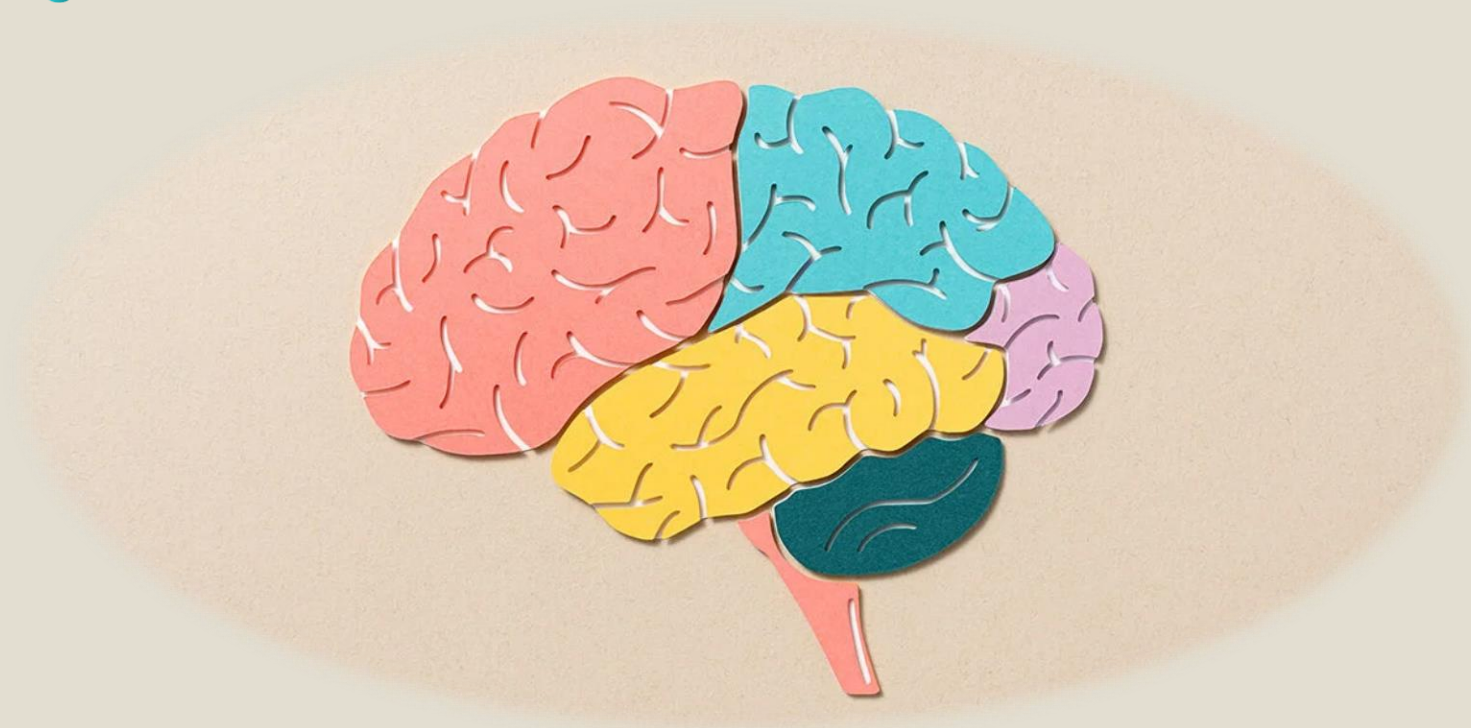
DETECTION OF PARKINSON'S DISEASE

USING

MACHINE LEARNING

An Analysis of Symptoms, Preprocessing, and Model Selection



Sreyashree Pati

What is Parkinson’s Disease

•**Definition**: Parkinson’s Disease (PD) is a progressive disorder of the nervous system affecting movement, often leading to tremors and stiffness.

•**Characteristics**: Tremors, slowed movement (bradykinesia), rigid muscles, speech problems, and impaired balance.

•**Impact**: Affects millions globally and is the second most common neurodegenerative disorder.

Symptoms of Parkinson’s Disease

•**Motor Symptoms**:

- Tremors (shaking, usually begins in a limb, often hand or fingers).
- Bradykinesia (slowness of movement that reduces spontaneity of movement).
- Rigid muscles (muscle stiffness that can limit movement and cause pain).

