Red Wine Quality Dataset Practice Question

Section A: Basic Data Handling

1. Load the dataset winequality-red.csv using pandas and display the first 10 rows.
2. Find:
    - Number of rows and columns
    - Column names
    - Data types of each column
3. Check if the dataset contains:
    - Missing values
    - Duplicate rows
4. Remove duplicate rows and display the new shape of the dataset.

Section B: Statistical Analysis

5. Display the statistical summary of the dataset.
6. Find the mean, median, and standard deviation of the alcohol column.
7. Find the correlation between:
    - Alcohol and Quality
    - Volatile acidity and Quality
8. Which feature has the highest positive correlation with quality?

Section C: Data Visualization

9. Plot a bar chart showing the distribution of wine quality.
10. Plot a histogram of all numerical features.
11. Create a heatmap of the correlation matrix.
12. Draw a boxplot showing the relationship between alcohol and quality.

Section D: Feature-Based Questions

14. Create a new column called quality_label:

- If quality ≥ 7 → "Good"
- If quality = 5 or 6 → "Average"
- If quality ≤ 4 → "Poor"

15. Count how many wines fall into each quality label.

16. Find the average alcohol content for each quality category.

Section E: Intermediate Level

17. Identify the top 3 features most correlated with quality.

18. Normalize the dataset using Min-Max Scaling.

19. Split the dataset into:

- Features (X)
- Target (y)

20. Perform train-test split (80-20 ratio).

Section F: Advanced / Model-Based

21. Build a Linear Regression model to predict wine quality.

22. Build a Logistic Regression model to classify wine quality (Good vs Not Good).

23. Evaluate the model using:

- Accuracy
- Confusion Matrix
- Classification Report

24. Handle class imbalance using:

- SMOTE
- OR Class Weights