A Report on Predicting and Interpreting Student Performance Using Machine Learning Techniques

Institution:
Zaio Institute of Technology

GitHub Repository (Proof of Work & Progress):
https://github.com/Pelly-DS/student_data_analysis.git

Mbuso Pellican Mhlongo

Table of Contents

Executive Summary

The aim of this project is to study how student habits affect their academic performance using data and machine learning. The dataset used is called student_habits_performance.csv. It includes information about study hours, sleep time, class attendance, and other daily habits.

The data was cleaned, explored, and used to build models that can predict how well a student might perform. Two models were tested: Decision Tree and Logistic Regression. Their performance was compared to find which one gives the most accurate results. The analysis showed that students who study regularly, sleep between six and eight hours, and attend most of their classes usually perform better. Students with poor attendance or too much screen time tend to have lower grades.

The best model gave strong and reliable predictions. Tools like SHAP and LIME were used to explain the results and show which habits have the biggest effect on student success.

Introduction

Student success is influenced by many habits such as study time, sleep patterns, attendance, and lifestyle choices. Understanding how these habits affect academic performance can help both students and educators make better decisions to improve learning outcomes.

In this project, we use data from the student_habits_performance.csv dataset to explore and analyse these relationships. The goal is to identify which habits have the strongest impact on student performance and to predict academic results using machine learning models.

Through data cleaning, visualisation, and predictive modelling, this project provides insights into how students can improve their study routines and how schools can better support learning and development.

Objectives

This project aims to use data to understand how different student habits affect academic performance. The main goals are:

1. Explore the Data Thoroughly
Understand the main patterns and relationships between study habits, attendance, sleep time, and student grades.

2. Identify Key Factors
Find out which habits have the biggest influence on academic performance, and which ones may lead to lower results.

3. Test Ideas with Statistics
Use basic statistical methods to check if factors such as study hours, sleep duration, and attendance really affect student outcomes.

4. Build Predictive Models
Develop machine learning models, such as Decision Tree and Logistic Regression, to predict student performance based on their habits.

5. Create Clear Visuals
Use charts and graphs to show patterns and insights in a way that is easy to understand and useful for both students and educators.

6. Provide Practical Recommendations
Give clear, data-based suggestions to help students improve their habits and help educators design better academic support strategies.

Methodology

Data Collection

For this project, the dataset named *student_habits_performance.csv* was used as the main source of data. It contains information about students' daily habits such as study hours, sleep duration, attendance, screen time, and overall academic performance.

The dataset was imported into Jupyter Notebook for analysis. Before starting any modelling, the data was carefully reviewed to ensure accuracy and completeness.

To make the data easier to work with, the following steps were taken:

- The dataset was saved in Excel (.xlsx) format for initial review.
- Column names were checked for clarity and consistency.
- Missing or incorrect values were identified and prepared for cleaning.
- Each variable was reviewed to understand its meaning and importance in predicting student performance.

Data Preparation

1. Checking and Cleaning Missing Values:
The dataset was carefully reviewed for missing or incomplete values. Rows with major gaps or errors were corrected or removed to keep the data accurate and reliable.

2. Fixing Formatting and Data Types:
Some variables were in the wrong format, such as numeric values stored as text. These were corrected to make sure all calculations and model training worked properly.

3. Removing Unnecessary Columns:
Certain columns that did not directly help with the main analysis were deleted. One of these was parental education. While parental education can have some influence on a student's background, it is not always the main factor that determines academic success.

Before focusing on parental background, it is more important to look at the student's well-being, such as whether the student feels safe, supported, and free from any form of abuse or neglect. A healthy and positive environment has a stronger and more direct effect on learning than a parent's level of education. As the saying goes, *"You can take a horse to the water, but you can't make it drink."* This means that no matter how good the school or background is, the student's own motivation and environment play the biggest roles in success.

4. Renaming Columns for Clarity:
All column names were changed to short, clear, and meaningful titles such as Study_Hours, Sleep_Hours, Attendance, Screen Time, and Performance_Level. This made the dataset easier to understand and work with.

