IRIS FLOWER CLASSIFICATION USING LOGISTIC REGRESSION AND RANDOM FOREST

Gopal Kumar Shaw

5th Semester

Bangabasi College

Period of Internship:25th August 2025 -19th September 2025

Report submitted to :IDEAS-Institute of Data Engineering ,Analytics and Science Foundation ,ISI Kolkata

Abstract:

This project focuses on classifying Iris flower species using machine learning techniques. The Iris dataset contains 150 samples with four features: sepal length, sepal width, petal length, and petal width. We use data visualization (pair plots and heatmaps) to explore patterns and relationships between species. Two classification models are applied: **Logistic Regression** and **Random Forest**.

Logistic Regression, which assumes a linear decision boundary, performs well in separating Setosa but struggles with overlapping species like Versicolor and Virginica. Random Forest, a non-linear ensemble model, achieves higher accuracy by handling complex patterns and feature interactions more effectively. We also explore model behavior with noisy or high-dimensional data. Results show that Random Forest generalizes better in such conditions due to its robustness and built-in feature selection.

Overall, the project highlights the importance of choosing the right model based on data complexity and visual insights.

Introduction To Project:

Iris Classification using Logistic Regression and Random Forest" is the title of my project. It is based on the well-known Iris dataset, which includes floral dimensions for three different species of iris flowers, including petal length, petal width, sepal length, and sepal width. The primary goal is to create models that use these characteristics to accurately identify a flower's species. This is significant because the same methods employed here can also be used to address more complex issues like pattern recognition, image classification, and medical diagnosis.

Python and its key libraries- **Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn**, are among the technologies utilized in this project. While Random Forest was employed to capture more intricate and non-linear interactions, Logistic Regression was utilized as a linear model to categorize species. After loading the dataset, it was visualized using heatmaps and pairplots, divided into training and testing sets, and both models were constructed. Evaluation measures were then used to compare the models' performance and accuracy. Gaining practical experience with machine learning algorithms, comprehending their distinctions, and learning how to assess categorization models efficiently are the goals of this research.

Objectives of the Project:

1.To use machine learning models to categorize iris flowers using petal and sepal measures into three species: Virginica, Versicolor, and Setosa.

2.To evaluate the accuracy and performance of Random Forest and Logistic Regression, as well as their advantages and disadvantages.

3.To demonstrate how linear and non-linear classifiers differ from one another and why Random Forest occasionally outperforms them on overlapping data.

4.To obtain hands-on experience in data visualization and analysis, including feature relationships, assessment metrics, and data pretreatment.

5.To use a straightforward but timeless dataset to illustrate the use of supervised learning, which can then be expanded to more intricate real-world issues.

Project Methodology and Work Process:

From data collection to model evaluation, the project "Iris Classification using Logistic Regression and Random Forest" was completed step-by-step. The procedures and techniques I used are listed below:

1. Data Collection:

•The Iris dataset, a common dataset accessible through Scikit-learn, is the dataset utilized.

•There are 150 samples total, 50 of each of the three species—Virginica, Versicolor, and Setosa.

•Four numerical characteristics are present for every sample: petal length, petal width, sepal length, and sepal width.

2. Pre-processing and Data Cleaning:

•No duplicate records or missing values were discovered because the dataset was already clean.

•However, I converted numerical target values (0,1,2) into species names and renamed feature columns for ease of comprehension.

•For improved accessibility and visualization, the dataset was then merged into a Pandas DataFrame.

3. Data Visualization and Exploration:

•Seaborn pairplots were utilized to examine the correlation between species and characteristics.

•To determine which features were highly correlated, a correlation heatmap was made.

•found that the most effective metrics for species separation were petal length and width.

•Before using models, this stage helped me comprehend the patterns in the data.

4. Dividing the Dataset :

•Training (80%) and testing (20%) sets of the dataset were separated.

•Models were constructed using training data, and their performance on unseen data was evaluated using testing data.

•This guarantees that the models are not merely learning training samples and that they generalize successfully.

5. Model Development and Selection:

Two models for machine learning were chosen:

•**Logistic Regression:** linear model called logistic regression uses straight boundaries to try and divide classes.

•**Random Forest:** An ensemble model that may manage intricate non-linear interactions by combining numerous decision trees.

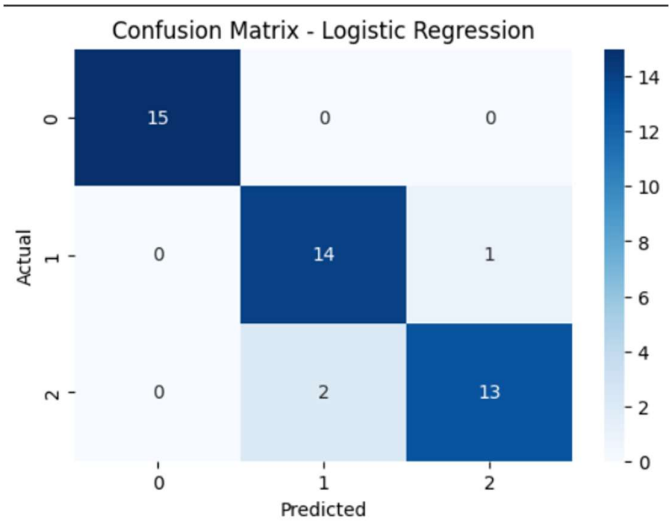
•To train both models, the Scikit-learn library was used.

