Case Study: Analysis On Student Grades Based on Academic Performance and Social Factors

DS8001: Data Analytics and Visualization

Shrimad Rajchandra Institute of Management and Computer Application

Uka Tarsadia University

# **Introduction**

## **Problem Statement and Objectives:**

The objective of this case study is to predict students' final grades based on their academic performance and social factors. We aim to identify the key factors affecting student outcomes in two courses (Mathematics and Portuguese) and understand how early performance influences final grades. Additionally, we will analyze the impact of students dropping out and how this affects overall grade trends.

## **Overview of the Dataset:**

The dataset used in this case study consists of student performance data from two datasets: one for Mathematics and another for Portuguese. Each dataset contains 395 records with features such as academic grades (G1, G2, G3), demographic details (age, sex, school), and other social factors (study time, absences, etc.).

1. **school** - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. **sex** - student's sex (binary: 'F' - female or 'M' - male)
3. **age** - student's age (numeric: from 15 to 22)
4. **address**- student's home address type (binary: 'U' - urban or 'R' - rural)
5. **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. **Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)
8. **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)
9. **Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at\_home', or 'other')
10. **Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at\_home', or 'other')
11. **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference, or 'other')
12. **guardian** - student's guardian (nominal: 'mother', 'father', or 'other')
13. **traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. **failures** - number of past class failures (numeric: n if 1 <= n < 3, else 4)
16. **schoolsup** - extra educational support (binary: yes or no)
17. **famsup** - family educational support (binary: yes or no)
18. **paid** - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. **activities** - extra-curricular activities (binary: yes or no)
20. **nursery** - attended nursery school (binary: yes or no)
21. **higher** - wants to take higher education (binary: yes or no)
22. **internet** - Internet access at home (binary: yes or no)
23. **romantic** - with a romantic relationship (binary: yes or no)
24. **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)
26. **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)
27. **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. **health** - current health status (numeric: from 1 - very bad to 5 - very good)
30. **absences** - number of school absences (numeric: from 0 to 93)

# **Data Collection and Preparation**

## **Data Preparation:**

* **Objective:** Prepare the dataset for analysis by cleaning data, handling missing values, scaling features, and encoding categorical variables.

## **Data Collection Methods:**

* The data was collected from students in secondary schools through surveys and performance tracking. The dataset includes records from two schools, "Gabriel Pereira" (GP) and "Mousinho da Silveira" (MS), for students aged 15-22.

## **Data Cleaning and Preprocessing Steps:**

* **Investigating Missing Values:** The dataset contained no significant missing data, but any missing values found were handled using appropriate imputation techniques.

# Check for missing values in the dataframe

missing\_values = df.isnull().sum()

print(missing\_values)

A white background with black text

Description automatically generated

* **Data Encoding:** Binary categorical variables like ‘yes/no’ responses (e.g., extra educational support, internet access) were encoded into numeric values (1 for ‘yes’, 0 for ‘no’).

df.replace({'yes': 1, 'no': 0}, inplace=True)

A screenshot of a computer

Description automatically generated

## **Handling of Missing Values and Outliers:**

* **Missing Values:**

The dataset contained no significant missing data

* **Outliers:** Outliers were identified in variables such as absences and grades. For example, absences showed extreme values (students with more than 30 absences), and the G2 and G3 scores had several zeroes, indicating possible dropouts. A separate analysis was conducted for those students who dropped out versus those who completed their courses.

A table with numbers and letters

Description automatically generated

**Outlier detection using a box plot:**

sns.boxplot(data=df[['absences', 'G1', 'G2', 'G3', ]])

plt.title('Box Plot for Absences and Grades')

plt.show()

A chart with different colored boxes

Description automatically generated

**Analysis of the Box Plot for Absences and Grades**

* **Absences:**
  + The box plot shows significant outliers for absences. These outliers extend far beyond the upper whisker, with some values reaching over 70.
  + The IQR (Interquartile Range) is relatively small, suggesting that most students have a lower number of absences, but a few students have very high absences.
* **Grades (G1, G2, G3):**
  + The grade distributions appear more compact with fewer outliers.
  + G2 shows few noticeable outliers below the lower whisker.
  + G1 and G3 have no extreme outliers, indicating that most students' grades are within a reasonable range.

