GCET G H PATEL COLLEGE OF ENGINEERING & TECHNOLOGY	G H Patel College of Engineering and Technology	CVM UNIVERSITY

COMPUTER ENGINEERING DEPARTMENT

Project Report

on

Parkinsons-Disease-Predictor

Submitted By

Name of Student-1: Akshat Mistry

Enrolment Number: 12202040501006

Name of Student-2: Aryika Patni

Enrolment Number: 12202040501009

Guided By: Dr. Priyang Bhatt

A.Y. 2024-25 EVEN TERM

Objective

The primary objective of this project is to develop a machine learning model that predicts Parkinson’s disease using vocal and signal features extracted from voice recordings. By analyzing attributes such as frequency (MDVP:Fo), jitter, shimmer, and pitch period entropy (PPE), the model aims to classify individuals as healthy or affected, aiding in early diagnosis. The project also seeks to deploy an accessible web application for real-time predictions.

Dataset Used

For this project, we utilized the **UCI Parkinson’s Disease Classification Dataset**, available at: UCI Machine Learning Repository

Key Features in Dataset

- MDVP:Fo(Hz): Average vocal fundamental frequency
- MDVP:Jitter(%): Variation in fundamental frequency
- MDVP:Shimmer: Variation in amplitude
- NHR: Noise-to-harmonics ratio
- HNR: Harmonics-to-noise ratio
- RPDE: Recurrence period density entropy
- DFA: Detrended fluctuation analysis
- spread1, spread2: Nonlinear measures of fundamental frequency variation
- PPE: Pitch period entropy

Target Variable:

Status: Binary classification

- 0: Healthy
- 1: Parkinson’s Disease

Model Chosen:

After exploring various classification algorithms, we selected the Random Forest Classifier for its high accuracy and ability to handle complex feature interactions. The model was trained on the dataset to identify patterns in voice measures, achieving robust performance in distinguishing between healthy and Parkinson’s-affected individuals.

How Random Forest Classifier Works

Step 1: Select random samples from the dataset.

Step 2: Construct a decision tree for each sample, predicting an outcome.

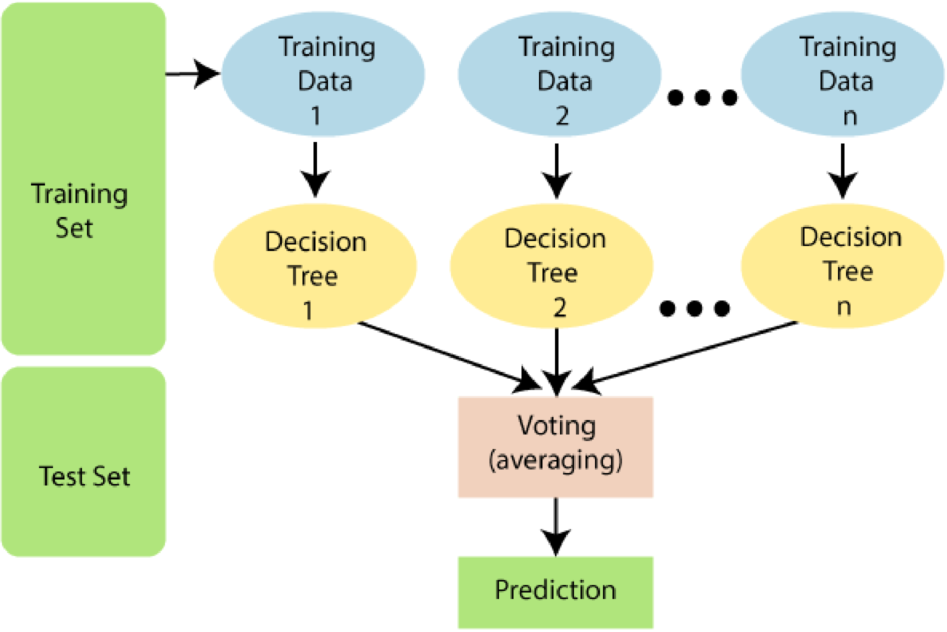
Step 3: Collect votes from each decision tree for the predicted outcomes.

Step 4: Choose the outcome with the most votes as the final prediction.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Why Random Forest?

- Effectively manages numerical features with varying scales
- Mitigates overfitting through ensemble averaging of multiple trees
- Delivers superior accuracy compared to single decision tree models



