DATA MINING AND ANALYTICS

18CSE355T

Diagnosis of Parkinson's Disease via Speech Analysis

NAME: Meghna Sharma
SECTION: C1
REGISTRATION NO: RA1911003010189
FACULTY NAME: Dr. G. ABIRAMI

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be **University** *u/s 3 of UGC Act, 1956)*

FACULTY OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
Kattankulathur, Chennai
NOVEMBER 2021

Diagnosis of Parkinson's Disease via Speech Analysis

Introduction

- Parkinson’s Disease is the second most prevalent neurodegenerative disorder after Alzheimer’s, affecting more than 10 million people worldwide. Parkinson’s is characterized primarily by the deterioration of motor and cognitive ability.
- There is no single test which can be administered for diagnosis. Instead, doctors must perform a careful clinical analysis of the patient’s medical history.
- Speech is very predictive and characteristic of Parkinson’s disease; almost every Parkinson’s patient experiences severe vocal degradation (inability to produce sustained phonations, tremor, hoarseness), so it makes sense to use voice to diagnose the disease. Voice analysis gives the added benefit of being non-invasive, inexpensive, and very easy to extract clinically.

Background

Parkinson's Disease

- Parkinson’s is a progressive neurodegenerative condition resulting from the death of the dopamine containing cells of the substantia nigra (which plays an important role in movement).
- Symptoms include: “frozen” facial features, bradykinesia (slowness of movement), akinesia (impairment of voluntary movement), tremor, and voice impairment.
- Typically, by the time the disease is diagnosed, 60% of nigrostriatal neurons have degenerated, and 80% of striatal dopamine have been depleted.

Performance Metrics

- TP = true positive, FP = false positive, TN = true negative, FN = false negative
- Accuracy: (TP+TN)/(P+N)
- Matthews Correlation Coefficient: 1=perfect, 0=random, -1=completely inaccurate

Algorithms Employed

- Logistic Regression (LR):

  Uses the sigmoid logistic equation with weights (coefficient values) and biases (constants) to model the probability of a certain class for binary classification. An output of 1 represents one class, and an output of 0 represents the other. Training the model will learn the optimal weights and biases.

- Linear Discriminant Analysis (LDA):

  Assumes that the data is Gaussian and each feature has the same variance. LDA estimates the mean and variance for each class from the training data, and then uses properties of statistics (Bayes theorem , Gaussian distribution, etc) to compute the probability of a particular instance belonging to a given class. The class with the largest probability is the prediction.

- k Nearest Neighbors (KNN):

  Makes predictions about the validation set using the entire training set. KNN makes a prediction about a new instance by searching through the entire set to find the k “closest” instances. “Closeness” is determined using a proximity measurement (Euclidean) across all features. The class that the majority of the k closest instances belong to is the class that the model predicts the new instance to be.

- Decision Tree (DT):

  Represented by a binary tree, where each root node represents an input variable and a split point, and each leaf node contains an output used to make a prediction.

- Neural Network (NN):

  Models the way the human brain makes decisions. Each neuron takes in 1+ inputs, and then uses an activation function to process the input with weights and biases to produce an output. Neurons can be arranged into layers, and multiple layers can form a network to model complex decisions. Training the network involves using the training instances to optimize the weights and biases.

- Naive Bayes (NB):

  Simplifies the calculation of probabilities by assuming that all features are independent of one another (a strong but effective assumption). Employs Bayes Theorem to calculate the probabilities that the instance to be predicted is in each class, then finds the class with the highest probability.

- Gradient Boost (GB):

  Generally used when seeking a model with very high predictive performance. Used to reduce bias and variance (“error”) by combining multiple “weak learners” (not very good models) to create a “strong learner” (high performance model). Involves 3 elements: a loss function (error function) to be optimized, a weak learner (decision tree) to make predictions, and an additive model to add trees to minimize the loss function. Gradient descent is used to minimize error after adding each tree (one by one).

