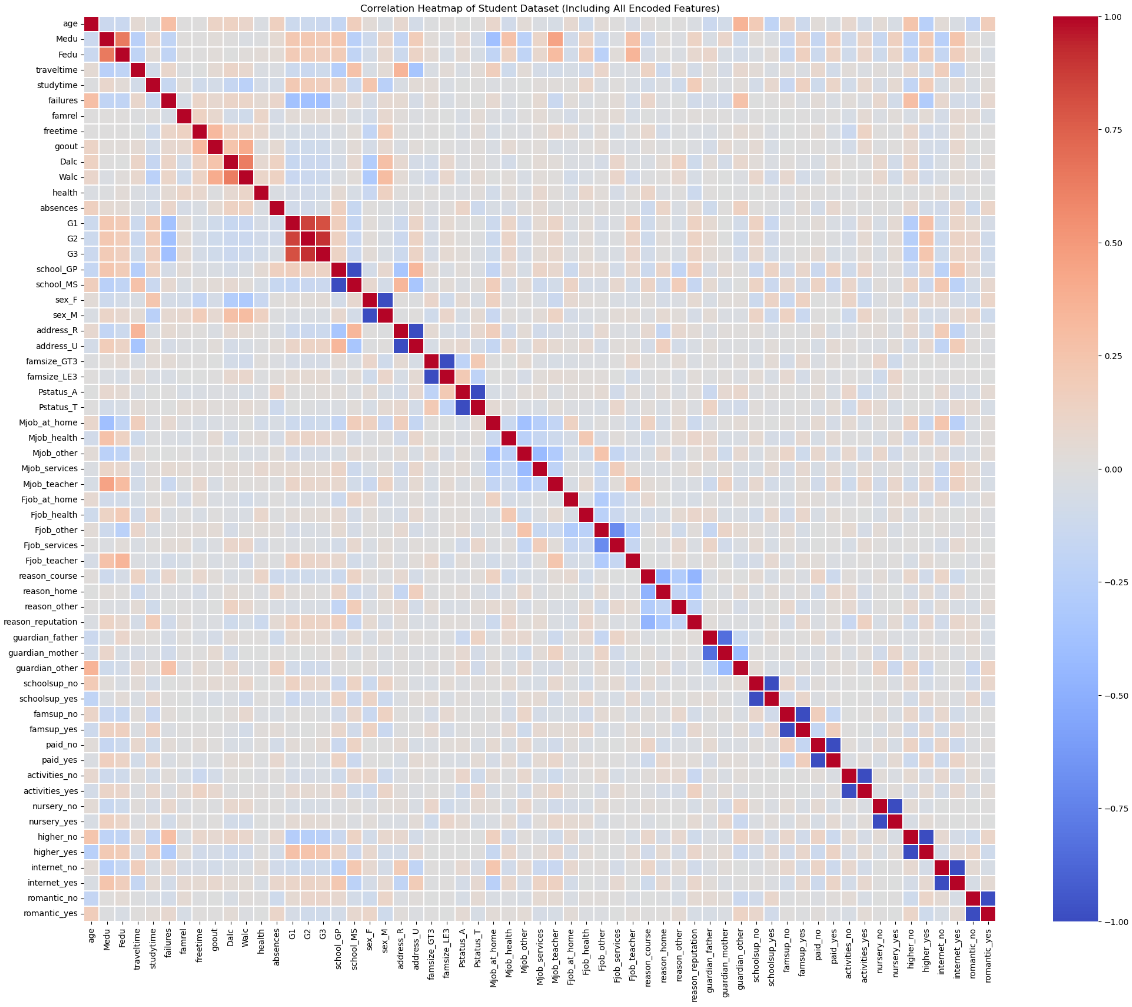
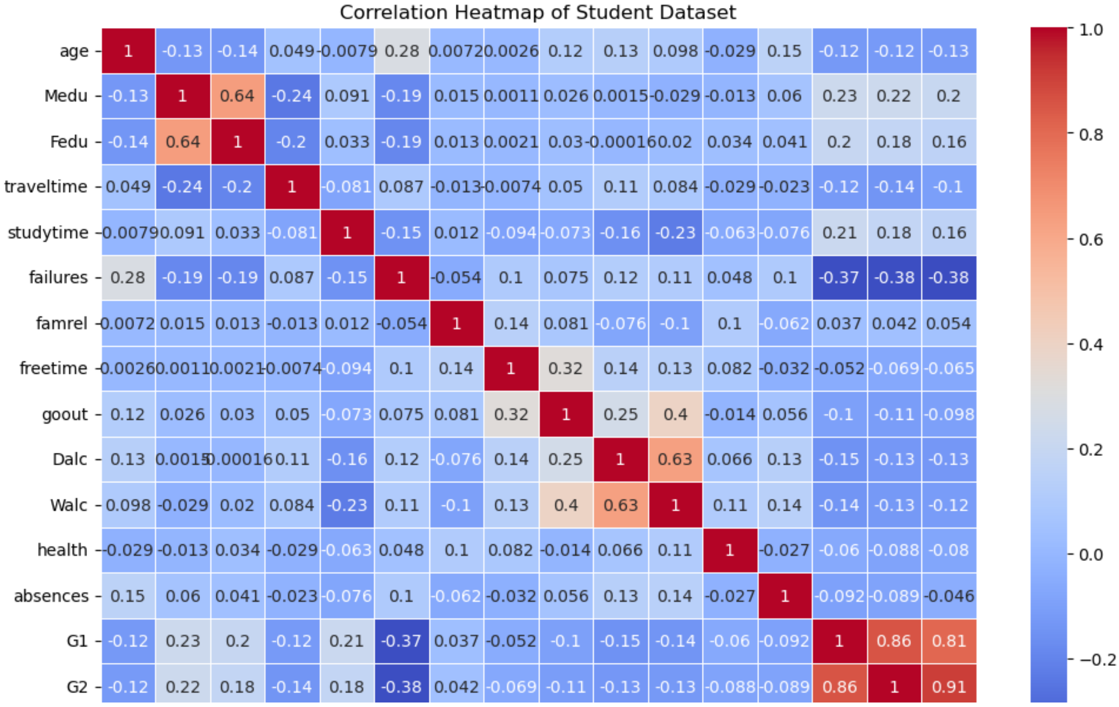
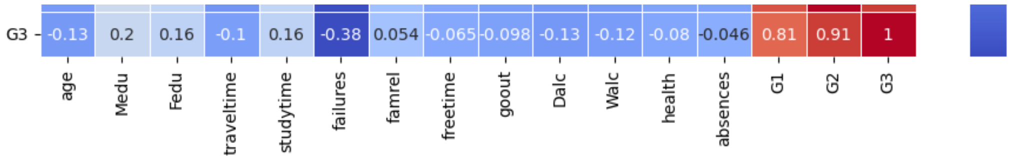