Performance Metrics

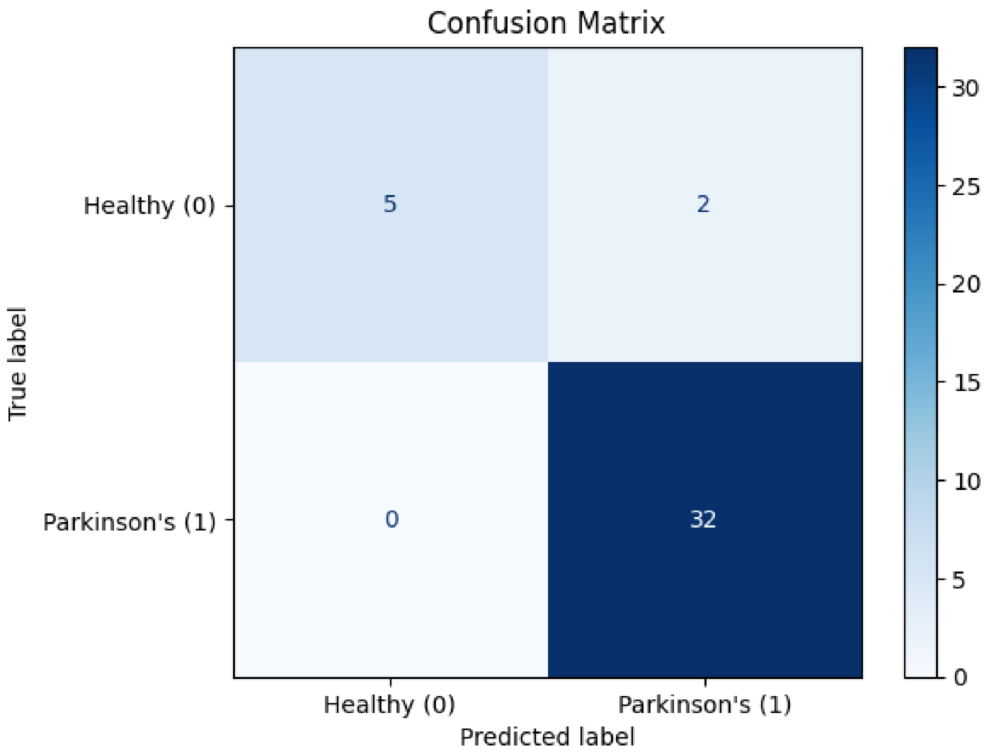
To assess the model's effectiveness, we employed the following metrics:

- Accuracy: Percentage of correct predictions
- Precision, Recall & F1-Score: Detailed evaluation of classification performance
- Confusion Matrix: Visualizes prediction distribution across classes

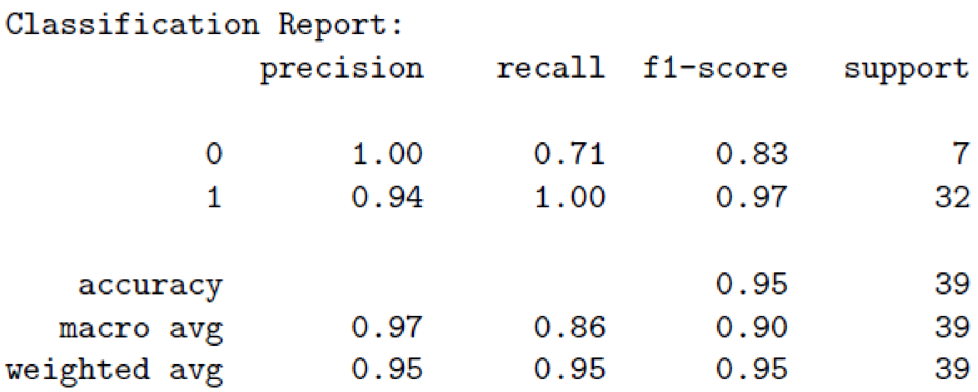
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN}, \qquad F1\ score = \frac{2 * Precision * Recall}{Precision + \ Recall}$$

Confusion Matrix



Model Performance



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.71      0.83         7
           1       0.94      1.00      0.97        32

    accuracy                           0.95        39
   macro avg       0.97      0.86      0.90        39
weighted avg       0.95      0.95      0.95        39
```

Challenges & Learnings

Challenges Faced:

- Feature Selection: Identifying the most impactful vocal features for classification
- Class Imbalance: Addressing the skewed distribution (147 Parkinson's vs. 48 Healthy)
- Model Deployment: Ensuring the model and scaler work seamlessly in a web environment

Key Learnings:

- Feature importance analysis (e.g., PPE, MDVP:Fo) enhanced model interpretability
- Random Forest's robustness improved performance despite limited data
- Visualization tools like confusion matrices aided in understanding model behavior

Conclusion

This project successfully demonstrates the application of machine learning to predict Parkinson's disease using voice measures, achieving an accuracy of **94.87%** with a Random Forest Classifier. The deployed Streamlit web app provides an intuitive interface for real-time diagnosis, making it a valuable tool for early detection. Future enhancements could include integrating larger datasets or exploring deep learning techniques for even higher precision.

Tools Used: Python, Streamlit, Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn

Repository: Github Repository

References:

1. T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins and R. H. Ghomi, "Parkinson's Disease Diagnosis Using Machine Learning and Voice," *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Philadelphia, PA, USA, 2018, pp. 1-7, doi: 10.1109/SPMB.2018.8615607.

2. Ilias Tougui, Mehdi Zakroum, Ouassim Karrakchou, Mounir Ghogho, "PD-VOST: Parkinson's Disease Voice Spectrogram Transformer", *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1-5, 2025.