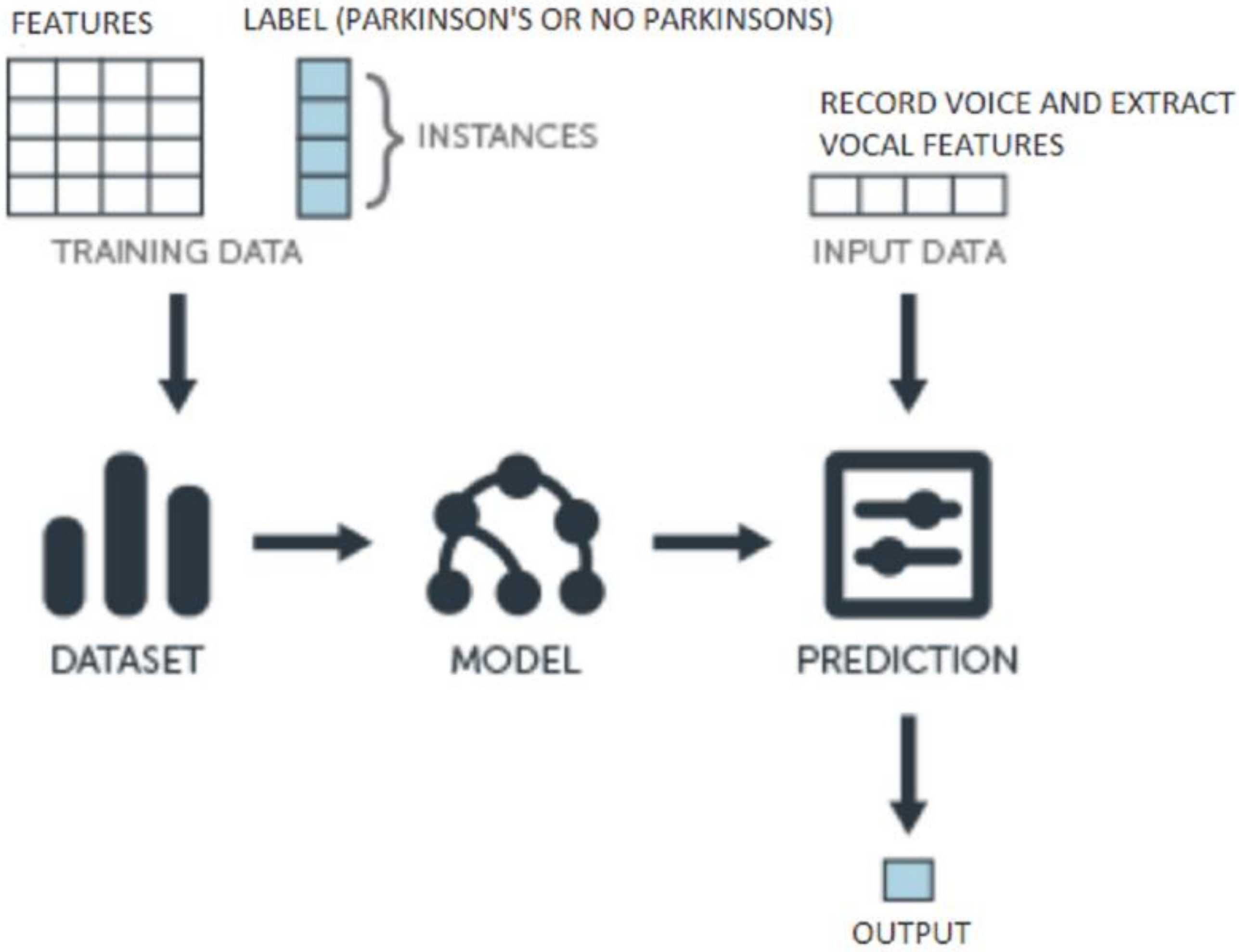
Engineering Goal

Produce a machine learning model to diagnose Parkinson’s disease given various features of a patient’s speech with at least 90% accuracy and/or a Matthews Correlation Coefficient of at least 0.9. Compare various algorithms and parameters to determine the best model for predicting Parkinson’s.

Dataset Description

- Source: the University of Oxford
- 195 instances (147 subjects with Parkinson’s, 48 without Parkinson’s)
- 22 features (elements that are possibly characteristic of Parkinson’s, such as frequency, pitch, amplitude / period of the sound wave)
- 1 label (1 for Parkinson’s, 0 for no Parkinson’s)