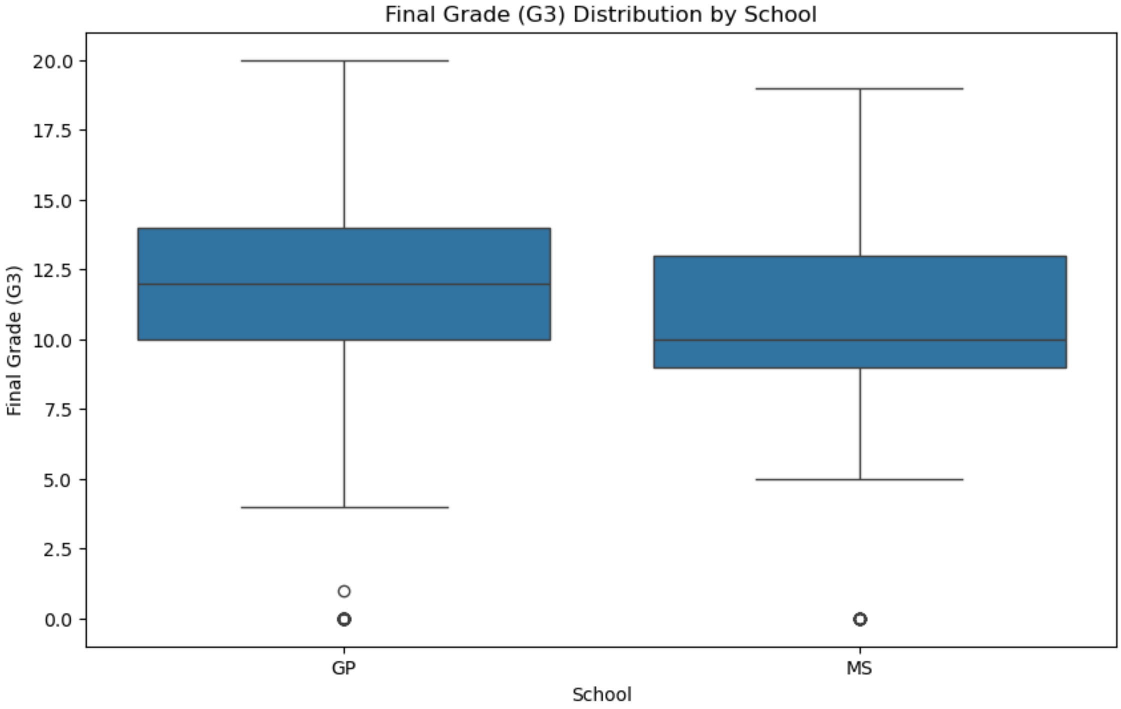
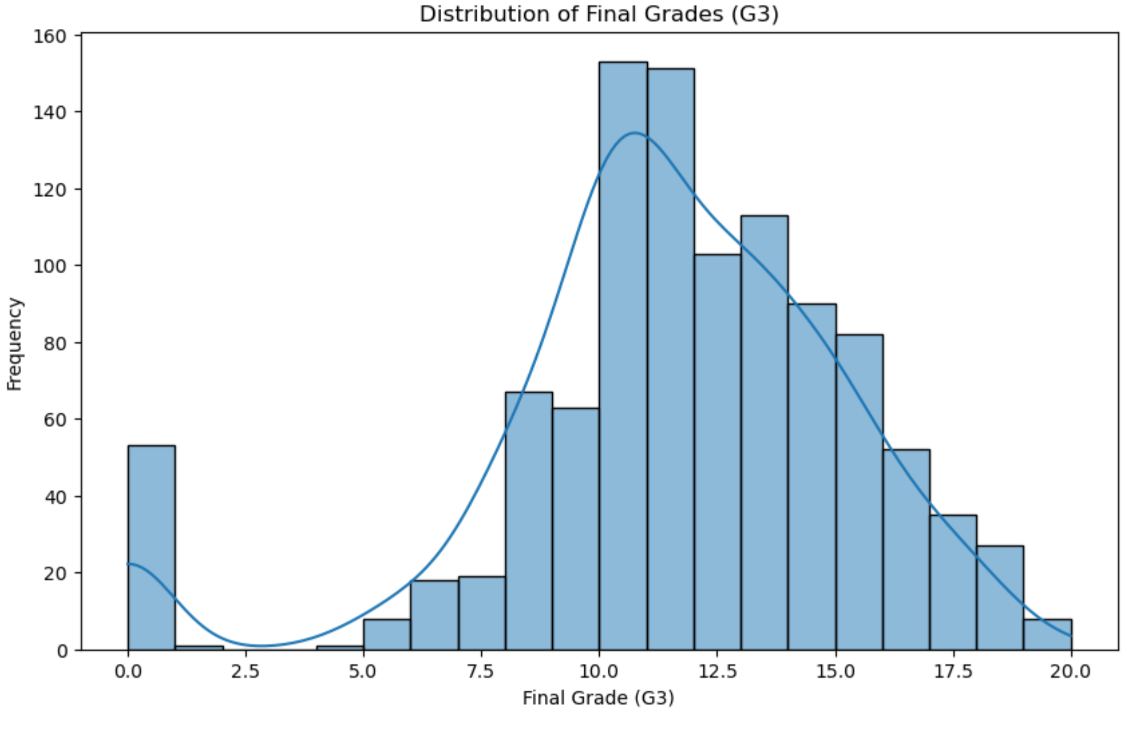
Exploratory Data Analysis


