Below is a brief description of each column in your dataset. These definitions are consistent with the **Student Alcohol Consumption** dataset commonly referenced from UCI Machine Learning Repository.

1. **school**
   * **Definition:** The school the student attends.
   * **Possible Values:** Typically 'GP' (Gabriel Pereira) or 'MS' (Mousinho da Silveira).
2. **sex**
   * **Definition:** The student’s gender.
   * **Possible Values:** 'F' (female) or 'M' (male).
3. **age**
   * **Definition:** The student’s age in years.
   * **Typical Range:** 15 to 22.
4. **address**
   * **Definition:** The type of home address.
   * **Possible Values:** 'U' (urban) or 'R' (rural).
5. **famsize**
   * **Definition:** The size of the student’s family.
   * **Possible Values:**
     + 'LE3' (less or equal to 3)
     + 'GT3' (greater than 3).
6. **Pstatus**
   * **Definition:** The cohabitation status of the student’s parents.
   * **Possible Values:**
     + 'T' (parents living together)
     + 'A' (parents apart).
7. **Medu**
   * **Definition:** Mother’s education level.
   * **Scale (commonly):**
     + 0: none
     + 1: primary education (4th grade)
     + 2: 5th to 9th grade
     + 3: secondary education
     + 4: higher education.
8. **Fedu**
   * **Definition:** Father’s education level.
   * **Scale:** Same interpretation as **Medu**.
9. **Mjob**
   * **Definition:** Mother’s job.
   * **Possible Values (examples):** 'teacher', 'health', 'services', 'at\_home', 'other'.
10. **Fjob**
    * **Definition:** Father’s job.
    * **Possible Values (examples):** Same set as **Mjob**.
11. **reason**
    * **Definition:** The reason why the student chose this school.
    * **Possible Values (examples):** 'home', 'reputation', 'course', 'other'.
12. **guardian**
    * **Definition:** The student’s primary guardian.
    * **Possible Values:** 'mother', 'father', 'other'.
13. **traveltime**
    * **Definition:** The time it takes for the student to travel from home to school.
    * **Scale (commonly):**
      + 1: <15 minutes
      + 2: 15 to 30 minutes
      + 3: 30 minutes to 1 hour
      + 4: >1 hour.
14. **studytime**
    * **Definition:** The weekly study time.
    * **Scale (commonly):**
      + 1: <2 hours
      + 2: 2 to 5 hours
      + 3: 5 to 10 hours
      + 4: >10 hours.
15. **failures**
    * **Definition:** The number of past class failures.
    * **Typical Range:** 0 to 3 (with 4 often indicating 4 or more failures).
16. **schoolsup**
    * **Definition:** Extra educational support provided by the school.
    * **Possible Values:** 'yes' or 'no'.
17. **famsup**
    * **Definition:** Family educational support.
    * **Possible Values:** 'yes' or 'no'.
18. **paid**
    * **Definition:** Extra paid classes within the course subject (e.g., math or Portuguese).
    * **Possible Values:** 'yes' or 'no'.
19. **activities**
    * **Definition:** Involvement in extracurricular activities.
    * **Possible Values:** 'yes' or 'no'.
20. **nursery**
    * **Definition:** Whether the student attended nursery school.
    * **Possible Values:** 'yes' or 'no'.
21. **higher**
    * **Definition:** Plans to pursue higher education.
    * **Possible Values:** 'yes' or 'no'.
22. **internet**
    * **Definition:** Access to the internet at home.
    * **Possible Values:** 'yes' or 'no'.
23. **romantic**
    * **Definition:** Whether the student is in a romantic relationship.
    * **Possible Values:** 'yes' or 'no'.
24. **famrel**
    * **Definition:** Quality of family relationships.
    * **Scale (commonly):** 1 (very bad) to 5 (excellent).
25. **freetime**
    * **Definition:** Amount of free time after school.
    * **Scale:** 1 (very low) to 5 (very high).
26. **goout**
    * **Definition:** Frequency of going out with friends.
    * **Scale:** 1 (very low) to 5 (very high).
27. **Dalc**
    * **Definition:** Workday (weekday) alcohol consumption.
    * **Scale:** 1 (very low) to 5 (very high).
28. **Walc**
    * **Definition:** Weekend alcohol consumption.
    * **Scale:** 1 (very low) to 5 (very high).
29. **health**
    * **Definition:** Current health status.
    * **Scale:** 1 (very bad) to 5 (very good).
30. **absences**
    * **Definition:** The number of school absences.
    * **Range:** Typically 0 to 93.
31. **G1**
    * **Definition:** First period grade (intermediate result).
    * **Range:** 0 to 20.
32. **G2**
    * **Definition:** Second period grade (intermediate result).
    * **Range:** 0 to 20.
33. **G3**
    * **Definition:** Final grade (target variable in many studies).
    * **Range:** 0 to 20.

