

Design and Development of a Decision Support System for Predicting Student Academic Performance

By

OJONUGWA WADA

Abstract

In recent times, ML has emerged as a significant tool in education, delivering data-driven insights into student performance and enabling proactive interventions. This study uses supervised machine learning models to investigate how behavioral habits, demographic traits, and lifestyle choices influence academic attainment. This study investigates the association between non-academic activities and test results using a dataset of 1,000 students that includes variables such as study hours, sleep duration, mental health assessment, internet quality, and others. Four regression models were trained and evaluated using RMSE and R^2 Score measures: linear regression, decision tree regression, random forest regression, and gradient boosting regression. The Linear Regression model fared better than others, The RMSE of 5.14 and R^2 Score of 0.90 indicate good predictive power. A special interactive prediction interface was created to put these findings into practice for educators, parents, and students. This work contributes to the emerging field of educational data mining by including holistic behavioral aspects into academic performance modeling, hence facilitating the creation of individualized, context-aware educational assistance systems.

Keywords: academic performance, behavioral habits, machine learning, regression models, student analytics, mental health, predictive modeling.

1 Introduction

1.1 Background of Study

Machine learning and predictive analysis have become important educational tools in recent years, providing novel methods for forecasting academic achievements, tailoring learning routes, and analyzing student behavior. With the growing complexity of educational settings and the proliferation of digital learning platforms, it is critical to investigate how students' everyday routines affect their academic achievement. To predict performance outcomes, this study applies supervised machine learning models to a dataset combining behavioral, demographic, and academic information about students.

The rationale for this study stems from the belief that student performance is influenced not only by cognitive ability but also by behavioral patterns such as study time, sleep duration, social activity, and overall health. By identifying significant behavioral indicators of academic performance, this study hopes to help educators, school counselors, and lawmakers build more effective, data-driven interventions and individualized support systems.

1.2 Research Problem

While ML models have demonstrated potential in predicting academic achievement using traditional measures, they frequently ignore non-academic behavioral aspects. Key lifestyle factors such as sleep duration, study routines, internet engagement, and mental health are still underexplored in performance prediction models. This divide hampers educational institutions' ability to provide holistic and tailored help.

This work tackles the difficulty by adding a variety of behavioral and demographic factors into supervised machine learning models, with the goal of improving prediction accuracy and providing a fuller knowledge of student performance.

1.3 Research Questions

This project aims to solve the following research questions:

- How do behavioral habits influence academic performance?
- Which behavioral features are the strongest predictors of academic success?

- Which supervised regression model best predicts academic outcomes based on behavioral and demographic data?

2 Literature Review

ML has proven itself as a powerful tool in education, revolutionizing the way learning processes are studied and developed, notably for predicting student academic achievement. As institutions strive to improve learning outcomes and execute timely interventions, data-driven techniques have become critical for identifying at-risk students and customizing educational experiences.

A growing corpus of research has examined the impact of ML models to forecast academic achievement, leveraging a variety of datasets that include demographic information, past academic records, behavioral indicators, and engagement measures. Sarker et al. (2024) stressed the importance of data mining approaches in detecting hidden patterns in student performance data, therefore establishing the groundwork for predictive modeling in educational contexts.

Recent advancements have resulted in more powerful methods, such as Random Forests, Support Vector Machines, and Neural Networks, that give more accuracy in predicting academic achievements. Ensemble approaches, in particular, have demonstrated significant potential. Tang, Li, and Zhao (2024) showed that deep ensemble learning outperforms standard models in capturing complicated, nonlinear connections in student data. Similarly, Butt et al. (2023) found that multi-model ensembles improve prediction accuracy in higher education, emphasizing the importance of feature selection and model interpretability.

Deep learning has also gained popularity, with Baniata et al. (2024) demonstrating how these structures can represent complicated connections in educational datasets, enhancing analytical capacities.

Beyond computational advances, there is a rising trend of adding behavioral and interactional data from learning management systems like Moodle and Canvas. Metrics like as login frequency, time spent on activities, and assignment submission timeframes give real-time insights on student involvement, allowing for early warning systems and proactive help.

Despite these gains, significant problems remain, notably in terms of data quality, model generalizability across institutions, and ethical issues about privacy and prejudice. Nonetheless, the research continually emphasizes machine learning's transformational potential for developing individualized, egalitarian, and effective educational methods.

