MASTERING SUCCESS:
HOW TO BE A GOOD STUDENT

JOHAN VEENPERE, HENDRIK ARUOJA, ANNA SULG

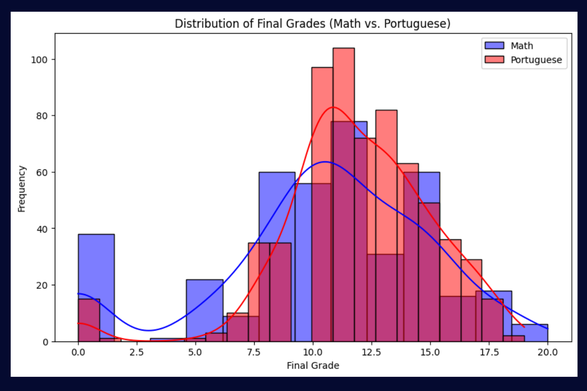
INTRODUCTION

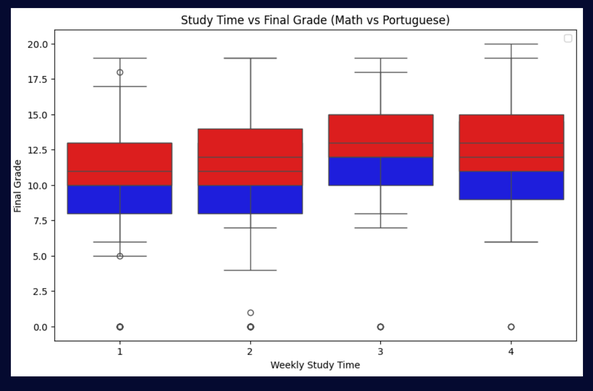
As students ourselves we know how sometimes despite putting in what feels like efficient effort, your test scores end up indicating otherwise. Our goal is to analyze the effect of different factors on a student's performance in school, develop a predictive model that students can use to learn the potential test score they would achieve based on their lifestyle, studying habits and effort. We also want to give students personalised tips based on data analysis on how to improve their test scores. We will conduct a deeper dive into the data and connections between different factors. We wish to determine the main factors that contribute to a student's success, debunk or back up some common suggestions for improving scores and making our work easily accessible to all who are interested in gaining insight on how to be a better student.

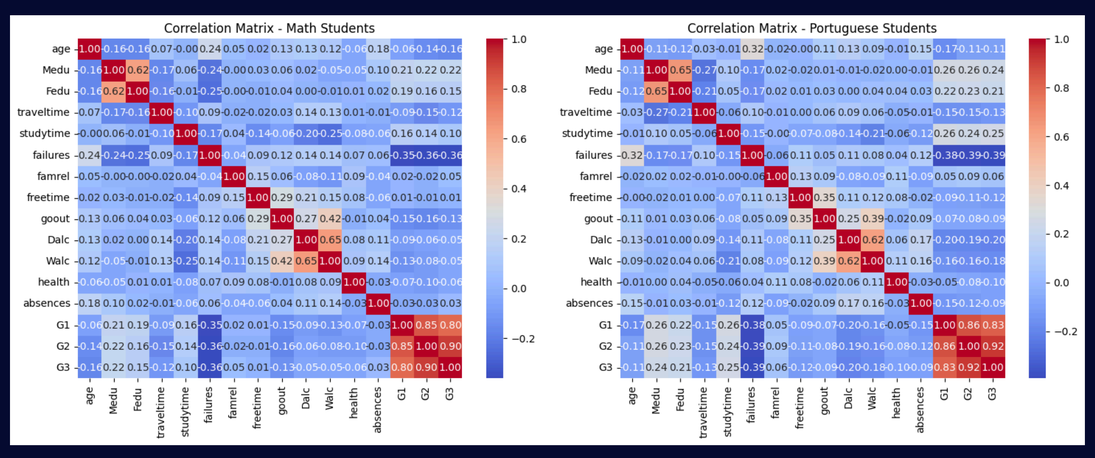
DATA UNDERSTANDING

We started our research out with a total of 4 different datasets - (overall student performance factors dataset, math exam dataset, Portuguese language exam dataset, Sleep dataset). The main was "Student Performance Factors", a dataset on Kaggle that provides insight into student performance and contributing factors. Our analysis yielded these three leading factors as the most significant:

- Hours spent studying in a week
- Attendace
- Sleep hours







OUR APPROACH

The solution is to train different machine learning models that can predict the exam results based on other features and choose the best model based on MSE[1].

