Prediction of Parkinson

**Algorithms Used:**

* Logistic Regression
* XGBClassifier
* Support Vector Classifier
* Random Forest

**Colab Notebook:** <https://colab.research.google.com/drive/101GJi0iAiYk2b9oTbZWi3pOnT08lxztG?usp=sharing>

**Source:**

<https://www.geeksforgeeks.org/parkinson-disease-prediction-using-machine-learning-python/>

**Dataset:**

<https://www.kaggle.com/code/naveenkumar20bps1137/parkinson-s-disease-detection-using-ml-algorithms/input>

**Rows, columns:** (195, 23)

**Initial Data Exploration Summary**

* **Objective**: Develop a machine learning model to predict Parkinson's disease based on voice recordings.
* **Dataset Overview**:
* Biomedical voice measurements from 31 subjects, including 23 with Parkinson's disease (PD).
* Features include acoustic markers related to frequency, amplitude, noise, and nonlinear dynamics.
* status column indicates health status:
* 1: Parkinson's disease
* 0: Healthy

**Key Features in Dataset**

**Frequency Variations:**

* **MDVP:Fo(Hz)**: Average fundamental frequency
* **MDVP:Fhi(Hz)**: Maximum fundamental frequency
* **MDVP:Flo(Hz)**: Minimum fundamental frequency

**Jitter Measures (Cycle Variability):**

* **MDVP:Jitter(%)**, **MDVP:Jitter(Abs)**, **MDVP:RAP**, **MDVP:PPQ**, **Jitter:DDP**

**Shimmer Measures (Amplitude Variability):**

* **MDVP:Shimmer**, **MDVP:Shimmer(dB)**, **Shimmer:APQ3**, **Shimmer:APQ5**, **MDVP:APQ**, **Shimmer:DDA**

**Noise Ratios:**

* **NHR**: Noise-to-harmonics ratio
* **HNR**: Harmonics-to-noise ratio

**Nonlinear Dynamics:**

* **RPDE**, **D2**, **DFA**

**Fundamental Frequency Variations:**

* **spread1**, **spread2**, **PPE**

**Analysis and Techniques**

**Data Preprocessing:**

* Removed the status column for feature selection.

**Visualization:**

* **t-SNE**: Applied for dimensionality reduction to visualize relationships in high-dimensional data.

**Classification Models:**

* **K-Nearest Neighbors (K-NN)**:
* Predicts based on labels of k nearest neighbors.
* Optimized hyperparameters using GridSearchCV.
* **Scoring Metric**: Recall, to maximize identification of true positive cases.
* **Logistic Regression**:
* Used for feature importance analysis via coefficients.
* Larger coefficients indicate more significant features.
* **Support Vector Machines (SVM)**:
* Separates classes using hyperplanes with maximum margin.
* Parameters:
* C: Regularization (controls margin width).
* kernel: Transformation function (e.g., linear, RBF).
* gamma: Radius of influence for support vectors.
* degree: Degree for polynomial kernels.
* **Random Forest**:
* Handles complex patterns with high accuracy.
* Supports interpretability of feature importance.

**Interpretability and Insights**

* **Machine Learning Interpretability**:
* Techniques like t-SNE and feature importance scores reveal insights into model decisions.
* Logistic regression coefficients indicate the influence of features on Parkinson's detection.
* **SVM Insights**:
* **Pros**: Effective for high-dimensional data, handles non-linear separability, robust to overfitting.
* **Cons**: Computationally expensive, harder to interpret, requires careful hyperparameter tuning.
* **Random Forest Benefits**:
* High accuracy, interpretable results, versatile for various tasks.

**Algorithms Model and Accuracy**

|  |  |
| --- | --- |
| **Models** | **Scores** |
| Random\_forest | 0.8983 |
| Support Vector Machine | 0.8323 |
| K-Nearest Neighbors | 0.8644 |
| Logistic Regression | 0.8474 |

Best: **Random Forest**