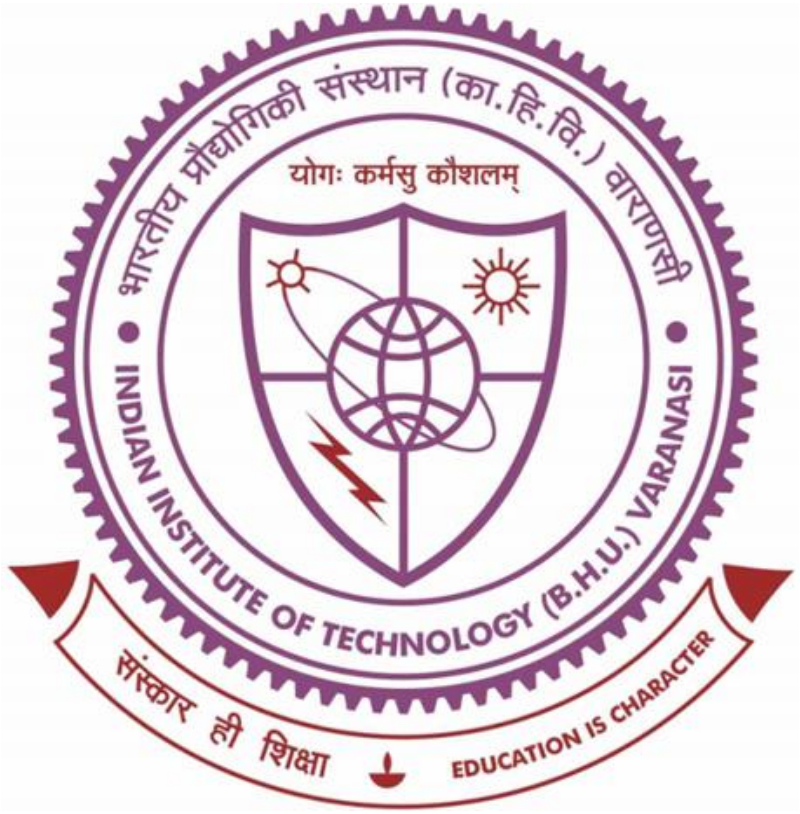
Parkinson Disease Detection using Machine Learning



Name – Anuj Achintya Biswas

Roll no. - 19164005

Department - Pharmaceutical Engineering and Technology (IDD)

Course name – UG Project-I

Course code - PH-392

Date of submission – 5th May 2022

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

Parkinson's disease(PD) is a degenerative disease that affects the central, peripheral, and enteric nervous systems of humans. (Braak, 2000) Multiple neural systems are involved in the underlying pathogenic process, which moves slowly but steadily. (Braak, 2000) Only a few types of nerve cells are susceptible to the disease, which is caused by alterations in the neuronal cytoskeleton. (Braak, 2000) Afflicted neurons eventually produce Lewy bodies in their perikarya and Lewy neurites in their neuronal processes. (Braak, 2000) Treatment options have rapidly expanded, both in the early and later stages of the illness, as well as a greater awareness of non-motor consequences. (Davie, 2008)

Parkinson's disease (PD), the second most common neurodegenerative condition, has a number of unknown origins. (De Lau, 2006) According to current thought, significant gene mutations are responsible for just a tiny percentage of all instances, and that in the majority of cases, non-genetic variables play a role, most likely in conjunction with susceptibility genes. (De Lau, 2006) Many epidemiological studies have been conducted in order to discover non-genetic risk variables, however the majority of them were small and methodologically constrained. (De Lau, 2006) Larger, well-designed prospective cohort studies have just lately reached the point where they have enough incident patients and person-years of follow-up to look into potential risk variables and their interactions. (De Lau, 2006)