5. Normalising and Encoding Data:
Numeric values were standardised to keep all measurements on the same scale, and categorical variables were converted into numeric form to be compatible with machine learning algorithms.

6. Creating the Clean Dataset:
After all cleaning and formatting steps, the final dataset was saved as students_clean.csv, ready for further exploration and modelling.

Missing Values

While reviewing the dataset, some records were found to have missing information in a few important columns such as Study_Hours, Sleep_Hours, Attendance, and Screen Time. These missing values required special attention to make sure the dataset remained accurate and reliable.

Before filling in the missing values, all column names and formats were standardised. For example, text entries were cleaned by removing unnecessary spaces and converting everything to lowercase or uppercase where needed, ensuring consistency across the dataset.

For numeric columns such as study hours and sleep duration, missing values were filled using the median of other students' data. The median was chosen instead of the mean because it is less affected by extreme values and gives a more balanced estimate. This made the data more realistic and reduced bias.

For categorical columns such as attendance or performance level, missing values were filled using the most common category (mode) among students with similar study patterns. This approach helps maintain data consistency while preserving the overall trends.

In some cases, missing values could not be reasonably estimated — for example, when key information about a student's background or habits was completely absent. These rows were removed to avoid errors in later stages of analysis.

This careful handling of missing values ensured that the dataset remained clean, complete, and ready for accurate data exploration and machine learning modelling.

Handling Missing Values

During data cleaning, the dataset was checked for any missing or incomplete information. The only column with missing data was parental_education_level. This column was removed from the dataset because it contained too many gaps and was not essential for the core analysis. Before making this decision, it was considered that student performance is often influenced more by personal habits, well-being, attendance, and mental health than by parental education alone. Removing this column helped simplify the dataset and focus on factors that directly reflect students' lifestyles and learning behaviour.

Tools Usage

Dataset was analysed using a range of tools to ensure accuracy and efficiency throughout the project. Jupyter Notebook served as the main environment for running Python code and documenting each step.

Key Metrics and Insights

This section presents the most important findings from the student habits dataset. It shows how factors such as study hours, sleep duration, attendance, screen time, and mental health relate to students' overall academic performance. By examining these key indicators, we can understand common patterns and identify which habits have the strongest influence on exam results.

Performance Indicators

The dataset includes several performance indicators that show how well students are performing academically. These include the exam score, which represents overall academic achievement, and other key factors such as attendance percentage, study hours per day, and sleep duration. Additional indicators like mental health rating, exercise frequency, and screen time (social media and Netflix hours) also play an important role in understanding students' well-being and learning balance. Together, these indicators help identify which habits and lifestyle choices are most closely linked to strong academic performance.

Student Segmentation

Student segmentation was done by analysing patterns in study habits, lifestyle behaviours, and academic performance. Different groups of students were identified based on factors such as study hours, sleep duration, attendance percentage, mental health, and screen time. For example, some students showed strong academic performance with balanced study and rest habits, while others struggled due to poor sleep, low attendance, or excessive screen time. Understanding these groups helps identify which habits lead to better results and allows for more focused support and improvement strategies.

Appendix

- Students who study more than 4 hours per day generally achieve higher exam scores than those who study less.
- Those with good sleep (7–8 hours) show better focus and performance compared to students who sleep less than 5 hours.
- High attendance (above 85%) strongly links to better academic results.
- Students who spend too much time on social media or Netflix tend to have lower scores, showing the impact of distractions.
- A balanced lifestyle, including regular exercise and good mental health, contributes to improved academic performance.

Model Accuracy

Several machine learning models were tested to predict student performance based on factors such as study hours, attendance, sleep, and lifestyle habits. The results showed that all models performed well, with accuracy scores ranging from 78% to 89%. The Decision Tree achieved an accuracy of 83%, while the Random Forest model slightly improved performance to 84%. The Logistic Regression model performed the best, reaching an accuracy of 89%, meaning it made the most reliable predictions. The Support Vector Machine (SVM) had a lower accuracy of 78%, showing some difficulty in balancing both performance groups. Lastly, the XGBoost model achieved 85% accuracy, providing strong and consistent results. Overall, Logistic Regression proved to be the most effective model for predicting student academic outcomes.

