Parkinson Disease Prediction

using Machine Learning Algorithms

1. Introduction

Parkinson's disease is an escalating neurological problem that occurs in elder people usually. It impacts their motor as well as non-motor characteristics. It is the second most common neurological disease after Alzheimers. Most people with Parkinson develop speech impairment disorders which cause slurred speech or even trailing off at the end of the statement they speak. Hence speech measurements become a chief method in the prediction of Parkinson's disease. In this work, various supervised machine learning classification algorithms are juxtaposed to obtain precise results and select a suitable model which can be used in the early detection of the disease. When SVM is used using the polynomial type kernel, the highest accuracy of about 97.97% is attained. Unsupervised machine learning algorithms such as clustering are also performed on unlabelled speech measurements to analyze how the patients can be segregated into groups when labelled data is absent, or when we do not know if the person is suffering from Parkinson's disease or not. Grouping of the clusters obtained eventually leads to two main groups that are somewhat away and uncorrelated from each other.

2. Problem Definition and Algorithms

2.1 Task Definition

The main focus of this study is to classify the speech measurements present in the Voice dataset for Parkinson Disease that contains features such as jitter and shimmer in the speech of the person. Supervised ML algorithms are used to classify a person as positive or negative for Parkinson's disease based on the labelled features. Finally unsupervised ML algorithms will be also used to make sense of the information by clustering when labelled data is not present. This can help in the early diagnosis of the disease and help the doctors suggest necessary precautions and methods in reducing the effects of this disease.

2.2 Algorithms Used

The following Supervised Machine Learning Algorithms are used:

Table 1: Supervised ML Algorithms used

1.	Decision Tree Algorithm
2.	Random Forest Algorithm
3.	Naive Bayes
4.	K – nearest neighbours
5.	SVM
6.	Logistic Regression

The following Unsupervised Clustering Algorithms have been used:

Table 2: Unsupervised ML Algorithms used

1.	K- means Clustering
2.	Agglomerative Hierarchal Clustering

In addition to the above algorithms, Artificial Neural Networks (ANN) has also been implemented on the data to achieve and compare the accuracy results.

3. Dataset Description

The dataset is taken from [1]. It was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals.

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recordings from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient; the name of the patient is identified in the first column.

Table3: Dataset Description

Data Set Characteristics	Multivariate

Number of Instances (or rows)	197
Attribute Characteristics	Real Numbers
Number of Attributes (or features/columns)	23

Table 4: Description of Features

Feature	**Description**
name	ASCII subject name and recording number
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
MDVP:Flo(Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	Several measures of variation in fundamental frequency
MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA	Several measures of variation in amplitude
NHR,HNR	measures of ratio of noise to tonal components in the voice
status	Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE,D2	Two nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1, spread2, PPE	Three nonlinear measures of fundamental frequency variation

4. Implementation Details

The machine learning problems used in this study have been implemented on Jupyter Notebook, using Python Programming Language. Libraries such as numpy, pandas, matplotlib and sklearn have been imported and used to attain the desired results. Tensorflow and keras have also been used for the implementation of Artificial Neural Networks.

Table 5: Technologies used

1.	Numpy	Python library used for working with mathematical expressions such as arrays, matrices etc
2.	Pandas	It is a python library which is useful in data manipulation and is best for the analysis of data.
3.	Scikit & SKLearn	Library Used to implement machine learning algorithms as well as laying out the accuracy results
4.	Matplotlib	to analyze and visualize the data efficiently and in a better way
5.	Tensorflow	A library used for applications of machine learning such as neural networks
6.	keras	A library which is open source and is used for neural networks. It is written in Python

5. Experimental Evaluation

5.1 Methodology

The voice dataset that has been collected from the UCI repository, has been analyzed and supervised as well as unsupervised machine learning algorithms have been used on this data to get classify the patients into having Parkinson disease of not.

For the classification of features, that is the speech measurements, the column containing the names of the patient has been dropped.

- The features contain all the columns except the column ‘Status’ which contains labels. These features are stored in a matrix ‘x’.
- The labels stored in matrix ‘y’ contain only the column named ‘Status’ which contains labels 0 and 1. Labels are set 0 for healthy patients and 1 for patients with Parkinson’s Disease.