1.1 Symptoms of Parkinson Disease

Before the cardinal motor aspects of PD show, it is estimated that up to 80% of dopaminergic cells in the nigro-striatal pathway are destroyed. (Sveinbjornsdottir, 2016) The earliest motor signs are usually used to diagnose the condition. The diagnosis is made using criteria established by the UK PD Brain Bank. (Sveinbjornsdottir, 2016) Slow voluntary movement start with increasing reduction in speed and amplitude of repeating acts (bradykinesia) plus one additional symptom, such as muscle rigidity, resting tremor, or postural instability, is required for diagnosis. (Sveinbjornsdottir, 2016) Step 2 in the diagnosis is to rule out symptoms that could indicate other aetiologies, such as parkinsonian syndromes, which have their own neuropathological changes, and Step 3 is to establish at least three supportive criteria for Parkinson's disease, such as unilateral onset of symptoms, persistent asymmetry of clinical symptoms, good response to levodopa treatment, and induction of dyskinesias by dopaminergic treatment. (Sveinbjornsdottir, 2016) In most cases, symptoms begin on one side of the body and progress to the opposite side within a few years. (Sveinbjornsdottir, 2016) When walking,

the body posture becomes stooped, there is axial and limb rigidity with or without cogwheel phenomena, a shuffling gait, and a lack of arm movement. (Sveinbjornsdottir, 2016) The bradykinesia can cause an expressionless face (hypomimia) and reduced handwriting amplitudes (micrographia). (Sveinbjornsdottir, 2016) Around 80% of people have limb tremor, which is most typically described as a resting pill-rolling tremor of the hands. (Sveinbjornsdottir, 2016) The tendency of the thumb and index finger to come into touch and move in a circular motion is known as pill rolling. (Sveinbjornsdottir, 2016) The tremor can sometimes affect the legs, and other tremor kinds can develop. (Sveinbjornsdottir, 2016) Other gait disorders besides shuffling include blocking, hesitating, and gait festination, in which steps get smaller and faster over time, potentially leading to loss of balance and falls. (Sveinbjornsdottir, 2016) After several years of beginning, a quarter to 60 percent of individuals report freezing of movements. (Sveinbjornsdottir, 2016)

