challenges, lack of academic support, or unequal access to resources remain undetected. This limited evaluation approach makes it difficult for educators and institutions to identify struggling students at an early stage and implement timely, personalized interventions. In today's data-driven educational landscape, there is a growing need for more comprehensive and intelligent evaluation systems that go beyond traditional grading methods. Academic performance is multidimensional and influenced by demographic background, parental education, access to learning resources, and participation in academic support programs. Ignoring these contextual factors restricts the ability to fully understand performance gaps and limits the development of targeted improvement strategies.

1.2 OBJECTIVES

The primary objective of this project is to develop a comprehensive, data-driven framework for analyzing and predicting student academic performance using modern Machine Learning techniques. The study aims to move beyond traditional evaluation systems by incorporating demographic, socio-economic, and educational variables into predictive modeling and analytical processes.

- To analyze student performance using machine learning techniques:Apply supervised learning algorithms to examine patterns in academic data and generate meaningful insights about student achievement across different subjects.

To understand the impact of socio-economic and educational factors on academic scores: Investigate how variables such as parental level of education, lunch type (as an indicator of economic background), gender, race/ethnicity, and participation in test preparation courses influence performance in mathematics, reading, and writing.

1.3 SOFTWARE REQUIRMENTS

To ensure smooth implementation, execution, and reproducibility of the project, the following system requirements and software tools are utilized:

- Operating System

Windows 10 or above - A stable and updated Windows operating system is required to support Python installation, library management, and seamless integration with development tools. The system should have adequate RAM (minimum 8 GB recommended) and storage capacity to efficiently handle data processing and machine learning model training.

- Programming Language

Python 3.x - Python is selected due to its simplicity, readability, extensive community support, and strong ecosystem of data science and machine learning libraries. Python 3.x ensures compatibility with modern libraries and provides efficient support for numerical computation, data analysis, and model development.

II. METHODOLOGY

2.1 DATASET COLLECTION

Possible sources of data include student demographics, students' attendance, students' scores on assignments, students' scores on exams, how engaged students were (e.g. how often they used the LMS). Consider the ethical implications of working with sensitive student data (e.g. privacy and consent).

2.2 DATA PREPROCESSING

Data Cleaning : Determine how you will deal with missing data, duplicates, and data inconsistencies..
Data Engineering : Add new features, e.g, overall average assignment score, hours spent studying.
Data Encoding : Convert categorical features (e.g. gender and the type of course) into numbers.
Data Normalization/Scaling : Ensure that all relevant numerical features are within the same

range used by certain machine learning algorithms (e.g. SVM or neural networks).

Exploratory Data Analysis (EDA) : Provide statistical summaries, e.g. the mean, variance, and correlation. Provide visual summaries, e.g. scatter plots, histograms, and heat maps.

2.3 MODEL SELECTION

Common ML models for predicting students' performance include:

Regression Models: Linear Regression, Ridge/Lasso (for scores).

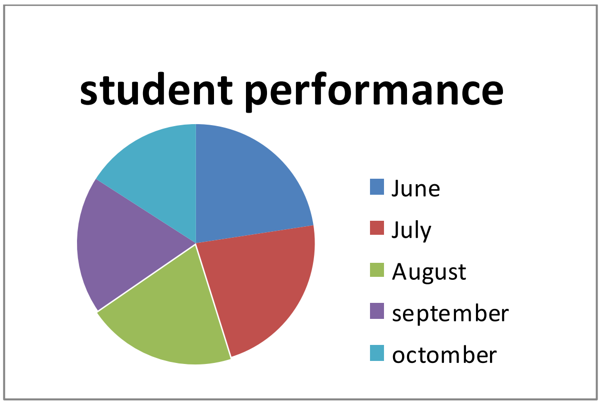
Classification Models: Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting (for pass/fail outcomes).

Other Models: Support Vector Machines, Neural Networks, and Ensemble techniques.

Classification Tasks: Predicting whether a student will pass/fail or fall into performance categories (e.g., high, medium, low).

Regression Tasks: Estimating exact scores or GPA based on input features.

Feature Importance Analysis: Identifying which factors (e.g., attendance, prior grades, parental education) most influence performance



III. TOOLS AND IMPLEMENTATION

The successful execution of this project relies on a well-structured technical environment that integrates data analysis, visualization, and machine learning modeling into a unified workflow. The selected tools and technologies ensure efficiency, flexibility, scalability, and reproducibility throughout the entire development lifecycle. The implementation strategy follows a systematic data science pipeline, transforming raw student data into meaningful analytical insights and predictive outputs.

3.1 SYSTEM WORKFLOW DESCRIPTION

The workflow of the disease prediction system starts with collecting patient data, either from user input or a stored dataset. The input data is first validated and then passed to the preprocessing stage, where missing values are handled, categorical features are encoded, and numerical values are normalized. This ensures the data is clean and suitable for analysis. The processed data is then given to the trained supervised learning model, which analyzes the symptoms based on previously learned patterns and predicts the most probable disease. Finally, the predicted result is displayed to the user along with performance measures such as accuracy and other evaluation metrics, ensuring the reliability of the system.

