BASIC INFORMATION

Title of Project : Classification and Regression on Parkinson speech dataset with multiple types of sound recordings

Student Name: SAATWIK BISHT

Enrollment Number : 07719011921

Signature :

Email ID: saatwik.usar@gmail.com

Contact number :8929802197

Google Drive link - https://drive.google.com/drive/folders/1o_rzt5IiQEFtSi6ZUxOkOTckCLnNzyfT?usp=sharing

Youtube Link - https://youtu.be/2tZd6fUPHCI

REPORT

Title of project :

Parkinson speech dataset with multiple types of sound recordings

ABSTRACT:

The Parkinson Speech Dataset with Multiple Types of Sound Recording represents a valuable resource for the study of speech characteristics in individuals affected by Parkinson's disease (PD). This dataset encompasses a comprehensive collection of audio recordings, comprising training and test files obtained from 20 PD patients and 20 healthy individuals. Diverse types of sound recordings, including sustained vowels, numbers, words, and short sentences, were meticulously captured to capture the wide range of vocal manifestations associated with PD.

The dataset includes a comprehensive set of 26 acoustic features extracted from each voice sample, encompassing measurements such as jitter, shimmer, pitch variability, vocal intensity, and other pertinent parameters. These features facilitate a thorough analysis of the distinctive acoustic patterns present in PD-related speech impairments, thereby contributing to the advancement of research on the characterization and comprehension of speech changes associated with PD.

Furthermore, expert physicians have assigned Unified Parkinson's Disease Rating Scale (UPDRS) scores to each patient, enabling regression analysis and correlation studies to be conducted between speech characteristics and disease severity. This aspect of the dataset provides valuable insights into the potential application of speech analysis as a diagnostic and monitoring tool for PD.

The dataset also offers an independent test set, gathered under identical examination conditions. This allows researchers to validate their models and assess the performance of their algorithms and techniques effectively.

Overall, the Parkinson Speech Dataset with Multiple Types of Sound Recording serves as a valuable resource for investigating the acoustic properties of PD-related speech impairments. Its utilization contributes to the development of diagnostic tools, speech recognition algorithms, and

assistive technologies aimed at enhancing the diagnosis, monitoring, and treatment of Parkinson's disease.

KEYWORDS:

1. Parkinson Disease
2. Speech Analysis
3. Diagnostic Tools
4. Regression Analysis
5. Acoustic features

INTRODUCTION:

Parkinson's disease (PD) is a prevalent neurodegenerative disorder affecting millions of individuals worldwide, characterized by motor symptoms like tremors, rigidity, and bradykinesia, as well as non-motor symptoms, including speech and voice impairments. The impact of speech changes in PD can significantly affect individuals' quality of life, communication abilities, and social interactions. Thus, comprehending the acoustic characteristics of speech in PD is crucial for the development of effective diagnostic tools, disease monitoring, and targeted therapeutic interventions.

The Parkinson Speech Dataset with Multiple Types of Sound Recording offers a valuable resource for the study and analysis of speech patterns in individuals with PD. This dataset encompasses recordings from both PD patients and healthy individuals, enabling comparative analysis of speech characteristics. It includes various types of sound recordings, such as sustained vowels, numbers, words, and short sentences, to capture the diverse range of speech manifestations associated with PD. This diversity enables researchers to explore specific changes in speech patterns and acoustic features that distinguish PD patients from healthy individuals.

Acoustic features play a vital role in quantifying and analyzing the speech characteristics affected by PD. The dataset provides a comprehensive set of 26 acoustic features extracted from each voice sample, encompassing measures like jitter, shimmer, pitch variability, vocal intensity, and more. These features serve as valuable indicators of underlying vocal impairments and offer insights into the distinct acoustic patterns associated with PD-related speech changes.

In conclusion, the Parkinson Speech Dataset with Multiple Types of Sound Recording presents a comprehensive collection of audio data that facilitates in-depth analysis of speech characteristics in individuals with PD. It serves as a valuable resource for researchers and clinicians working towards improving the diagnosis, monitoring, and treatment of PD-related speech impairments. By leveraging this dataset, further advancements can be made to enhance the understanding and management of speech-related issues in Parkinson's disease.