Moving forward, recent research emphasizes the incorporation of lifestyle-related characteristics such as study habits, sleep patterns, digital behavior, and mental health into prediction frameworks. By including such diverse information, this study adds to a more comprehensive understanding of the elements that influence academic achievement, hence facilitating the creation of context-aware, individualized treatments. This effort, which sits at the crossroads of data science and education, is consistent with the larger drive toward explainable, ethical, and effective AI in learning contexts.

3 Methodology

This study employed data-driven approaches, such as supervised ML, to forecast student academic progress based on behavioral, demographic, and lifestyle factors. The approach is divided into five primary stages: gathering information and the preprocessing phase, exploring data analysis, feature engineering, model creation and evaluation, and predictive interface implementation.

3.1 Data Gathering and Preparation

The data for this investigation was taken from Kaggle. It has 1,000 rows and 16 columns that indicate factors such as student behavior, demographics, and academic performance. The key features include:

Variable Name	Description	Data Type
Student ID	Unique identifier for each student.	Identifier
Age	Age of the student.	Numeric
Gender	Gender of the student (Male, Female).	Categorical
Study Hours	Average number of study hours per day.	Numeric
Social Media Hours	Daily social media usage in hours.	Numeric
Netflix Hours	Daily streaming (e.g., Netflix) usage in hours.	Numeric
Part-time Job	Whether the student has a part-time job (Yes, No).	Categorical
Attendance Percentage	Class attendance rate.	Numeric
Sleep Hours	Average nightly sleep duration.	Numeric

Diet Quality	Quality of diet (Poor, Fair, Good).	Categorical
Exercise Frequency	Number of days the student exercises per week.	Numeric
Parental Education Level	Highest education level of parents (None, High School, Bachelor, Master).	Categorical
Internet Quality	Quality of home internet connection (Poor, Average, Good).	Categorical
Mental Health Rating	Self-reported mental health rating (1–10).	Numeric
Extracurricular Participation	Participation in extracurricular activities (Yes, No).	Categorical
Exam Score	Final exam score.	Numeric

Before analysis and modeling, the dataset underwent a comprehensive pretreatment procedure to verify data quality, consistency, and applicability for machine learning. Initial checks indicated that the dataset was clean and well-structured no missing values or duplicate entries were detected, considerably decreasing the requirement for major data repair. A more trustworthy and significant exploratory data analysis was made possible by this excellent foundation, enabling a detailed examination of demographic and behavioral trends.

3.2 Exploratory data analysis (EDA)

Investigating and comprehending the underlying patterns and relationships in the dataset is essential before developing predictive models. Important patterns and connections pertaining to student behavior and academic achievement were discovered during this experimental phase,

which comprised examining both numerical and category variables using a range of visualization tools. The analysis's conclusions influenced feature selection and the modeling methodology, providing a strong basis for creating precise and significant predictive models.

How Do Behavioral and Demographic Features Distribute Across Students?

The distributions of important numerical variables and categorical components are examined in this section. In addition to guiding feature selection for predictive modeling, these visualizations highlight trends in student demographics and behaviors that assist explain their impact on academic success.



Figure 1: Feature Distribution

The majority of students had moderate study and sleep habits, according to the data, and their exam results were typically in the middle to high range. Attendance was high while social media and Netflix usage was low. The distribution of genders was well equal, and A large proportion of students did not engage in social events or pursue part-time employment. Parental education levels were usually high school or bachelor's degrees, diet quality was generally fair or good, and internet quality was generally average or higher. These trends offer crucial background information for comprehending the potential effects of student conduct and personal circumstances on academic achievement.

Which Behavioral Features Are Most Strongly Correlated with Academic Performance?

Using a correlation heatmap, this part investigates the connections between test results and numerical behavioral characteristics. Finding these relationships aids in highlighting the important elements that affect academic achievement and guides the choice of predictors for machine learning models.