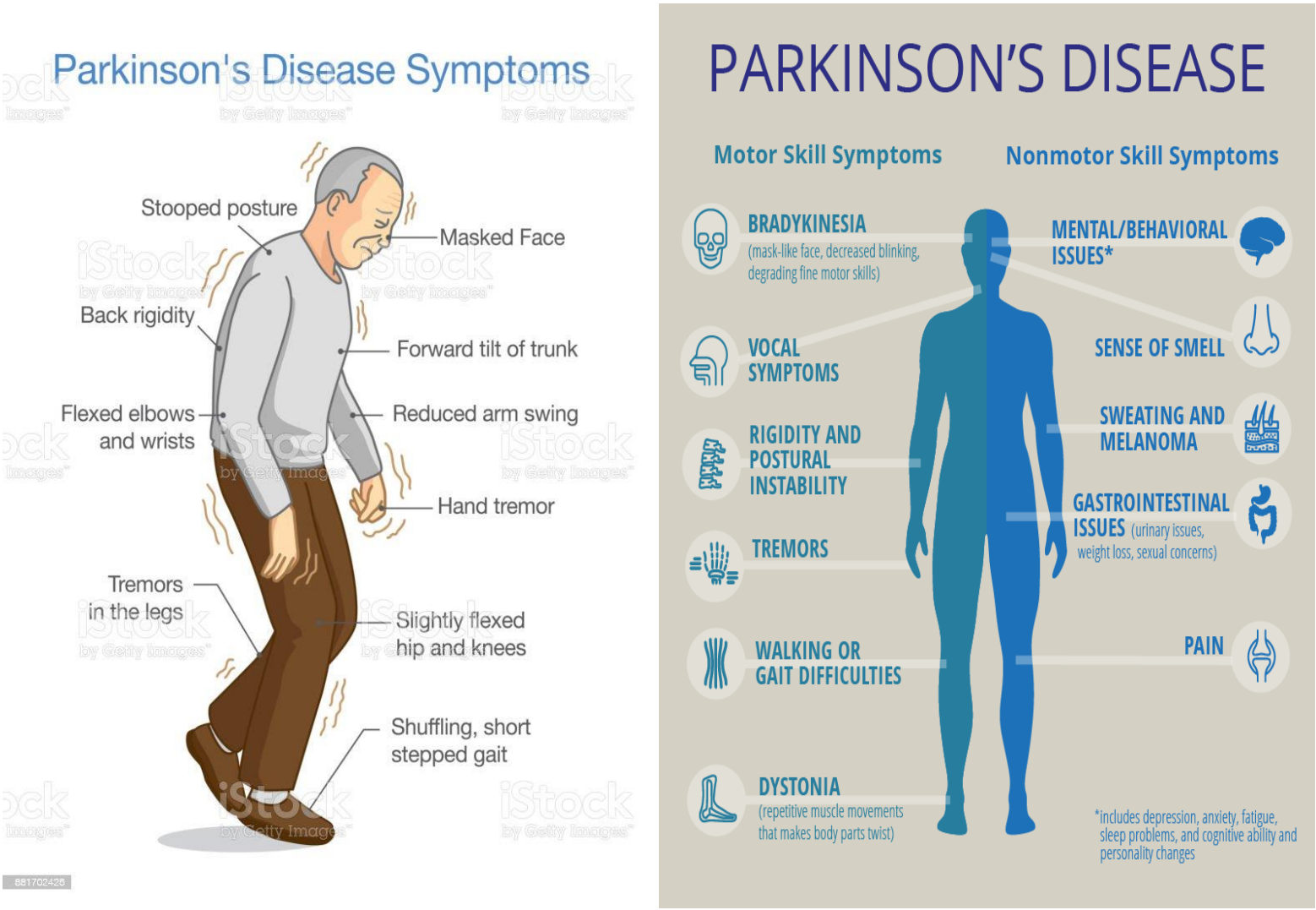


Figure 1: Parkinson disease symptoms

1.2 Treatment of Parkinson Disease

The cause of Parkinson's disease is likely complex, and there is currently no treatment that can slow or stop the disease's course. (Sveinbjornsdottir, 2016) The use of dopaminergic medications is symptomatic, with the goal of resolving the motor abnormalities. (Sveinbjornsdottir, 2016) Standard and most usual initial medication for patients is levodopa, a prodrug to dopamine. (Sveinbjornsdottir, 2016) The initial response is usually positive. The majority of patients experience shorter duration of response to individual doses (wearing-off symptoms), alternate phases of good and poor response to medication (on-off symptoms), involuntary movements of the head, trunk, or limbs (dyskinesias), and other motor complications as the disease progresses and the system's capacity to store dopamine decreases. To control these swings, other dopaminergic drugs are employed. (Sveinbjornsdottir, 2016) Monoamine oxidase type B inhibitors, catechol-O-methyltransferase inhibitors, amantadine, an NMDA receptor antagonist, and dopamine receptor agonists are among them. (Sveinbjornsdottir, 2016) When medicinal therapy fails to manage the motor symptoms, surgical therapy, usually with deep brain electrical stimulation, is possible for a small percentage of patients. (Sveinbjornsdottir, 2016)

2 What is Machine Learning?

Machine learning is an area of artificial intelligence (AI) and computer science that focuses on using data and algorithms to mimic the way humans learn, with the goal of steadily improving accuracy. (Education, 2020) Machine learning is a crucial part of the rapidly expanding discipline of data science. (Education, 2020) Algorithms are trained to generate classifications or predictions using statistical approaches, revealing crucial insights in data mining initiatives. Following that, these insights drive decision-making within applications and organisations, ideally influencing key growth metrics. (Education, 2020)

2.1 Types of Machine Learning

There are mainly two types of Machine Learning:

2.1.1 Supervised Machine Learning

The use of labelled datasets to train algorithms that reliably classify data or predict outcomes is characterised as supervised learning, often known as supervised machine learning. As more data is introduced into the model, the weights are adjusted until the model is properly fitted. (Education, 2020)

2.1.2 Unsupervised Machine Learning

Unsupervised learning, also known as unsupervised machine learning, analyses and clusters unlabeled datasets using machine learning techniques. Without the need for human intervention, these algorithms uncover hidden patterns or data groupings. (Education, 2020)

3 Parkinson Disease Detection using Machine Learning

I had used Supervised Machine Learning algorithm to detect Parkinson Disease. For this, we need some training labelled data to train the model. After that I made a model using machine learning algorithm which can predict that a person has Parkinson disease or not by passing some voice measurement of that person into our model. We will take the data from https://archive.ics.uci.edu/ml/datasets/parkinsons. The dataset provided in that website 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. The original study published the feature extraction methods for general voice disorders. The data was accessed from from the University of California Irvine's (UCI) machine learning repository archive.