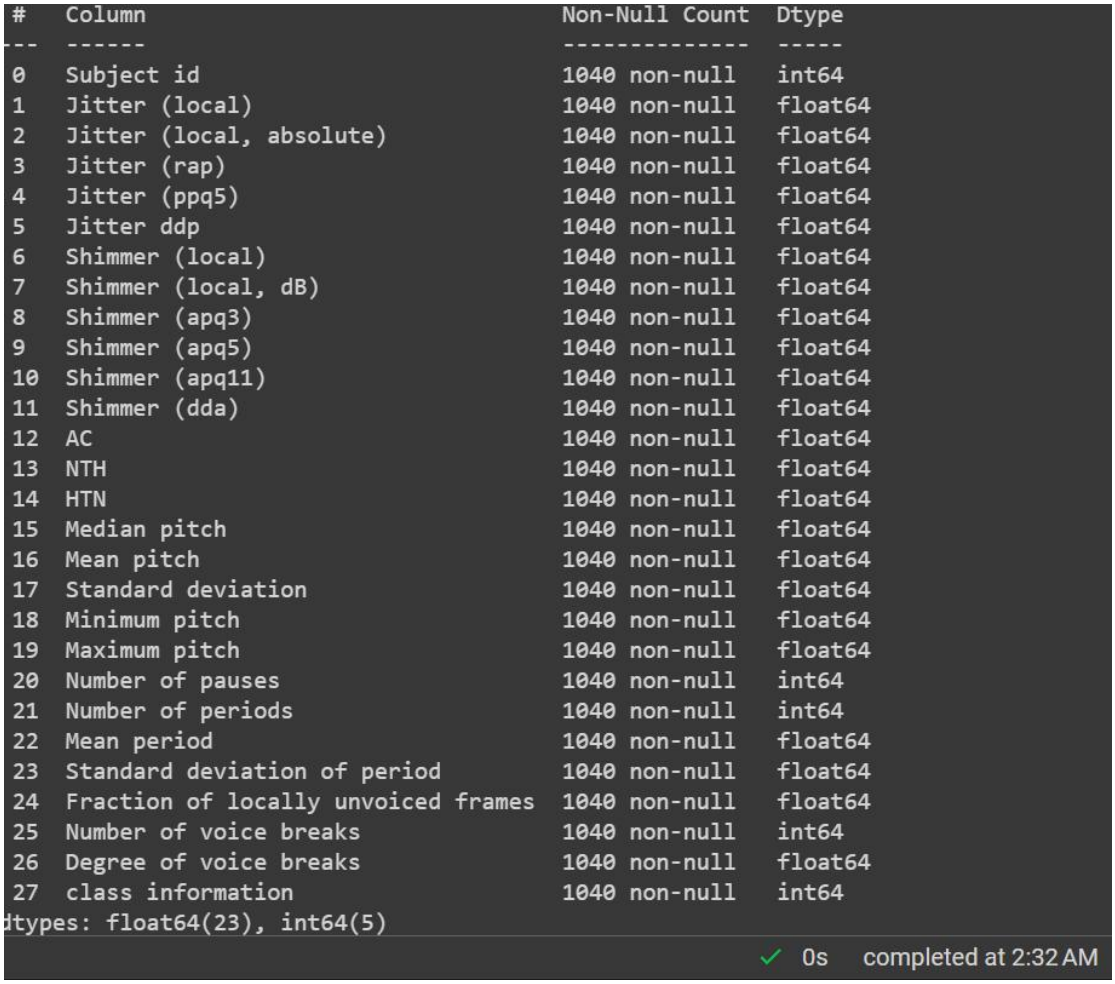
Proposed Methodology:

1) **Datasets:** We have 2 datasets Training and Testing.

Training Dataset: The training dataset comprises recordings from 20 PD patients and 20 healthy individuals. Each subject's recordings encompass multiple types of sound recordings, including sustained vowels, numbers, words, and short sentences. These diverse recordings capture the range of vocal patterns and speech samples affected by PD. Furthermore, 26 acoustic features were extracted from each voice sample, quantitatively measuring characteristics like pitch, intensity, and variations in vocal quality. The dataset also includes Unified Parkinson's Disease Rating Scale (UPDRS) scores assigned by expert evaluators, which assess the severity of PD in each patient.

Testing Dataset: The testing dataset includes recordings from 28 PD patients, with a specific focus on sustained vowel sounds 'a' and 'o'. Each patient was requested to produce these vowel sounds three times, resulting in a total of 168 recordings. Similar to the training dataset, the same set of 26 acoustic features extracted from the training data were calculated for these voice samples. The testing data serves as an independent set, allowing for the validation and verification of the results obtained from the training data.