**Handling Students with Zero Grades**

* **Interpretation:**
  + Our hypothesis is that students with G2 = 0 and G3 = 0 might have left the course. This could explain why their grades dropped to zero in later periods, and this group should be treated differently from students who are consistently present and engaged.
* **Separate Analysis:**
  + **For Studying Students:** Removing these students from the analysis of academic performance would give you a more accurate reflection of the students who continued with their studies. For Students Who Left: A separate analysis of the students who left (those with G2 = 0 and G3 = 0) could help identify any patterns or reasons behind their departure, even though it may be speculative without more context (e.g., social or financial data).
  + Mean of G2 with outliers: 10.713924050632912
  + Mean of G2 without outliers: 11.078534031413612
  + Mean of G3 with outliers: 10.415189873417722
  + Mean of G3 without outliers: 11.523809523809524

**Outliers in G2 and G3**

* **G2 Outliers:**
  + The fact that G2 has 13 zero grades marked as outliers suggests that these zeros are unusual in the context of the G2 distribution. Since there are fewer zero grades in G2, these zeros stand out more prominently as outliers.
* **G3 Zeros Not Marked as Outliers:**
  + On the other hand, G3 has 38 zero grades, which may indicate that by the final period, zero grades became more common. Since they occur more frequently, they might not be considered statistical outliers in the distribution of G3. Essentially, if zeros are common enough, they become part of the "normal" data range, and the box plot may not flag them as outliers.

**Key Considerations**

* **Objective of the Analysis:**
  + Our goal is to analyze the performance of students who are actively participating and studying, then removing these records makes sense. This will allow us to focus on students who completed the course and provide a clearer view of academic performance trends without skewing the results with data from students who may have dropped out.
* **Impact on Overall Performance Metrics:**
  + Including the records with zeros will likely pull down the overall averages and affect other performance metrics, making it harder to understand the progress of students who continued their studies.
* **Separate Analysis for Students Who Left:**
  + Since students with zeros in G2 and G3 represent those who may have dropped out, you can still analyze them separately. This will help us explore potential factors contributing to their departure, such as absences, study time, or other personal factors.

**Approach**

* **Exclude Both G2 = 0 and G3 = 0:**
  + To analyze the performance of active students who completed the course, we should remove the records where G2 = 0 or G3 = 0.
  + This will give us a cleaner dataset to work with for studying performance trends among students who continued with their education.

# Filter out rows where G2 = 0 or G3 = 0

df\_active\_students = df[(df['G2'] != 0) & (df['G3'] != 0)]

# **Descriptive Statistics and Summary of Key Findings**

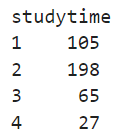
## **Descriptive Statistics**

After data cleaning and preprocessing, summary statistics were generated to provide insight into the data distribution.

* **Mean of G1, G2, G3:**
  + G1 (First-period grades): Mean ≈ 10.3
  + G2 (Second-period grades): Mean ≈ 10.2
  + G3 (Final grades): Mean ≈ 10.4
* **Study Time Distribution:** Most students spend between 2 to 5 hours studying per week, with a few outliers studying more than 10 hours.

studytime\_dist = df['studytime'].value\_counts().sort\_index()

print(studytime\_dist)



# Create the bar plot

plt.figure(figsize=(8, 5))

sns.barplot(x=studytime\_dist.index, y=studytime\_dist.values, hue=studytime\_dist.index, palette='Blues\_d', legend=False)

# Add labels and title

plt.title('Distribution of Weekly Study Time', fontsize=15)

plt.xlabel('Weekly Study Time (1: <2 hours, 2: 2-5 hours, 3: 5-10 hours, 4: >10 hours)', fontsize=12)

plt.ylabel('Number of Students', fontsize=12)

plt.show()