3.1 Dataset Information

This dataset contains a variety of biological voice measurements from 31 individuals, 23 of whom have Parkinson's disease (PD). Each column in the table refers to a specific voice measure, and each row corresponds to one of the 195 voice recordings made by these people ("name" column). The primary goal of the data is to distinguish healthy persons from those with Parkinson's disease using the "status" column, which is set to 0 for healthy people and 1 for those with PD.

3.1.1 Attributes Information

There are 24 columns also known as attributes in the dataset. The description of all columns are given below:

- **Name** - 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** – Two measures of ratio of noise to tonal components in the voice.
- **RPDE, D2** – Two nonlinear dynamical complexity measures.
- **DFA** – Signal fractal scaling exponent.
- **Spread1, Spread2, PPE** – Three nonlinear measures of fundamental frequency variation.
- **Status** - Health status of the subject, (one) - Parkinson's, (zero) – Healthy.

3.2 Measures Explanation

3.2.1 Fundamental Frequency

The lowest frequency of a sound wave is a Fundamental frequency. (Matthews, 2016) Our brain interprets sound waves as a single pitch when we speak and listen to each other, even though they actually contain various frequencies (source). For this dataset, the lowest frequency of the subject's voice is captured. (Matthews, 2016)

3.2.2 Ratio to Noise and Tonal Components

The tone refers to one frequency and a noise refers to multiple frequencies. (Matthews, 2016)

3.2.3 Dynamical Complexity Measure

Entropy measure in an individual’s voice. A low score means less entropy. (Matthews, 2016)

3.2.4 Signal Fractal Scaling Component

The measure looks at the pattern of a time-ordered series of real numbers that repeats itself. A low score indicates that a person's voice has more entropy. (Matthews, 2016)

4 Explanation of all Algorithm used to detect Parkinson Disease

4.1 Importing libraries and dataset

I used pandas, numpy, matplotlib, seaborn libraires to explore the data. I also used sklearn library to make a model. After importing this, I set a default resolution at which a graph will display during data exploration. Then I imported our dataset using pandas.

4.1.1 Pandas

Pandas has long been one of the most widely used data cleaning and analysis tools in Data Science and Machine Learning. Using pandas Series and DataFrame to modify data is highly fast and effective. These two pandas data structures will allow you alter data in many ways.

We can say that pandas is the best for **processing data** based on the features available. It can **manage missing data, clean up data, and work with a variety of file types**. This implies it can read or write data in a variety of formats, including CSV, Excel, SQL, and others.

4.1.2 NumPy

NumPy (numerical Python) is a library that consists of multidimensional array objects and a collection of functions for manipulating them. NumPy allows you to **conduct mathematical and logical operations on arrays**. NumPy is a Python scripting language. **'Numerical Python'** is what it stands for. It is a library that contains multidimensional array objects as well as a collection of array processing routines.

4.1.3 Matplotlib

Matplotlib is a Python package that allows you to create static, animated, and interactive visualisations. Matplotlib makes simple things simple and difficult things possible.

4.1.4 Seaborn

Seaborn is a matplotlib-based Python data visualisation package. It has a high-level interface for creating visually appealing and instructive statistics visuals. Seaborn assists you in exploring and comprehending your data.

4.1.5 Sklearn or Scikit Learn

Scikit-learn is undoubtedly Python's most helpful machine learning library. Classification, regression, clustering, and dimensionality reduction are just a few of the useful capabilities in the sklearn toolkit for machine learning and statistical modelling. Machine learning models are built with sklearn.