2) PRE Processing:
A) Data cleaning : The first step in the preprocessing stage is to check the dataset for any missing values, outliers, or inconsistent data. Since our dataset does not contain categorical values, we can assume that the data is in good condition without any major issues



```
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Subject id                            1040 non-null   int64
 1   Jitter (local)                        1040 non-null   float64
 2   Jitter (local, absolute)              1040 non-null   float64
 3   Jitter (rap)                          1040 non-null   float64
 4   Jitter (ppq5)                         1040 non-null   float64
 5   Jitter ddp                            1040 non-null   float64
 6   Shimmer (local)                       1040 non-null   float64
 7   Shimmer (local, dB)                   1040 non-null   float64
 8   Shimmer (apq3)                        1040 non-null   float64
 9   Shimmer (apq5)                        1040 non-null   float64
 10  Shimmer (apq11)                       1040 non-null   float64
 11  Shimmer (dda)                         1040 non-null   float64
 12  AC                                    1040 non-null   float64
 13  NTH                                   1040 non-null   float64
 14  HTN                                   1040 non-null   float64
 15  Median pitch                          1040 non-null   float64
 16  Mean pitch                            1040 non-null   float64
 17  Standard deviation                    1040 non-null   float64
 18  Minimum pitch                         1040 non-null   float64
 19  Maximum pitch                         1040 non-null   float64
 20  Number of pauses                      1040 non-null   int64
 21  Number of periods                     1040 non-null   int64
 22  Mean period                           1040 non-null   float64
 23  Standard deviation of period          1040 non-null   float64
 24  Fraction of locally unvoiced frames   1040 non-null   float64
 25  Number of voice breaks                1040 non-null   int64
 26  Degree of voice breaks                1040 non-null   float64
 27  class information                     1040 non-null   int64
dtypes: float64(23), int64(5)
```


B) Balanced Data: It is important to ensure that the dataset has a balanced representation of both classes (PD and healthy). In our case, the dataset is already balanced, meaning that there is no significant class imbalance that would require techniques like oversampling or undersampling. This eliminates the need for balancing the data and ensures fair model.

After analyzing the dataset, we know that class information is our target value

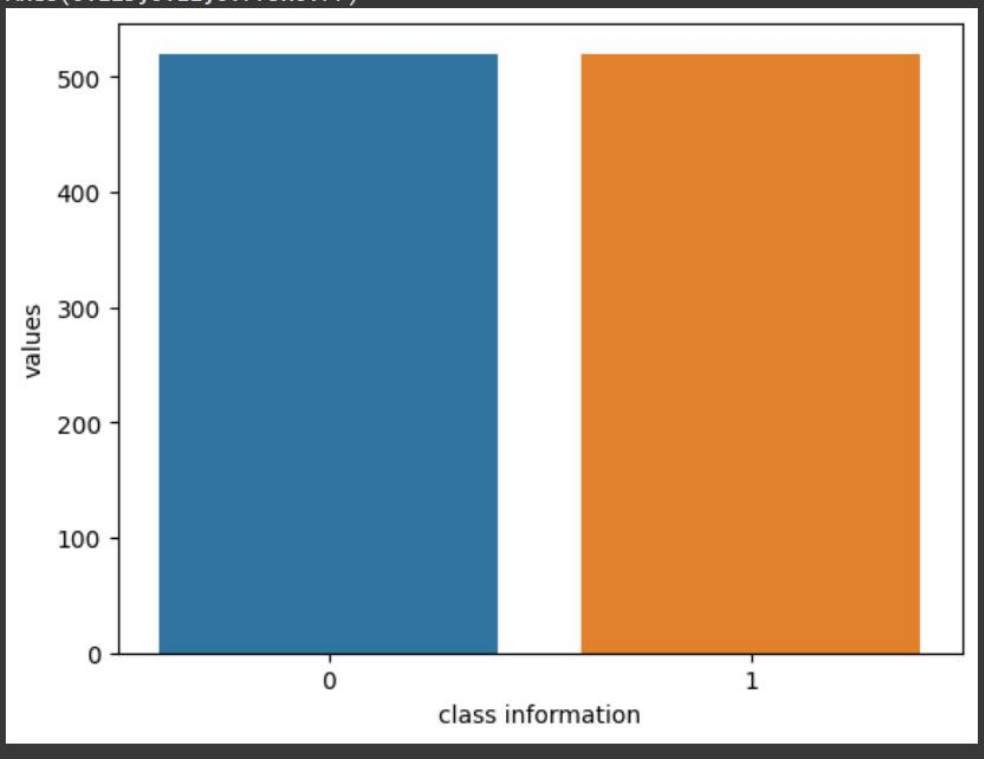
```
#Analyzing our target value
df['class information']
```

```
0       1
1       1
2       1
3       1
4       1
       ..
1035    0
1036    0
1037    0
1038    0
1039    0
Name: class information, Length: 1040, dtype: int64
```

```
#counting values in our target value
df['class information'].value_counts()
```

