**Data Exploration**

**Basic Information and Data Structure**

* **Dataset Info:** The dataset consists of 649 entries and 33 columns, with no missing values.
* **Types of Variables:**
  + 16 numerical columns.
  + 17 categorical columns.

**Numerical Features Distribution**

The histograms and KDE plots show the distribution of each numerical feature:

1. **Age:** Most students are aged between 15 to 19, with peaks at ages 16 and 17.
2. **Medu (Mother's Education):** Education levels are fairly distributed, with the highest frequency at 2 (second level of education).
3. **Fedu (Father's Education):** Similar distribution to mother's education, but slightly lower frequencies at higher education levels.
4. **Failures:** Majority of students have no past failures, indicating a positively skewed distribution.
5. **Famrel (Family Relationship):** Most students rate their family relationship as good (4 or 5).
6. **Freetime:** Distributed mostly between 2 to 4, indicating moderate to high free time.
7. **Dalc (Workday Alcohol Consumption) & Walc (Weekend Alcohol Consumption):** Most students have low alcohol consumption, with peaks at 1.
8. **Health:** Health status is mostly average to good (3 to 5).
9. **Absences:** Absences are highly skewed with most students having few or no absences.
10. **G1, G2, G3 (Grades):** Grades are normally distributed with a peak around 10 to 14.

**Correlation Analysis**

The correlation matrix and heatmap indicate relationships between numerical features:

* **Strong Positive Correlations:**
  + G1 and G2 with G3 (grades from previous periods strongly predict the final grade).
  + Higher education aspiration (higher\_yes) shows a positive correlation with G3.
  + Studytime, Medu, and Fedu also show positive correlations with G3.
* **Strong Negative Correlations:**
  + Failures have a significant negative correlation with G3, indicating that more failures lead to lower grades.
  + Alcohol consumption (Dalc, Walc) negatively correlates with G3.
  + Age, freetime, and travel time show weak negative correlations with G3.

**Categorical Features Analysis**

Count plots for categorical features provide insights into their distribution and relationship with the target variable (G3):

* **School Type:** Students from school GP tend to have higher grades compared to MS.
* **Sex:** Both males and females show similar distributions of grades.
* **Address:** Urban students have slightly higher grades compared to rural students.
* **Family Size:** Students from larger families (GT3) tend to have higher grades.
* **Parent's Job and Guardian:** Distribution is fairly even, but there are nuances such as students with mothers as their guardian or teachers as parents showing better grades.
* **Support Systems:** Students with family support or no extra paid classes generally perform better.
* **Extra Activities:** Participation in extracurricular activities or internet access shows a positive correlation with higher grades.

**Feature Importance from Random Forest Model**

The Random Forest model highlights the most important features for predicting G3:

1. **G2:** Strongest predictor of G3.
2. **Absences:** Higher absences negatively impact G3.
3. **G1:** Also a strong predictor, but less than G2.
4. **Age, freetime, and family relationships:** Have moderate impacts.
5. **Dalc and Medu:** Indicate that lower alcohol consumption and higher mother’s education lead to better grades.
6. **School type (MS):** Indicates that attending school MS may negatively impact grades.

**Outlier Detection**

Boxplots help identify outliers and distributions within the top numerical features and the target variable G3:

* **G3:** Shows a few outliers on the lower end.
* **Other features like absences, G1, G2:** Outliers are present but distributions are generally consistent.

**Conclusion**

From the data exploration:

* **Key Predictors:** G2, G1, absences, mother's education, alcohol consumption, and participation in higher education plans are significant predictors of student grades.
* **Interventions:** To improve grades, interventions could focus on reducing absences, addressing alcohol consumption, and enhancing parental involvement in education.
* **Target Groups:** Specific attention might be needed for students from school MS, older students, and those with higher family responsibilities or more free time.

### Data Cleaning:

### Data Information

1. **Initial Data Overview**:
   * **Total Entries**: 649
   * **Columns**: 33
   * **Data Types**: 16 integer columns, 17 object (categorical) columns
2. **Summary Statistics**:
   * **Numerical Features**: The mean, standard deviation, min, 25th percentile, median, 75th percentile, and max for each numerical feature.
   * **Categorical Features**: The count, unique values, most frequent value (top), and frequency of the top value for each categorical feature.