Below is a brief analysis summarizing the steps we have implemented in our code:

1. **Data Loading and Inspection:**
   * We began by importing essential libraries (Pandas, NumPy, Matplotlib, Seaborn, and scikit-learn) and loaded our dataset using Pandas.
   * We inspected the data using methods like .head(), .info(), and .describe() to understand its structure and identify any missing values.
2. **Data Cleaning and Preparation:**
   * Missing values were addressed (in our example, rows with missing values were dropped).
   * Categorical variables (such as 'schoolsup', 'famsup', etc.) were converted into numeric format using mapping and one-hot encoding with pd.get\_dummies().
   * This step ensured that all data fed into the models was numeric.
3. **Exploratory Data Analysis (EDA):**
   * We visualized the distribution of the target variable (final grade, **G3**) using a histogram.
   * A correlation matrix was generated via Seaborn’s heatmap to identify relationships between numeric features.
4. **Model Preparation and Feature Selection:**
   * A subset of relevant features was selected (e.g., 'studytime', 'failures', 'absences', 'Dalc', 'Walc', 'G1', and 'G2').
   * The dataset was split into features (X) and the target variable (y), and then further divided into training and testing sets.
5. **Linear Regression Modeling and Evaluation:**
   * A Linear Regression model was trained on the data.
   * The model’s performance was evaluated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R².
   * We then implemented k-fold cross-validation (5-fold) to assess model robustness, calculating average R² and MAE across folds.
6. **Experimenting with Model Complexity using Random Forests:**
   * A Random Forest Regressor was introduced as an alternative to capture non-linear relationships.
   * We applied 5-fold cross-validation to evaluate the Random Forest model, obtaining improved R² and lower MAE scores compared to the linear model.
   * Feature importance was extracted to understand which predictors most influenced the model’s predictions—showing that early academic performance (G2) was the most influential feature.

**Analysis of Model Performance and Feature Importance**

**1. Linear Regression Model Results**

* **MAE: 1.34**  
  On average, the linear regression predictions differ from the actual final grade (G3) by 1.34 points.
* **MSE: 4.44 and RMSE: 2.11**  
  The Root Mean Squared Error (RMSE) of 2.11 indicates that larger errors are penalized more. In context, this suggests that while most predictions are close, some may have larger deviations.
* **R²: 0.78**  
  About 78% of the variability in final grades is explained by the linear regression model. This is a decent fit, but there is room for improvement, particularly in capturing non-linear relationships.

**2. Random Forest Model Results**

* **R² Scores:**  
  The fold-specific R² scores are:  
  [0.875, 0.877, 0.837, 0.911, 0.741]  
  The average R² is ~0.848, meaning the Random Forest model explains approximately 84.8% of the variance in the final grade. This is a significant improvement over the linear model, suggesting that the Random Forest is better at capturing complex, non-linear relationships in the data.
* **MAE Scores:**  
  The MAE for each fold is around:  
  [1.047, 1.007, 1.026, 0.875, 1.242]  
  With an average MAE of ~1.039, the Random Forest predictions are off by about 1.04 grade points on average—again an improvement over the linear regression model.
* **Overall MAE from Cross-Validation:**  
  The overall MAE from cross-validation matches the average (1.04), confirming that the error metric is consistent across different subsets of the data.

**3. Feature Importances from the Random Forest**

The Random Forest model provides insights into which features are most influential:

* **G2 (0.7897):**  
  This is the most important feature by far, indicating that the second period grade is a very strong predictor of the final grade. It shows that early academic performance is critical.
* **Absences (0.1251):**  
  The number of school absences is the next most important factor. Higher absences may negatively affect student performance.
* **Studytime (0.0260) and G1 (0.0238):**  
  These features contribute to the model but are less influential compared to G2 and absences.
* **Walc (0.0133), Failures (0.0127), Dalc (0.0093):**  
  Both workday (Dalc) and weekend (Walc) alcohol consumption have relatively low importance. This suggests that while alcohol consumption does have some impact, its effect on the final grade is much smaller compared to academic performance metrics like G1, G2, and attendance (absences).

**Overall Interpretation**

* **Model Comparison:**  
  The Random Forest model outperforms the linear regression model with higher R² (0.848 vs. 0.78) and lower MAE (1.04 vs. 1.34). This improvement indicates that the Random Forest is better at capturing the nuances and non-linear relationships in the data.
* **Key Predictors:**  
  Early academic performance, especially the second period grade (G2), dominates the prediction of final grades. Absences also play a significant role, while the influence of alcohol consumption is comparatively minor.
* **Implications:**  
  The analysis suggests that interventions aimed at improving early performance and reducing absences may have the most significant impact on final grades. While monitoring alcohol consumption is important, its direct influence on academic performance seems limited when compared to these other factors.