Correlation Heatmap of Student Dataset (Including All Encoded Features)
age
Medu
Fedu
traveltime
studytime
failures
famrel
freetime
goout
Dalc
Walc
health
absences
G1
G2
G3
school_GP
school_MS
sex_F
sex_M
address_R
address_U
famsize_GT3
famsize_LE3
Pstatus_A
Pstatus_T
Mjob_at_home
Mjob_health
Mjob_other
Mjob_services
Mjob_teacher
Fjob_at_home
Fjob_health
Fjob_other
Fjob_services
Fjob_teacher
reason_course
reason_home
reason_other
reason_reputation
guardian_father
guardian_mother
guardian_other
schoolsup_no
schoolsup_yes
famsup_no
famsup_yes
paid_no
paid_yes
activities_no
activities_yes
nursery_no
nursery_yes
higher_no
higher_yes
internet_no
internet_yes
romantic_no
romantic_yes
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation Heatmap of Student Dataset
age 1 -0.13 -0.14 0.049 -0.0079 0.28 0.0072 0.0026 0.12 0.13 0.098 -0.029 0.15 -0.12 -0.12 -0.13
Medu -0.13 1 0.64 -0.24 0.091 -0.19 0.015 0.0011 0.026 0.0015 -0.029 -0.013 0.06 0.23 0.22 0.2
Fedu -0.14 0.64 1 -0.2 0.033 -0.19 0.013 0.0021 0.03 -0.00016 0.02 0.034 0.041 0.2 0.18 0.16
traveltime 0.049 -0.24 -0.2 1 -0.081 0.087 -0.013 -0.0074 0.05 0.11 0.084 -0.029 -0.023 -0.12 -0.14 -0.1
studytime -0.0079 0.091 0.033 -0.081 1 -0.15 0.012 -0.094 -0.073 -0.16 -0.23 -0.063 -0.076 0.21 0.18 0.16
failures 0.28 -0.19 -0.19 0.087 -0.15 1 -0.054 0.1 0.075 0.12 0.11 0.048 0.1 -0.37 -0.38 -0.38
famrel -0.0072 0.015 0.013 -0.013 0.012 -0.054 1 0.14 0.081 -0.076 -0.1 0.1 -0.062 0.037 0.042 0.054
freetime -0.0026 0.0011 0.0021 -0.0074 -0.094 0.1 0.14 1 0.32 0.14 0.13 0.082 -0.032 -0.052 -0.069 -0.065
goout 0.12 0.026 0.03 0.05 -0.073 0.075 0.081 0.32 1 0.25 0.4 -0.014 0.056 -0.1 -0.11 -0.098
Dalc 0.13 0.0015 0.00016 0.11 -0.16 0.12 -0.076 0.14 0.25 1 0.63 0.066 0.13 -0.15 -0.13 -0.13
Walc 0.098 -0.029 0.02 0.084 -0.23 0.11 -0.1 0.13 0.4 0.63 1 0.11 0.14 -0.14 -0.13 -0.12
health -0.029 -0.013 0.034 -0.029 -0.063 0.048 0.1 0.082 -0.014 0.066 0.11 1 -0.027 -0.06 -0.088 -0.08
absences 0.15 0.06 0.041 -0.023 -0.076 0.1 -0.062 -0.032 0.056 0.13 0.14 -0.027 1 -0.092 -0.089 -0.046
G1 -0.12 0.23 0.2 -0.12 0.21 -0.37 0.037 -0.052 -0.1 -0.15 -0.14 -0.06 -0.092 1 0.86 0.81
G2 -0.12 0.22 0.18 -0.14 0.18 -0.38 0.042 -0.069 -0.11 -0.13 -0.13 -0.088 -0.089 0.86 1 0.91
1.0
0.8
0.6
0.4
0.2
0.0
−0.2




