Exploratory Data Analysis (EDA) Project Report

Project Title:

Exploratory Data Analysis on the Iris Flower Dataset

Objective:

The objective of this project is to perform Exploratory Data Analysis (EDA) on the classic Iris dataset to understand its structure, detect patterns, test assumptions, and prepare it for future machine learning tasks such as classification.

Dataset Information:

- Dataset Name: Iris Dataset
- Source: Built-in dataset from seaborn or sklearn.datasets
- Number of Instances: 150
- Number of Features: 4 numeric features + 1 categorical target
- Features:
- sepal_length
- sepal_width
- petal_length
- petal_width
- species (Target: Setosa, Versicolor, Virginica)

Key Questions Asked:

- What is the distribution of each numeric feature?
- Are there any missing values or anomalies?
- Which features are most correlated?
- Are the species well-separated by these features?

- Are numerical features normally distributed?

- Any outliers present?

Data Exploration & Structure:

- All 4 features are of type float64.

- The target column species is a categorical feature.

- No missing values or null entries.

- No duplicate records found in the dataset.

- Summary statistics (mean, median, std) were generated for each feature.

Visual Analysis Performed:

- Histograms and Boxplots showed the spread of feature values.

- Pairplots revealed that:

- Setosa is clearly distinguishable from other species.

- Petal measurements are more discriminative than sepal measurements.

- Heatmap of correlation:

- petal_length and petal_width: strongly correlated.

- sepal_length and sepal_width: weakly correlated.

- Outliers: Slight outliers seen in sepal_width, no extreme values.

Statistical Testing:

- Shapiro-Wilk test applied to check normality for numeric features.

- ANOVA tested mean differences between species groups for each feature.

- Results showed statistically significant differences in all features across species ($p < 0.05$).

Data Issues Identified:

- No missing data or formatting problems.

- No major outliers.

- Slight skew in sepal_width, but not critical.

Conclusions & Insights:

- The dataset is clean, well-balanced, and suitable for classification tasks.

- Petal-based features are strong predictors for species.

- Species are linearly separable, especially Setosa.

- EDA confirms that the dataset is ready for machine learning.

Next Steps (Recommended):

- Encode the categorical variable species into numeric labels.

- Train a machine learning model (e.g., Decision Tree, SVM).

- Evaluate the model with accuracy, precision, recall.

- Visualize decision boundaries or feature importances.

Tools Used:

- Python

- Libraries: pandas, numpy, matplotlib, seaborn, scipy

Project Completion Status:

EDA Task Completed Successfully

Dataset Analyzed

Insights Documented

Ready for Model Building