For the implementation of Unsupervised Machine Learning algorithms, the labelled data has been removed and only the features have been analyzed. Clusters of features have been plotted to group similar type of features together and make sense of the data.
Following is the flow chart that represents the working and implementation of this model.

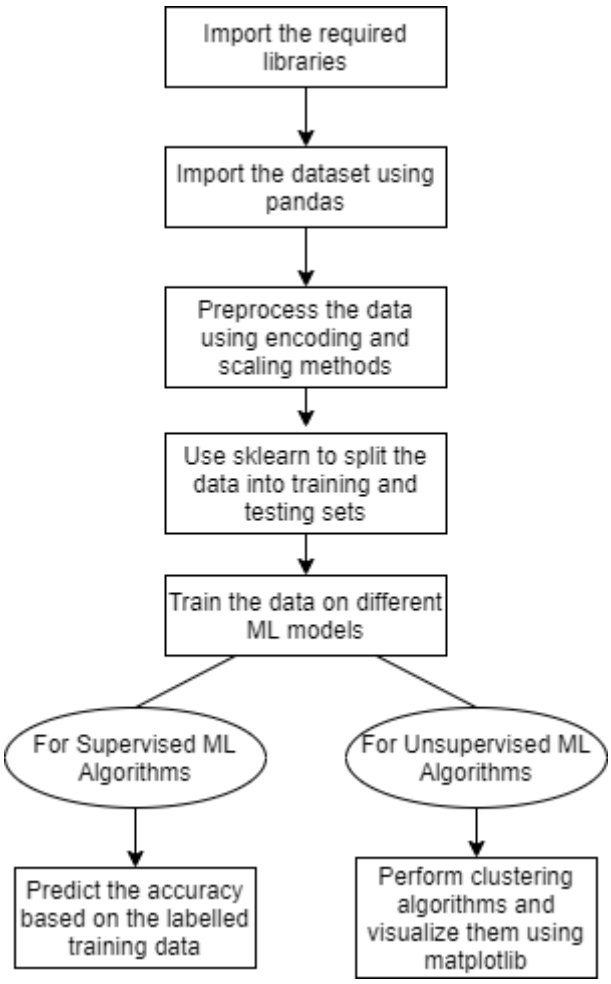

Figure 1: Flow of the model

The flow of steps presented in the above model has been used to implement classification of Parkinson's Disease using Supervised and Unsupervised Machine Learning models.
Dataset is imported using the Pandas library. It is pre-processed using the Standard Scaler which scales the data within a range for better classification. Sklearn library is used to split the data into training and testing sets according to the requirements. Supervised classification machine learning models such as SVM, Logistic Regression, Decision Tree, k- NN are implemented on this training and testing dataset to predict the output whether a person is healthy or not. This is done based on the labelled data. Further, for the analysis of an

unsupervised machine learning approach, Labelled data has been removed and clusters based on the features have been formed.

5.2 Results and Discussion

5.2.1 Implementation of Supervised Machine Learning Algortihms:

Table 6: Results obtained from supervised ML Algorithms

	Supervised Machine Learning Algorithms		**Accuracy**
1.	Decision Tree Algorithm	Using Entropy	93.87
		Using Gini Index	89.79
2.	Naive Bayes (splitting data using K-Fold)		74.22
3.	Logistic Regression		89.74
4.	k- nearest neighbours (k-NN)	Using Euclidean Distance, k=3	94.87
		Using Manhattan Distance, k=5	94.87
		Using Minowski Distance, k=7	89.74
5.	SVM	Using Polynomial type kernel	**97.95**
		Using Gaussian type kernel	85.71
		Using Sigmoid type kernel	73.46
6.	Artificial Neural Networks (ANN)		77.55
7.	Random Forest Classifier		87.75

The accuracy results after implementation of the algorithms have been presented above in a tabular form. From the results, it is observed that the highest accuracy was achieved when SVM model using the Polynomial kernel type was used. This gave an accuracy of **97.95%.** Whereas, using SVM with a Sigmoid type kernel gives a low accuracy of about **73.46%.** The main function of the kernel is to transform the data into the required form that has been taken as input.

Another observation is that when Naive Bayes was implemented and data was randomly split into training and testing sets, a very low accuracy of about 30% was obtained. But when data was split using the k-fold method, an accuracy of of ~75% was obtained using the Naive

Bayes classifier. The poor performance of the Naive Bayes algorithm can be accounted for its poor weight decisions when one class has more training data than the other. Thus k-fold helps to improve the prediction accuracy by dividing the dataset into k folds of approximately equal size.