Final Grade (G3) Distribution by School


Distribution of Final Grades (G3)

We used the student data from the UCI Machine Learning Repository to generate a heatmap and distribution plot of **G3**. As observed, **G1**, **G2**, and **G3** have a high correlation with each other. Additionally, there are correlations between the education levels of the mother and father. The box plot reveals that most values for **G3** lie between 10 and 15. By setting a lower and upper bound (around 5 to 20), we identified a significant number of outliers with a grade of 0. This is also reflected in the distribution, which shows a large number of outliers at 0 that deviate from the bell curve.

1. Distribution of Final Grades (G3)

The histogram of **G3** indicates a slightly left-skewed distribution. Most students score in the range of 10 to 15, with a notable peak around 12. Few students achieve scores below 5, representing outliers, likely corresponding to those with multiple prior failures. This distribution highlights that the majority of students maintain average performance, while only a small proportion achieve very high or very low grades.

2. Comparison of Final Grades by School

The boxplot comparing **G3** distributions between the two schools (GP and MS) reveals:

- Students from **GP** tend to have slightly higher median grades.
- The interquartile ranges (IQRs) of both schools overlap significantly, indicating similar overall performance.
- A few outliers are present in both schools, corresponding to very low-performing students.

This difference may suggest that factors like school resources or teaching quality slightly favor GP students, though further analysis is needed to confirm this.

3. Correlation Heatmaps

The correlation heatmaps offer insights into relationships among features and their connection to **G3**:

- **G3** strongly correlates with **G1** (0.81) and **G2** (0.91), as earlier grades directly contribute to final performance.
- Moderate positive correlations are observed with parental education levels (**Medu: 0.22**, **Fedu: 0.16**), reflecting the importance of an academically supportive environment at home.
- Negative correlations exist with **failures (-0.38)**, indicating that students with repeated academic failures perform poorly in their final grades.
- **goout** shows a weaker negative correlation (-0.10), suggesting that excessive social outings might slightly detract from academic focus.

