Iris Flower Classification - Internship Project

Task Description

The challenge was to build a Machine Learning model to classify Iris flowers into three species — Setosa, Versicolor, and Virginica — based on sepal and petal length and width measurements. This task is part of the AICTE Oasis Infobyte Data Science Internship, focusing on supervised learning and model evaluation.

Model Selection: Logistic Regression

For this task, I used the Logistic Regression model from Scikit-learn. Logistic Regression is one of the simplest yet effective algorithms for classification problems. It is well-suited for this dataset because it handles multi-class classification efficiently and provides clear decision boundaries among classes.

What is Logistic Regression?

Logistic Regression is a supervised machine learning algorithm used for **classification tasks**.
Unlike linear regression, which predicts continuous values, logistic regression predicts **categorical outcomes** — that is, it classifies input data into discrete classes (e.g., "Setosa", "Versicolor", "Virginica" in this case).

It works by estimating probabilities using a **logistic (sigmoid) function**, which outputs values between 0 and 1. These probabilities are then used to assign class labels to the data points.

Why I Chose Logistic Regression

I chose **Logistic Regression** for this task because:

- It is **simple yet effective** for classification problems like the Iris dataset.
- It provides **clear interpretability** — we can easily understand how each feature affects the prediction.
- It works well with **small to medium-sized datasets**, which suits the Iris dataset perfectly.
- It supports **multi-class classification** using the "one-vs-rest" (OvR) strategy.
- It's a **foundational model** that helps understand more complex algorithms later.

In this project, Logistic Regression successfully classified the iris flowers into their respective species with **high accuracy**, making it a reliable and efficient choice.

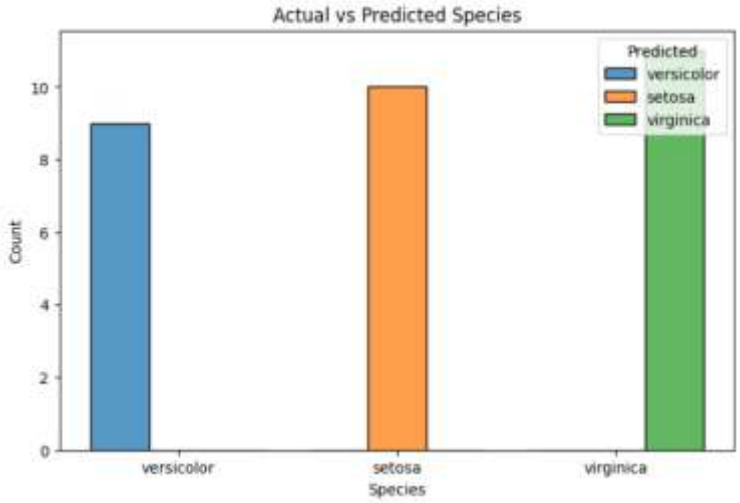
Approach

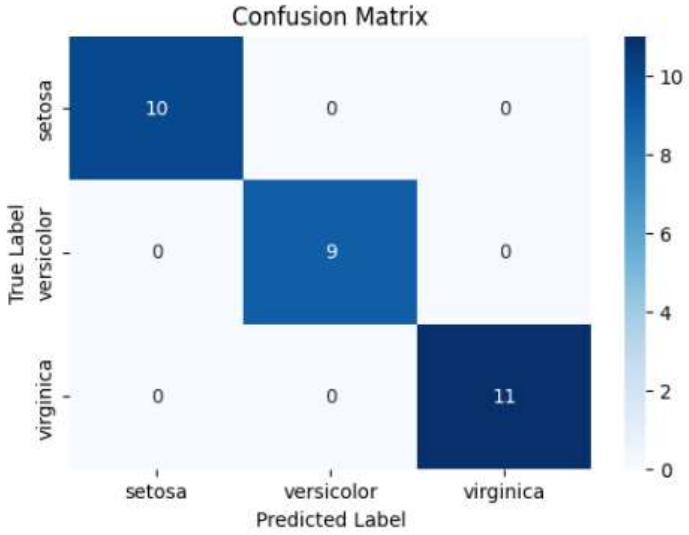
1. Loaded the Iris dataset from Scikit-learn.
   2. Split the data into training and testing sets (80%-20%).
   3. Trained the Logistic Regression model on the training data.
   4. Made predictions on the test data.
   5. Evaluated performance using accuracy, classification report, and confusion matrix.
   6. Visualized the results using Matplotlib and Seaborn.

Results

The model achieved an outstanding accuracy of 100%. The classification report shows precision, recall, and F1-score of 1.0 for all three classes, indicating perfect classification. The confusion matrix further confirms that all test samples were predicted correctly with no misclassifications.

An example of the result visualization is shown below.





Key Insights

- Logistic Regression performed exceptionally well on the Iris dataset.
- The dataset is well-balanced and linearly separable, making it ideal for this model.
- Visualization of actual vs predicted results confirms the model's perfect accuracy.
- This project strengthened my understanding of supervised learning, model evaluation metrics, and result visualization.

Conclusion

This project provided hands-on experience in building, training, and evaluating a classification model. It also demonstrated the power of Logistic Regression in solving real-world classification problems.