Report

Major things are evident in the notebook

I will just walk through the process
I initially encoded some columns(these columns either had nan value or had mor than 2 string type classes) myself then the rest encoding was done by label encoder
Imputation

For imputation I found out the accuracy of predictive imputation by performing accuracy analysis using 80-20 split on the non nan values and lso found the accuracy of knn imputer then accordingly made choice how to impute values

Analysis of EDA

1. School and Final Grades Correlation

The bar chart reveals a noticeable difference in average final grades (G3) between schools. Students from GP consistently achieve higher average grades compared to MS, suggesting potential institutional disparities. Factors such as resource allocation, teaching quality, or socioeconomic demographics may contribute to this gap. Addressing these differences could help bridge the performance divide and promote equitable educational outcomes.

---

2. Weekday Alcohol Consumption and Health

The violin plot indicates that students with higher weekday alcohol consumption (Dalc ≥3) tend to report poorer health scores. While moderate drinkers show stable health distributions, excessive use correlates with wider variability and lower median health. This highlights the need for interventions targeting alcohol education and mental health support to mitigate long-term health risks.

---

3. Urban vs. Rural Travel Time Accessibility

The stacked bar chart demonstrates urban students predominantly experience shorter travel times (≤1 hour), whereas rural students face longer commutes (≥2 hours). This disparity underscores infrastructural challenges in rural areas, potentially affecting attendance and academic engagement. Policymakers could prioritize transportation improvements to enhance accessibility for rural communities.

---

4. Family Support and Academic Resilience

The grouped bar chart shows students with family support (famsup=yes) have fewer past failures compared to unsupported peers. This underscores the role of familial encouragement in fostering academic resilience. Schools might consider programs to engage families in student learning, particularly for at-risk groups lacking home support.

---

5. Higher Education Aspirations and Internet Access

The heatmap reveals a strong association between internet access and aspirations for higher education. Over 70% of students aiming for higher education have home internet, suggesting socioeconomic factors influence aspirations. Expanding digital access could democratize opportunities for students from disadvantaged backgrounds.

---

6. Maternal Education and Student Performance

The scatter plot illustrates a positive trend: higher maternal education (Medu) correlates with improved student grades (G3). Gender differences are minimal, though female students slightly outperform males at higher Medu levels. Parental education may indirectly boost academic success through enriched home learning environments.

---

7. Free Time, Absences, and Academic Outcomes

Students with moderate free time (2–3 on the scale) achieve the highest grades, while excessive leisure (≥4) correlates with lower G3 and more absences. This implies balanced time management is critical. Schools could promote structured extracurriculars to optimize student engagement and attendance.

---

8. Gender Disparity in Extracurricular Participation

The pie charts show a gender gap: 65% of non-participants in activities are male, while female participation is nearly equal (52.7% vs. 47.3%). Cultural norms or accessibility barriers may deter male students. Encouraging inclusive programs could foster broader participation and skill development across genders.

---

9. Romantic Relationships and Academic Performance

The box plot suggests no significant grade difference between students in romantic relationships and those single. However, single students exhibit slightly tighter grade distributions, while daters show more variability. Schools should focus on holistic support systems rather than attributing performance solely to relationship status.

---

10. Absences and Grades Across Schools

Regression plots indicate absences negatively impact grades in both schools, but the effect is steeper in MS. This implies consistent attendance is more critical for MS students, possibly due to less academic flexibility. Targeted attendance policies and early intervention could mitigate performance declines in vulnerable cohorts.

Romantic prediction analysis

Logistic Regression achieved an accuracy of around 56%, suggesting that while it captures some patterns, linear relationships alone are insufficient to model the complexity of romantic relationships among students. The model performs well in identifying students who are not in a relationship, with a recall of 0.80 for class 0, but it struggles significantly with class 1, where the recall drops to just 0.16. This means the model misses the vast majority of students who are actually in a relationship. Precision for class 1 is also low at 0.33, leading to an F1-score of only 0.22 for this class. The confusion matrix reveals that only 8 of the 49 students in relationships were correctly identified, highlighting a severe class imbalance issue. Overall, the model appears to underfit, and its performance is dominated by the majority class. Although it provides a starting point, Logistic Regression is not suitable for applications where identifying relationships accurately is essential. Introducing class weights or resampling techniques may help balance its predictions and improve recall for the minority class.

Random Forest, despite being a more complex model capable of capturing non-linear patterns, performed slightly worse with an accuracy of about 54.6%. Like Logistic Regression, it demonstrates a heavy bias toward class 0, with a recall of 0.84, but it performs very poorly on class 1 with a recall of just 0.06. This means the model is effectively unable to identify students in romantic relationships, with only 3 correct predictions out of 49 in the test set. The precision for class 1 is very low at 0.19, and the corresponding F1-score is just 0.09, indicating near-total failure in that class. The confusion matrix reinforces this imbalance, and suggests that the model may be overfitting to the majority class while failing to generalize for minority class patterns. Despite Random Forest's potential, it struggles in this setting without class balancing methods. Its poor performance may also suggest that behavioral or social predictors are subtle and not strongly differentiated in the current feature set. Applying class weights, feature engineering, or synthetic oversampling could help improve its performance.

XGBoost performed the best among the three models, achieving an accuracy of approximately 58.5%. It shows more balanced predictive behavior, with a class 1 recall of 0.41 and precision of 0.44, resulting in a relatively strong F1-score of 0.43 for students

in relationships. This is a notable improvement over the other models and suggests that XGBoost was better at extracting meaningful patterns from the data. The confusion matrix confirms this, showing 20 correct predictions for class 1, which is significantly higher than the other two models. XGBoost likely benefits from its ability to model interactions and subtle non-linear effects in the feature space. Though the overall accuracy is still modest, the model is more reliable for both classes, making it a better starting point for further optimization. With additional steps such as hyperparameter tuning, SMOTE or other balancing techniques, and perhaps more nuanced behavioral features, this model could be significantly improved. It demonstrates the most promise in effectively predicting romantic involvement among students based on the available data.