•**Non-Motor Symptoms**:

- Cognitive impairment, sleep disturbances, and emotional changes (like depression).

Understanding the Perspective

- **Goal of Analysis**: Early detection of Parkinson’s disease is crucial for improving quality of life. Machine learning models can help in predicting the onset of PD based on specific clinical and vocal features.
- **Importance of Accuracy**: A highly accurate prediction model helps in early diagnosis, leading to earlier intervention and better management of symptoms

Problem Statement

- **Research Problem**: To create a model that can accurately predict Parkinson's disease based on medical and vocal features of patients.
- **Dataset**: The dataset consists of clinical features of people with and without Parkinson’s disease.
- **Objective**: Predict if a patient has Parkinson’s disease based on these features, using various machine learning algorithms.

Data Preprocessing

•**Steps Taken**:

- **Data Cleaning**: Removed missing values, handled outliers, and ensured dataset consistency.
- **Normalization**: Applied scaling techniques to normalize feature ranges, improving algorithm performance.
- **Feature Selection**: Used methods like PCA or correlation analysis to select the most important features.
- **Splitting the Data**: Divided the data into training and testing sets (e.g., 80% training, 20% testing) for model evaluation.

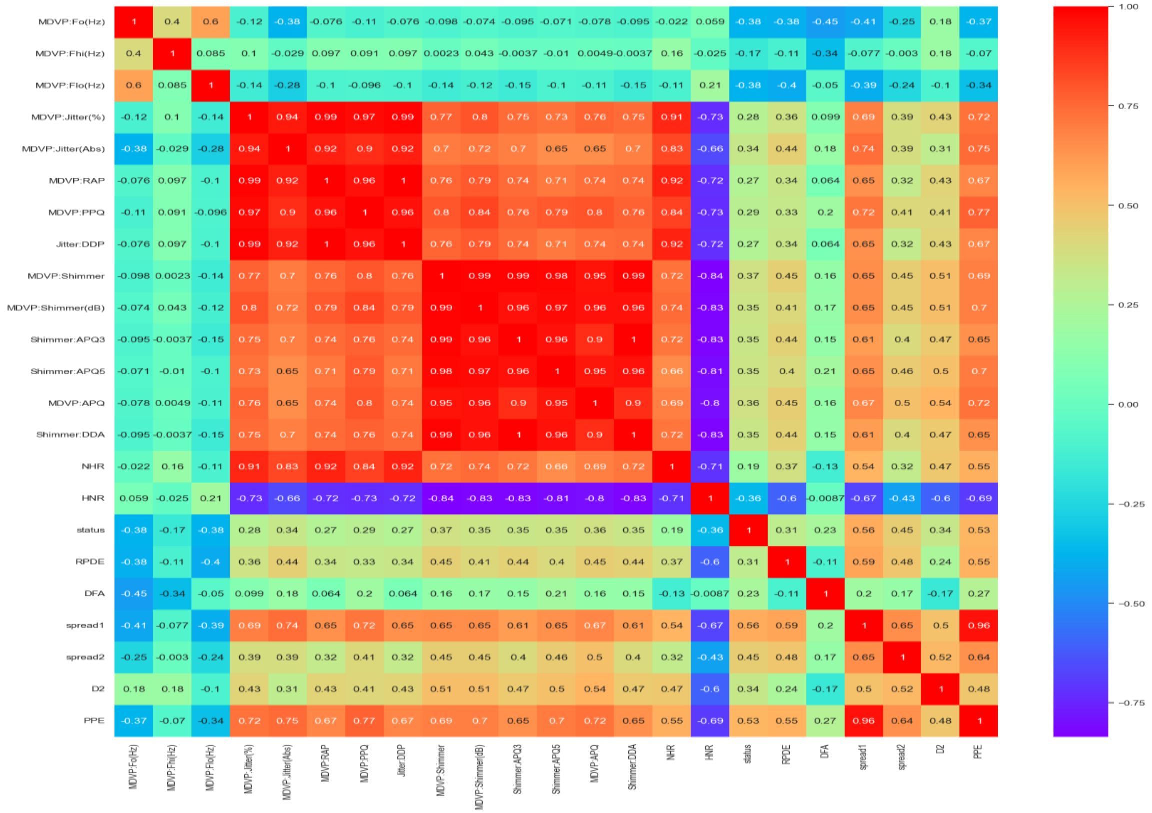
Data Visualization and Observations

1.Distribution Plot (Displot): Visualizes the distribution of key features, helping identify data spread and potential outliers.

