Mini-project 1

Student performance in secondary education

by I.Font

“To miss out on mathematics is to live without an opportunity to play with beautiful ideas and see the world in a new light….”

Francis Su, Mathematics for human flourishing

Aim

The aim of this analysis is to explore and understand how various academic and social factors influence students' performance in Mathematics.

The dataset

Source: UCI Machine Learning Repository.

Dataset: A record of student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related attributes collected through school reports and questionnaires. The dataset contains 30 features for 395 students, and relates to performance in Mathematics.

Categorical

- school
- sex
- address
- family size
- parent's cohabitation status
- mother's job
- father's job
- reason to choose this school
- guardian

Categorical: Yes/No

- extra educational support
- family educational support
- extra paid classes
- extra-curricular activities
- attended nursery school
- wants to take higher education
- Internet access at home
- with a romantic relationship

Categorical: number

- mother's education level
- father's education level
- home to school travel time
- weekly study time
- number of past class failures
- quality of family relationships
- free time after school
- going out with friends
- workday alcohol consumption
- weekend alcohol consumption
- current health status

Numerical

- age
- number of school absences
- first period grade
- second period grade
- final grade

Method

Exploratory Data Analysis

Data inspected for integrity and missing values

Value counts for categorical variables

Summary statistics of numerical variables

Visualisations of final marks against all other variables

Correlation of final mark vs. absences and other marks

Further exploration of distribution of final marks vs. five categorical variables

Comments & observations

Data quality was high with no missing values

Several categorical variables in the dataset exhibit a high degree of imbalance, where the majority of observations fall into a single category (ie. 83% had internet access, 89% from two parent households, etc)

Students ages were between 15 to 22 and marks were between 0 and 20.

Maximum number of absences was 75 days, but 75% of students missed 8 days or less

The final mark showed a strong positive correlation with previous marks as expected

The variables that seemed to have an impact on final results were: living in urban areas, internet access, study time, **desire of accessing higher education** and **past failures**

Method

Hypothesis testing

Question: Are the mathematical final results of students who have never failed a class higher than those who have failed one or more classes?

> H_N: Students without past failures don't achieve higher final marks (difference is due to chance)
>
> H_A: Students without past failures score higher on final marks.

One sided two-sample z-test with unequal variances and significance level of 5% or 0.05

Results

Z-statistic: 7.0156

p-value: 0.0000

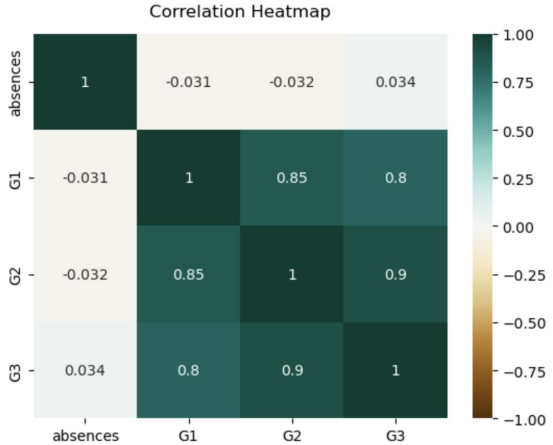
The results provide strong statistical evidence to reject the null hypothesis in favour of the alternative hypothesis: Students who have not failed a class tend to achieve significantly higher final marks than those who have failed in the past.