**Data Cleaning Steps**

1. **Handling Missing Values**:
   * Numerical features were imputed with the mean value. This ensures that there are no missing values that could potentially cause errors during analysis or model training.
2. **Encoding Categorical Variables**:
   * One-hot encoding was applied to categorical features, converting them into binary columns. This is necessary for machine learning models that require numerical input.
3. **Outlier Removal**:
   * Outlier removal using the IQR method was attempted but encountered issues due to boolean data types after one-hot encoding. This step needs to be revisited if outlier removal is critical.
4. **Normalization**:
   * Numerical features were normalized using StandardScaler. This ensures that all features have a mean of 0 and a standard deviation of 1, preventing features with larger scales from dominating the model training process.

**Data Insights**

1. **Balanced Data**:
   * The data appears to be well-balanced with no missing values after imputation.
   * The distribution of categorical variables was handled, and their conversion to binary ensures consistency.
2. **Numerical Data**:
   * The standardized numerical data has mean values very close to zero and standard deviations of one, indicating successful normalization.
3. **Outliers**:
   * Outlier removal wasn't fully executed due to boolean type issues after one-hot encoding. This should be reconsidered if outliers significantly impact the analysis.
4. **Data Saved**:
   * The cleaned dataset was successfully saved as cleaned\_student\_data.csv, ready for further analysis or machine learning model training.

**Data Analysis**

 **Distribution of Features**:

* The distribution plots show the frequency of each value for the relevant features. Most features like age, Medu, Fedu, studytime, failures, goout, Dalc, Walc, health, and absences have a varied distribution.
* Dalc and Walc show skewed distributions, indicating most students have low daily and weekly alcohol consumption.

 **Boxplots**:

* The boxplots for each relevant feature help in identifying the presence of outliers.
* Features like failures, Dalc, Walc, and absences have noticeable outliers, which might need further investigation or removal depending on their impact on the analysis.

 **Scatter Plots**:

* The scatter plots between the relevant features and the target variable G2 show varied relationships.
* Features like failures and absences have a negative relationship with G2, indicating that higher failures and absences lead to lower grades.
* Medu and Fedu show a positive relationship with G2, suggesting that higher parental education levels are associated with better grades.

 **Correlation Heatmap**:

* The correlation matrix provides an overview of the linear relationships between the features.
* Strong positive correlations are observed between G2, G3, and G1, indicating consistency in students' grades over different periods.
* Negative correlations are observed between G2 and features like failures, absences, and Walc.

 **Descriptive Statistics**:

* The summary statistics provide insights into the central tendency, dispersion, and shape of the dataset's distribution.
* Most features have a mean close to zero after normalization, with a standard deviation of one.

 **Correlation Analysis with Target**:

* The top positively correlated features with G2 include G3, G1, higher\_yes, Medu, studytime, and Fedu.
* The top negatively correlated features with G2 include failures, school\_MS, Dalc, Walc, and traveltime.

 **Feature Importance (Random Forest)**:

* The Random Forest model identifies G3, G1, Mjob\_other, famrel, health, internet\_yes, traveltime, age, Medu, and Dalc as the top important features.
* G3 and G1 are the most important features, indicating that past grades are strong predictors of future grades.

 **Model Evaluation**:

* The Linear Regression model has a training RMSE of 0.344 and a testing RMSE of 0.335, with R² scores of 0.879 (training) and 0.894 (testing).
* The Random Forest model has a training RMSE of 0.148 and a testing RMSE of 0.344, with R² scores of 0.978 (training) and 0.889 (testing).
* Both models perform well, but Random Forest shows slight overfitting.

 **Correlation Heatmap and Pairplot for Top Features**:

* The heatmap for the top features shows strong correlations between grades and some features like G3 and G1.
* The pairplot for the top features visualizes the relationships between these features, providing a more granular view of the interactions.

 **Detailed Correlation Analysis for Top 3 Features**:

* Detailed correlation heatmaps for G3, G1, and Mjob\_other with their top positive and negative correlated features provide further insights into the dataset's structure and feature interactions.

**Bias Detection**

**Mean Absolute Error (MAE) by Group**

The Mean Absolute Error (MAE) varies significantly across different groups:

* Higher errors in some groups suggest that the model's predictions are less accurate for those specific combinations of age\_bin and Dalc\_bin.