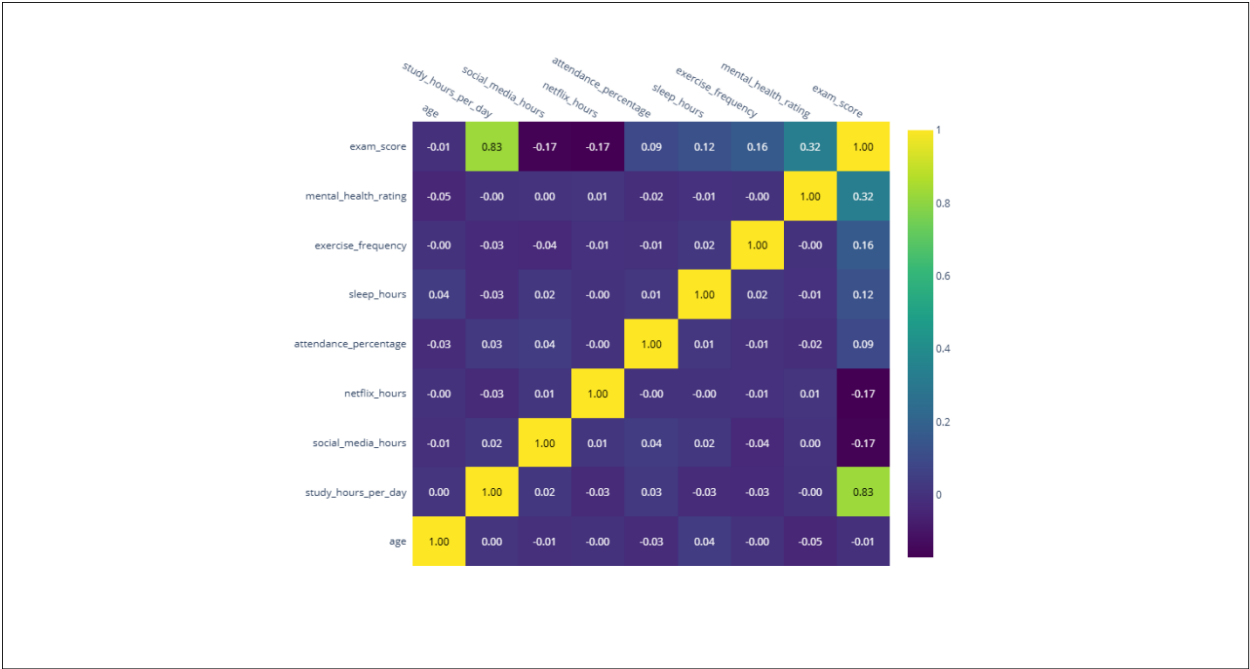


Figure 2: Correlation Heatmap

Consistent study habits have a major influence on academic achievement, as evidenced by the correlation analysis, which shows that the number of study hours per day has the largest positive link with exam results (0.83). While sleep duration and exercise frequency have lesser but still beneficial associations, mental health rating also has a somewhat positive connection (0.32). On the other hand, there are slight negative correlations (-0.17) between social media and Netflix use, indicating that more screen time may have a detrimental impact on performance. There is little association between other factors like age and attendance. Overall, the results highlight how important study practices and mental health are in determining academic success.

How Do Key Behavioral and Contextual Factors Relate to Academic Performance?

This section looks at how test grades relate to environmental elements like internet quality and mental health status as well as significant behavioral variables like study hours and sleep length.

These illustrations offer complex perspectives on the various factors influencing student performance.



Figure 3: Plots Showing Factors Influencing Exam Scores

All genders show a substantial positive correlation between study hours and exam results in the scatter plot, suggesting that students who put in more study time typically perform higher on tests. The scatter plot of sleep duration vs test results, on the other hand, indicates a smaller but still positive correlation, indicating that getting enough sleep may have a minor impact on academic achievement. The combined significance of digital access and well-being is highlighted by the grouped bar chart, which shows that pupils with higher mental health ratings and better internet quality typically have higher average exam results. Collectively, these findings highlight the significant influences that study practices, sleep patterns, internet quality, and mental health have on academic performance.

Summary of Key Findings

Important new information on the demographic and behavioral traits of the student body and how they relate to academic achievement was revealed by the exploratory data analysis. While characteristics like sleep length and digital media consumption showed more modest patterns, key predictors like the number of study hours per day and mental health rating shown significant connections with exam grades.

Building on these discoveries, the next stage involves model creation, in which the discovered traits are utilized to predict academic achievements using supervised machine learning techniques.

3.3 Model Development

Four supervised regression algorithms LR, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor were implemented throughout the model construction phase in order to forecast student academic success based on behavioral, demographic, and lifestyle characteristics. From straightforward linear connections to ensemble-based non-linear pattern recognition, these models were chosen to reflect a variety of complexity and learning techniques. This made it possible to compare their prediction skills in detail.

Label Encoding was used to transform categorical variables such as gender, part-time employment status, food quality, parental education level, internet quality, and extracurricular involvement into numerical form appropriate for machine learning models prior to model training. StandardScaler was then used to standardize numerical characteristics in order to guarantee scale consistency and enhance model convergence.

