DA331 Lab 1 Report

Name: Aryan Gupta
Roll Number: 230150003
Course: Big Data Analytics – Tools & Techniques
Date: July 30, 2025

1. Objective

The goal was to load the Iris dataset, clean it, explore its properties, then apply three classification models—Logistic Regression, K-Nearest Neighbors, and Decision Tree—to compare their test accuracies.

2. Part A – Data Cleaning

- Loaded Iris via `sklearn.datasets`.
- Displayed first five rows, data types, null counts, and class distribution.
- Found no missing values or duplicates; data ready for analysis.

3. Part B – Exploratory Data Analysis

- Descriptive stats: computed mean, median, standard deviation for each feature.
- Correlation: generated and reviewed the feature correlation matrix.
- Plots:
  - Histograms of petal/sepal lengths and widths
  - Boxplots comparing distributions across species
  - Scatter plot colored by species to visualize class separability

4. Part C – Classification Models

1. Data split: 70% train / 30% test

2. Models & Test Accuracies:

Model	Accuracy
Logistic Regression	1.00
K-Nearest Neighbors (k=5)	1.00
Decision Tree	1.00

3. Discussion:
All three models achieved perfect accuracy on the test set. This suggests the Iris classes are linearly separable in this split. Adjusting the KNN `k` value or logistic regression's regularization could affect these results on different splits.

5. Conclusion

The dataset’s clear class boundaries led every model to 100% test accuracy. In practice, tuning hyperparameters (e.g., `k` in KNN, regularization strength in Logistic Regression) and validating on multiple splits would guard against overfitting.

6. Google Colab Link

Google colab notebook