4. Social and Family Factors

Failures vs. G3

A detailed analysis of **failures** reveals:

- Students with no prior failures have the highest median **G3** scores.
- A steep drop in performance is observed as the number of failures increases.

Goout vs. G3

While most students have moderate social activity levels:

- Those with higher **goout** scores exhibit greater variability in **G3**.
- Many students with frequent socialization perform below average, suggesting that excessive socialization might lead to academic distractions.

Romantic Relationships (romantic_yes)

Students in relationships show slightly lower **G3** medians:

- The difference is not substantial, implying potential distractions but not to a degree that severely impacts academic performance.

Interaction Between Features

The interaction between features, such as **Medu** and **failures**, demonstrates compounded effects on **G3**:

- Students with higher parental education but no failures consistently outperform others.
- This combination of supportive home environments and academic discipline is a critical factor for success.

Key Findings

The analysis highlights the critical role of:

- Parental education
- Past failures
- Social habits

Students with strong academic histories, supportive family backgrounds, and balanced social lives tend to excel. Outliers, such as students with multiple failures or very low grades, suggest areas where targeted interventions may significantly improve outcomes.

Visualizations such as histograms, boxplots, and heatmaps support these findings, providing actionable insights for feature selection and modeling.

Dataset Description

Feature selection is a critical step in the modeling process, as it determines which variables provide the most meaningful contributions to predicting the target variable, **G3 (final grade)**. The selection was guided by exploratory data analysis (EDA), statistical correlations, and domain knowledge. This section explores each selected feature in detail and justifies its inclusion in the predictive model.

The datasets used in this analysis contain student performance data for two subjects: mathematics and Portuguese language. Each dataset includes 33 features capturing demographic, social, and academic aspects of students. For this project, we combined relevant features from both datasets to understand the factors influencing the final grade (**G3**).

We selected the following features for our analysis based on their observed correlations with **G3**:

- **failures**: Number of past class failures.
- **Medu**: Mother’s education level (0 = none to 4 = higher education).
- **Fedu**: Father’s education level (0 = none to 4 = higher education).
- **higher_yes**: Whether the student plans to pursue higher education (binary).
- **age**: Student’s age in years.
- **goout**: Frequency of going out with friends (1 = low to 5 = high).
- **romantic_yes**: Whether the student is in a romantic relationship (binary).
- **travel_time**: Time it takes to travel to school (1 = <15 min to 4 = >60 min).

These features were chosen because they showed statistically significant correlations with **G3**, either positively or negatively. For example, higher parental education levels (**Medu**, **Fedu**) and planning for higher education (**higher_yes**) correlated positively, while frequent going out (**goout**) or multiple past failures (**failures**) correlated negatively.

1. Failures

- **Definition**: The number of past academic failures.
- **Correlation with G3**: Strong negative correlation (-0.38).
- **Importance**: Failures emerged as one of the most impactful features. Students with multiple failures consistently had lower grades, as shown in the boxplot analysis. This feature captures the cumulative effect of academic struggles over time, making it a significant predictor of **G3**. Including this variable ensures that the model accounts for students’ academic history and resilience.

2. Mother’s Education Level (Medu)

- **Definition**: Education level of the student’s mother, scaled from 0 (none) to 4 (higher education).
- **Correlation with G3**: Positive correlation (0.22).
- **Importance**: **Medu** reflects the home environment’s academic support and resources. A higher level of maternal education often corresponds to better academic outcomes, as mothers with advanced education may provide stronger guidance and encouragement. This feature is particularly significant in identifying students who benefit from enriched educational support at home.

3. Father’s Education Level (Fedu)

- **Definition**: Education level of the student’s father, scaled from 0 (none) to 4 (higher education).
- **Correlation with G3**: Positive correlation (0.16).
- **Importance**: Similar to **Medu**, **Fedu** represents the father’s role in shaping the student’s academic aspirations and performance. Although the correlation with **G3** is slightly weaker than **Medu**, the combined influence of both parents’ education levels amplifies their predictive power. Including **Fedu** ensures a holistic view of parental support.

4. Higher Education Aspiration (higher_yes)

- **Definition**: A binary variable indicating whether the student plans to pursue higher education.