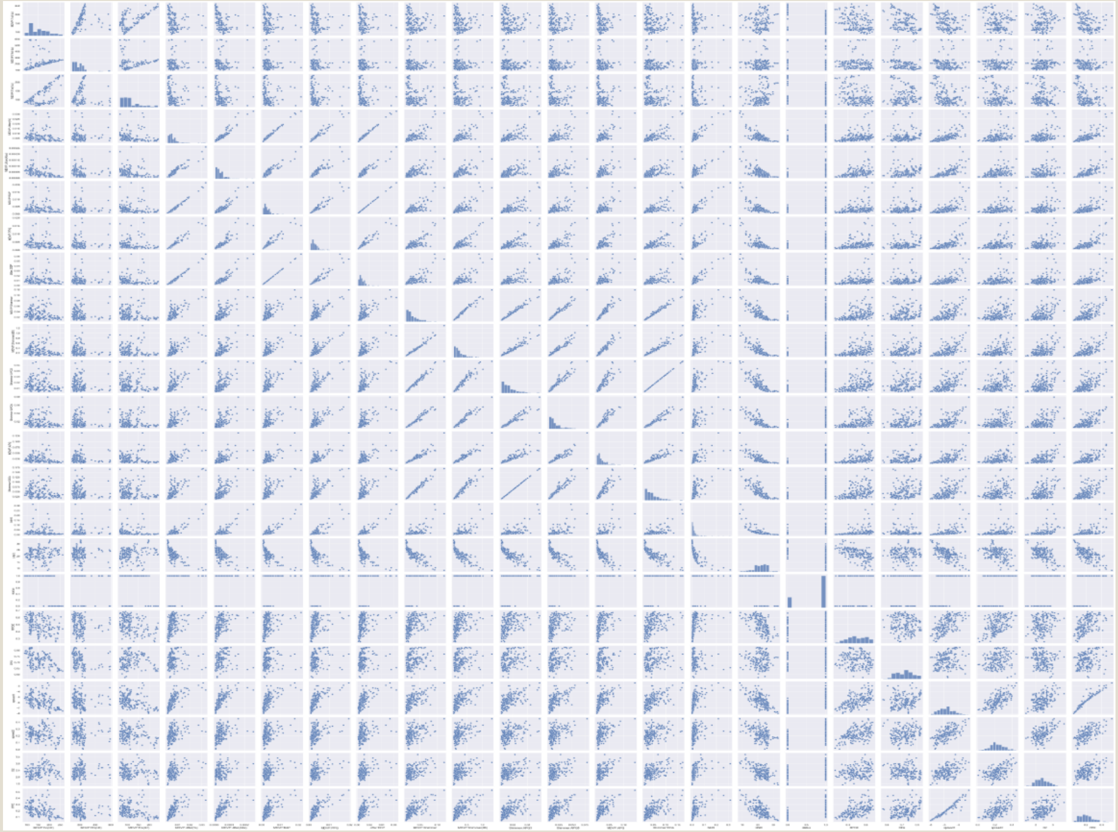
2.Boxplot: Displays the median, interquartile range, and outliers for various features, highlighting data variability and anomalies.

3.Pairplot: Examines relationships between different features, revealing interactions and correlations that may inform model performance.

4.Heatmap: Illustrates the correlation matrix among features, aiding in the identification of highly correlated variables for feature selection.



This heatmap shows the correlation between different features. Strong positive correlations (in red) are present between many variables, especially shimmer and jitter metrics, indicating multicollinearity. Meanwhile, HNR and status show negative correlations with shimmer-related variables. The highly correlated features should be considered for dimensionality reduction techniques like PCA to avoid redundancy in the model.



This pairplot visualizes the distribution and relationships between pairs of features. The diagonal represents the distribution of individual features, while the scatter plots between different variables show patterns or potential linear/non-linear relationships. A few features exhibit clustering or linear trends, suggesting they may be useful for classification or regression tasks.