**Root Mean Squared Error (RMSE) by Group**

The Root Mean Squared Error (RMSE) also shows substantial variation:

* This indicates that the model's performance in terms of prediction error is inconsistent across different subgroups.

**R^2 Score by Group**

The R^2 Score is negative or very low across all groups:

* A negative R^2 indicates that the model is performing worse than a horizontal line (mean of the data) for those groups.
* This suggests that the model is not capturing the variance in the data well for any subgroup.

**Selection Rate by Group**

The Selection Rate is uniformly 1 across all groups:

* This means the model predicts all instances as positive (binary target), indicating potential overfitting or a threshold issue.

**False Positive Rate (FPR) by Group**

The False Positive Rate (FPR) is uniformly 1 across all groups:

* This aligns with the uniform selection rate, showing that all predictions are false positives, except where no true negatives exist.

**False Negative Rate (FNR) by Group**

The False Negative Rate (FNR) is uniformly 0 across all groups:

* This complements the FPR, indicating no false negatives because every instance is predicted as positive.

**True Positive Rate (TPR) by Group**

The True Positive Rate (TPR) is uniformly 1 across all groups:

* This aligns with the previous observations, showing that every instance is correctly identified as positive when it is indeed positive.

**Additional Metrics by 'age\_bin'**

* **Accuracy, Precision, Recall, F1 Score**: These metrics are relatively consistent within age groups, with precision and recall being particularly high, indicating a potential imbalance or threshold issue in the prediction.

**Additional Metrics by 'Dalc\_bin'**

* **Accuracy, Precision, Recall, F1 Score**: There is variability in these metrics across different levels of daily alcohol consumption. For instance:
  + Dalc\_bin 0, 1, and 3 show high precision, recall, and F1 scores, indicating good performance.
  + Dalc\_bin 2 shows lower performance, suggesting the model struggles more with this group.
  + Dalc\_bin 4 has the worst performance, with accuracy and precision at 0, suggesting severe issues in prediction for this subgroup.

**Fairness Metrics**

* **Demographic Parity Difference (DPD)**: 0.0
* **Equalized Odds Difference (EOD)**: 0.0
* **False Positive Rate Difference (FPRD)**: 0.0
* **False Negative Rate Difference (FNRD)**: 0.0
* **Selection Rate Difference (SRD)**: 0.0

These fairness metrics suggest that, across the sensitive feature age\_bin, there is no observed disparity in terms of prediction rates. However, the uniformity in selection rate and the consistent high error metrics indicate potential underlying issues with model generalization.

**Conclusion**

1. **Model Performance**: The RandomForestRegressor model's predictions are consistent but show high error rates and negative R^2 scores, indicating poor generalization and accuracy across all groups.
2. **Bias and Fairness**: While fairness metrics show no disparity across age groups, the consistent selection rate and false positive rate across all groups suggest a potential issue with the model's threshold or an imbalance in the dataset.
3. **Group-Specific Performance**: The model performs variably across different levels of daily alcohol consumption (Dalc\_bin), indicating a need for potential group-specific tuning or re-evaluation of feature importance and data balancing.

Overall, the analysis reveals significant performance issues across all groups, and while fairness metrics indicate no group is unfairly treated, the model's overall accuracy and predictive power are lacking. Revisiting model training with adjusted parameters, additional feature engineering, or balancing techniques might be necessary.

**Bias Mitigation:**

**Overall Metrics:**

**Base Model:**

* Accuracy: 0.9462
* Precision: 0.9412
* Recall: 0.9552
* F1 Score: 0.9481

**Reweighed Model:**

* Accuracy: 0.9385
* Precision: 0.9155
* Recall: 0.9701
* F1 Score: 0.9420

Both models perform well overall, with high accuracy, precision, recall, and F1 scores. The reweighed model has a slightly lower accuracy and precision but higher recall compared to the base model.