- **Correlation with G3**: Positive correlation (0.23).
- **Importance**: This feature serves as a proxy for the student’s motivation and long-term goals. Students who aspire to higher education are more likely to perform better academically, reflecting their dedication and focus. Including this feature captures the motivational aspect of academic success, which cannot be directly measured through grades alone.

5. Age

- **Definition**: The student’s age in years.
- **Correlation with G3**: Weak negative correlation (-0.13).
- **Importance**: While the correlation is weak, age is an essential demographic feature. Older students in the dataset often had slightly lower grades, possibly due to repeating years or external responsibilities that distract from academics. Including age allows the model to account for subtle age-related trends and their impact on academic performance.

6. Going Out with Friends (goout)

- **Definition**: Frequency of going out with friends, scaled from 1 (rarely) to 5 (frequently).
- **Correlation with G3**: Moderate negative correlation (-0.18).
- **Importance**: This feature reflects social behavior and time management. Students with higher **goout** scores often had lower grades, suggesting that excessive socialization might detract from academic focus. Including **goout** captures the trade-off between social activities and academic performance, providing insights into students’ priorities and habits.

7. Romantic Relationship (romantic_yes)

- **Definition**: A binary variable indicating whether the student is in a romantic relationship.
- **Correlation with G3**: Weak negative correlation (-0.10).
- **Importance**: While the correlation is not strong, the presence of a romantic relationship may influence academic performance through emotional or time-related distractions. Including this feature ensures that the model accounts for relational factors that might subtly impact grades.

8. Travel Time (travel_time)

- **Definition**: The time it takes for the student to travel to school, scaled from 1 (<15 minutes) to 4 (>60 minutes).
- **Correlation with G3**: Weak negative correlation (-0.10).
- **Importance**: **Travel_time** is included as a proxy for the accessibility of educational resources. Students with longer commutes might experience fatigue or reduced study time, which could negatively impact their academic performance. While the effect is minimal, including this feature ensures that logistical barriers are considered in the model.

Feature Selection Justification

The selected features were chosen based on their correlations with **G3**, interpretability, and relevance to academic performance. Each feature provides unique insights:

- **failures** is a direct indicator of academic struggles, while **Medu** and **Fedu** capture parental support.
- **higher_yes** and **age** reflect student motivation and demographic trends.
- Behavioral variables like **goout** and **romantic_yes** highlight the influence of social factors on performance.
- **travel_time** ensures that logistical barriers are accounted for, even if their impact is small.

By including a diverse set of features, the model balances academic, behavioral, and demographic factors, providing a comprehensive approach to predicting **G3**. This selection was informed by statistical analysis and domain knowledge, ensuring that the features are both meaningful and actionable for educational interventions.

Proposed Methodology

We applied three regression models to predict **G3**: linear regression, polynomial regression, and decision tree regression. Each method was evaluated based on its **Mean Squared Error (MSE)** and **R² score**. Below are the results:

1. Linear Regression

- **Cross-Validated Mean Squared Error (MSE)**: 11.93
- **Mean Squared Error (MSE)**: 14.44
- **R²**: 0.066

Linear regression showed limited predictive power, with an **R²** close to zero, indicating it could not capture the complex relationships between the features and **G3**.

2. Polynomial Regression

Polynomial regression was tested to capture non-linear patterns in the data. However, this model performed worse than linear regression, likely due to overfitting and noise in the data. Detailed results are omitted for brevity.

3. Decision Tree Regression

Depth	MSE	R²	Regularized Risk
3	4.70	0.77	5.50
4	6.20	0.70	7.80
5	6.09	0.70	9.18
6	4.27	0.79	9.17
7	4.30	0.79	11.99

The decision tree model at depths 3 and 6 performed best, achieving an **R²** of approximately **0.79**. This indicates that decision trees can capture non-linear relationships more effectively than the other models.

Key Insights

The decision tree model emerged as the most effective for predicting **G3**. Key factors influencing predictions include:

1. **Parental Education Levels**:
   - Higher parental education levels positively influenced **G3**, reflecting the supportive academic environment provided by educated parents.
2. **Failures**:
   - Multiple failures strongly reduced **G3** predictions, emphasizing the cumulative impact of academic struggles.
3. **Social and Behavioral Factors**:
   - Frequent social activities (**goout**) and romantic relationships (**romantic_yes**) were associated with lower **G3** values, possibly due to distractions from academic priorities.

The decision tree model's ability to highlight feature importance and capture non-linear patterns makes it a strong choice for predicting **G3**.