Tools Used

In this project, several tools and techniques were used to ensure accurate analysis and reliable results. The data was first cleaned by correcting errors and handling missing information to improve quality. Mathematical and statistical methods were then applied to discover important patterns and relationships between student habits and academic performance. Different machine learning models were tested to find the most effective one for predicting student success. New features were also created from the data to improve model accuracy and understanding. Finally, graphs and visual reports were developed to present the findings in a clear and professional way, helping support better decision-making. The entire system was designed to be efficient, reliable, and easy to evaluate.

Applied Statistical Analysis

Applied Statistical Analysis was used to understand and interpret the student data in depth. The process began with cleaning the dataset by correcting errors, handling missing values, and removing outliers to ensure accuracy. New features were then created to improve the analysis and model performance. Statistical tests were conducted to determine whether factors such as study hours, attendance, or sleep had a significant effect on academic performance. Correlation analysis was also applied to identify which habits were most closely linked to exam scores. Techniques like data normalization and feature scaling helped ensure fair comparisons between variables. Overall, the statistical analysis provided meaningful insights into the key factors that influence student success and supported the development of accurate prediction models.

Advanced Analysis

Advanced Analysis was used to explore the student data in greater depth using powerful machine learning techniques. This included building predictive models to estimate student performance based on habits such as study hours, sleep duration, attendance, and lifestyle factors. Clustering methods were also used to group students with similar behaviour and performance patterns, helping to identify at-risk learners and high achievers. Each model was fine-tuned and tested using cross-validation to ensure reliability and accuracy on new data. These advanced techniques helped uncover hidden relationships between student habits and academic success, providing valuable insights for educators and learners to make data-driven decisions.

Summary of Findings

The analysis showed that several key factors strongly influence student academic performance. Students who study for more hours per day, maintain good sleep habits, and have high attendance rates generally achieve better exam results. Excessive use of social media or streaming platforms, such as Netflix, was found to negatively affect performance. Regular exercise and good mental health were also linked to improved academic outcomes. Among the models tested, Logistic Regression performed the best with an accuracy of 89%, showing strong predictive ability. Overall, the project demonstrated that consistent study habits, balanced well-being, and responsible time management play a major role in student success.

Reflection

This project provided valuable experience in understanding how data can be used to study and improve student performance. One of the key lessons learned was the importance of data cleaning, as even small errors or missing values can lead to inaccurate results. By carefully handling and preparing the data, the analysis became more reliable and meaningful.

We also learned how different habits such as study time, sleep quality, attendance, and screen time affect academic outcomes. It became clear that balanced habits and consistent effort lead to better performance, while excessive time spent on social media or entertainment can lower results.

Using advanced tools like machine learning models helped predict student performance with high accuracy. These models not only identified key patterns but also provided insights that can help educators and students make better decisions.

Overall, this project showed how data-driven approaches can be used to support education. It emphasized the value of careful preparation, testing, and interpretation to ensure that the findings are both real and useful. It also highlighted the importance of using technology responsibly to improve learning outcomes and student well-being.

Advanced Data Quality Assessment

A detailed study of student habits and performance was carried out using advanced data analysis techniques such as statistics, machine learning, and data visualization. The dataset was thoroughly cleaned to remove errors, handle missing values, and ensure consistency. Each variable was carefully reviewed to confirm accuracy and reliability. After cleaning, statistical tests and correlation checks were performed to identify key relationships between study habits, lifestyle choices, and academic outcomes. This advanced data quality assessment ensured that the information used for modelling and insights was valid, trustworthy, and ready for deeper analysis using machine learning tools.

Data Quality Achievements

Several data quality issues were successfully identified and fixed to make the dataset more accurate and reliable. Missing values were filled based on logical patterns and relationships within the data. Outliers — unusual or extreme values — were detected and removed using the Interquartile Range (IQR) method to prevent distortion in analysis. Additionally, new useful features were created, such as calculated averages and combined metrics, to better explain relationships between student habits and performance. These improvements strengthened the overall quality of the dataset and made it ready for advanced modelling and analysis.