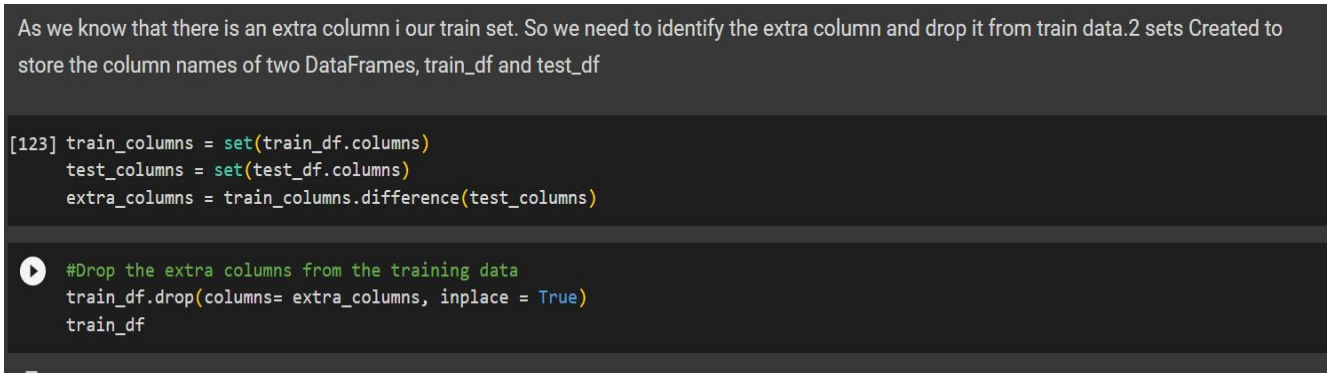
```
1    520
0    520
Name: class information, dtype: int64
```

```
Axes(0.125,0.11;0.775x0.77)
```



C) Splitting Training and Testing Data: In our scenario, the dataset is already divided into two separate sets: the training dataset and the test dataset. However, it's worth noting that these two datasets have different numbers of columns. The training dataset has 29 columns, while the test dataset has 28 columns.