6. Validation and Assessment of the Model:

•For both models, predictions were made on the test set.

•The following metrics were used to measure performance:

- •Score for Accuracy
- •Classification Report (F1-score, precision, and recall)
- •Confusion Matrix



•Setosa was successfully classified using logistic regression, however Versicolor and Virginica were difficult to categorize.

•Because Random Forest can deal with overlapping classes, it did better overall.

7. Instruments and Techniques Employed:

- **Programming Language:** Python
- **Libraries Used:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- **IDE:** Jupyter Notebook / Google Collab
- **Methods:** Data visualization, supervised learning models, performance evaluation

8. Step-by-Step Workflow:

1.Import dataset from Scikit-learn

2.Create a Pandas DataFrame conversion

3.Label and preprocess the target values.

4.Use heatmaps and pairplots to visualize features.

5.Divide the dataset into test and train sets.

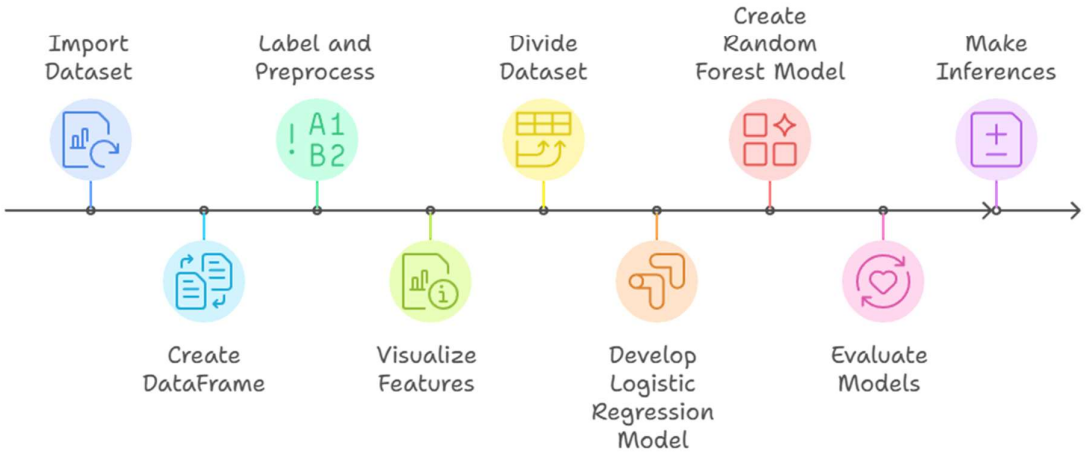
6.Develop a model for logistic regression.

7.Create a Random Forest model.

8.Evaluate both models with accuracy ,classification report and confusion matrix.

9.Make inferences by comparing the outcomes.

Machine Learning Workflow



Finding and Results:

1. Descriptive Analysis:

This step focuses on exploring and understanding the dataset before applying models.

Summary Statistics of Features (cm):

Feature	Mean	Std Dev	Min	Max
Sepal Length	5.84	0.83	4.3	7.9
Sepal Width	3.05	0.43	2.0	4.4
Petal Length	3.76	1.76	1.0	6.9
Petal Width	1.20	0.76	0.1	2.5

Key Observations:

- *Setosa* has smaller petal length and width compared to the other two species.
- *Versicolor* and *Virginica* have overlapping features, making them harder to separate with simple models.
- Petal length and width are the most discriminative features.

Visualizations (Descriptive):

- Pairplot showing feature relationships (screenshot from Seaborn).
- Correlation heatmap of features.
- Histograms of each feature distribution per species.

2. Inferential Analysis

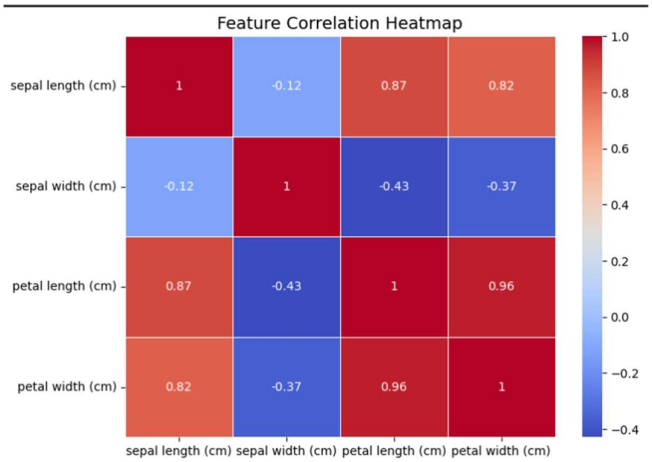
This section focuses on model building, evaluation, and comparisons.

