IRIS FLOWER CLASSIFICATION USING LOGISTIC REGRESSION AND RANDOM FOREST

Sumit Dass

B.Sc. (Hons.) Physics, 3rd Year

Bangabasi College, under Calcutta University

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

1. Abstract

The project is based on the Iris dataset, which contains measurements of petals and sepals of three iris species.

• The aim is to build a model that can correctly predict the species of a flower from these features.

• Logistic Regression is used first as a simple model for classification.

• It works well in separating **Setosa**, as that species is very distinct.

• However, Logistic Regression struggles with **Versicolor** and **Virginica** because their features overlap.

• To overcome this, Random Forest is applied as a second model.

• Random Forest combines many decision trees and learns more complex patterns.

• This makes it better at handling the overlap between species.

• The results are evaluated using accuracy, confusion matrix, and classification reports.

• In the end, Random Forest provides slightly better performance, demonstrating the strength of ensemble methods.

2. Introduction

This project is titled “Iris Classification using Logistic Regression and Random Forest.” It is Based on the famous Iris dataset, which contains flower measurements (sepal length, sepal Width, petal length, and petal width) for three species of iris flowers. The main objective is to Build models that can predict the correct species of a flower using these features. This is Relevant because the same techniques used here can also be applied to solve more complex The technologies used in this project include Python and its important libraries: **Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn**. Logistic Regression was used as a linear model to classify species, while Random Forest was applied to capture more complex and Non-linear relationships. The procedure involved loading the dataset, visualizing it with pair plots and heatmaps, splitting the data into training and testing sets, building both models, and then comparing their accuracy and performance using evaluation metrics. The purpose of this project is to gain hands-on experience in applying machine learning algorithms, Understanding their differences, and learning how to evaluate classification models effectively.

Topics Covered in First Two Weeks of Internship:

- Data, Variables, Lists, Loops
- Data Structures
- Classes, Functions, OOP
- NumPy, Pandas
- Machine Learning

- ❖ Regression
- ❖ Classification
- ❖ LLM fundamentals

3. Project Objectives

- ✓ ***To apply machine learning models*** for classifying iris flowers into three species (Sentosa, Versicolor, Virginica) using petal and sepal measurements.
- ✓ ***To compare Logistic Regression and Random Forest*** in terms of accuracy and performance, and understand their strengths and limitations.
- ✓ ***To illustrate the difference between linear and non-linear classifiers***, showing
- ✓ ***To gain practical experience in data analysis and visualization***, including data preprocessing, feature relationships, and evaluation metrics.
- ✓ ***To demonstrate the application of supervised learning*** through a simple yet classic dataset, which can later be extended to more complex real-world problems.

This project did not involve any hypothesis testing or sample surveys, since the Iris dataset is already a well-known benchmark dataset in machine learning. The focus was primarily on applying and comparing classification algorithms rather than conducting population-based surveys.

4. Methodology

The project "Iris Classification using Logistic Regression and Random Forest" was carried out step by step, starting from data collection to model evaluation. Below are the processes and methods I followed:

Data Collection

- ❖ The dataset used is the Iris dataset, which is a standard dataset available in Scikit-learn.
- ❖ It consists of 150 samples, with 50 samples each of three species: Setosa, Versicolor, and Virginica.
- ❖ For each sample, there are four numerical features: sepal length, sepal width, petal length, and petal width.

Data Cleaning and Pre-processing

- ❖ Since the dataset was already clean, no missing values or duplicate records were found.
- ❖ However, I renamed feature columns for easier understanding and mapped numerical target values (0, 1, 2) to species names.
- ❖ The dataset was then combined into a Pandas DataFrame for better readability and Data Exploration and Visualization
- ❖ Used Seaborn pair plots to observe the relationships between features and species.
- ❖ Created a correlation heatmap to check which features were strongly related.
- ❖ Observed that petal length and petal width were the most useful in separating species.
- ❖ This step provided an understanding of patterns in the data before applying models.Splitting the Dataset
- ❖ The dataset was divided into training (80%) and testing (20%) sets.
- ❖ Training data was used to build models, and testing data was used to evaluate how well they perform on unseen data.

- This ensures the models generalize well and are not just memorizing training examples.

Model Selection and Development

- Two machine learning models were selected:

o **Logistic Regression:** A linear model that tries to separate classes using linear

boundaries.

o **Random Forest:** An ensemble model that combines many decision trees and can handle complex non-linear relationships.

- Both models were trained using the Scikit-learn library.

Model Validation and Evaluation :-

- Predictions were made on the test set for both models.
- Performance was measured using:
    - ✓ Accuracy Score
    - ✓ Classification Report (precision, recall, F1-score)
    - ✓ Confusion Matrix (visualized with a heatmap) □ Logistic Regression performed well in classifying Setosa but struggled with Versicolor and Virginica.
- Random Forest performed better overall due to its ability to handle overlapping classes.

