Student Performance Analysis

Edgardo Caceres

26 october 2025

Introduction

This report analyzes the "tudent-mat.csv" dataset to understand factors affecting student performance. Part A: provides an Exploratory Data Analysis (EDA) with visualizations. Part B: builds a decision tree to predict if a student will pass (get a final grade of 10 or higher).

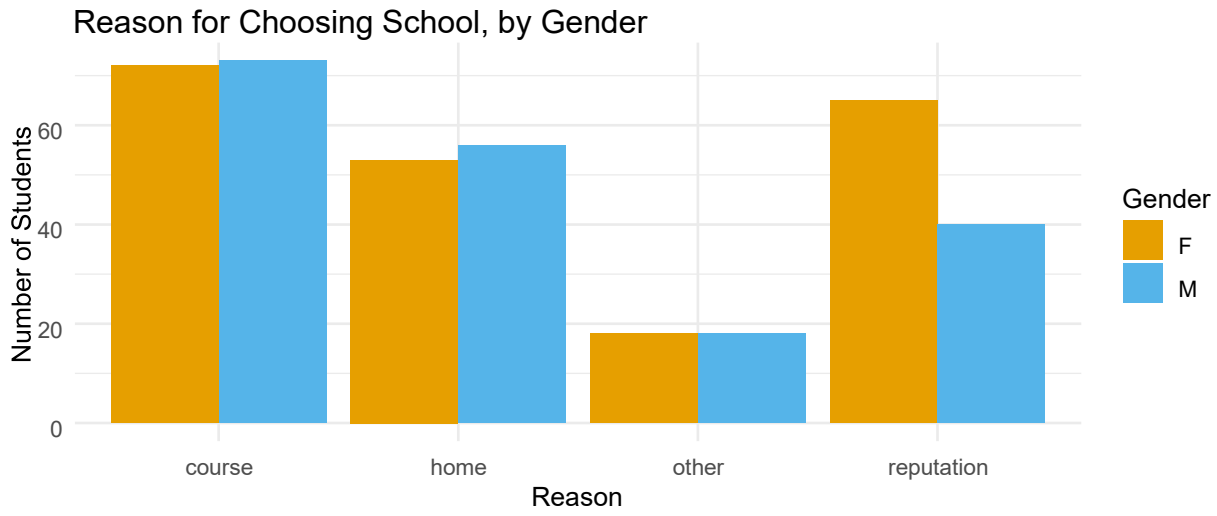
Part A: Exploratory Data Analysis (EDA)

First, we need to load the data. The CSV file uses semicolons (;) as separators.

Now, we will ask and answer questions using visualizations.

1. Bar Chart: Reason for Choosing School

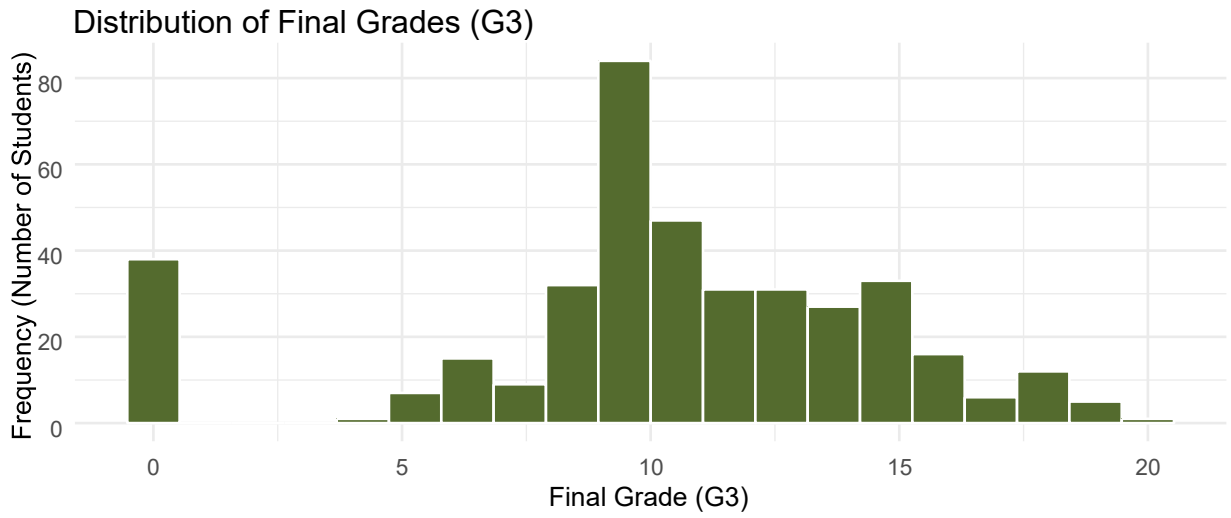
Question: What are the most common reasons students give for choosing their school, and how does this differ by gender?