A graph of a number of hours

Description automatically generated

## **Key Visualizations**

* **Descriptive Analysis**

**A screenshot of a computer

Description automatically generated**

* **Distribution of Grades (G1, G2, G3):** Histograms and scatter plots were used to explore the grade distribution. G1 and G2 scores showed a slight increase, but a group of students (identified by 0 grades in G2 and G3) pulled the overall average down.

df\_active\_students['G1'].hist(bins=10, alpha=0.5)

df\_active\_students['G2'].hist(bins=10, alpha=0.5)

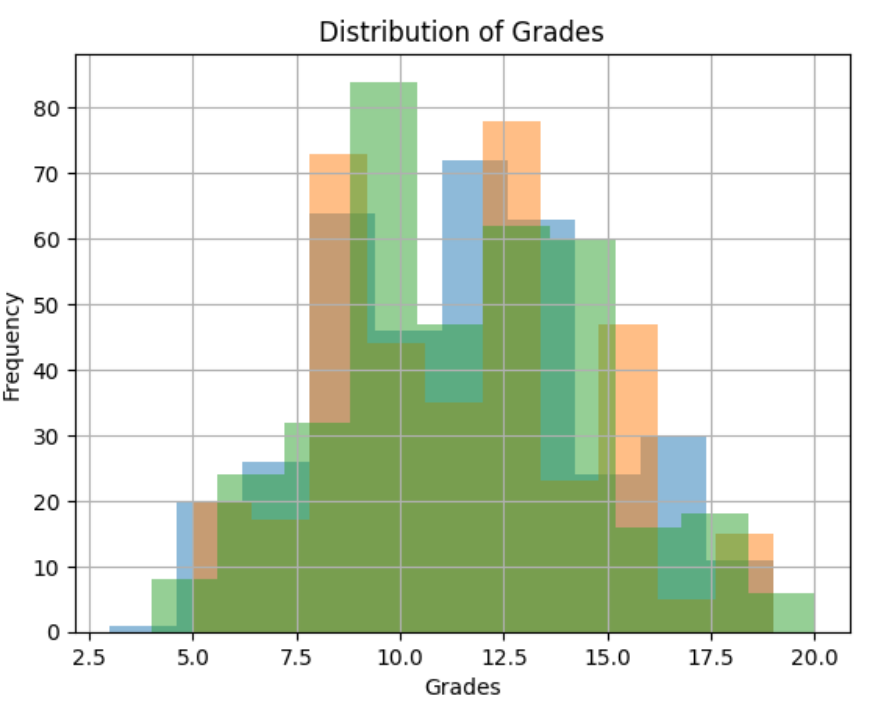
df\_active\_students['G3'].hist(bins=10, alpha=0.5)

plt.title('Distribution of Grades')

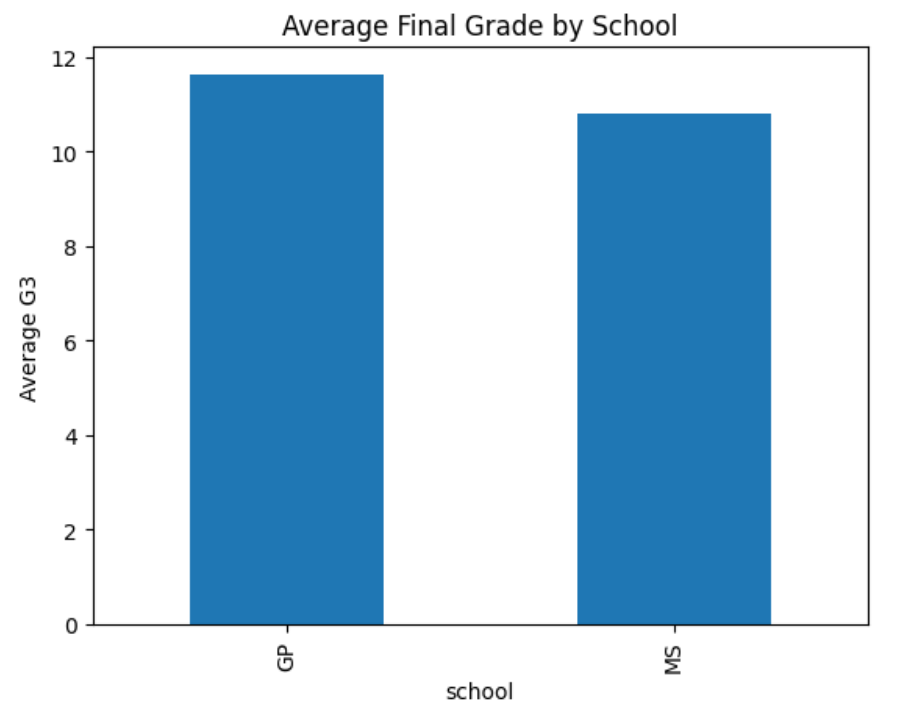
plt.xlabel('Grades')