DATA ANALYSIS AND TRAINING

We trained 4 different models and compared their mean squared errors.

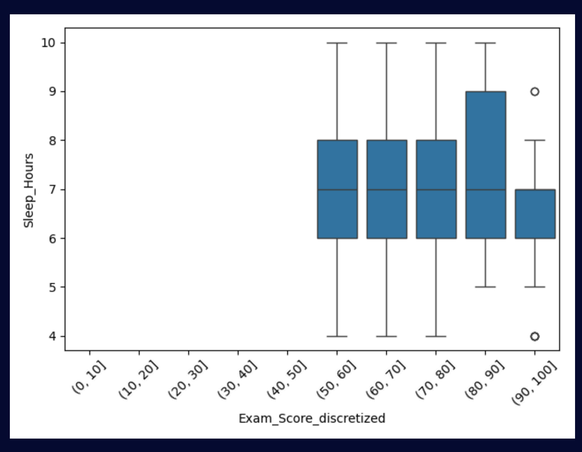
KNeighborsRegressor - 9,304
DecisionTreeRegressor - 14.659
RandomForestRegressor - 5.805
SupportVectorRegressor - 4.318

Best models for our problem were RandomForestRegressor and SupportVectorRegressor, both of which had a MSE[1] of around 5.

Most important features were Hours Studied, Attendance and Sleep Hours.

COMMON MISCONCEPTION?

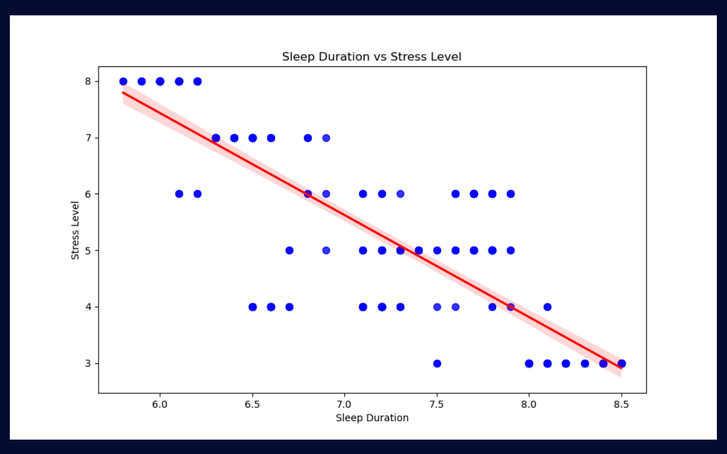
It is often said that students need 7-9 hours of sleep every night, but does an adequate amount of sleep mean higher exam scores? Let's find out. If we want to find correlation between sleep hours and exam scores we should discretize[2] one of the attributes to avoid visualising the relationship of two continuous attributes.



Here we can see that people with very good exam scores don't tend to sleep less than 5 hours or over 8 hours, this is the close to the recommended 7-9 hours of sleep. Could this mean that sleep does play a big role in exam scores? To be able to say for sure we needed to conduct a t-test[3].

All of the tests produced a p-value above our significance threshold, meaning we cannot say for sure that there is a correlation between sleep hours and exam scores.

But then why is the recommendation to sleep 7-9 hours so common? To analyze this we will have to look at an additional dataset to our original one. Students usually have a lot of stress so let's analyze if insufficient sleep could further amplify stress.



There is a clear linear relationship between the two attributes. Fewer sleeping hours are associated with higher stress levels.

Stress has numerous negative effects on the body and mind such as irritability, depression, insomnia (leading to even less sleep), and high blood pressure. Lowering stress levels is essential for students to be able to focus on their studies.

RESULTS

We have a model which can predict a student's exam score with a deviation of around 2 points.

REFERENCES

Student Performance Factors. URL : *https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data*
Sleep Health and Lifestyle Dataset. URL : *https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset*
https://www.statology.org/what-is-a-strong-correlation/

Definitions:
[1] Mean Squared Error (MSE) is calculated by finding the average of squares of differences between predicted and actual values.
[2] *Discretization* means making a continuous attribute into a categorical one. This makes it easier to analyze relationships.
[3] *T-test* is a tool that helps you decide if two groups are really different, or if their differences could just be due to random chance.