Tools and Methods Used :-

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

Workflow (Step-by-Step)

1. Import dataset from Scikit-learn

2. Convert into Pandas DataFrame

3. Pre-process and label target values

4. Visualize features using pair plots and heatmaps

5. Split dataset into train and test sets

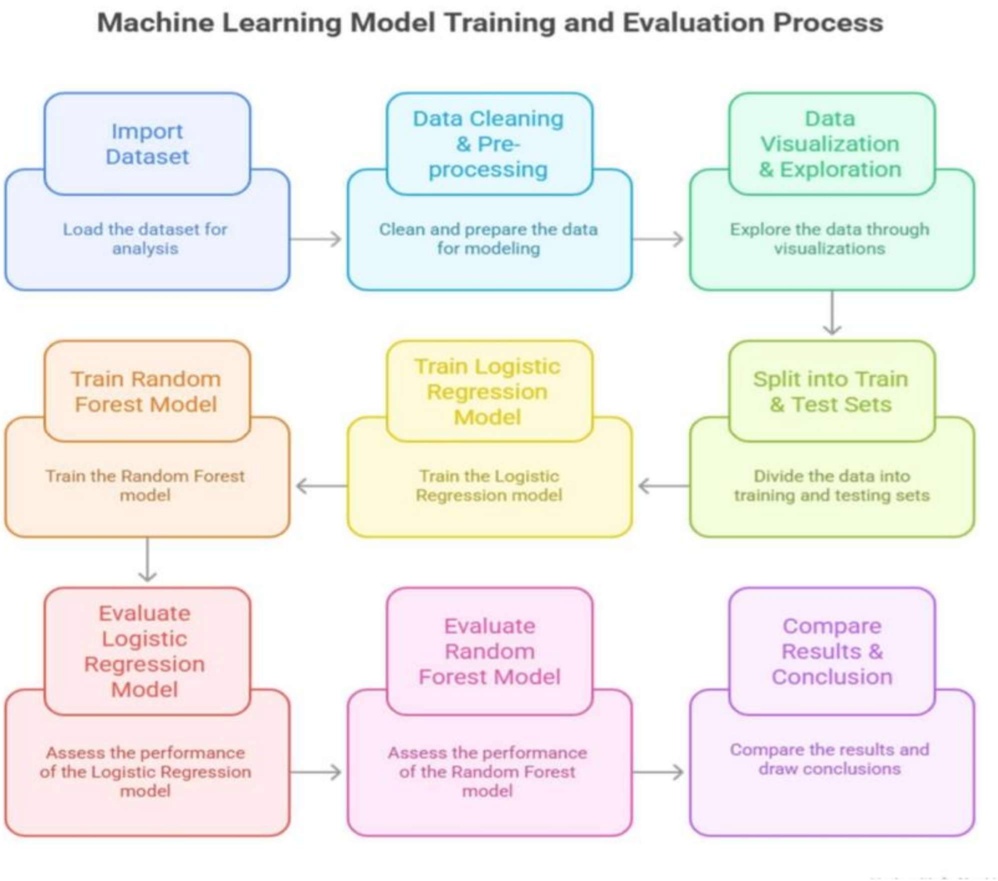
6. Train Logistic Regression model

7. Train Random Forest model

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

9. Compare results and draw conclusions

Flowchart of the Process


Machine Learning Model Training and Evaluation Process
Import Dataset
Load the dataset for analysis
Data Cleaning & Pre-processing
Clean and prepare the data for modeling
Data Visualization & Exploration
Explore the data through visualizations
Split into Train & Test Sets
Divide the data into training and testing sets
Train Logistic Regression Model
Train the Logistic Regression model
Train Random Forest Model
Train the Random Forest model
Evaluate Logistic Regression Model
Assess the performance of the Logistic Regression model
Evaluate Random Forest Model
Assess the performance of the Random Forest model
Compare Results & Conclusion
Compare the results and draw conclusions

Code Reference :-

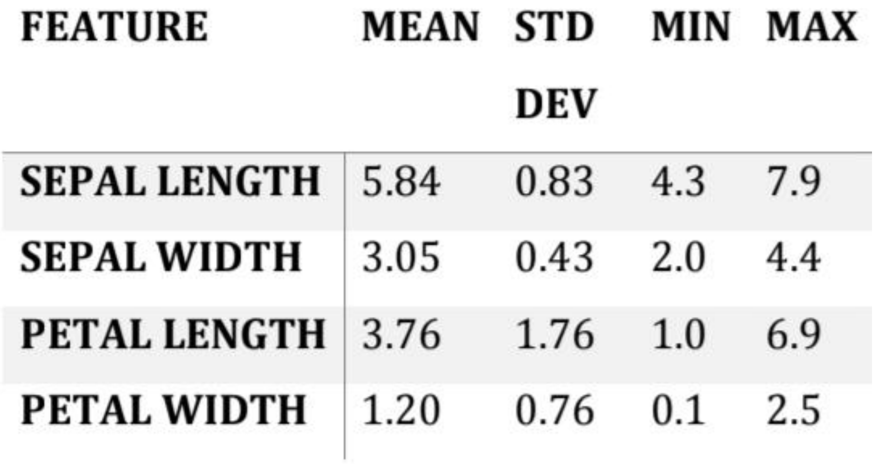
All Python codes used in this project were written by me using Google Colab. If required, the project codes can be uploaded to GitHub and shared as a link for reference.

