Project Report

Detecting Parkinson’s Disease using Machine Learning

1.INTRODUCTION:

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python.

1.1 Project Overview:

Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

1.2 Purpose :

Parkinson's disease (PD) is a long-term degenerative disorder of the central nervous system that mainly affects the motor system.Where the symptoms usually emerge slowly. Because of which the detection and treatment is done at later stage of the diseasae. The main purpose of this project is early detection of the disease and early diagnosis of Parkinson's disease using machine learning techniques and SVM algorithm with voice input.

2. LITERATURE SURVEY:

Find the Best Possible Accuracy on Different Kernel Values for The Given Dataset. They Have Downloaded the Dataset From [10], where 197 Voice Recording Samples Are There Of 31 People from Which 23 Are Having the Parkinson Disease. On Changing the Split Ratio and Repeating the Test They Achieved Better Result. On The Random Split of Dataset, They Concluded That the Best Accuracy Achieved was 65.2174%.

Samavedham et al. [12] proposed an innovative and effective approach for monitoring the disease progression and clinical diagnosis, which is based on the combination of Self-Organizing Maps and Least Squares Support Vector Machines. The proposed approach can achieve an accuracy of up to 97% concerning the differential diagnosis of PD using the PPMI dataset. The same group of authors used unsupervised learning techniques to identify reliable biomarkers to aid the diagnosis of neurodegenerative diseases .

Segovia et al. [7] demonstrated a new method based on SVMs and Bayesian networks to separate IPS from APS (atypical parkinsonian syndromes) that makes use of the 18FFDG PET dataset, that allows assessing the glucose metabolism of the brain. Their methodology achieved an accuracy rate over 78%, a reasonable result between sensitivity and specificity, suggesting the proposed method is suitable to assist the diagnosis of PD.

Wahid et al. [3] presented a study with two main contributions: firstly, they used a multiple regression normalization strategy to identify differences in spatial-temporal gait features between PD patients and control (healthy) individuals. Secondly, they evaluated the effectiveness of machine learning strategies in classifying PD gait after multiple regression normalization. The authors argued the study has important implications for the analysis of spatial-temporal gait data concerning the diagnosis of PD, as well as the evaluation of its severity with five machine learning strategies employed to classify PD gait: kernel Fisher Fiscriminant (KFD), Naïve Bayesian Approach (NB), k-nn, SVM, and Random Forest (RF).

Shamir et al. [9] proposed an approach called Clinical Decision Support Systems (CDSS) to examine the results of the incorporation of patient-specific symptoms and medications into three key functions: (i) information retrieval; (ii) visualization of treatment; and (iii) recommendation on expected effective stimulation and drug dosages. In order to fulfil this purpose, the authors used Naïve Bayes, Support Vector Machines and Random Forest to predict the treatment outcomes. The combined machine learning algorithms were able to accurately predict 86% of the motor improvement scores at one year after surgery.

Salama A. Mostafa et al. [9] proposed (i) Multiple Feature Evaluation Approach (MFEA) of a multi-agent system (ii) Implementation of five classification schemas which are Decision Tree, Random Forests, Neural Network, Naïve Bayes and Support Vector Machine on the Parkinson’s diagnosis before and after applying their approach, and (iii)Author approach witnessed the following average rate of accuracies : Decision Tree achieved accuracy of 10.51%, Naïve Bayes shown 15.22%, Neural Network is found with 9.19%, Random Forests and SVM performed with 12.75% and 9.13% respectively.

Mohammad S Islam et al. [7] Conducted A Comparative Analysis To Detect Parkinson’s Disease Using Various Classifiers. Support Vector Machine (SVM), Feedforward Back-Propagation Based Artificial Neural Network (FBANN) And Random Tree (RT) Classifiers Were Used and A Comparison Between Them Is Made to Differentiate Between PD And Healthy Patients. The Study Has Utilized the UCI Machine Learning Repository From [8],[9]. The Dataset Consists Of 195 Voice Samples From 31 Individuals Comprising of Both Males and Females. From The Taken Subjects 23 Were Determined with PD And 8 Were Healthy. To Improve the Classification Accuracy with Minimal Error Rate A 10-Fold Cross Validation Which Was Repeated 100 Times Has Been Implemented for All the Three Classifiers. The FBANN Classifier Has Achieved A 97.37% Recognition Accuracy Thus Outperforming the Other Two Classifiers.