To maintain the distribution of important variables in both sets, Using stratified sampling, the dataset was separated into two subsets: training (80%) and testing (20%). To provide for fair comparison in the testing phase that followed, each model was trained independently on the scaled training data with no previous evaluation.

This method established the foundation for a thorough performance evaluation in the following stage by guaranteeing that the models retained their capacity to generalize to new situations while learning patterns from the data.

3.4 Evaluation

After Following the training of the four regression models, the predictive performance of each model was separately assessed using the test set.

The accuracy and dependability of the model were assessed using two primary assessment metrics:

- The Root Mean Squared Error (RMSE) measures prediction accuracy, with lower values indicating better results.
- The model's ability to explain exam score variation is shown by the R^2 Score (Coefficient of Determination); values nearer 1 signify a better match.

An unbiased assessment of the model's efficacy was made possible by the evaluation's combination of both measures, which struck a compromise between the requirement for low prediction error and robust explanatory power. This thorough evaluation made it easy to choose the best model, opening the door for its implementation and real-world use in forecasting student academic performance.

3.5 Model Selection and Deployment

Four supervised regression models were trained, and their performance on the test set was used to choose a model. Because it produced the most accurate exam score predictions, The model that had the lowest Root Mean Squared Error (RMSE) was chosen for deployment.

The chosen model and related feature scaler were stored using the joblib package to guarantee repeatability and future usage. As a result, throughout time, the system can continue to exhibit consistent preprocessing and prediction behavior.

A user-friendly, interactive prediction tool was created, allowing users to enter essential behavioral and demographic information for specific kids. The system runs this input through the trained model and returns a projected exam score, along with a confidence estimate based on the model's error margin. This deployment stage bridges the gap between model creation and real-world implementation by providing educators, parents, and students with accessible insights for data-driven academic assistance.

4 Result Interpretation

This section provides a comparative examination of the four regression models used to predict student academic success. By comparing each model to the test dataset using conventional metrics. The findings assist in determining the best model for deployment while also highlighting the benefits and limits of each strategy.

4.1 Model Performance Comparison

To predict student academic achievement, four supervised regression algorithms were trained and assessed using RMSE and R^2 Score.

Model Performance Summary

Model	RMSE	R^2 Score
Linear Regression	5.14	0.9
Gradient Boosting	5.61	0.88
Random Forest	6.28	0.85
Decision Tree	9.85	0.62

Figure 4: Model Performance Comparison

Linear Regression outperformed other models, with the lowest RMSE (5.14) and greatest R^2 Score (0.90), suggesting good accuracy and explanatory power. Gradient Boosting followed closely, with somewhat more error but still producing strong results. Random Forest had reasonable predictive capacity, but Decision Tree did poorly, with the biggest inaccuracy and the lowest explanatory capability. These findings demonstrate the efficacy of Linear Regression and Gradient Boosting in

capturing the intricate interactions between behavioral, demographic, and academic performance factors, making them the best models for practical use in educational prediction tasks.

4.2 Prediction Accuracy Analysis

The prediction ability was further demonstrated by showing real against projected exam scores using the Linear Regression model.

