Iris Flower Classification: A Data Analysis Project

By: [Janhavi Bhosale]
Course: [Data Analysis]
Date: [13/07/2025]

Introduction
The goal is to analyze the Iris dataset and classify flowers into three species based on petal and sepal measurements using machine learning techniques.

Dataset Description
The dataset is sourced from sklearn's built-in load_iris(). It contains 150 samples with 4 features: sepal length, sepal width, petal length, and petal width. The target variable has three classes representing different iris species.

Data Preprocessing
Loaded dataset using sklearn, checked for missing values (none found), added species labels, and ensured data cleanliness.

Exploratory Data Analysis
Used descriptive statistics and visualizations like pairplots and countplots to understand relationships between features and class distribution. Found petal features are more useful for class separation.

Feature Selection
Used all four numerical features. Target (Species) was separated from the features.

Data Splitting & Scaling
Split dataset into training (80%) and testing (20%) sets. Used StandardScaler for feature normalization.

Model Training
Applied K-Nearest Neighbors classifier with k=3. Trained the model using scaled training data.

Model Evaluation
Evaluated model using confusion matrix and classification report. Achieved approximately 96% accuracy. Plotted a heatmap to visualize the confusion matrix.

Conclusion
The Iris dataset is ideal for beginners. KNN performed well without complex tuning. Visualizations and analysis confirmed strong separability between classes, especially using petal features.

Future Scope
Experiment with other models (e.g., SVM, Random Forest), try hyperparameter tuning, and consider deploying the model as a web app.