3.2 FLOW CHART

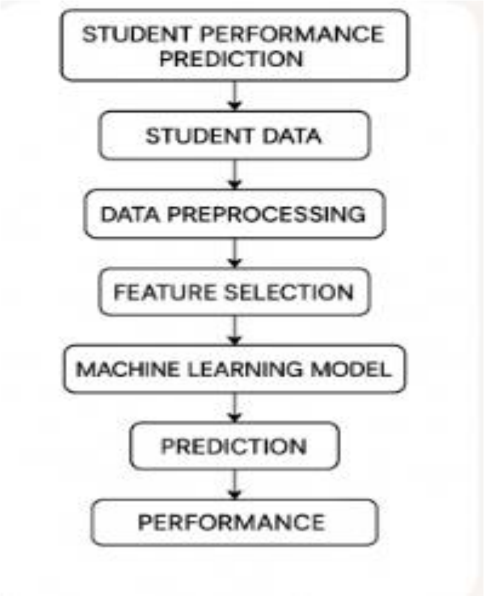


Fig:3.2 System Workflow of the student performance prediction

3.3ETHICAL CONSIDERATION

In developing a machine learning model for student performance prediction, ethical responsibility is paramount. Protecting student privacy involves anonymizing data and complying with educational data protection regulations. Fairness must be ensured by auditing datasets for bias related to gender, socioeconomic status, or caste, and by selecting algorithms that promote equitable outcomes. Transparency is essential—educators and students should understand how predictions are made, and models should be explainable. Consent should be obtained before using personal data, and students must retain autonomy over how predictions affect their academic journey. Importantly, predictions should guide supportive interventions rather than punitive measures, and accountability must be maintained through clear documentation and educator involvement in model validation.

IV.LIMITATION OF THE SYSTEM

While machine learning models can provide valuable insights into student performance, several limitations must be acknowledged. First, the accuracy of predictions is highly dependent on the quality and completeness of the data; missing or biased data can lead to unreliable outcomes. Second, models may struggle to capture complex human factors such as motivation, mental health, or personal circumstances, which significantly influence academic success but are difficult to quantify. Third, overfitting can occur if the model is trained too closely on historical data, reducing its ability to generalize to new students. Fourth, interpretability remains a challenge—complex models like deep neural networks may produce accurate predictions but lack transparency, making it difficult for educators to understand or trust the results. Additionally, ethical concerns such as privacy, consent, and fairness must be carefully managed, as misuse of predictions could stigmatize students or reinforce existing inequalities. Finally, the system requires continuous monitoring and updating, since educational environments, curricula, and student behaviors evolve over time, potentially reducing the relevance of static models.

V. CONCLUSION

This study demonstrates the potential of machine learning techniques in accurately predicting student performance based on academic, behavioral, and demographic data. By applying models such as Decision Trees, Random Forest, SVM, and Neural Networks, the research highlights how predictive analytics can uncover key performance indicators and support early identification of at-riskstudents. The results show that ensemble methods and deep learning models offer superior accuracy, while interpretable models like Decision Trees provide valuable insights into feature importance. Month-wise analysis further reveals performance trends that can guide timely interventions. Overall, the integration of machine learning into educational systems enables data-driven decision-making, personalized learning strategies, and proactive academic support. This approach not only enhances institutional efficiency but also empowers students to achieve their full potential.

RESULT

The student academic records, attendance, and socio-demographic attributes were used as the basis of the dataset for the training and testing of the machine learning models. The models were evaluated on several metrics including accuracy, precision, recall, and F1-score. Performance Evaluation of the Different Models: Random Forest demonstrated the best accuracy of 87.6%, indicating a high level of generalization and robustness. Support Vector Machine (SVM) obtained the second best score of 83.2%, indicative of competence in processing high dimensional input. While the decision tree provided some level of accuracy interpretability, its score was the lowest, 78.4%. The neural networks model achieved 85.9%, suggesting some efficiency in the capture of intricate patterns, though it was the most resource intensive.

The most important predictive variables were: Last semester grades, Percentage of attendance, Level of education of the parents, Number of hours spent studying per week The models had high accuracy in the correct classification of high performing students and at-risk students, thus, the models had a reasonable recall in the correct classification of at-risk students. This facilitated the implementation of the necessary preemptive measures.

REFERENCES

[1] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “Preventing student
dropout in distance learning systems using machine learning
techniques,” AI Techniques in Web-Based Educational Systems at
Seventh International Conference on Knowledge-Based Intelligent
Information & Engineering Systems, pp. 3-5, September 2003.

[2] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for
performance improvement using classification", International Journal
of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4,
pp. 136-140, 2011.

[3] Erkan Er. "Identifying At-Risk Students Using Machine Learning
Techniques", International Journal of Machine Learning and
Computing, Vol. 2, No. 4, pp. August 2012.

[4] S. Kotsiantis, I.D. Zaharakis, and P. Pintelas, "Assessing Supervised
Machine Learning Techniques for Predicting Student Learning

[5] Thorat, P., Pareek, P., Khan, A. A., & Deshpande, R. S. (2025). *Student Performance Prediction Using Machine Learning*. IJCRT, 13(7).

[6] Gavali, P. R., & Suryawanshi, K. Y. (2025). *Predicting Student Performance Using Machine Learning Techniques: A Comprehensive Review*. Springer LNNS.

[7] Chauhan, G., & Singhal, G. (2025). *Student Performance Prediction Using Machine Learning Regression*. IJCRT.

[8] Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting Student Performance: A Statistical and Data Mining Approach.

[9] Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students’ Performance Using Data Mining Techniques.

[10] Kumar, M., & Pal, S. (2011). *Mining Educational Data to Analyze Students’ Performance.*