4.2 Data Exploration

By using " .shape" I came to know that the data contains 195 rows and 24 columns. Then I checked Data types of each columns by using '.info'. After that I checked the null values. Null values are nothing but the missing data. .isnull() is basically used to check that are there missing

data in our dataset or not. I found no missing values are there in dataset. Then using .describe(), I observe some statistical calculation of our dataset like mean, median, count, etc.

4.3 Data Visualization

In this section, I visualized how many rows (or person) have Parkinson disease. I found that 147 rows values (person) indicates Parkinson disease while 48 indicates Healthy. I also made a pie graph on this observation to understand in more efficient way. After this, I made a heat map indicating the corelation between all features/columns of dataset.

4.4 Data Pre-processing

Data Pre-processing is done to improve the performance or accuracy of model which will predict that the person has Parkinson disease or not.

I first dropped the name column as it won't be useful for making model because it doesn't providing any useful information. Then I separated the status column which indicate that the person is healthy or diseased. Then using MinMaxScaler, I normalized the data.

4.4.1 Normalization of Data

By establishing new values and preserving a broad distribution as well as a ratio in data, normalisation prevents raw data and numerous dataset difficulties. It also uses numerous strategies and algorithms to increase the performance and accuracy of machine learning models.

After this, I split the data into training data and testing data. I choosed to split as 70% training data while 30% as testing data. After splitting, training data contains 136 rows and 22 columns (As status column separated before) while test data contains 59 rows and 22 columns.

4.4.2 Training and Testing Data

Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate your model's accuracy.

4.5 Model Making

I made Logistic Regression model, Decision tree model, Random Forest model, bagging model, Adaboost model, Support vector Machine (SVM), XGBoost model, and kNN model. Among these, kNN model showed highest accuracy and AUC/ROC score.

4.5.1 AUC (Area under curve) /ROC (Receiver operating characteristic)

The Area Under the Curve (AUC) is a summary of the ROC curve that measures a classifier's ability to distinguish between classes. (aniruddha, 2020) The AUC indicates how well the model distinguishes between positive and negative classes. (aniruddha, 2020) The greater the AUC, the better. The AUC-ROC curve is a visual representation of how effectively our machine learning classifier performs. (aniruddha, 2020)

4.5.2 Logistic Regression Model

Under the Supervised Learning approach, one of the most prominent Machine Learning algorithms is logistic regression. It's a method for predicting a categorical dependent variable from a set of independent variables. A categorical dependent variable's output is predicted using logistic regression. As a result, the result must be a discrete or categorical value. It can be Yes or No, 0 or 1, true or false, and so on, but instead of giving exact values like 0 and 1, it delivers probabilistic values that are somewhere between 0 and 1.

4.5.3 Decision Tree Classifier Model

Decision Tree is a supervised learning technique that may be used to solve both classification and regression problems, however it is most commonly employed to solve classification issues. Internal nodes represent dataset features, branches represent decision rules, and each leaf node represents the outcome in this tree-structured classifier.

4.5.4 Random Forest Classifier Model

Random Forest is a well-known machine learning algorithm that uses the supervised learning method. In machine learning, it can be utilised for both classification and regression issues. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complex problem and increase the model's performance.

"Random Forest is a classifier that contains a number of decision trees on various subsets of a given dataset and takes the average to enhance the predicted accuracy of that dataset," according to the name. Instead than relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions.

4.5.5 XGBoost Model

XGBoost is a gradient boosting-based decision-tree-based ensemble Machine Learning technique.

4.5.6 Support Vector Machine

The Support Vector Machine, or SVM, is a popular Supervised Learning technique that may be used to solve both classification and regression issues. However, it is mostly utilised in Machine Learning for Classification difficulties. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorising n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future. A hyperplane is the name for the optimal choice boundary.

4.5.7 K-Nearest Neighbour (KNN)

The K-Nearest Neighbour algorithm is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms. The KNN method assumes that the new case/data and existing cases are similar and places the new case in the category that is most similar to the existing categories. The KNN method stores all available data and classifies a new data point based on its similarity to the existing data. This means that new data can be quickly sorted into a well-defined category using the KNN method.