Statistical Analysis Results

The statistical analysis provided clear insights into how different habits affect student academic performance. Hypothesis testing showed that factors such as study hours per day, attendance percentage, and sleep duration have a significant relationship with exam scores. This means that students who study regularly, attend classes consistently, and get enough rest tend to perform better.

In contrast, factors such as parental education level or screen time showed weaker statistical relationships with performance. This suggests that while background and environment play some role, a student's personal habits and daily routines are the strongest predictors of success.

Overall, the results highlight that consistent study habits, good sleep, and class participation are the most influential factors in achieving better academic outcomes.

Hypothesis Testing Results

The statistical analysis revealed clear relationships between student habits and academic performance. Hypothesis testing showed that factors such as study hours, attendance percentage, and sleep duration have a strong and significant effect on exam scores. Students who spend more time studying, attend classes regularly, and maintain healthy sleep patterns generally perform better academically.

On the other hand, variables such as screen time (social media and Netflix hours) showed a negative impact on performance when used excessively. Similarly, diet quality, exercise frequency, and mental health rating were found to contribute positively to overall academic outcomes, highlighting the importance of balance and well-being.

Overall, the statistical analysis confirmed that consistent effort, healthy routines, and good time management are key drivers of strong student performance.

Machine Learning Model Performance

The machine learning models performed very well in predicting student academic performance. Each model analysed key factors such as study hours, attendance, sleep duration, screen time, and lifestyle habits. Among all models tested, Logistic Regression achieved the highest accuracy of 0.89 (89%), showing strong prediction ability. Random Forest and XGBoost also performed well with accuracies of 0.84 and 0.85, respectively, while Support Vector Machine (SVM) had a moderate accuracy of 0.78.

These results show that the models can reliably predict whether a student is likely to perform well based on their habits and behaviours. The high accuracy scores demonstrate that the approach is effective, practical, and useful for identifying students who may need academic support or habit improvement.

Category Analysis & Key Insights

The analysis of student lifestyle categories revealed clear patterns that influence academic performance. The most common group of students studied between 2 to 4 hours per day, while a smaller portion studied for more than 5 hours. Students with high attendance rates (above 85%) and adequate sleep (7–8 hours) consistently achieved better exam scores.

In contrast, students who spent more than 3 hours per day on social media or Netflix generally performed below average, showing that excessive screen time negatively affects learning outcomes. Additionally, those with regular exercise and balanced diets demonstrated higher concentration and motivation levels.

These insights show that healthy routines and consistent study habits are key drivers of strong academic performance, while poor time management and excessive leisure screen time can significantly reduce success.

Most Influential Factors on Performance

The analysis found that some habits have a much stronger effect on student academic results than others. Students who study for more hours per day and maintain high attendance levels achieved the highest exam scores. Those who sleep well (7–8 hours per night) also performed better, showing that rest plays an important role in learning.

On the other hand, students who spend too much time on social media or streaming platforms had noticeably lower results, suggesting that distractions negatively affect focus and productivity. Additionally, students with good diet quality, regular exercise, and positive mental health ratings tended to perform better overall.

These findings show that focusing on productive study habits and personal well-being can significantly improve academic performance, helping students reach their full potential.

Data-Driven Insights

To improve student performance, schools should focus on the key factors that have the strongest impact on results such as consistent study habits, regular attendance, and sufficient sleep. Providing academic support programs, mentorship, and wellness initiatives can help students maintain focus and motivation. Data shows that external factors like parental education have less influence than students' own efforts, so interventions should prioritize personal learning habits and time management. Encouraging balance between study and rest can lead to more sustainable academic success. Analysis revealed clear patterns about what influences student performance the most. Study hours, attendance, and sleep quality were the strongest predictors of academic success. Students who studied consistently, attended classes regularly, and maintained healthy sleep habits achieved higher scores. On the other hand, factors such as parental education and income had less direct impact once personal habits were considered. These insights highlight the importance of focusing on students' daily learning behaviours and well-being rather than external background factors. Schools can use this information to design targeted support programs that encourage consistent study routines and improve overall student outcomes.