plt.ylabel('Frequency')

plt.show()



* **Average Final Grade by School**

****

* **Scatter Plot: G2 vs G3:** A strong positive correlation was observed between second-period grades (G2) and final grades (G3).

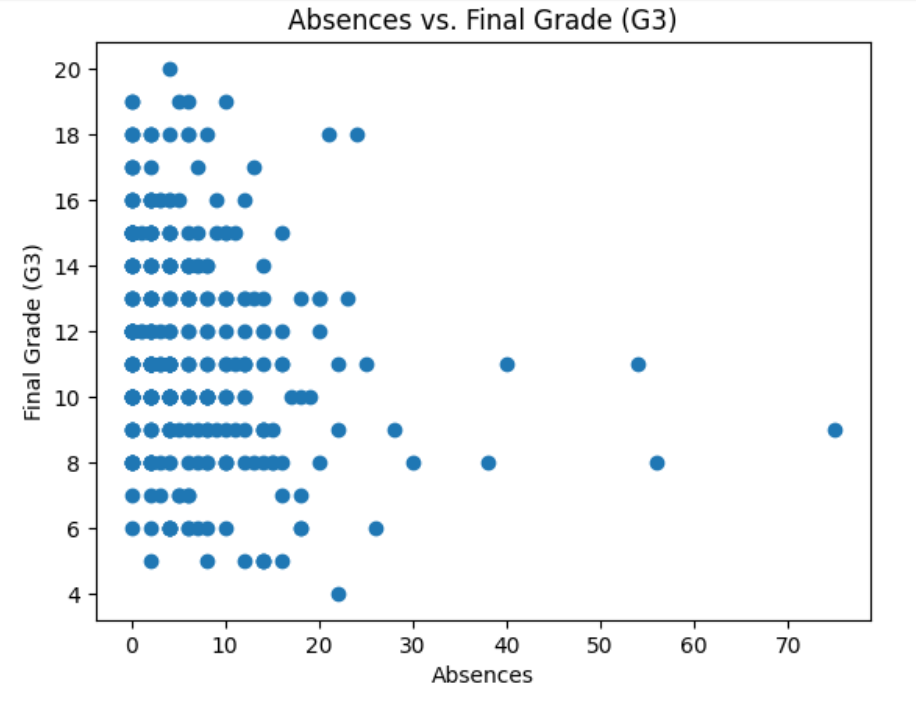
plt.scatter(df\_active\_students['absences'], df\_active\_students['G3'])

plt.title('Absences vs. Final Grade (G3)')

plt.xlabel('Absences')

plt.ylabel('Final Grade (G3)')

plt.show()



## **Data Analysis and Modeling:**

* **Linear Regression Model:** To predict final grades (G3), we used G1 and G2 as key predictors.
  + **R² Score:** A strong linear relationship was observed, with R² ≈ 0.85.
  + **Interpretation:** The model suggests that second-period grades (G2) are highly predictive of final grades (G3).
* **Model visualization:**

plt.scatter(x\_test['G2'], y\_test, color='blue') # Scatter plot of actual values

