Classification with Logistic Regression and Random Forest

Sayantani Porel

Course-Statistics Honours , Section 1

Sister Nivedita GOVT. General Degree College for Girls

Period of Internship:25th August2025-19th September2025

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

1. Abstract

This project explores two datasets: the Iris flower dataset and the Parkinson’s disease dataset. The Iris dataset, introduced by R.A.Fisher, contains measurements of Iris flowers to classify species. The Parkinson’s dataset, developed by Max Little, includes voice features to tell apart healthy individuals and those with the disease. Both datasets undergo preprocessing, exploratory data analysis, visualization and machine learning models like Logistic Regression, Decision Trees, and Random Forest. The study shows how data science techniques can find patterns, create predictive models and support decision-making in biological and medical fields.

2. Introduction

This project looks at the Iris flower dataset and the Parkinson's disease dataset to show how data science is used in biology and medicine. It highlights how machine learning can help classify plant species and aid in the early detection of neurological disorders. The tools used include Python, Jupyter Notebook, Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. The background review covers R.A. Fisher's work on the Iris dataset and Max Little's input on biomedical voice data. The process involves data cleaning, exploratory data analysis (EDA), visualization, and using machine learning methods like Logistic Regression, Decision Trees, and Random Forest. The goal of the project is to understand how organized datasets can be modeled to find patterns, improve predictions, and assist in decision-making.

Training topics:

- Basics of Python programming
- Data handling with Pandas and NumPy
- Introduction to Jupyter Notebook
- Exploratory Data Analysis
- Basics of machine learning
- Data preprocessing and cleaning techniques
- Evaluation metrics for classification

3. Project Objective

- To illustrate the application of data preprocessing, visualization, and analysis techniques on real-world datasets.
- To build and compare machine learning models for classification in both biological (Iris) and medical (Parkinson's) domains.

- To demonstrate how statistical methods and hypothesis testing can validate patterns observed in data.
- To show the practical relevance of AI and data science in supporting decision-making for scientific and healthcare purposes.
- To enhance understanding of structured datasets and predictive modeling as part of data science training.

4. Methodology

1.Data Collection

Iris Dataset: Collected from the UCI Machine Learning Repository; contains 150 flower samples across three species (Setosa, Versicolor, Virginica), each with 4 features (sepal length, sepal width, petal length, petal width).

Parkinson’s Dataset: Also sourced from the UCI Repository; contains 195 records with 23 voice-related features from both healthy individuals and Parkinson’s patients.

Survey: No separate survey was conducted for this project. The datasets used are publicly available benchmark datasets, ensuring authenticity and standardization

2. Data Preprocessing and Cleaning

Checked for missing values, duplicates, and inconsistencies.

Renamed/standardized column names for readability.

Normalized/standardized numerical features where necessary (important for models like Logistic Regression and SVM).

Encoded categorical variables (where applicable)

3. Exploratory Data Analysis (EDA)

Used Matplotlib and Seaborn for visualization.

Created histograms, box plots, pair plots, and heatmaps to identify relationships among variables.

In Iris dataset: observed how petal length and width are strong discriminators of species.

In Parkinson's dataset: analyzed jitter, shimmer, and frequency features for separation between patients and healthy subjects.

4. Tools and Technologies Used

Python (3.x)

Jupyter Notebook (for interactive coding)

Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

GitHub (for code sharing and version control)

5. Modeling and Analysis

Train-Test Split: Both datasets were split into 80% training and 20% testing.

Models Applied:

Logistic Regression

Decision Tree

Random Forest

Support Vector Machine (for Parkinson's dataset)

Model Selection:

Initial models were trained and compared using accuracy, precision, recall, and F1-score.

Cross-validation was used to reduce bias and overfitting.

Results:

Iris dataset achieved >95% accuracy with Random Forest.

Parkinson’s dataset performed best with SVM and Random Forest (>90% accuracy).

6. Steps of Work Done

1. Dataset collection (from UCI ML Repository).

2. Understanding dataset structure (features, target variable).

3. Data cleaning (handling missing/duplicate values).

4. Data preprocessing (scaling, encoding).

5. Exploratory Data Analysis (visualizations, feature correlations).

6. Splitting into train-test sets.

7. Applying ML models (Logistic Regression, Decision Tree, Random Forest, SVM).

8. Evaluating performance (accuracy, confusion matrix, classification report).

9. Comparing models to select best-performing one.

10. Documenting findings and conclusions.

5. Data Analysis and Results

1. Descriptive Analysis

Iris Dataset

Feature	Mean	Median	Min	Max	Std. Dev.
Sepal Length	5.84	5.8	4.3	7.9	0.83
Sepal Width	3.05	3.0	2.0	4.4	0.43
Petal Length	3.76	4.35	1.0	6.9	1.76
Petal Width	1.20	1.3	0.1	2.5	0.76

Key Observations:

- Setosa species clearly separates from the other two by petal length and width.
- Overlap occurs between Versicolor and Virginica in sepal dimensions.

Parkinson’s Dataset

Feature	Mean	Median	Min	Max	Std. Dev.
MDVP:Fo(Hz)	154.23	148.0	88.3	260.1	41.4
MDVP:Fhi(Hz)	197.1	166.1	102.1	592.0	91.5
MDVP:Flo(Hz)	116.3	110.0	65.4	239.2	43.3
MDVP:Jitter(%)	0.0078	0.006	0.001	0.034	0.0048
MDVP:Shimmer(dB)	0.038	0.032	0.009	0.119	0.022

Key Observations:

Higher jitter and shimmer values are more common in Parkinson’s patients.

Distribution shows clear differences between healthy vs. affected groups.

2. Inferential Analysis

Hypothesis Testing (Parkinson’s Dataset)

Null Hypothesis (H_0): There is no significant difference in jitter between healthy individuals and Parkinson’s patients.

Alternative Hypothesis (H_1): There is a significant difference.

Test Used: Independent sample t-test.

Result: p-value < 0.05 → Reject H_0.

Conclusion: Jitter values significantly differ, confirming its role as a biomarker.

3. Machine Learning Models – Comparative Analysis

Iris Dataset

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	96%	0.95	0.95	0.95
Decision Tree	94%	0.94	0.94	0.94
Random Forest	97%	0.97	0.97	0.97
SVM	96%	0.96	0.96	0.96

Random Forest performed best with 97% accuracy.

Parkinson’s Dataset

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	87%	0.86	0.87	0.86
Decision Tree	84%	0.83	0.84	0.83
Random Forest	91%	0.90	0.91	0.90
SVM	93%	0.92	0.93	0.92

SVM performed best with 93% accuracy, followed closely by Random Forest.

4. Flow of Analysis

Start

│

Collect Datasets

│

Data Cleaning

│

Data Preprocessing

│

Exploratory Data Analysis

│

Train-Test Split (80-20)

│

Apply ML Models (LR, DT, RF, SVM)

│

Evaluate Model Accuracy

│

Compare & Select Best Model

│

Conclusion

│

End

- Synthetic data was successfully generated and visualized, replicating the structure of the original dataset. It can be used for analysis and model testing without exposing real data.

6. Conclusion

The project showed how data science can be applied to biological and medical datasets.

In the Iris dataset, petal length and width were the most important features, and Random Forest gave the highest accuracy (97%).

In the Parkinson’s dataset, jitter and shimmer were significant indicators, and SVM performed best with 93% accuracy.

Overall, the study highlights the role of preprocessing, analysis, and model selection in building accurate predictive systems. Future work can involve larger datasets, advanced models like deep learning, and real-time healthcare applications.