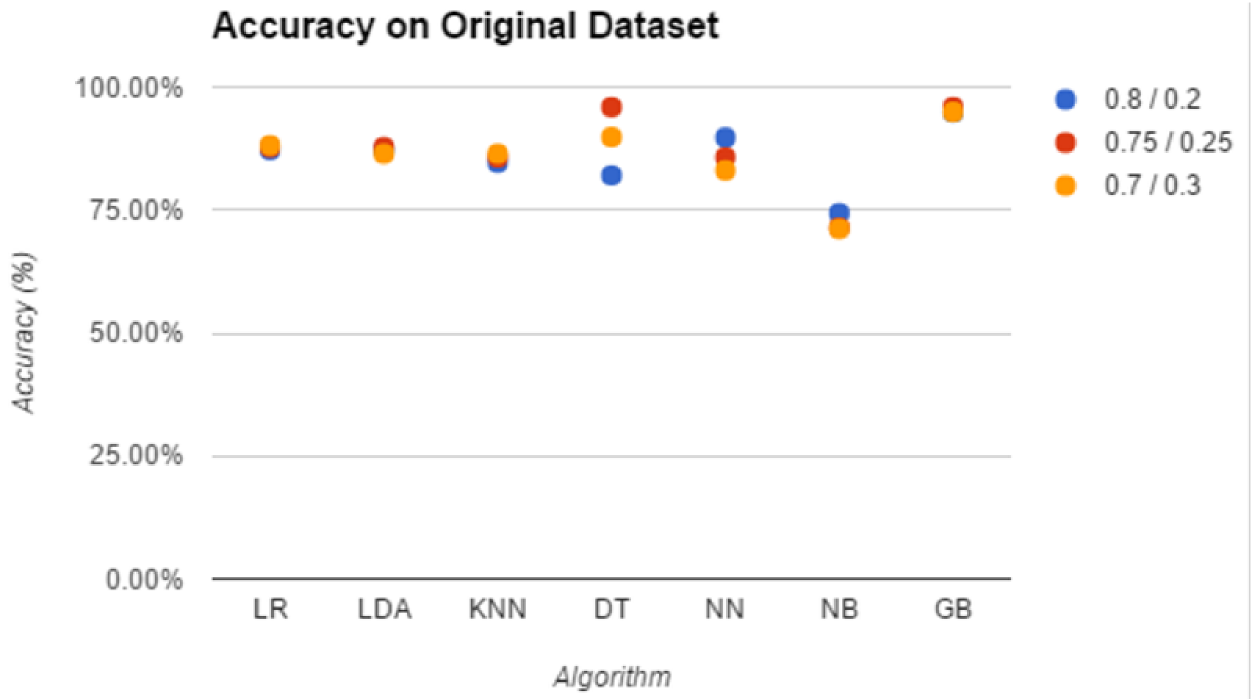
Project Pipeline



Summary of Procedure

- Split the Oxford Parkinson's Dataset into two parts: one for training, one for validation (evaluate how well the model performs)
- Train each of the following algorithms with the training set: Logistic Regression, Linear Discriminant Analysis, k Nearest Neighbors, Decision Tree, Neural Network, Naive Bayes, Gradient Boost
- Evaluate results using the validation set
- Repeat for the following training set to validation set splits: 80% training / 20% validation, 75% / 25%, and 70% / 30%
- Repeat for a rescaled version of the dataset (scale all the numbers in the dataset to a range from 0 to 1: this helps to reduce the effect of outliers)
- Conduct 5 trials and average the results