The KNN approach can be used for both regression and classification, but it is more commonly utilised for classification tasks. The KNN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data. It's also known as a lazy learner algorithm since it doesn't learn from the training set right away; instead, it saves the dataset and performs an action on it when it comes time to classify it. During the training phase, the KNN algorithm simply stores the dataset, and when it receives new data, it classifies it into a category that is quite similar to the new data.

5 Deployment of our model

5.1 What is Model Deployment? And why it is important?

Deployment is the process of integrating a machine learning model into an existing production environment in order to make data-driven business decisions. It's one of the last steps in the machine learning process, and it's also one of the most time-consuming. Model deployment is one of the most difficult processes of gaining value from machine learning. In simple terms, model deployment is used and important to make model more easily accessible with some nice looking graphics and interface.

5.2 How to Deploy our KNN Model?

The Simplest way to deploy a machine learning model is to create a web service for prediction. Below is the step wise explanation to create web service:

1. Save the model as pickle file (pickle the model). By doing this, we can easily access model and make prediction. But doing this only isn't sufficient as we can't access pickle file directly. So I first saved my model code file by using pickle.

2. Secondly, I made an python file where I used **Flask** to make an link between the python file and saved model pickle file. Not only this but also I linked python file with **index.html** file. Index.html file is 'html' file. Using html language, I made this file which is linked to **style.css** file. Style.css file is used here to give our web service page a design (to make a website look better).

This all pickle file, html file, css file, coding file should present in one folder, otherwise it won't execute. So, using different types of algorithm, I made a link between all of these files. Basically, app file (python file) is our main app which will take input and give output by using model pickle file. Html and css file is for webpage. Html file is used to create a basic input and output visualization in webpage while css file is used to give webpage a better look.

3. Similarly I made a python file name as 'scaling' which will normalize the input data. Also I created one txt file name as 'requirement.txt' which contains the name and version of all libraries we used. It will required to Heroku in deployment as a webpage. Not only this txt file but also one more important thing required to Heroku is Procfile which is of none file type which contains some one line code. That one line states that it is for webservice and it link python file to webpage.

4. Then I made a webpage using **Heroku**. Steps for using Heroku as :

    i. Type heroku login in the terminal window of VS code or any other IDE you're working with, or in the terminal window of your operating system. You'll be taken to Heroku's login page after running this command, where you must log in. Account details will appear in a terminal window after successful login. (LENDAVE, 2022)

    ii. After login in, we must set up the current directory as a working git repository and add all of the files in the directory; to do so, run the scripts below one by one. A) *git init* B) *git add* C) *git commit -m 'First version'*

    iii. After creating the git repository, we need to tell Heroku to construct our app by giving the app's name, as seen below. In the event that the name you want is not available, simply change it to something else. (LENDAVE, 2022)
    Heroku Parkinson-detector-knn

    After successful deployment completion of this command, Heroku will give the URL where our app is going to be hosted. For this deployment the URL is
    https://parkinson-detector-knn.herokuapp.com/

    iv. Now all we have to do is push our repository to the server, which will generate our application with all of the dependencies from the requirements.txt file, which you can track in your terminal window. The following is the command to push the application over the internet:
    git heroku push master
    Our application will be live at the above-mentioned URL after this command has been successfully executed. (LENDAVE, 2022)

5.2.1 Pickle file

The pickle module keeps track of the objects it's already serialised so that subsequent references to the same item aren't serialised twice, allowing for speedier execution. Basically, it allows you to save a model in a short amount of time.

5.2.2 Flask

Flask is a python online application framework that allows end users to interact with your python code (in this case our machine learning models) directly from their web browser without the need for any libraries, code files, or other software. (Sharma, 2021) Flask makes it simple to build web apps, allowing you to spend your efforts on more critical aspects of the ML lifecycle, like as EDA and feature engineering. (Sharma, 2021)

5.2.3 HTML

HTML stands for Hyper Text Markup Language, which is a programming language used to create online pages and apps. (S., 2022) HTML, or HyperText Markup Language, allows online users to employ elements, tags, and attributes to create and organise sections, paragraphs, and connections. (S., 2022) HTML code is used by web developers to define how a browser renders web page elements including text, hyperlinks, and media files. (S., 2022)

5.2.4 CSS

CSS, or Cascading Style Sheets, is a simple design language designed to make the process of making web pages presentable easier. The style and feel of a web page is handled by CSS. You can use CSS to manage the colour of the text, font style, paragraph spacing, how columns are scaled and laid out, what background pictures or colours are used, layout designs, display variants for different devices and screen sizes, and a multitude of other effects. CSS is simple to learn and understand, but it gives you a lot of power over how an HTML document looks. CSS is frequently used in conjunction with the markup languages HTML or XHTML. (Anon., n.d.)