Interpretation: The bar chart shows the frequency of each reason, split by gender. "course" remains the most popular reason for both females (F) and males (M). "home" and "reputation" are also nearly equal in popularity among females and males. Although preferences are generally similar, we see that "Reputation" is slightly more common among females.

2. Histogram: Distribution of Final Grades (G3)

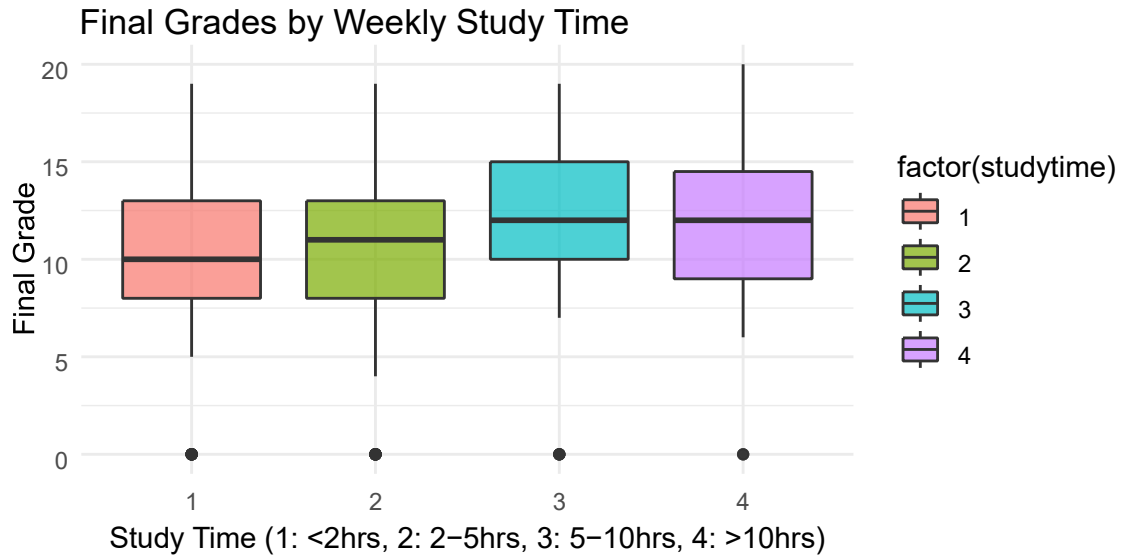
Question: What is the distribution of students' final grades (G3)?



Interpretation: This histogram shows that many students have final grades between 8 and 14. There is a large peak of students who scored 0. This might be missing data, or it could represent students who failed or dropped the course. The distribution is not perfectly normal.

