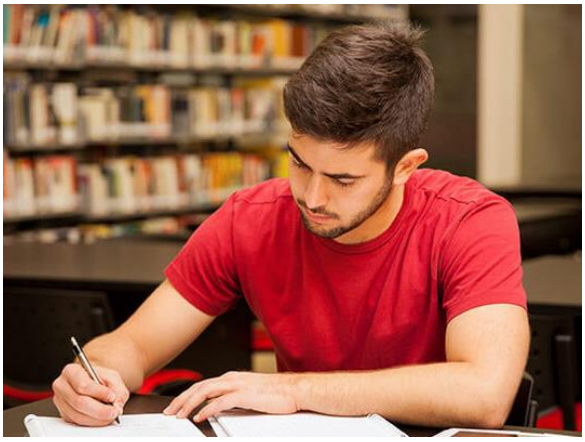
Prediction of Student Exam Performance

By Nnaemeka Ukpa
Student Number: 23035351
GitHub Link: GitHub
Data Source: Kaggle



Introduction

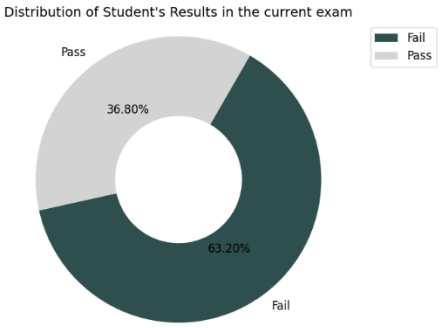
Exam have just been concluded and as we all know it, some students are going to pass and others will fail. What if we are able to predict if a student will pass their current exams based on their previous records?

The dataset given contains records for 500 students and it contains fields such as the number of hours the student uses to study, their score in their previous exam and a “Pass/Fail” column which tells us if the student passed their current exam. The main goal of this analysis is to explore the data, observe patterns and correlations in students performance and build a model that predicts the outcomes of student exams.

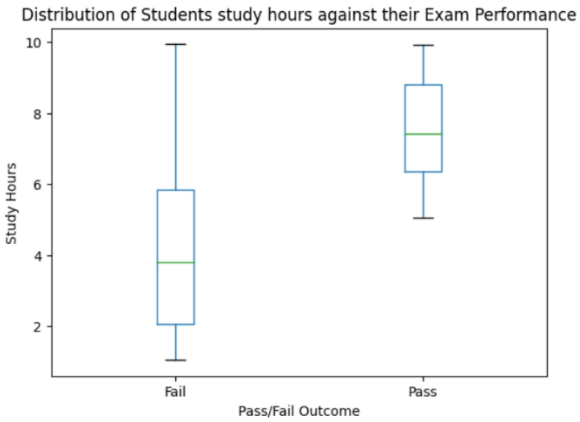
This analysis was done using Python.

Exploratory Analysis

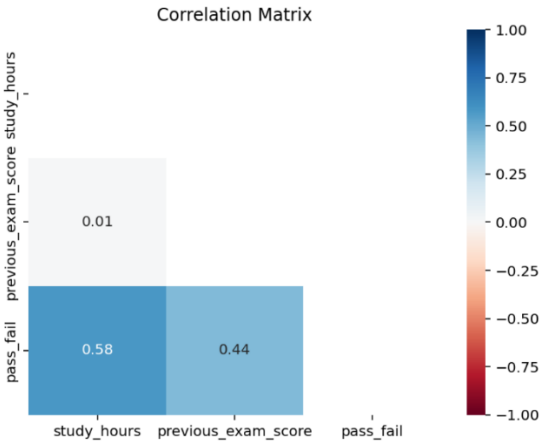
The first step was to study the data for the percentage of students who passed their current exams to have an understanding of the general student’s performance. From the image below, it is showing that only 36.8% of students passed their exams and the remaining failed.

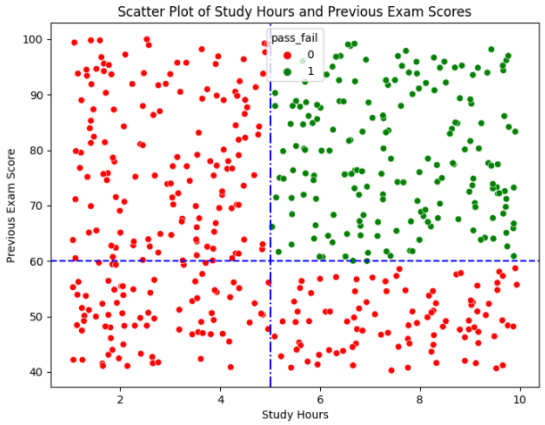


I presumed that students who passed their exams had more study hours compared to students who failed. I tested this hypothesis by checking the distribution of student study hours for those who passed and those who failed. The image below shows that students who passed have more study hours on average than the student who failed.



The whiskers of the boxplot on the left also indicates that there are students who had more study hours but still failed. Does this mean that your study hours alone isn’t enough to Pass the current exam? To prove this, I had to check for correlation between the study hours, previous exam performance and the students performance in exam.





The correlation between the study hours and the student performance is 0.58 which is a moderate correlation. This means that students who study more tend to get higher scores on average, but there will still be variability in scores among students who study the same amount and this was clearly seen in previous charts. The same is observed in the correlation between previous exam score and student performance.

A scatter plot was created for further observation and hue was included to differentiate between students who passed and failed. A pattern was observed that only students who scored minimum of 60 points and studied a minimum of 5 hours, passed their exams.

Logistic Regression

A logistic regression model was built to accurate predict if a student will succeed or fail in their current exam based on the number of hours they study and their performance in their last exam. The model has an accuracy of 85% and the model was evaluated using the classification report and the confusion matrix. More details about can be found in the GitHub Repository.