5.2.5 Heroku

Heroku is a Platform as a Service (PaaS) platform for hosting serverless applications. This means that scripts can be written to fulfil either of these requirements. (LENDAVE, 2022) The Heroku platform, which is an infrastructure as a service solution, is hosted by AWS (Amazon Web Services). (LENDAVE, 2022) Heroku is a free platform with a limit of 500 hours of uptime. The apps run on a dyno, which goes to sleep after 30 minutes of inactivity. (LENDAVE, 2022) This ensures that while you're not using your app, it doesn't eat up all of your free time. On the platform, Ruby, Java, PHP, Python, Node, Go, and Scala are all supported. (LENDAVE, 2022) The majority of Data Science beginners utilise this platform to obtain real-world experience running and deploying models. (LENDAVE, 2022)

6 Github Link for my project code and Website link

Github Link of my project code which includes coding file, pickle file, app file (python file: contains flask), dataset, index.html, style.css, etc :

https://github.com/Abis47/Parkinson_Detection-KNN_WebApp

Link of Webpage which I have created as a part of deployment of model is:
https://parkinson-detector-knn.herokuapp.com/

Here is a screenshot (taken at 50% zoom level) of webpage created:

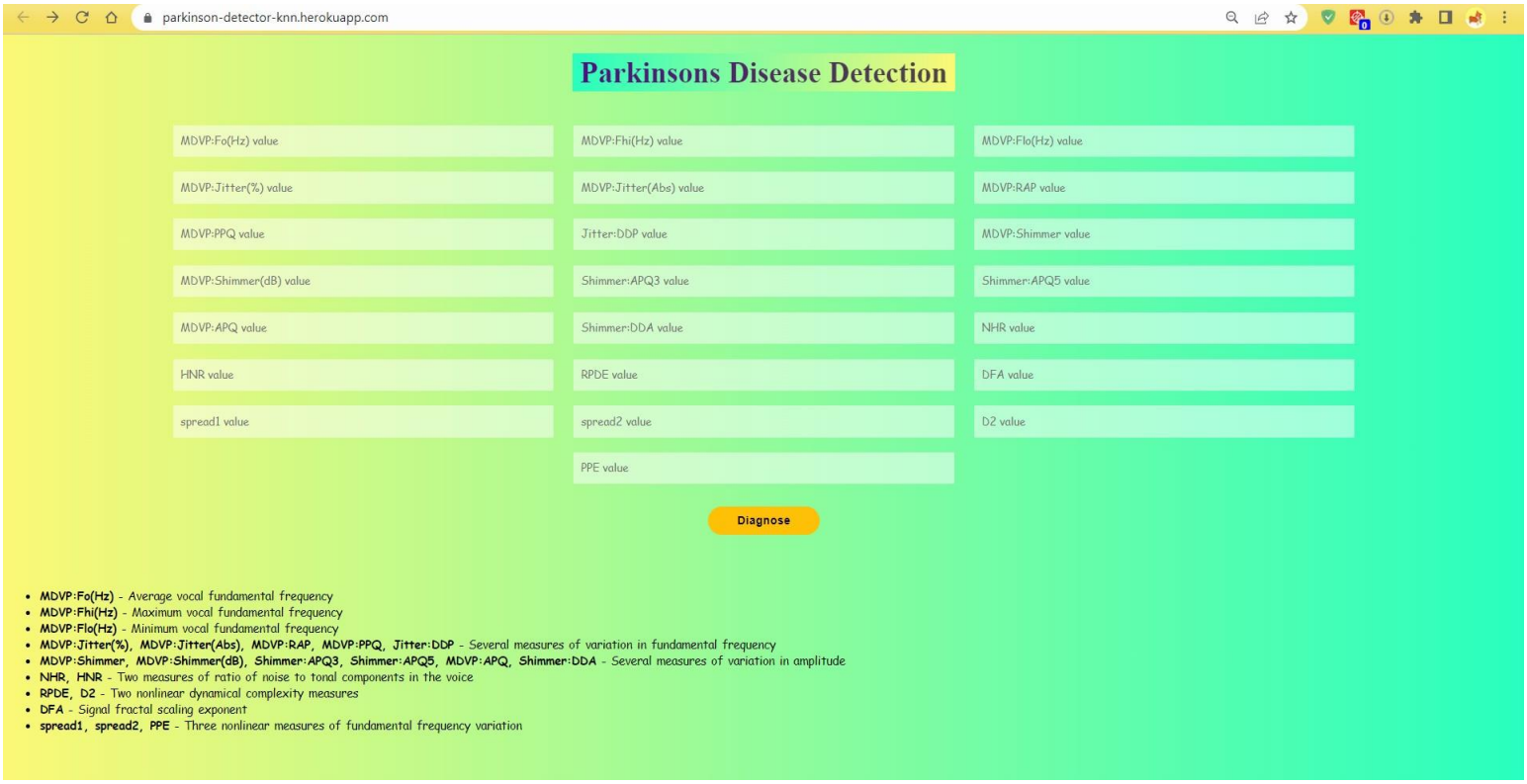


Figure 2: Screenshot of webpage created as a part of Model Deployment

7 Conclusion

I have successfully made a machine learning model which can detect Parkinson disease in a person by providing certain information. With digitalization taking over every industry, including healthcare, Machine Learning helping healthcare industry and normal people to diagnose the diseases (like Parkinson disease) easily and at cheaper cost. Not only this, there are numerous application of machine learning in Pharmaceutical and Healthcare industry, detection of disease is just a one part/one application of it.