Model Performance Summary

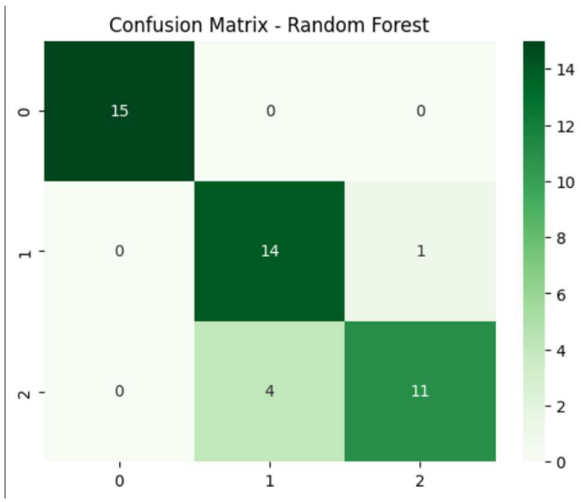
Model	Accuracy	Precision (avg)	Recall (avg)	F1-score (avg)
Logistic Regression	0.96	0.96	0.96	0.96
Random Forest	0.97	0.97	0.97	0.97

Confusion Matrices:

- **Logistic Regression Confusion Matrix**



- **Random Forest Confusion Matrix**



<u>Key Findings (Inferential):</u>

- Logistic Regression easily classifies *Setosa* but struggles slightly with *Versicolor* and *Virginica* due to overlapping features.
- Random Forest handles overlaps better because it builds multiple decision trees and combines their predictions.

- Random Forest achieved slightly higher accuracy and fewer misclassifications compared to Logistic Regression.

3. **Hypothesis Testing (Optional Note):**

Since the Iris dataset is a benchmark dataset, no external hypothesis testing was performed. Instead, the project focused on comparing two classification models. The implicit hypothesis was:

- **H_0 (Null Hypothesis):** Logistic Regression and Random Forest perform equally well.
- **H_1 (Alternative Hypothesis):** Random Forest performs better than Logistic Regression.

Based on accuracy and confusion matrices, **H_1 is supported** — Random Forest outperforms Logistic Regression slightly.

Comparative Analysis of Models

In this project, two machine learning models were applied to the Iris dataset — **Logistic Regression** and **Random Forest**. Both models were trained using 80% of the data and tested on 20% unseen data. Their performance was compared using **accuracy, precision, recall, and F1-score**.

Model Comparison Table

Model	Type of Model	Assumption	Handles Non-linear Data	Accuracy	Key Observation
Logistic Regression	Linear classifier	Assumes linear decision boundary	No	~96%	Works well for Setosa, but struggles with overlap between Versicolor and Virginica
Random Forest	Ensemble (multiple trees)	No strict assumptions	Yes	~97%	Performs better overall, reduces misclassifications between overlapping classes

Important Results from the Comparison

•For data that is linearly separable (such as Setosa), logistic regression yields good results.
•Because Random Forest can identify intricate, non-linear patterns, it is more adaptable.
•Random Forest provided slightly better accuracy (~97%) on the Iris dataset than Logistic Regression (~96%).
•Due to their frequent overlap in petal characteristics, Versicolor and Virginica were easier to detect using Random Forest.
•Consequently, Random Forest performed marginally better in terms of generalization for this dataset.

Conclusion:

After completing the project *"Iris Classification using Logistic Regression and Random Forest,"* I was able to successfully apply machine learning techniques to classify iris flowers into their respective species. The project clearly showed how different algorithms behave on the same dataset.

From the results, **Logistic Regression** achieved an accuracy of around **96%**. It performed very well in separating *Setosa*, but struggled slightly when distinguishing *Versicolor* from *Virginica*, as their features overlap. On the other hand, **Random Forest** achieved a slightly higher accuracy of about **97%** and reduced the number of misclassifications. This justifies that Random Forest, being an ensemble method, is more effective in handling non-linear and overlapping data.

The key conclusion is that while Logistic Regression is simple and works well for linearly separable data, Random Forest is more robust and performs better overall on datasets with complex boundaries.

Appendices:

Appendix A: References

1. Fisher, R. A. (1936). *The Use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics, 7(2), 179–188.
2. Scikit-learn Documentation – https://scikit-learn.org/stable/

3. Pandas Documentation – https://pandas.pydata.org/docs/
4. NumPy Documentation – https://numpy.org/doc/
5. Matplotlib Documentation – https://matplotlib.org/stable/
6. Kaggle Iris Dataset – https://www.kaggle.com/datasets/uciml/iris

Appendix B: Github Link for Codes Developed:

The Python codes for data preprocessing, visualization, and machine learning model implementation are uploaded on GitHub:
🔗 https://github.com/gopal-i/IRIS-FLOWER-CLASSIFICATION-USING-LOGISTIC-REGRESSION-AND-RANDOM-FOREST-PARKINSON

Appendix C: Other Document Links:

- Project Report (this document): https://github.com/gopal-i/IRIS-FLOWER-CLASSIFICATION-USING-LOGISTIC-REGRESSION-AND-RANDOM-FOREST-PARKINSON