5.2.2 Implementation of Unsupervised Machine Learning Algorithms.

- **K-means clustering:**

  This algorithm is based on the method of centroid based clustering. Euclidean Distance is used as a measure to segregate the data into clusters. It is also very important to know the correct amount of clusters that is the value of k which can be used in the algorithm. The elbow method, or the *Within clusters sum of squares method (WCSS)* helps to determine the correct number of clusters.

  Below is the graph obtained after using the elbow method.

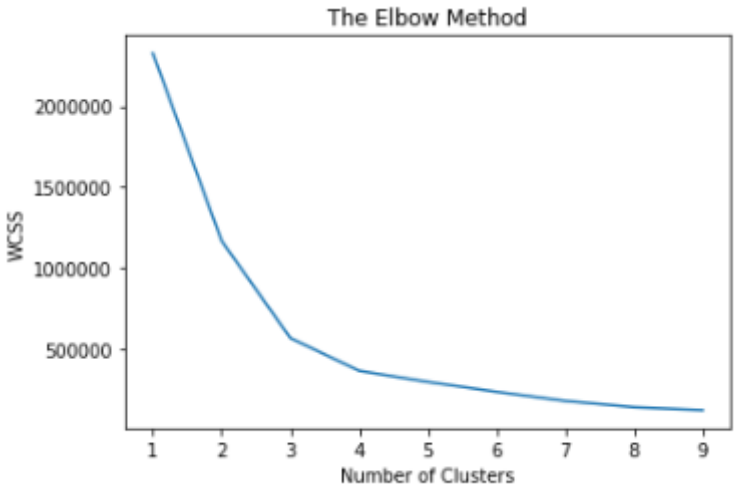

Figure 2: Graph from the implementation of the Elbow Method

Using the elbow method, we see the right number of clusters for this dataset can be 3 or 4. So we choose 3 clusters and implement K-means algorithm to segregate our features into clusters.

Further k means clustering is done on only the features to get the results which are visualized using the matplotlib library.

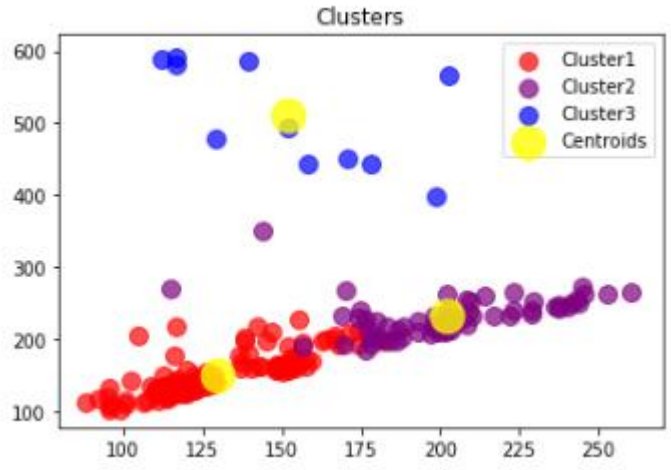

Figure 3: Clusters obtained after k-means

From the above results, it can be seen that we have got our features grouped into three clusters, with their respective centroids.

- Cluster 1 (Red) and Cluster 2 (Purple) are near to each other, so they are healthy people since more features belonged to people who were healthy.
- Cluster 3 (Blue) is away from the rest of the two clusters and hence it probably belongs to people suffering from Parkinson's disease.

- **Agglomerative Hierarchal Clustering:**

  Another unsupervised machine learning method, when implemented gave the following results:

  A dendogram is plotted to obtain the correct amount of clusters which in this case was calculate to be 4.