C. Okan Saka et al. [19] Used Support Vector Machine for Building A Classification Model And A Cross Validation Scheme That Is Called Leave-One-Individual-Out Is Used For Testing. This Scheme Fits with The Dataset Better Than the Traditional Bootstrapping Methods. Parkinson’s dataset Consists Of 195 Voice Samples From 31 Individuals Comprising of Both Males and Females. From The Taken Subjects 23 Were Determined with PD And 8 Were Healthy. The Dataset Was Taken from UCI Machine-

Learning Repository [8]. They Optimized the SVM Parameters as Suggested by the work In [20], [21] So as To Build an SVM Model Capable of achieving where Reported Results Of 91% with Only 4 Features And 90% with a Greater Set Consisting Of 10 Features.

Athanasios Tsanas et al. [25] Adopted Four Algorithms for Feature Selection to Diagnose PD. They Computed 132 Dysphonia Measures from Sustained Vowels. Then, Four Subsets of These Dysphonia Measures Which Are Parsimonious Are Selected Using the Feature Selection Algorithms. These Subsets Are Mapped to A Binary Classification Response Using Two Statistical Classifiers: Random Forests and Support Vector Machines. The NCVS Database Comprises Of 263 Phonations From 43 Subjects (17 Females And 26 Males; 10 Healthy Individuals And 33 PWP), An Extension of The Database Used In [25] Is Used in The Paper. It Gives 99% Overall Classification Accuracy Using Just Ten Dysphonia Features.

Carmen Camara et al. [13] Proposed A New Method for Stimulation, Detection Of Tremor Is Based On The Subtype Of Tremor The Patient Has. Electrophysiology Is the Study of Electrical Activity the Body [14],[15]. Extracellular Physiology Is the Best Method to Detect the Neurons. Measure The Electrical Potential with Microelectrodes. The Signal Is Filtered and The Lfp Signal Is Represented. The Dataset Diagnosed with Tremor-Dominant PD, And Who Underwent Surgery for The Implantation of a

Neurostimulator. Clustering And Detection Are Combined in The Proposed System. Back Propagation MultiLayer Perceptron is the Training Algorithm Used. From Their Experimentation and As a Result they Showed the Existence Of Two Subgroups Of Patients Within The Group-1 Of Patients According To The Consensus Statement Of The Movement Disorder Society On Tremor [16].

Ahmadlou et al. [6] presented an Enhanced Probabilistic Neural Networks (EPNN), a machine learning technique that make use of local decision circles surrounding training samples to control the spread of the Gaussian kernel. Using the Parkinson's Progression Markers Initiative dataset, the proposed approach obtained an accuracy of 98.6% when classifying healthy people from PD patients, and 92.5% of recognition rate when dealing with data of six clinical exams and functional neuro imaging data for two regions of interest of the brain.

Cook et al. [8] proposed to employ a combination between smart home and machine learning technologies to observe and quantify the behavioral changes of PD patients. The main focus is to aid the clinical assessment and a better understanding of the differences between healthy older adults (HOA) and older adults with cognitive and physical impairments, also classified by the authors as mild cognitive impairment (MCI). The results indicated that smart homes, wearable devices and ubiquitous computing technologies can be useful for monitoring the activity of PD patients, as well as to pinpoint the differences between HOAs and older adults with PD or MCI. However, the authors described some limitations concerning the devices, such as to operate in settings with multiple residents and interrupted activities.

Tucker et al. [10] predict a PD patient's adherence to medication protocols based on variations in their gait. Using whole-body movement data readings from the patients, it is possible to discriminate PD patients that are “on” or “off” medication with accuracy of 97% customized model, and an accuracy of78% considering a generalized model containing multiple patient gait data.

2.1 Existing problem

Betala E. et al. (2014) proposed a SVM and k-Nearest Neighbour (k-NN) Tele-monitoring of PD patients remotely by taking their voice recording at regular interval.

Shahbakhi et al. (2014) presented that a Genetic Algorithm (GA) and SVM were used for classification between healthy and people with Parkinson.

2.2. References

1. S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
2. J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
3. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999
4. M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in Proc. ECOC’00, 2000, paper 11.3.4, p. 109.
5. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997
6. Shen, D., Zhang, D., Young, A., and Parvin, B.: ‘Machine Learning and Data Mining in Medical Imaging’, IEEE journal of biomedical and health informatics, 2015
7. B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, O. Kursun, "Collection And Analysis Of A Parkinson Speech Dataset With Multiple Types Of Sound Recordings", Ieee Journal Of Biomedical And Health Informatics, Vol. 17, No. 4, Pp. 828-834, July 2013.

8. G. E. Hinton And R. R. Salakhutdinov, "Reducing The Dimensionality Of Data With Neural Networks," Science, Vol. 313, Pp. 504-507, 2006.