Model 1 – Logistic Regression

•**Overview**: Logistic regression is a simple yet effective method for binary classification.

•**Why Logistic Regression?**: It is interpretable and works well when the relationship between the features and outcome is linear.

•**Performance**: Mention the accuracy, precision, recall, or F1-score achieved with logistic regression on the dataset.

Model 2 – Random Forest

•**Overview**: Random Forest is an ensemble method that builds multiple decision trees and merges them to get more accurate and stable predictions.

•**Advantages**: Handles large datasets with higher dimensionality and reduces overfitting by averaging predictions from different trees.

•**Performance**: Share performance metrics and how it compared to Logistic Regression.

Bagging

- Technique: Bootstrap aggregating reduces variance by combining predictions from multiple models trained on different subsets of data.
- Example: Random Forest is a bagging algorithm.

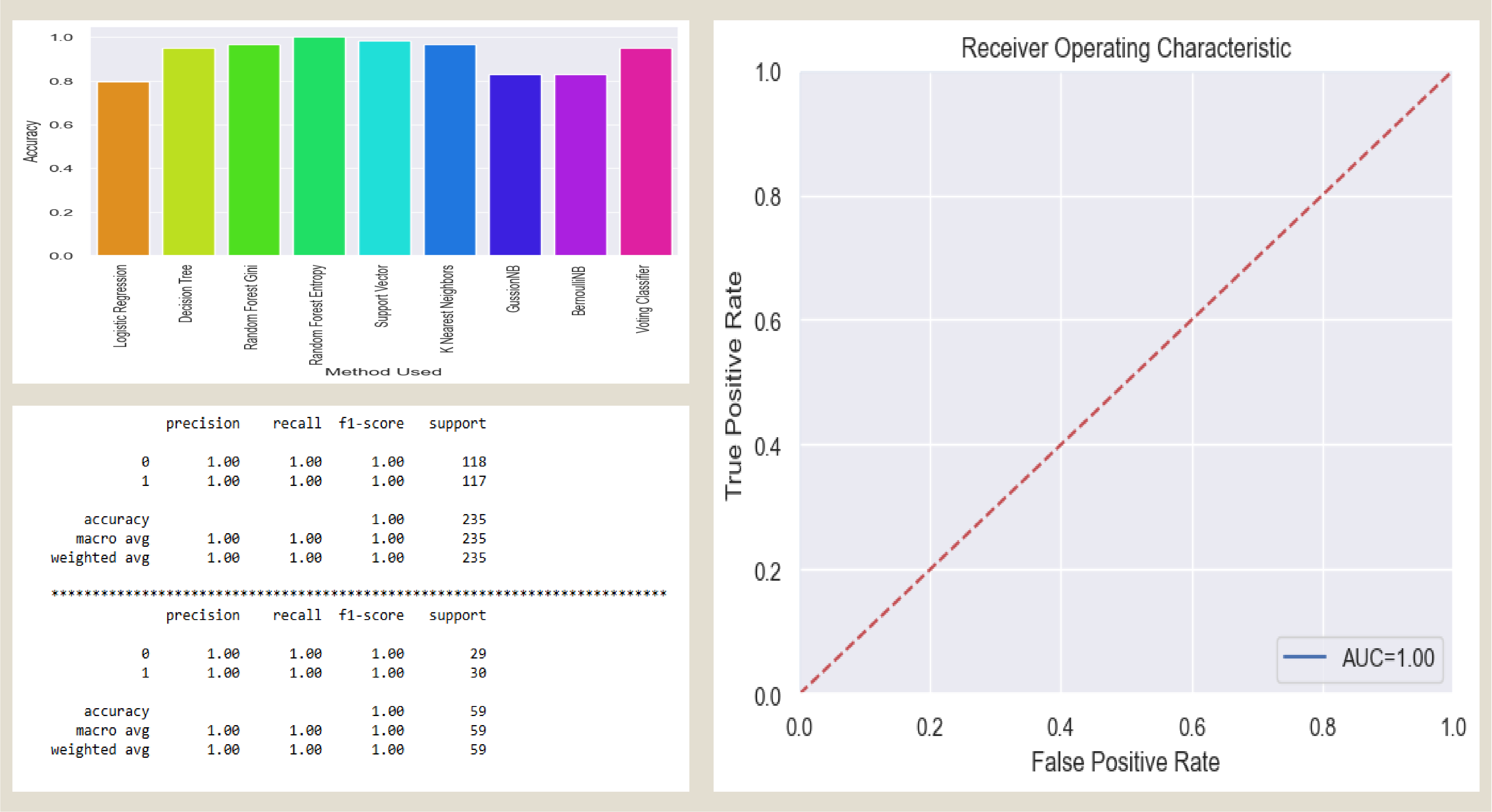
•**Performance**: Mention if you implemented these techniques and how they improved the model.

