{AUTUMN INTERNSHIP PROJECT REPORT FORMAT}

**Project Title :Classification with logistic regression and random forest for Parkinsons**

Debanjan Samanta

BTech (CSE) RCC Institute of Information Technology

Period of Internship: 25th August 2025 - 19th September 2025 (Do not change the dates)

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

**1. Abstract**

This project focuses on applying machine learning techniques to classify Parkinson’s disease patients using biomedical voice measurements. The work is inspired by a given Colab notebook originally based on the Iris dataset and extended to the Parkinson’s dataset from the UCI Machine Learning Repository. The dataset contains biomedical voice features of 195 subjects, including both Parkinson’s patients and healthy controls. The main objective was to preprocess the dataset, perform exploratory data analysis (EDA), and apply classification models such as Logistic Regression and Random Forest. Accuracy, precision, recall, and F1-scores were used to evaluate performance. Logistic Regression achieved ~87% accuracy, while Random Forest achieved ~94%, showing superior performance. The results highlight the importance of machine learning in supporting medical diagnosis. The project demonstrates how supervised learning can effectively distinguish between healthy individuals and Parkinson’s patients.

**2. Introduction**

Parkinson’s disease is a progressive neurological disorder that severely affects movement and speech. Traditional diagnosis relies on clinical observation, which may lead to late or subjective detection. Machine learning provides a way to support diagnosis using biomedical data, enabling early intervention and accurate predictions.

This project uses the **Parkinson’s dataset** from the UCI Machine Learning Repository. It contains voice features derived from sustained phonations. The dataset was selected because voice disorders are common in Parkinson’s disease, making it a good candidate for ML-based classification.

The project is based on a machine learning pipeline originally applied to the Iris dataset. The same process (data loading, preprocessing, visualization, train-test split, model training, and evaluation) was replicated and adapted for the Parkinson’s dataset.

**Relevance**:

* Shows the application of ML in healthcare.
* Highlights the importance of feature preprocessing and scaling.
* Demonstrates comparison between linear and ensemble ML models.

**Technology involved**:

* **Python 3**, Google Colab, **scikit-learn**, **pandas**, **seaborn**, **matplotlib**.

**Topics covered during internship training (first two weeks):**

1. Basics of Python and data handling using pandas.
2. Exploratory data analysis (EDA).
3. Data visualization with matplotlib and seaborn.
4. Fundamentals of machine learning (classification models).
5. Train-test split and evaluation metrics.
6. Logistic Regression and Random Forest algorithms.
7. GitHub usage for code storage and version control.

**3. Project Objectives**

The main objectives of the project were:

* To understand and preprocess the Parkinson’s dataset for machine learning applications.
* To perform exploratory data analysis (EDA) and visualize feature relationships.
* To apply Logistic Regression and Random Forest models for classification.
* To evaluate and compare the performance of the models using accuracy, confusion matrices, and classification reports.
* To demonstrate how an ML pipeline can be adapted from a simple dataset (Iris) to a real-world medical dataset.

**4. Methodology**

**Step 1: Data Collection**

* The dataset was obtained from the UCI Machine Learning Repository.
* The file parkinsons.data was loaded into Colab using pandas.

**Step 2: Data Preprocessing**

* The name column (identifier) was dropped.
* The status column was set as the target variable (1 = Parkinson’s, 0 = Healthy).
* Features were standardized using StandardScaler to ensure comparability.

**Step 3: Exploratory Data Analysis (EDA)**

* Summary statistics were generated.
* Correlation heatmap was plotted to identify strongly correlated features.
* Pairplots were created for selected features to visualize separation between classes.

**Step 4: Train-Test Split**

* Data was split into 70% training and 30% testing sets.
* Stratified sampling was used to maintain class balance.

**Step 5: Model Training**

* **Logistic Regression**: A linear model used with max\_iter=200.
* **Random Forest**: An ensemble model trained with 100 decision trees.

**Step 6: Model Evaluation**

* Accuracy, classification reports, and confusion matrices were generated.
* Results were visualized using seaborn heatmaps.

**Step 7: Code Management**

* The Python notebook was modified in Colab.
* Codes can be uploaded to GitHub for version control and sharing.

**Flow of Work:**

Data Collection → Data Preprocessing → EDA → Train-Test Split → Model Training → Model Evaluation → Result Comparison

**5. Data Analysis and Results**

**5.1 Descriptive Analysis**

* The dataset contained **195 samples** and **22 features** (after removing the name column).
* Target distribution: Parkinson’s patients (147), healthy controls (48).
* Some features were highly correlated (e.g., jitter and shimmer).

**5.2 Visualizations**

* **Heatmap** showed strong correlations among voice measures.
* **Boxplots** highlighted differences in jitter/shimmer between classes.
* **Pairplots** (subset of features) showed partial separation between healthy and PD cases.

**5.3 Model Results**

| **Model** | **Accuracy** | **Precision (PD)** | **Recall (PD)** | **F1-Score (PD)** |
| --- | --- | --- | --- | --- |
| Logistic Regression | ~87% | 0.89 | 0.94 | 0.91 |
| Random Forest | ~94% | 0.95 | 0.96 | 0.95 |

**Confusion Matrices**

* Logistic Regression misclassified some healthy cases as PD.
* Random Forest had fewer errors and was more balanced.

**6. Conclusion**

This project demonstrated how a machine learning workflow can be applied to a medical dataset for disease classification. Logistic Regression performed reasonably well, but Random Forest achieved higher accuracy and more balanced performance.

**Key takeaways:**

* Data preprocessing and scaling are critical for model performance.
* Random Forest outperforms Logistic Regression in handling complex feature interactions.
* Machine learning holds potential for aiding early Parkinson’s detection.

**Recommendations for future work:**

* Hyperparameter tuning of Random Forest.
* Exploring other models like SVM, Gradient Boosting, or Neural Networks.
* Feature selection to reduce redundancy.
* Testing on larger datasets for generalization.

**7. Appendices**

**Appendix A: References**

1. UCI Machine Learning Repository – Parkinson’s Dataset: https://archive.ics.uci.edu/dataset/174/parkinsons
2. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” JMLR, 2011.
3. James et al., *An Introduction to Statistical Learning*, Springer, 2013.