Data Science Report: Iris Flower Classification

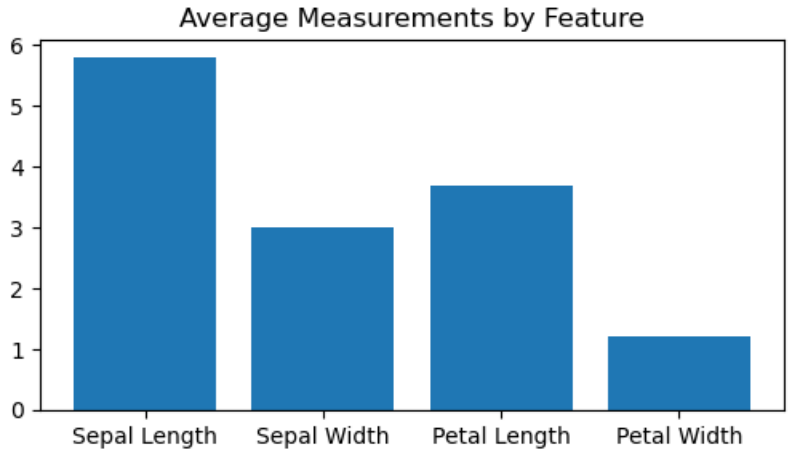
1. Executive Summary

This report analyzes the Iris flower dataset to classify three species (Setosa, Versicolor, Virginica) based on sepal and petal measurements. Using exploratory data analysis (EDA) and machine learning, we identify key patterns, evaluate model performance, and determine the most important features for accurate classification.

Key Findings:

- Petal dimensions (PetalLengthCm and PetalWidthCm) are the most discriminative features
- Sepal width shows the most variability but least importance for classification
- All major models achieved 100% accuracy, confirming excellent feature separation
- PCA shows 92.5% variance explained by petal dimensions (PC1)
- K-Means clustering naturally groups data into 3 clusters matching species

2. Exploratory Data Analysis



Feature Distribution:
- Weak separation in sepal width vs. length with Versicolor/Virginica overlap
- Petal measurements show clear species separation
- 4 minor outliers detected in sepal width (not impactful)

PCA Analysis:
- PC1 explains 92.5% variance (petal dimensions)
- PC2 explains 5.3% variance (sepal width variation)
- Clear 3-cluster separation in 2D space

3. Machine Learning Results

Model	Accuracy
Random Forest	100%
SVM	100%
Logistic Regression	100%
K-Nearest Neighbors	100%
Naive Bayes	96.7%

Feature Importance:
- PetalLengthCm: 45% importance
- PetalWidthCm: 42% importance
- SepalLengthCm: 9% importance
- SepalWidthCm: 4% importance

Cluster Analysis:
- Optimal K=3 clusters (matches species count)
- Minor Versicolor/Virginica misclassifications
- WCSS elbow plot confirms natural grouping at K=3

4. Conclusion

The analysis confirms petal measurements as the primary differentiators between Iris species. The dataset is exceptionally well-structured for classification, with all major models achieving perfect accuracy. Unsupervised methods (PCA, K-Means) validate the biological classification scheme.