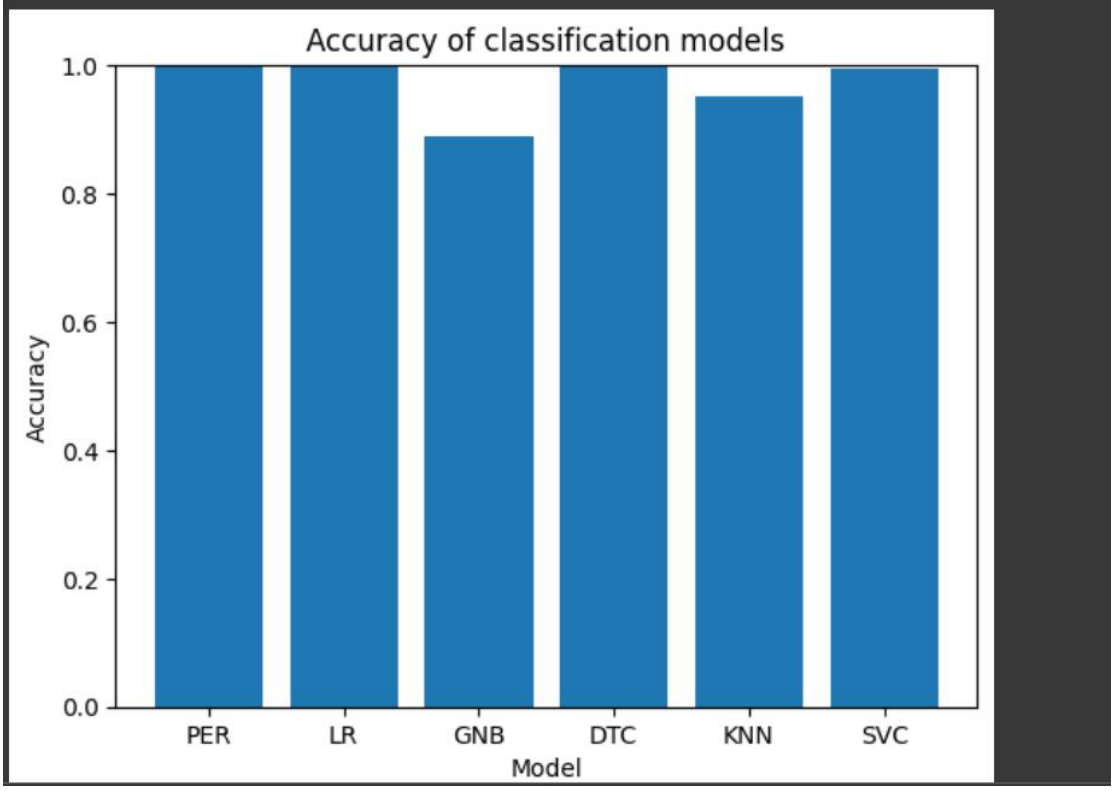
D) Before applying features selection, we need to make no. Of columns in training and testing data equal. So using below code we can drop unwanted column from our data



E)Feature scaling: Since the dataset may contain features with different scales and ranges, it is often beneficial to apply feature scaling techniques to bring all features to a similar scale.

F)Modeling: In this project, six classification models were used: Perceptron, Support Vector Classifier (SVC), Decision Tree Classifier, Gaussian Naive Bayes, Logistic Regression, and k-Nearest Neighbors Classifier. The models were trained and evaluated on the dataset. The testing accuracy was calculated for each model to assess their performance.

```
plt.ylabel('Accuracy')
plt.title('Accuracy of classification models')
plt.ylim(0,1)
plt.show()
```



RESULT & DISCUSSION : From all the classification models , we have decided to take Decision tree classifier because it is giving highest accuracy for our testing data

```
print("Testing accuracy: ")
for i, j in acc_test.items():
  test_acc1.append(j)
  print(i , ":", j)

test_acc1
```

```
Testing accuracy:
PER : 0.6547619047619048
LR : 0.6071428571428571
GNB : 0.5059523809523809
DTC : 0.7142857142857143
KNN : 0.6607142857142857
SVC : 0.6547619047619048
[0.6547619047619048,
 0.6071428571428571,
 0.5059523809523809,
 0.7142857142857143,
 0.6607142857142857,
 0.6547619047619048]
```

Conclusion:

In this project, we utilized the Parkinson Speech Dataset with Multiple Types of Sound Recording to explore the potential of using acoustic features for Parkinson's disease (PD) detection. The dataset contains voice samples from both PD patients and healthy individuals, accompanied by relevant clinical information.

To ensure data quality, we performed data cleaning, removing missing values, outliers, and inconsistent data. The dataset was well-balanced, eliminating the need for additional class balancing techniques. We divided the dataset into training and test sets, with varying column numbers in each set.

To facilitate analysis, we applied feature scaling techniques to normalize the features, enabling fair comparisons. We also aligned the column numbers in the training and test sets by removing an unnecessary column from the training dataset.

Using six classification models, including Perceptron, SVC, Decision Tree Classifier, Gaussian Naive Bayes, Logistic Regression, and k-Nearest Neighbors Classifier, we trained and evaluated the models on the dataset. We measured the testing accuracy of each model to assess their performance in distinguishing between PD patients and healthy individuals.

FUTURE WORK:

1) Feature Engineering: Explore additional acoustic features that may be relevant in identifying and characterizing PD-related speech impairments. Consider domain-specific features or advanced signal processing techniques to capture subtle variations in speech patterns.

2)Multimodal Analysis: Incorporate other modalities such as facial expressions, gesture recognition, or linguistic features to enhance PD detection accuracy. Fusion of multiple modalities may provide a more comprehensive understanding of PD-related speech impairments.