Student Performance Prediction Report

Author: Lucy Kinyanjui

ID: 672337

Course: DSA 1080VA FS2025

1. Dataset Summary

- Dataset: student-mat.csv (Kaggle Student Performance)
- Shape: 395 rows × 33 columns
- Target Variable: G3 (Final Math Grade, 0–20 scale)

Key features are grouped as follows:

- Demographics: age, sex, address, famsize, Pstatus
- Family: Medu (Mother's education), Fedu (Father's education), Mjob, Fjob
- Academic: studytime, failures, G1 (1st period grade), G2 (2nd period grade)
- Social: famrel, freetime, goout, Dalc/Walc (alcohol consumption), health, absences

The target distribution shows that G3 has a mean of approximately 10.4 with a standard deviation of about 4.6.

2. Data Exploration and Cleaning

Exploration steps included:

- Checking for missing values (none detected in the dataset).
- Identifying and removing outliers, especially extreme absences (e.g., more than about 75 days) and unrealistic grade values.
- Investigating multicollinearity using VIF (Variance Inflation Factor) and dropping features with VIF greater than 5 to reduce redundancy in predictors.

Outlier analysis flagged 1 student with unusually high age (22 years), 15 students with very high absences (e.g., 25–54 days), and 13 students with extreme G2 values (0). Depending on their impact, these observations were inspected and either retained or removed to reduce distortion in model training

3. Key Visualizations and Observations

In the notebook, several visualizations were created

Grade progression (G1 → G2 → G3):

A scatter plot of the average of G1 and G2 against G3 shows that as the average of G1 and G2 increases, G3 also increases almost linearly, with points forming a tight upward diagonal band. This visually confirms the very high correlation between earlier period grades (G1, G2) and the final grade G3 observed in the correlation analysis.

Study time vs. final grade:
This boxplot shows that students who report more study time (levels 1 to 4) tend to achieve slightly higher final grades (G3). As study time increases, the median G3 also rises, with the highest study-time group having a higher median grade than the lowest study-time group.

A Histogram grid

Shows that **G3** is roughly bell-shaped with most students scoring between 10 and 15, and relatively few very low scores, indicating a mix of performance levels but few outright failures. Absences are strongly right-skewed, with most students missing very few classes and only a small group with very high absence counts.

Study time is also right-skewed, as most students fall into the lower study time categories (1–2) and fewer report higher study levels (3–4), reflecting typical real-world behavior.

Age is concentrated between 15 and 18 years, with only a small number of older students.

Together, these patterns support the finding that higher G3 is associated with more favorable factors (such as higher parental education and fewer absences) and that accurate G3 prediction can help identify students at risk of scoring below 10 for early support.

4. Feature Engineering

The main features of engineering steps were:

1. **Categorical encoding:**
   One-hot encoding was applied to categorical variables such as school, sex, Mjob, Fjob, and other nominal features, producing additional binary columns.

2. **VIF-based filtering:**
   Variance Inflation Factor (VIF) was calculated for the features and variables with VIF greater than 5 were dropped to reduce multicollinearity among predictors.

3. **Scaling:**
   Numerical features such as age, absences, and studytime were standardised using a scaler so that models sensitive to feature scales could perform better.

4. **Feature selection:**
   Recursive Feature Elimination (RFE) was applied to select the top informative features.

In the analysis, I trained a Random Forest regression model to predict students' final grades (G3) based on several input features including prior grades (G1 and G2 average), social activities (goout), travel time to school (traveltime), academic failures, and school support (schoolsup_enc).

The model's feature importance results indicate the relative influence of each predictor in predicting final grades. The combined prior grades feature (G1_G2_avg) overwhelmingly dominates with an important score of 0.886, suggesting that past academic performance is the strongest indicator of final grade outcomes. Social activity levels (goout) hold the next largest influence, though much smaller (0.053), followed by the number of academic failures (0.026), travel time (0.021), and school support (0.014).

5. Modelling and Results

Random Forest Regression Performance:

RMSE: 2.355

MAE: 1.606

R^2: 0.730

Overall Performance: Good

I could not use classification metrics (accuracy_score, precision_score, recall_score, f1_score) with the Random Forest regression model. Those metrics are for classification (predicting

categories like "pass/fail") whereas our G3 grades are continuous numbers (0-20), so we need regression metrics.

6. Key Observations and Insights

- The model can predict the final grade G3 with reasonably high accuracy, with the Random Forest achieving an R^2 of about 0.73 and RMSE below 3.
- In practice, this means the model can predict final grades within about ±3 points for most students.
- Previous grades (G1 and especially G2) are by far the strongest predictors, confirming that continuous assessment strongly reflects final outcomes.
- Study time and parental education also play important roles and are actionable levers for improving performance.

8. Conclusion

The project successfully built and evaluated a machine learning models to predict students' final Math grades using demographic, academic, and social features. The Random Forest model achieved a good performance, with strong explanatory power driven mainly by prior grades, studytime, failures, and parental education. These results can help educators identify at risk students early and design targeted interventions to support improved academic outcomes.