Technical Implementation Highlights

The student performance project applied advanced data cleaning techniques to fix errors and handle missing information. Statistical methods were used to test relationships between study habits, attendance, sleep quality, and exam results, revealing key patterns that influence performance. Machine learning models were trained, tested, and fine-tuned to predict student outcomes accurately. New data features such as total study engagement and lifestyle balance were created to improve model performance. The system was built with reliable code and strong error handling to ensure smooth execution. Performance metrics and resource usage were monitored for efficiency, and clear visual reports and charts were developed to communicate the insights effectively.

References

GitHub Repository: https://github.com/Pelly-DS/student_data_analysis.git
Email: mhlongopelly@gmail.com
Tools Used: Visual Studio Code, Jupyter Notebook, Power BI

Useful Links for This Assignment

1. **Data Cleaning in Python with Pandas**
   https://realpython.com/python-data-cleaning-numpy-pandas/
   Explains handling missing values, fixing alignment issues, and data cleaning techniques.
2. **Introduction to Statistical Analysis in Python**
   https://www.geeksforgeeks.org/statistical-analysis-python/
   Covers hypothesis testing, data distributions, and fundamental statistical methods.
3. **Machine Learning with Scikit-learn**
   https://scikit-learn.org/stable/tutorial/basic/tutorial.html
   A detailed tutorial on model training, testing, cross-validation, and tuning.
4. **Power BI Reporting and Data Visualization**
   https://learn.microsoft.com/en-us/power-bi/fundamentals/desktop-getting-started
   Guide on creating dashboards and visual reports for presenting data insights.

Glossary

Student Performance means how well a student does in school, usually measured by exam scores or grades.
Exam Score is the mark a student receives for a test or assessment, often out of 100.
Study Hours are the number of hours a student spends studying each day.
Attendance Percentage shows how often a student attends class — high attendance usually improves performance.
Sleep Hours are the average number of hours a student sleeps per night, which affects focus and academic results.
Mental Health Rating measures how students feel emotionally and mentally, often linked to their performance.
Exercise Frequency is how often a student exercises per week, which can improve energy and concentration.
Diet Quality refers to how healthy a student's eating habits are, influencing their physical and mental performance.
Social Media Hours measure how much time a student spends online; too much can lower focus and grades.
Missing Values are gaps in the dataset where some student information is not available (like missing sleep hours).
Data Cleaning means fixing problems in the dataset — removing errors, filling in missing values, or making formats consistent.
Feature is any piece of information about a student (like study hours, attendance, or diet) used in analysis.
Predictive Model is a machine learning tool that guesses outcomes, such as predicting a student's exam score.
Trend Line is a line on a graph that shows the overall direction of data, like how study hours relate to exam scores.
Box Plot is a graph showing how exam scores are spread out — it helps find high and low performers.
Scatter Plot is a graph with dots showing the relationship between two things, such as study hours and performance.
Bar Plot is a chart with bars that shows numbers or averages, like the average score by gender or age group.
Clustering means grouping similar students together based on habits and performance, to find patterns.
Hypothesis Testing checks if differences in results (like between genders or age groups) are real or just by chance.
P-value is a number from hypothesis testing that shows if the result is meaningful — a small p-value means strong evidence.
Feature Importance shows which factors (like study hours or attendance) most affect

student performance.
Outliers are unusual data points — for example, a student with extremely high or low exam scores.
Median is the middle value in a set of numbers, often used to fill missing data fairly.
Correlation tells how two factors are related — for example, how more study hours often lead to higher grades.
Standardization means making all data consistent (like using the same units or formats) for fair comparison.
Machine Learning is when computers learn from student data to predict outcomes or find patterns without manual rules.
Cross-Validation tests if a model can make accurate predictions on new student data, ensuring it is reliable.