Iris Flower Classification using Machine Learning

1. Introduction

The Iris Flower dataset is a well-known benchmark dataset in machine learning and statistics. It consists of measurements of iris flowers belonging to three different species: **Setosa, Versicolor, and Virginica**. The objective of this project is to build a classification model that can accurately predict the species of a flower based on its **sepal length, sepal width, petal length, and petal width**.

This project demonstrates the process of:

- Data exploration and visualization,
- Feature analysis,
- Model training using classification algorithms,
- Evaluation of predictive performance.

---

2. Objective

The main goal of this project is to **classify iris flowers into one of the three species** using supervised machine learning algorithms.

---

3. Dataset Description

The dataset used in this project is the **Iris dataset**, first introduced by Ronald Fisher in 1936.

- **Number of Instances:** 150
- **Number of Features:** 4 (continuous variables)
- **Target Variable:** Species (Setosa, Versicolor, Virginica)

Features:

1. Sepal Length (cm)
2. Sepal Width (cm)
3. Petal Length (cm)
4. Petal Width (cm)

Target Classes:

- Iris Setosa (0)
- Iris Versicolor (1)
- Iris Virginica (2)

---

4. Methodology

The following steps were followed in the project:

4.1 Data Loading and Preprocessing

- The Iris dataset was loaded using Scikit-learn's inbuilt dataset.
- Checked for missing values (none were found).
- Target labels were mapped to their respective species names.

4.2 Exploratory Data Analysis (EDA)

- **Pairplots** were used to visualize relationships between features.
- **Heatmap** was plotted to examine feature correlations.
- Observations:
    - *Petal length and petal width* show strong correlation with species.
    - *Setosa* is easily separable, while *Versicolor* and *Virginica* overlap slightly.

4.3 Model Building

Three classification models were trained:

1. **Logistic Regression**
2. **Support Vector Machine (SVM)**
3. **K-Nearest Neighbors (KNN)**

The dataset was split into **80% training** and **20% testing** sets.

4.4 Model Evaluation

- Performance was evaluated using **Accuracy, Confusion Matrix, and Classification Report**.

---

5. Results

Model	Accuracy
Logistic Regression	0.97
Support Vector Machine	1.00
K-Nearest Neighbors	0.97

- **SVM (Support Vector Machine)** achieved the best performance with **100% accuracy** on the test dataset.
- **Confusion Matrix** confirmed that SVM correctly classified all instances.
- Logistic Regression and KNN also performed very well, with only minor misclassifications between *Versicolor* and *Virginica*.

---

6. Visualizations

1. **Pairplot:** Clear separation of Setosa, slight overlap of Versicolor & Virginica.
2. **Heatmap:** High correlation between Petal Length & Petal Width.
3. **Confusion Matrix:** Shows correct vs misclassified predictions.

---

7. Conclusion

The Iris Flower Classification project successfully demonstrated the application of machine learning in supervised classification tasks.

- The dataset is simple yet effective for beginners to practice data analysis and classification.
- Among the models tested, **Support Vector Machine (SVM)** provided the highest accuracy (100%).
- This confirms that linear boundaries are effective for this dataset due to well-separated feature spaces.

Future Work

- Experiment with deep learning models (Neural Networks).
- Apply cross-validation and hyperparameter tuning for more robust evaluation.

- Deploy the trained model as a simple **web application** for real-time predictions.

---

8. References

1. Fisher, R. A. (1936). *The use of multiple measurements in taxonomic problems.*
2. Scikit-learn Documentation: https://scikit-learn.org
3. UCI Machine Learning Repository – Iris Dataset.