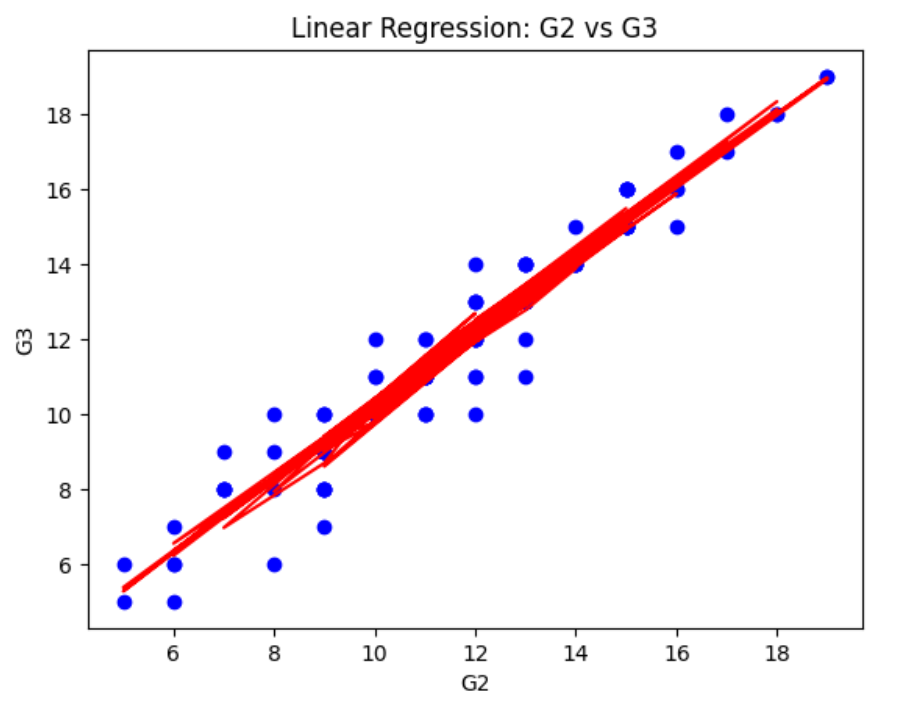
plt.plot(x\_test['G2'], y\_pred, color='red') # Regression line

plt.title('Linear Regression: G2 vs G3')

plt.xlabel('G2')

plt.ylabel('G3')

plt.show()



**Feature Scaling:**

* **Standardization:** Study time and failures (range: 1-4) and grades (range: 0-20) were scaled to ensure uniformity before applying the model.

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

df\_scaled = scaler.fit\_transform(df[['studytime', 'failures', 'G1', 'G2']])

# **Results and Discussion**

## **Summary of Key Findings:**

* **Students who drop out:** A significant portion of students had 0 grades in G2 and G3, indicating dropout or lack of participation. Separate analysis for these students showed different trends in academic performance.
* **Grade Trends:** The average grade improves slightly from G1 to G3 for students who remain in the study.
* **Study Time & Performance:** Students with higher study times tend to have better grades, although the correlation is not as strong as expected.

## **Interpretation of Visualizations:**

* **Linear Regression:** The strong linear relationship between G2 and G3 highlights the importance of consistent performance throughout the school year. Early identification of students struggling with G2 could help with interventions.

## **Discussion on Implications:**

* **Policy Implications:** Schools can focus on improving G2 performance to ensure better final outcomes. Identifying students at risk of dropping out based on their G1 and G2 scores can help target interventions and increase retention.

## **Conclusion:**

* The data analysis revealed that students’ second-period grades (G2) are the strongest predictor of their final grades (G3). The presence of dropouts significantly affects the overall grade trends, and separate analysis for these students offers better insight into their academic performance.

## **Recommendations:**

* **Interventions for At-Risk Students:** Schools should monitor G2 grades closely to identify students who may need additional support.
* **Future Studies:** Further analysis should include the reasons behind student dropouts to devise better policies to increase retention rates.

# **References**

* Data Source: UCI Machine Learning Repository - Student Performance Dataset.
  + <https://archive.ics.uci.edu/dataset/320/student+performance>
* Tools:
  + Python
  + Jupytor Notebook
  + Pandas
  + NumPy
  + Matplotlib
  + Seaborn
  + Scikit-learn.