8 References

aniruddha, 2020. *AUC-ROC Curve in Machine Learning Clearly Explained.* [Online] Available at: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/
[Accessed 01 05 2022].

Anon., n.d. *What is CSS?.* [Online] Available at: https://www.tutorialspoint.com/css/what_is_css.htm
[Accessed 01 05 2022].

Braak, H. a. B. E., 2000. Pathoanatomy of Parkinson's disease. *Journal of neurology,* 247(2), pp. II3--II10.

Davie, C. A., 2008. A review of Parkinson's disease. *British medical bulletin,* 86(1), pp. 109-127.

De Lau, L. M. a. B. M. M., 2006. Epidemiology of Parkinson's disease. *The Lancet Neurology,* 5(6), pp. 525-535.

Education, I. C., 2020. *Machine Learning.* [Online] Available at: https://www.ibm.com/in-en/cloud/learn/machine-learning
[Accessed 01 05 2022].

LENDAVE, V., 2022. *Step by step guide to deploy an ML model using Heroku.* [Online] Available at: https://analyticsindiamag.com/step-by-step-guide-to-deploy-an-ml-model-using-heroku/
[Accessed 01 05 2022].

Matthews, V., 2016. *An Analysis of Nonlinear and Fractal Measures in Vocal Frequencies for Developing Models to Predict Parkinson's Disease.* [Online] Available at: https://rstudio-pubs-static.s3.amazonaws.com/197257_4ecb7314bc1f4619b59ab44a1374ce7f.html
[Accessed 01 05 2022].

S., A., 2022. *What Is HTML? Hypertext Markup Language Basics Explained.* [Online] Available at: https://www.hostinger.in/tutorials/what-is-html
[Accessed 01 05 2022].

Sharma, R., 2021. *Model Deployment using Flask.* [Online] Available at: https://towardsdatascience.com/model-deployment-using-flask-c5dcbb6499c9 [Accessed 01 05 2022].

Sveinbjornsdottir, S., 2016. The clinical symptoms of Parkinson's disease. *Journal of neurochemistry,* Volume 139, pp. 318-324.