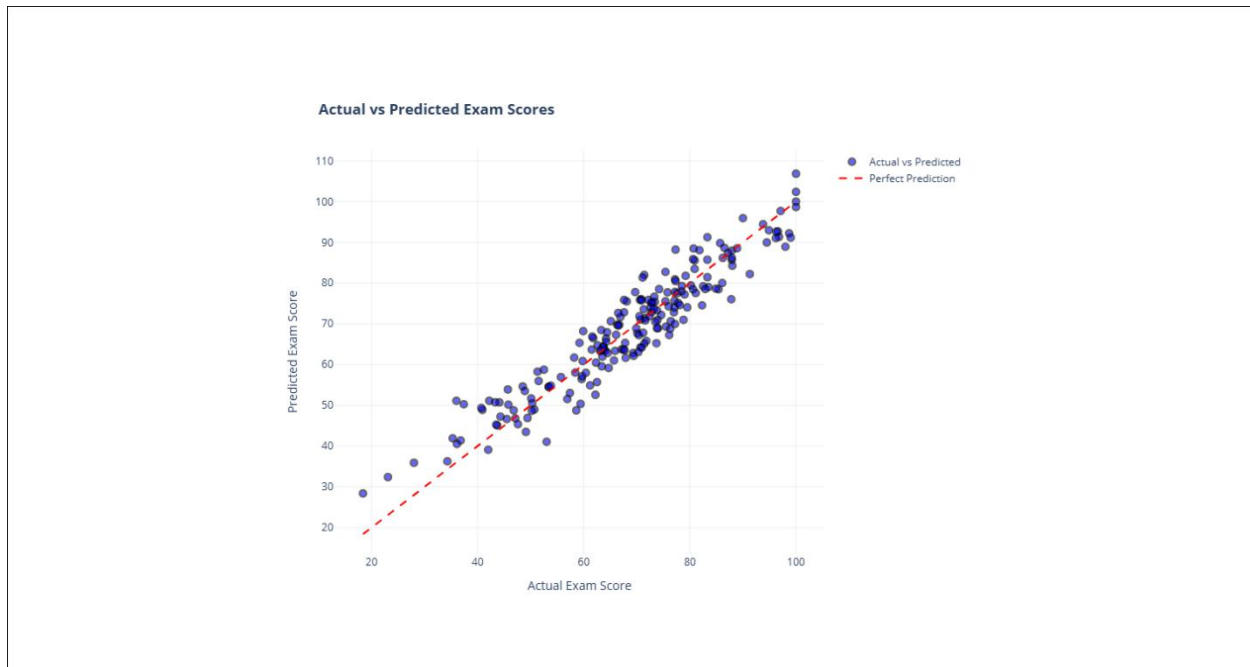



Figure 5: Actual vs Predicted Exam Scores Using Linear Regression

The scatter shows that the majority of data points cluster closely around the ideal prediction line, showing good alignment between actual and projected values. This shows that the model adequately captures the underlying trends in the data and makes accurate predictions about student academic achievement.


4.3 Practical Application via Interactive Prediction Tool

Aside from assessing model performance using conventional metrics, a special interactive prediction interface was created to demonstrate the actual use of the chosen model. This application allows users to enter specific student data such as behavioral, demographic, and lifestyle characteristics to get a projected exam score and an estimated confidence range. By combining model prediction with user input and visualization, the interface not only shows the


model's success in real-world circumstances, but it also improves interpretability and accessibility for educators, parents, and students.

 Student Achievement Prediction - Input Data


Field	Value
Age	18
Gender	Female
Study Hours Per Day	4
Social Media Hours	3
Netflix Hours	2
Part Time Job	No
Attendance Percentage	70
Sleep Hours	7
Diet Quality	Fair
Exercise Frequency	2
Parental Education Level	Bachelor
Internet Quality	Average
Mental Health Rating	6
Extracurricular Participation	Yes



Prediction Result



Predicted Exam Score: 70.04 / 100



Estimated Range: 64.9 – 75.2

Figure 6: Interactive Prediction Outcome

5 Conclusion and Future Work

This study used supervised machine learning to investigate the relationship between student behavioral patterns and academic achievement. Daily study hours and mental health evaluations

revealed large positive relationships with exam performance, but sleep length, screen usage, and internet quality all had an impact. Linear Regression outperformed the other models studied in terms of accuracy and interpretability when predicting performance from behavioral and demographic data.

The findings underline the importance of adding non-academic characteristics in educational prediction models, which provide more detailed insights than typical academic records alone. The creation of an interactive prediction tool highlights the practical possibilities for early intervention and support.

Future research should focus on incorporating the model into student dashboards to give tailored feedback and early warnings. Furthermore, adding live behavioral data from LMS platforms or wearable devices may enable real-time monitoring and adaptive learning tactics, resulting in more responsive, data-driven support systems.

References

Sarker, S., Paul, M.K., Thasin, S.T.H. and Hasan, M.A.M., 2024. Analyzing students' academic performance using educational data mining. *Computers and Education: Artificial Intelligence*, 7, p.100263.

Tang, B., Li, S. and Zhao, C., 2024. Predicting the Performance of Students Using Deep Ensemble Learning. *Journal of Intelligence*, 12(12), p.124.

Butt, N.A., Mahmood, Z., Shakeel, K., Alfarhood, S., Safran, M. and Ashraf, I., 2023. Performance prediction of students in higher education using multi-model ensemble approach. *IEEE Access*, 11, pp.136091-136108.

Baniata, L.H., Kang, S., Alsharaiah, M.A. and Baniata, M.H., 2024. Advanced deep learning model for predicting the academic performances of students in educational institutions. *Applied Sciences*, 14(5), p.1963.