**Fairness Metrics:**

**Base Model:**

* Demographic Parity Difference: 0.0532
* Equalized Odds Difference: 0.1034
* False Positive Rate Difference: 0.1034
* False Negative Rate Difference: 0.0667
* Selection Rate Difference: 0.0532

**Reweighed Model:**

* Demographic Parity Difference: 0.0217
* Equalized Odds Difference: 0.0667
* False Positive Rate Difference: 0.0265
* False Negative Rate Difference: 0.0667
* Selection Rate Difference: 0.0217

The reweighed model demonstrates significant improvements in fairness metrics:

* **Demographic Parity Difference:** Reduced from 0.0532 to 0.0217
* **Equalized Odds Difference:** Reduced from 0.1034 to 0.0667
* **False Positive Rate Difference:** Reduced from 0.1034 to 0.0265
* **Selection Rate Difference:** Reduced from 0.0532 to 0.0217

However, the False Negative Rate Difference remains the same for both models (0.0667).

**Metrics by Sensitive Feature Groups:**

**Age Groups (Base Model vs. Reweighed Model):**

**Base Model:**

* Age bin 0: Accuracy: 0.9180, Precision: 0.9091, Recall: 0.9375, F1: 0.9231
* Age bin 1: Accuracy: 1.0000, Precision: 1.0000, Recall: 1.0000, F1: 1.0000
* Age bin 2: Accuracy: 0.9286, Precision: 0.9333, Recall: 0.9333, F1: 0.9333

**Reweighed Model:**

* Age bin 0: Accuracy: 0.9344, Precision: 0.9118, Recall: 0.9688, F1: 0.9394
* Age bin 1: Accuracy: 0.9512, Precision: 0.9091, Recall: 1.0000, F1: 0.9524
* Age bin 2: Accuracy: 0.9286, Precision: 0.9333, Recall: 0.9333, F1: 0.9333

The reweighed model shows consistent performance across age groups with slight improvements in recall and F1 scores for age bins 0 and 1.

**Alcohol Consumption Groups (Base Model vs. Reweighed Model):**

**Base Model:**

* Dalc bin 0: Accuracy: 0.9886, Precision: 0.9778, Recall: 1.0000, F1: 0.9888
* Dalc bin 1: Accuracy: 0.8788, Precision: 0.9000, Recall: 0.9000, F1: 0.9000
* Dalc bin 2: Accuracy: 0.7143, Precision: 0.5000, Recall: 0.5000, F1: 0.5000
* Dalc bin 3: Accuracy: 1.0000, Precision: 1.0000, Recall: 1.0000, F1: 1.0000
* Dalc bin 4: Accuracy: 1.0000, Precision: 0.0000, Recall: 0.0000, F1: 0.0000

**Reweighed Model:**

* Dalc bin 0: Accuracy: 0.9659, Precision: 0.9362, Recall: 1.0000, F1: 0.9670
* Dalc bin 1: Accuracy: 0.9091, Precision: 0.9048, Recall: 0.9500, F1: 0.9268
* Dalc bin 2: Accuracy: 0.7143, Precision: 0.5000, Recall: 0.5000, F1: 0.5000
* Dalc bin 3: Accuracy: 1.0000, Precision: 1.0000, Recall: 1.0000, F1: 1.0000
* Dalc bin 4: Accuracy: 1.0000, Precision: 0.0000, Recall: 0.0000, F1: 0.0000

The reweighed model shows slight improvements in accuracy for Dalc bins 0 and 1 but still struggles with Dalc bin 4, where there are no true samples.

**Summary of Findings:**

* The **base model** performs very well overall but has noticeable disparities in fairness metrics across different sensitive groups.
* The **reweighed model** slightly sacrifices overall precision and accuracy but significantly improves fairness metrics, reducing disparities across sensitive groups.
* There are still challenges in some groups, such as Dalc bin 4, which may require further investigation or additional fairness interventions.

**Conclusion:**

Reweighing effectively balances the trade-off between overall performance and fairness, making the model fairer while maintaining high performance. This demonstrates the importance of incorporating fairness considerations into machine learning models to ensure equitable outcomes across different demographic groups.