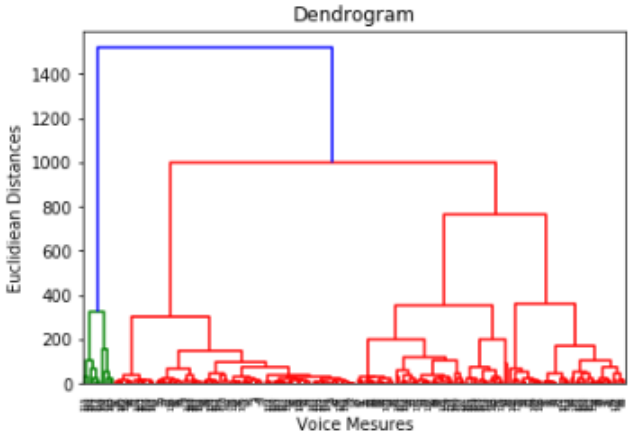

Figure 4: Dendogram obtained to determine the number of clusters

Then Agglomerative Hierarchal Clustering was implemented on the unlabelled data in order to obtain clusters of data as follows:

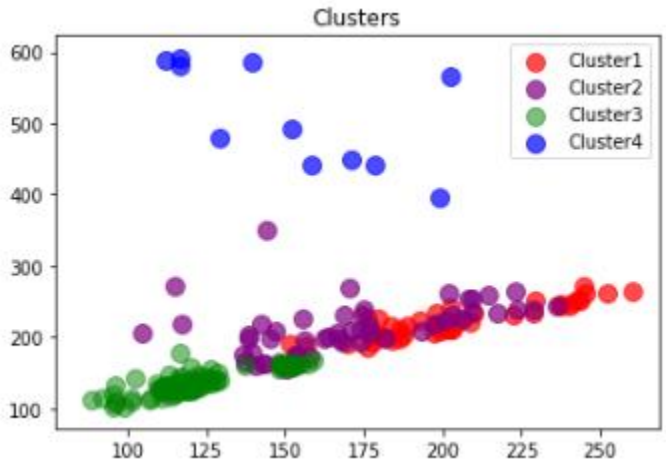

Figure 5: Clusters obtained after Hierarchal Clustering

From the plot, it is observed that clusters similar to those of K-means were obtained and hence again predict than one form of people belong to Cluster 4 while the other group of people belong to Clusters 1,2 and 3. This probably means that Cluster 4 belongs to patients who have Parkinson's disease, whereas the Cluster 1,2 and 3 belong to the group of healthy people who have similar voice features.

6. Related Work

In [2], Shiny et.al have compared the PD and healthy patients using their voice measures. They have used k-means and hierarchy clusters to show the results. They have shown high factored attribute values related to PD and healthy people's voice measurements. Statistical analysis have also been done by them.

In [3], Bhatia et.al have used various SVM classification methods to distinguish between healthy people and people with PD. They have also used a data mining tool named Weka for preprocessing the dataset before classification can be performed. On splitting the data randomly, they have achieved accuracy of 65.21%.

In [4], Ergin et.al have studied and compared different classification algorithms. Their main study is on the validation set using leave-one-out cross validation method. Non- linear SVM can predict all the samples correctly in their work. Logistic Regression as well as KNN give an accuracy of 92%.

7. Future Work

The major shortcomings of this study is that classification algorithms can be used more efficiently as well as data pre-processing can be done more efficiently to improve the results. A larger dataset can be used for a better and accurate analysis.

8. Conclusion

The voice measurement dataset was used to successfully implement both Supervised as well as Unsupervised machine learning algorithms. Each classification model was implemented and results compared. From the results, it is observed that the highest accuracy was achieved when SVM model using the Polynomial kernel type was used. This gave an accuracy of **97.95%.** Whereas, using SVM with a Sigmoid type kernel gives a low accuracy of about **73.46%.** For the implementation of Unsupervised Machine Learning algorithms, the labelled data has been removed and only the features have been analyzed. Clusters of features have been plotted and it has been observed that two major groups of clusters were found which show that people with Parkinson and healthy people have been majorly clustered into different groups based on the features extracted from their voice measurements.

Bilbiography

[1]https://archive.ics.uci.edu/ml/datasets/Parkinsons

[2] T. PanduRanga Vital, P. Shiny, S. E. Ashish, T. Sai Kumar, "Statistical and Unsupervised MLs Analysis on Parkinson's Disease Data set Acquired from A.P. India", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019

[3] Ipsita Bhattacharya, M.P.S Bhatia," SVM Classification to Distinguish Parkinson Disease Patients"

[4] Elcin Ergin1, Shu Hayakawa2, and Timardeep Kaur, "Classification Analysis of Parkinson Speech Dataset"