3. Boxplot: Study time vs Final Grades

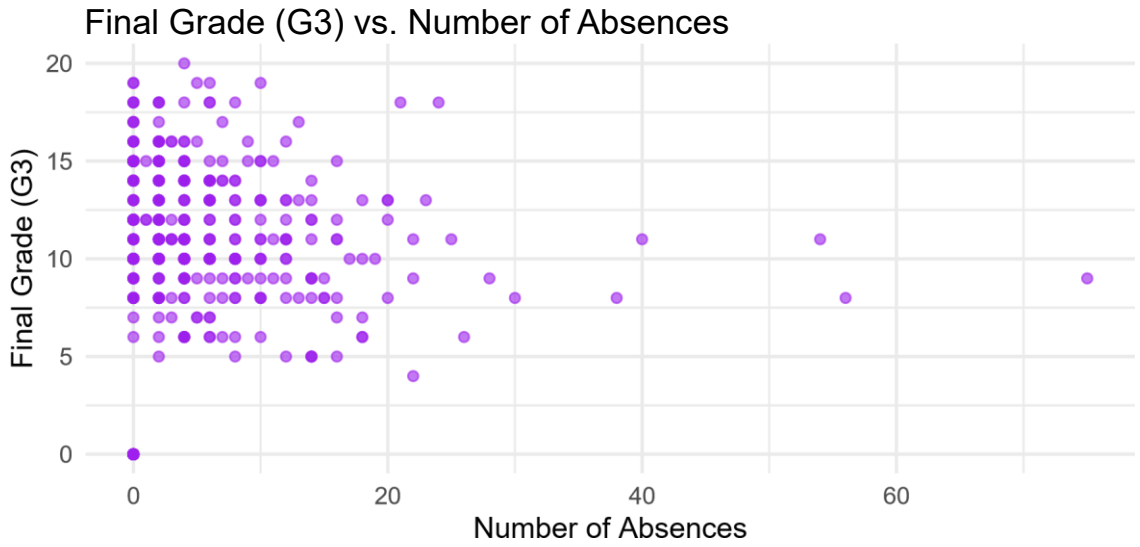
Question: How does study time affect final grades?



Interpretation: There is a positive relationship between study time and final grades. Students who study more than 10 hours per week (category 4) have the highest median grades, while those studying less than 2 hours (category 1) show lower performance. However, the relationship is not perfectly linear, suggesting that other factors also play important roles.

4. Scatter Plot: Absences vs. Final Grade

Question: Is there a relationship between the number of school absences and the final grade (G3)?



Interpretation: The scatter plot shows that most students have relatively few absences (less than 20). It does not show a very strong, clear relationship. However, it seems that students with a very high number of absences (more than 40) tend to have lower grades. Many students with 0 absences have a wide range of grades.

5. Descriptive Statistics: Summary of Study Time and Grades

Question: What are the summary statistics for the weekly study time "studytime" and the grades ("G1", "G2", "G3")?

```
## [1] "Summary of Study Time"

##
##   1   2   3   4
## 105 198 65 27

## [1] "Summary of G1, G2, G3 (First, Second, Thirds Period Grade
respectively)"
##    Min. 1st Qu. Median  Mean 3rd Qu.  Max.
##    3.00    8.00  11.00  10.91  13.00  19.00

##    Min. 1st Qu. Median  Mean 3rd Qu.  Max.
##    0.00    9.00  11.00  10.71  13.00  19.00

##    Min. 1st Qu. Median  Mean 3rd Qu.  Max.
##    0.00    8.00  11.00  10.42  14.00  20.00
```

Interpretation:

Study Time: The "Summary of Study Time" output above shows the count for each study time category (where 1 = <2 hrs, 2 = 2-5 hrs, 3 = 5-10 hrs, 4 = >10 hrs). We can see from the table that category "2" (2-5 hours/week) is the most common, followed by category '1' (<2 hours/week). Very few students are in category "4" (>10 hours/week).

Grades (G1, G2, G3): The "Summary of G1, G2, G3" outputs show the key statistics for all three grade periods. We can see that the average (Mean) and median (Median) grades are very similar across all three periods, staying around

10-11. This suggests that student performance is relatively consistent throughout the year. We can also see the full range of grades (Min to Max).

Part B: Decision Tree

Now, we will build a decision tree to predict whether a student scores 10 or higher in "G3".

1. Data Preparation

We need to create a new column, "pass", which will be "Yes" if "G3 >= 10" and "No" otherwise. This will be our target variable. We must convert it to a factor for classification.

2. Building the Tree

We will use the "rpart" function to build the tree. We will select a all variables to predict "pass" (Excluding G3).