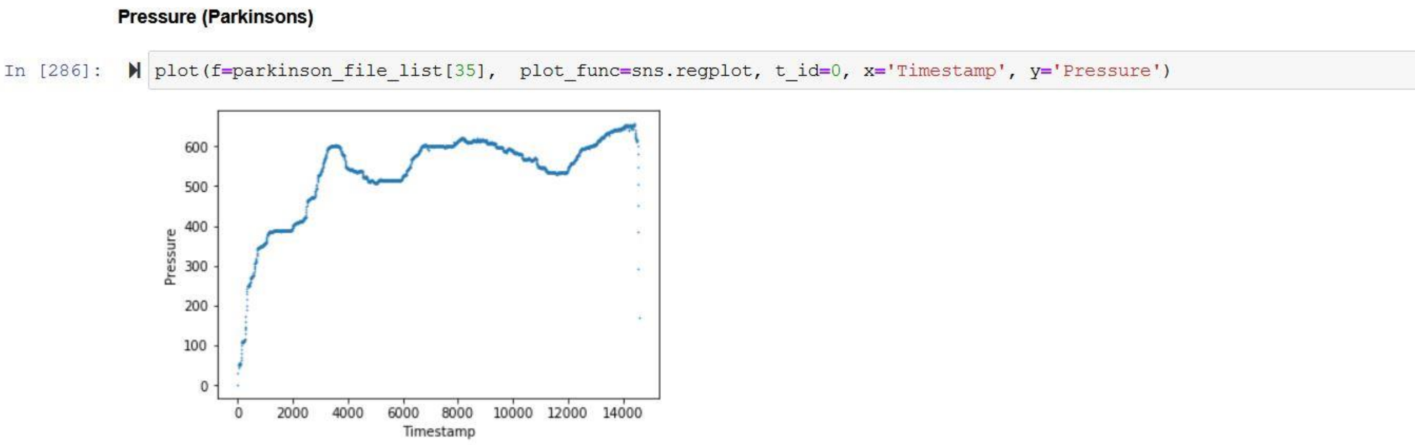
Data



Data Analysis

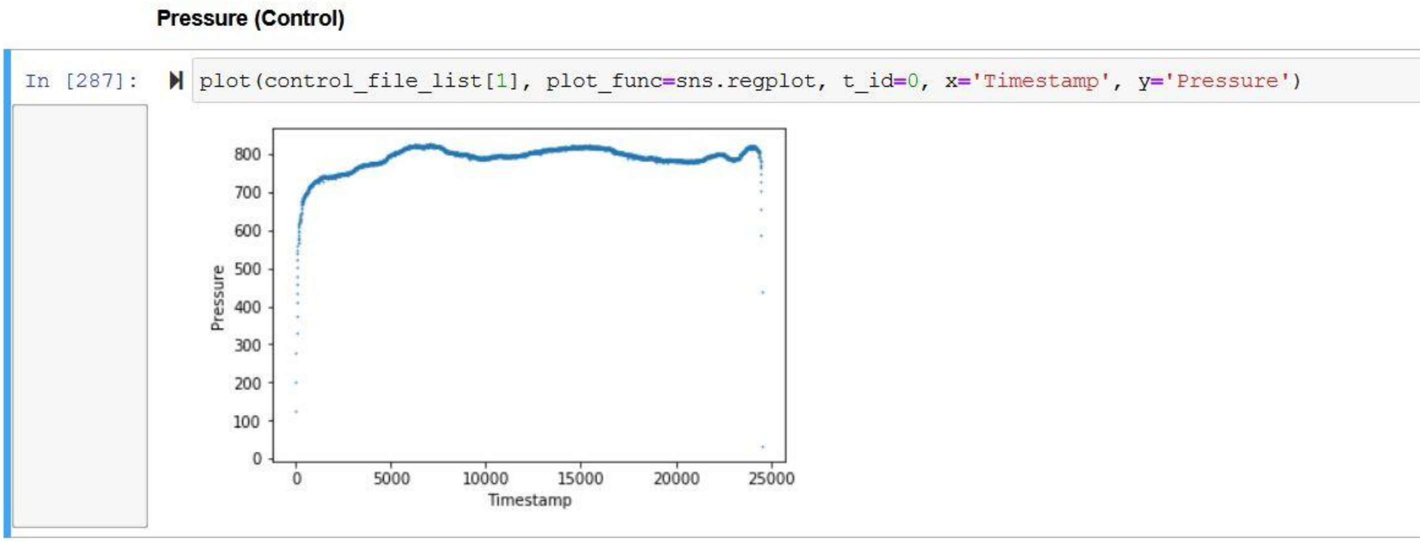
- In general, the models tended to perform the best (both in terms of accuracy and Matthews Correlation Coefficient) on the rescaled dataset with a 75-25 train-test split.
- The two highest performing algorithms, k Nearest Neighbors and the Neural Network, both achieved an accuracy of 98%. The NN achieved a MCC of 0.96, while KNN achieved a MCC of 0.94. These figures outperform most existing literature and significantly outperform current methods of diagnosis.

Screenshots From The Run



Pressure (Parkinsons)

```
In [286]:  plot(f=parkinson_file_list[35],  plot_func=sns.regplot, t_id=0, x='Timestamp', y='Pressure')
```



Pressure (Control)

```
In [287]:  plot(control_file_list[1], plot_func=sns.regplot, t_id=0, x='Timestamp', y='Pressure')
```

Conclusion and Significance

- These robust results suggest that a machine learning approach can indeed be implemented to significantly improve diagnosis methods of Parkinson’s disease. Given the necessity of early diagnosis for effective treatment, my machine learning models provide a very promising alternative to the current, rather ineffective method of diagnosis.
- Current methods of early diagnosis are only 53% accurate, while my machine learning model produces 98% accuracy. This 45% increase is critical because an accurate, early diagnosis is needed to effectively treat the disease.
- Typically, by the time the disease is diagnosed, 60% of nigrostriatal neurons have degenerated, and 80% of striatal dopamine have been depleted.
- With an earlier diagnosis, much of this degradation could have been slowed or treated.
- My results are very significant because Parkinson’s affects over 10 million people worldwide who could benefit greatly from an early, accurate diagnosis.
- Not only is my machine learning approach more accurate in terms of diagnostic accuracy, it is also more scalable, less expensive, and therefore more accessible to people who might not have access to established medical facilities and professionals.
- The diagnosis is also much simpler, requiring only a 10-15 second voice recording and producing an immediate diagnosis.

Results

We tested several classification algorithms such as Logistic Regression, Random Forest, SVM etc. The best results were obtained using SVM-

Accuracy	**100 %**
F1	0.66
Precision	0.5
Recall	1