**Summary Metrics Analysis**

**Comparison of Fairness Metrics between all Models:**

1. **Base Model**:
   * **Demographic Parity Difference**: 0.3
   * **Equalized Odds Difference**: 0.5
   * **False Positive Rate Difference**: 0.5
   * **False Negative Rate Difference**: 0.0476
   * **Selection Rate Difference**: 0.3
2. **Reweighed Model**:
   * **Demographic Parity Difference**: 0.1
   * **Equalized Odds Difference**: 0.0635
   * **False Positive Rate Difference**: 0.05
   * **False Negative Rate Difference**: 0.0635
   * **Selection Rate Difference**: 0.1
3. **Adversarial Debiasing Model**:
   * **Demographic Parity Difference**: 0.4797
   * **Equalized Odds Difference**: 0.75
   * **False Positive Rate Difference**: 0.75
   * **False Negative Rate Difference**: 0.2222
   * **Selection Rate Difference**: 0.4797
4. **Post-Processing Model**:
   * All fairness metrics: 0.0

**Observations:**

* The **Base Model** shows significant disparities across different fairness metrics, indicating a bias towards certain groups.
* The **Reweighed Model** improves significantly over the Base Model, reducing the disparities in all fairness metrics.
* The **Adversarial Debiasing Model** shows a large increase in fairness metric disparities, suggesting that the method might have introduced some issues, possibly due to overfitting or improper constraint handling.
* The **Post-Processing Model** achieves perfect fairness metrics (0.0 for all), indicating no disparities. However, this model might have performance trade-offs as seen in the accuracy metrics.

**Performance Metrics:**

1. **Base Model**:
   * Accuracy: 0.9462
   * Precision: 0.9412
   * Recall: 0.9552
   * F1 Score: 0.9481
2. **Reweighed Model**:
   * Accuracy: 0.9462
   * Precision: 0.9545
   * Recall: 0.9403
   * F1 Score: 0.9474
3. **Adversarial Debiasing Model**:
   * Accuracy: 0.7615
   * Precision: 0.7571
   * Recall: 0.7910
   * F1 Score: 0.7737
4. **Post-Processing Model**:
   * Accuracy: 0.4846
   * Precision: 0.0000
   * Recall: 0.0000
   * F1 Score: 0.0000

**Observations:**

* The **Base Model** and the **Reweighed Model** have similar performance metrics, indicating that reweighting did not significantly impact model performance.
* The **Adversarial Debiasing Model** shows a drop in performance, which might be a result of the model trying to balance fairness constraints at the cost of overall performance.
* The **Post-Processing Model** shows extremely poor performance metrics. This model achieved perfect fairness at the cost of predictive performance, essentially making it non-functional in terms of accuracy, precision, recall, and F1 score.

**Visualization Analysis:**

1. **Bar Plot Comparison**:
   * The **Base Model** and the **Adversarial Debiasing Model** have higher disparities across most fairness metrics.
   * The **Reweighed Model** shows reduced disparities, indicating an improvement in fairness.
   * The **Post-Processing Model** shows no disparity (all metrics are 0), confirming perfect fairness.
2. **Heatmap**:
   * The heatmap further highlights the disparities, with the **Post-Processing Model** having a uniform blue color indicating zero disparity.
   * The **Adversarial Debiasing Model** has the highest disparities, highlighted in red.
3. **Group-wise Accuracy Disparity Plot**:
   * The **Base Model** and **Reweighed Model** show better consistency in accuracy across groups compared to the **Adversarial Debiasing Model**.
   * The **Post-Processing Model** has significant performance issues, indicated by some groups having zero accuracy.
4. **CDF of Prediction Scores**:
   * The CDF plot for the **Post-Processing Model** shows a step function, indicating poor prediction score distribution across age groups.
5. **Confusion Matrix by Group**:
   * Confusion matrices highlight the performance of the **Base Model** across different groups, with group 0 performing better than groups 1 and 2.
6. **ROC Curves by Age Group**:
   * ROC curves indicate that the **Base Model** performs well across age groups, with high AUC values.

**Interpretation and Recommendations:**

* The **Base Model** has significant fairness issues, which are addressed to some extent by the **Reweighed Model**.
* The **Adversarial Debiasing Model** might need parameter tuning or reconsideration of constraints, as it shows high disparities and reduced performance.
* The **Post-Processing Model** achieves perfect fairness but at the cost of predictive performance, making it impractical for real-world use.
* Further improvements can be made by experimenting with different fairness constraints and mitigation techniques to balance fairness and performance better.
* It is important to regularly evaluate both performance and fairness metrics to ensure the model serves its intended purpose without introducing bias.