DA331 Lab 1 Assignment

Objectives:

- Load and explore a dataset using Python (pandas)
- Perform basic data cleaning and descriptive statistics
- Use classification models for prediction

Dataset: Use the Iris dataset (built-in in sklearn.datasets)

Part A: Data Loading & Cleaning

1. Load the dataset:

2. Display:
    - First 5 rows
    - Data types
    - Null values
    - Class distribution (species)
3. Check for duplicates and drop if any:

Part B: Exploratory Data Analysis

1. Compute:
    - Mean, median, std deviation of each numeric column
    - Correlation matrix
2. Plot:
    - Histogram
    - Boxplot to compare feature distributions
    - Scatter plot colored by class (species)

Part C: Classification Using ML Models

Step 1: Prepare the Data

Split the data into training and testing sets:

Step 2: Use ML models for classification

Model 1: Logistic Regression

Model 2: K-Nearest Neighbors (KNN)

Model 3: Decision Tree Classifier

Compare All Model Accuracies

Discussion Questions:

- Which model performed best in your test? Why might that be?
- How might the choice of `k` in KNN or regularization in logistic regression affect results

Submission Checklist:

Lab Report (in PDF format)

The report must include:

Title Page

- Name
- Roll Number
- Course: Big Data Analytics – Tools & Techniques
- Date

Report Structure

1. **Objective**
   A short paragraph explaining what the lab was about.
2. **Part A – Data Cleaning**
   - Brief summary of what you did
3. **Part B – Exploratory Data Analysis**
   - Describe the steps (e.g., histograms, scatter plots)
4. **Part C – Classification Models**
   - Describe the models used (Decision Tree, KNN, Logistic Regression)
   - Show metrics (accuracy, classification report)
   - Include a table comparing all three models
5. **Conclusion**
   - Reflect briefly on what you learned
   - Mention which model performed best and why
6. **Google Colab Link**
   - Add a clickable hyperlink at the **end of the report**, clearly labeled.

   Example: "**Google Colab Notebook:** Click here to view my notebook"

Important Notes:

- The Google Colab notebook must be public or shared with "Anyone with the link can view".
- The code must be complete, readable, and contain comments for each step.
- All outputs (plots, metrics) should be visible inside the notebook.