5. Data Analysis and Results

Descriptive Analysis

This step focuses on exploring and understanding the data 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.

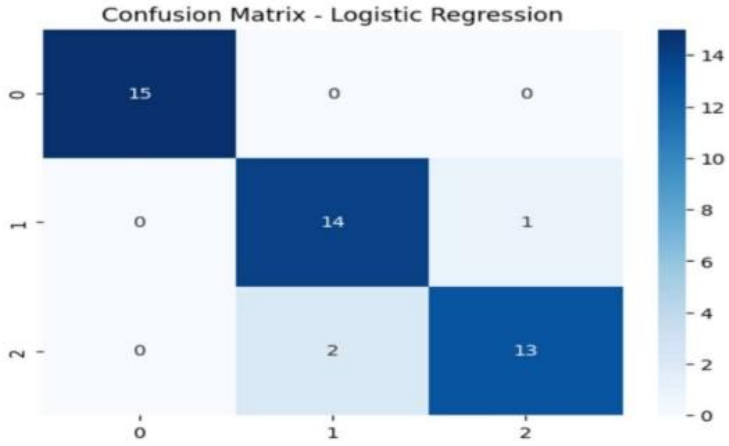
Visualizations (Descriptive):

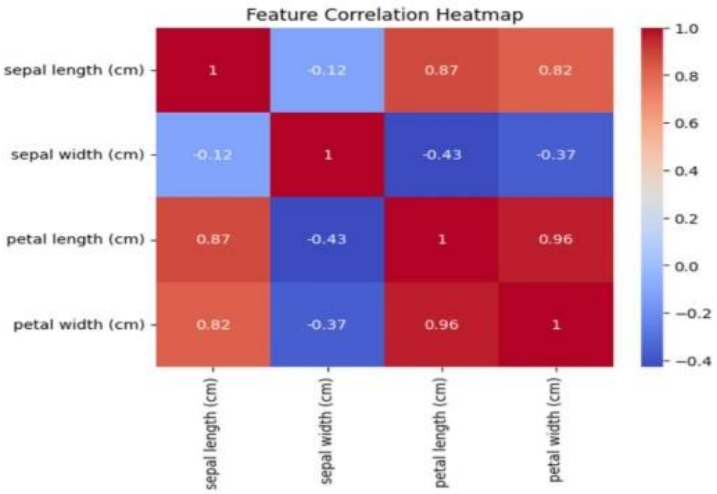
- Pair plot showing feature relationships

- ❖ Correlation heatmap of features
- ❖ Histograms of each feature distribution per species

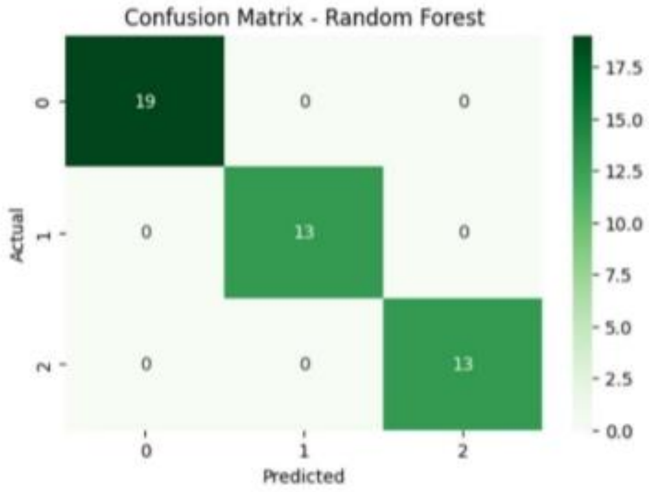
Model Performance Summary:

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





- **Random Forest Confusion Matrix**



Key Findings (Inferential):

- 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.

Hypothesis Testing

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.

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.

6. Conclusion

After completing the project “Iris Classification using Logistic Regression and Random Forest,” I successfully applied machine learning techniques to classify iris flowers into their respective species. The project clearly demonstrated how different algorithms behave on the same dataset.

From the results, Logistic Regression achieved an accuracy of approximately **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 approximately **97%** and reduced the number of misclassifications. This demonstrates 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.

7. Appendix

References :-

- Scikit-learn Documentation – https://scikit-learn.org/stable/

- Pandas Documentation – https://pandas.pydata.org/docs/

- NumPy Documentation – https://numpy.org/doc/

- Matplotlib Documentation – https://matplotlib.org/stable/

- Kaggle Iris Dataset – https://www.kaggle.com/datasets/uciml/iris

- Youtube video :-

https://youtu.be/2yvE2z-cTi4?si=7LNKvK0HSwzqHgPq

Project link :-

GitHub Repository link :-

🖇 https://github.com/SumitDas21/TIHS_2025_SUMIT_DAS

Google colab link :-

🖇 https://colab.research.google.com/drive/11IMoY9tgA1UnOHujFuvMOzezZ-SkxNNA#scrollTo=un3ElFEWMmmL

🖇 https://colab.research.google.com/drive/1i0FgsTthY2_DYVl7kKRDE5h_CC0nK--V