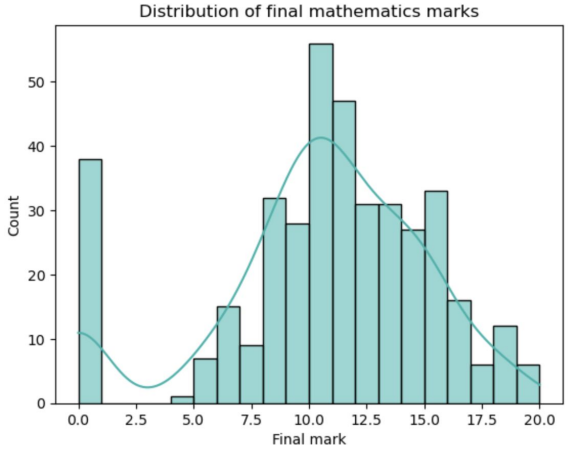
Future improvements

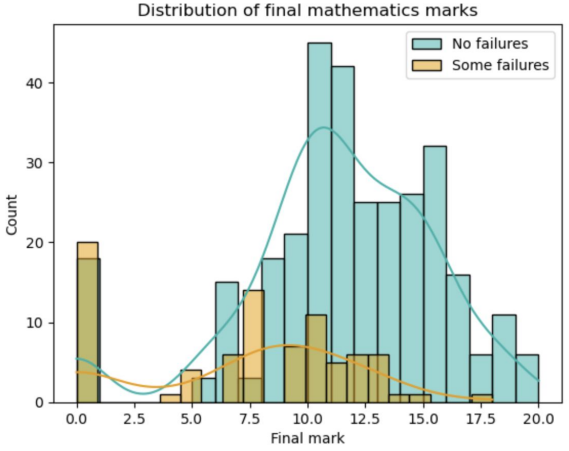
This analysis examined the effect of one factor at a time.

A more comprehensive approach could explore interactions between multiple factors — such as parents' education level, parental occupation, and residential area — to uncover deeper insights into student performance.

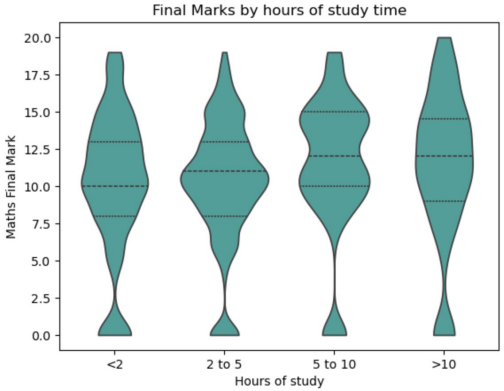
Additionally, developing a predictive model using these socioeconomic factors could help forecast student outcomes and support targeted interventions.

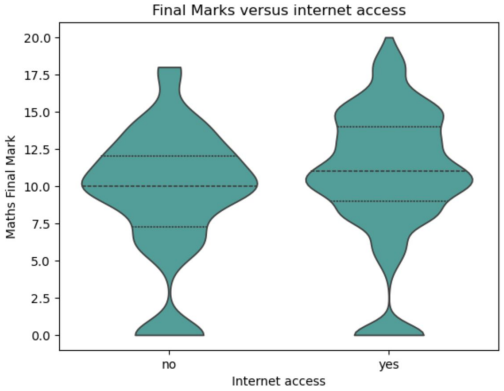


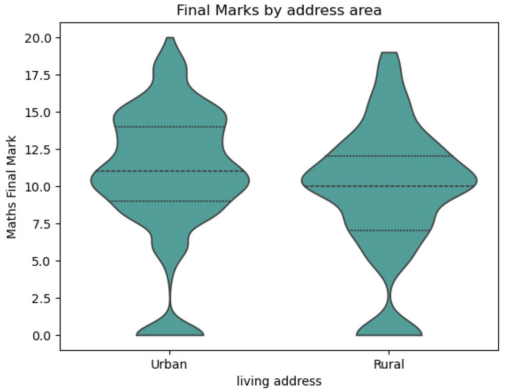


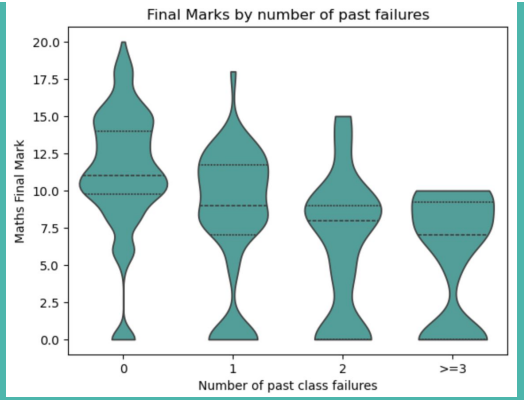


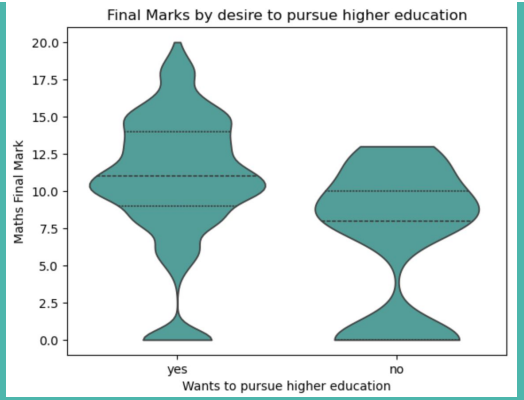
Appendix: Supporting Visuals











Appendix: Supporting Visuals