3. Cross-Validation Check

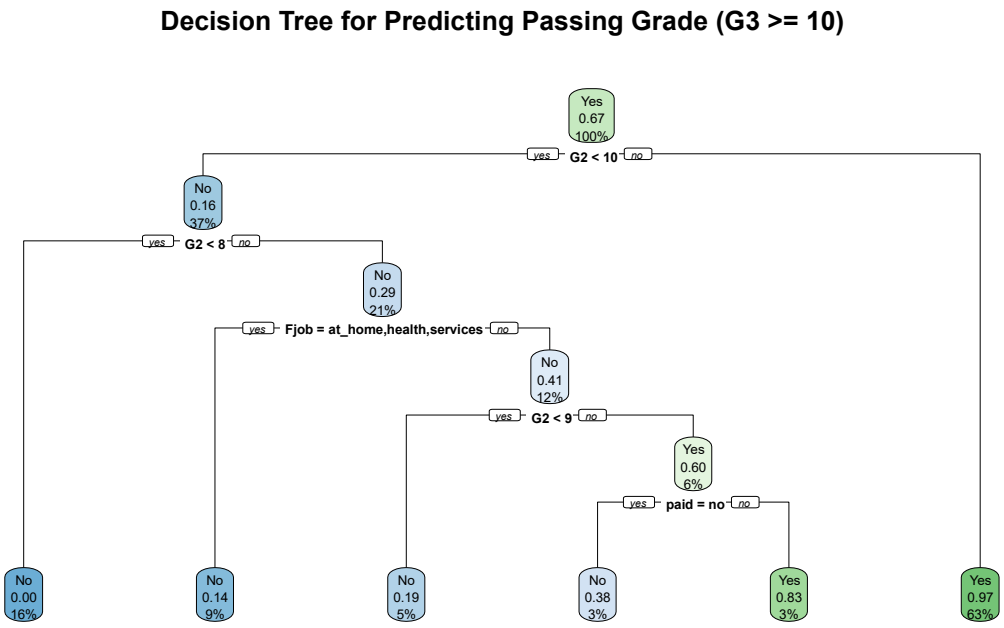
Let's check the cross-validation (CV) results, with "rpart".

```
##
## Classification tree:
## rpart(formula = pass ~ . - G3, data = student_data, method =
"class") ##
## Variables actually used in tree construction:
## [1] Fjob G2 paid
##
## Root node error: 130/395 = 0.32911
##
## n= 395
##
## CP nsplit rel error xerror xstd ## 1
0.753846 0 1.00000 1.00000 0.071838
## 2 0.012821    1 0.24615 0.24615 0.041714
## 3 0.010000    5 0.18462 0.27692 0.044000
```

Interpretation of CV: This table shows the result of cross-validation, which helps find the best tree size. xerror (Cross-Validation Error): most important metric. It estimates the model's error on new, unseen data. Lower is better. Row 2 (nsplit = 1): A tree with only one split has the lowest error (xerror = 0.24615). Row 3 (nsplit = 5): When the tree becomes more complex (5 splits), the error increases (xerror = 0.31538).

Conclusion: The optimal, most reliable model is the simple tree with just one split, as it performs best on new data. The extra variables (Fjob, paid) used in the complex tree are likely just noise.

4. Plotting and Interpreting the Tree



5. Tree Interpretation

G2 (Second Period Grade) is the Most Important Factor: The tree's first split is based on G2 < 10. If G2 is 10 or higher (the "no" branch on the right): The model predicts Pass (Yes). This branch accounts for 63% of all students (value of 0.97, meaning 97% of this group passed).

Students with G2 < 10 are At Risk: If G2 is less than 10 (the "yes" branch on the left, 37% of students), the prediction is generally Fail (No).

"At-Risk" Group:

If G2 < 8: The prediction is Fail (No). This group is 16% of the total students and has a pass rate of 0% (value 0.00).

If G2 is 8 or 9 (i.e., G2 < 10 is true, but G2 < 8 is false): The model looks at Fjob (father's job) and paid (extra paid classes).

If Fjob is "at_home," "health," or "services," the prediction is Fail (No) (9% of total).

If Fjob is not one of those and G2 is 8 (i.e., G2 < 9), the prediction is Fail (No) (5% of total).

If Fjob is not one of those and G2 is 9 (i.e., G2 < 9 is false), the model makes one final split based on paid classes:

paid = no (did not pay): Predict Fail (No) (3% of total). paid = yes (did pay): Predict Pass (Yes) (3% of total).