Stacking

•**Concept**: Stacking combines different machine learning models (e.g., logistic regression, random forest, SVM) to create a more powerful predictive model.

•**Implementation**: Trained multiple base learners and combined their predictions to get better results than individual models.

•**Performance**: Highlight any performance improvements through stacking.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       118
           1       1.00      1.00      1.00       117

    accuracy                           1.00       235
   macro avg       1.00      1.00      1.00       235
weighted avg       1.00      1.00      1.00       235

****************************************************************************
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        29
           1       1.00      1.00      1.00        30

    accuracy                           1.00        59
   macro avg       1.00      1.00      1.00        59
weighted avg       1.00      1.00      1.00        59
```

First Report (Training Set):

- **Precision, Recall, F1-Score**: All values are 1.00 for both classes (0 and 1), meaning the model perfectly classified all instances in the training set.
- **Support**: There are 118 samples of class 0 and 117 samples of class 1.
- **Accuracy**: 100% accuracy, indicating the model made no errors on the training set.

Second Report (Test Set):

- Similar to the training set, the model has 100% precision, recall, and F1-score on the test set.
- **Support**: There are 29 samples of class 0 and 30 samples of class 1.
- **Accuracy**: 100% accuracy on the test set as well, indicating no errors in classification.

❑ While the performance is perfect on both sets, further evaluation with more data or different validation techniques might be needed to ensure the model generalizes well in real-world scenarios.

Kernel Support Vector Machines (SVM)

•**Overview**: SVM is a powerful classification algorithm that works by finding the optimal hyperplane to separate different classes.

•**Kernel Trick**: Allows SVM to handle non-linear data by transforming the input space using functions like RBF.

•**Performance**: Showcase how well SVM performed compared to other models on Parkinson's dataset.

Conclusion

Summary

In this project, several machine learning models were applied to predict Parkinson's disease based on voice measurements. The dataset was thoroughly preprocessed, addressing any potential data quality issues such as missing values and normalizing features to improve model performance.

The models used included Logistic Regression, Random Forest, Bagging and Boosting (such as Gradient Boosting or AdaBoost), Stacking, and Kernel Support Vector Machines. These algorithms were selected based on their ability to handle classification tasks effectively, especially with relatively small and structured datasets.

Best Performing Model

Among all models tested, **Random Forest** emerged as the best-performing model. It outperformed other models based on key evaluation metrics such as accuracy, precision, and recall. The Random Forest model is particularly well-suited for this problem due to its ability to handle high-dimensional data and prevent overfitting by averaging over multiple decision trees. The robustness of the model and its resistance to overfitting contributed to its superior performance compared to other models like Logistic Regression and Support Vector Machines.

Future Work

There are several avenues for further improvement in this project:

1. **Fine-tuning models**: Hyperparameter optimization for Random Forest or Support Vector Machines could further boost model performance. Grid Search or Random Search techniques can be employed to fine-tune parameters such as the number of trees in the Random Forest or the kernel parameters in SVM.
2. **Incorporating Deep Learning**: Applying deep learning techniques like a Convolutional Neural Network (CNN) on a transformed feature set (such as spectrograms of voice signals) could improve model performance even further. Parkinson’s disease prediction from voice data can be complex, and deep learning models could capture non-linear relationships more effectively.
3. **Gathering more data**: Increasing the size and diversity of the dataset can help improve model generalization. Larger datasets tend to produce more reliable models, especially in the medical domain where patterns can vary across different demographics.