Title: Detecting Parkinson's Disease Using Vocal Features

1. Introduction

Parkinson's disease (PD) is a progressive neurological disorder that impairs movement and coordination. Early diagnosis is crucial for managing the disease effectively, but it remains a significant challenge. This project aims to develop an automated system using data science techniques in Python to accurately detect Parkinson's disease based on vocal features.

2. Objectives

- Develop an automated system for detecting Parkinson's disease using vocal features.
- Preprocess the dataset of voice samples to extract relevant features such as pitch, frequency, and formants.
- Train and evaluate machine learning models to distinguish between individuals with and without Parkinson's disease based on vocal features.
- Achieve high accuracy in detecting Parkinson's disease through rigorous testing and validation of the models.

3. Literature Review

Numerous studies have explored the use of vocal features for diagnosing Parkinson's disease. Changes in speech are one of the early symptoms of PD, and these changes can be detected through careful analysis of voice samples. Machine learning models have shown promise in distinguishing between healthy individuals and those with PD by analyzing these vocal features. Feature extraction techniques such as Mel-frequency cepstral coefficients (MFCC), jitter, and shimmer are commonly used in these studies.

4. Methodology

4.1. Data Collection

We will use a publicly available dataset of voice samples from individuals with and without Parkinson's disease. The dataset includes various vocal recordings and corresponding labels indicating the presence or absence of the disease.

4.2. Data Preprocessing

- **Feature Extraction**: Extract key vocal features such as pitch, frequency, formants, MFCC, jitter, and shimmer from the voice samples.
- **Normalization**: Normalize the extracted features to ensure consistency and improve model performance.
- **Labeling**: Ensure that each voice sample is accurately labeled as either PD or non-PD.

4.3. Model Selection and Training

- **Machine Learning Models**: Experiment with various machine learning models including Support Vector Machines (SVM), Random Forests, and Gradient Boosting.
- **Training**: Split the dataset into training, validation, and test sets. Train the models using the training set and validate their performance on the validation set.
- **Hyperparameter Tuning**: Perform hyperparameter tuning to optimize model performance.

4.4. Model Evaluation

- **Accuracy**: Measure the accuracy of the models in detecting Parkinson's disease.
- **Precision and Recall**: Evaluate precision and recall to understand the models' performance in terms of false positives and false negatives.
- **ROC-AUC Curve**: Plot the ROC-AUC curve to evaluate the models' ability to distinguish between PD and non-PD samples.

5. Expected Outcomes

- A machine learning model capable of accurately detecting Parkinson's disease based on vocal features.
- An automated system that can analyze voice samples and provide a diagnosis with high accuracy.
- Comprehensive evaluation metrics to assess the performance of the models.

6. Conclusion

This project aims to develop a reliable and efficient system for detecting Parkinson's disease using vocal features. By leveraging advanced machine learning techniques and feature extraction methods, we aim to provide an accurate diagnostic tool that can assist in the early detection and management of Parkinson's disease.