9. Ipsita Bhattacharya, M. P. S. Bhatia, “SVM Classification To Distinguish Parkinson Disease Patients”, Proceedings Of The 1st Amrita Acm-W Celebration On Women In Computing In India, 2010.

2.3. Problem Statement Definition

Betala E. et al. (2014) proposed a SVM and k-Nearest Neighbour (k-NN) Tele-monitoring of PD patients remotely by taking their voice recording at regular interval.

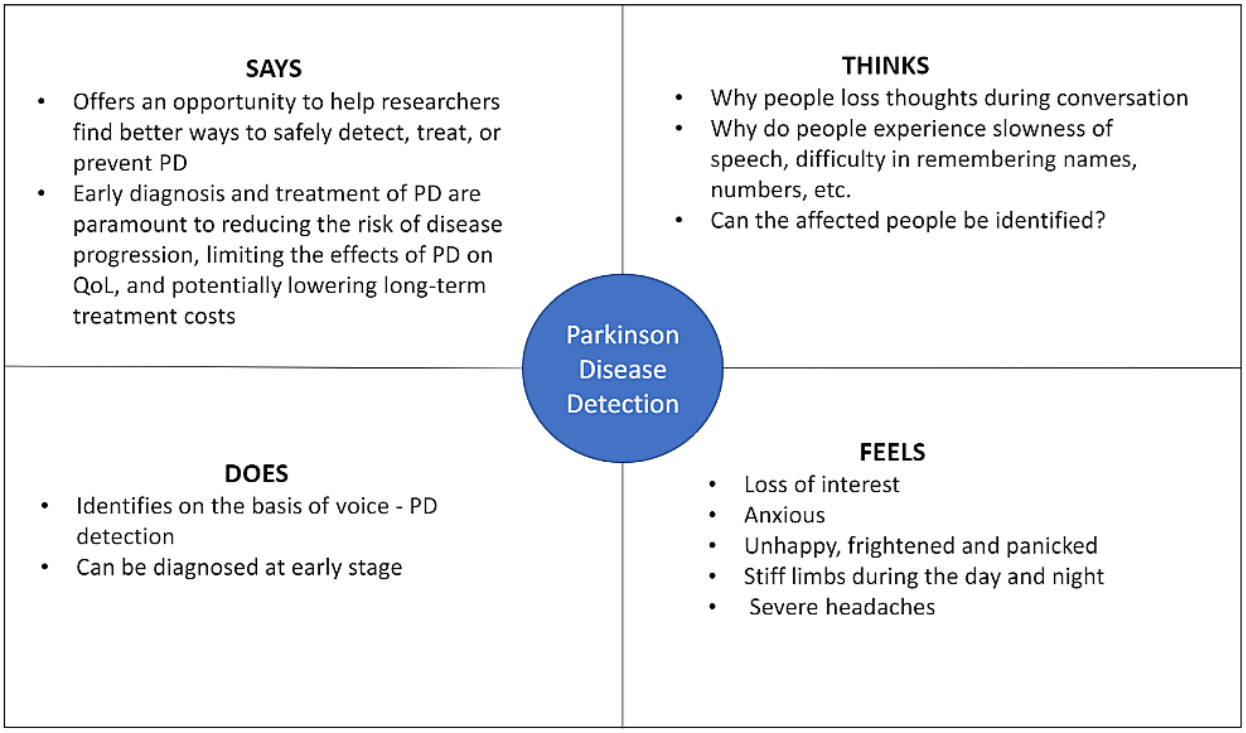
Shahbakhi et al. (2014) presented that a Genetic Algorithm (GA) and SVM were used for classification between healthy and people with Parkinson.

3. IDEATION & PROPOSED SOLUTION

Support vector machines classifier (SVM) in machine learning,is a set of related supervised learning methods widely used in pattern recognition, classification, voice activity detection and regression analysis .

A proposed system for classifying subtypes of Parkinson’s disease is strongly linked to disease duration and severity.

3.1.Empathy Map Canvas



Parkinson Disease Detection

SAYS

- Offers an opportunity to help researchers find better ways to safely detect, treat, or prevent PD
- Early diagnosis and treatment of PD are paramount to reducing the risk of disease progression, limiting the effects of PD on QoL, and potentially lowering long-term treatment costs

THINKS

- Why people loss thoughts during conversation
- Why do people experience slowness of speech, difficulty in remembering names, numbers, etc.
- Can the affected people be identified?

DOES

- Identifies on the basis of voice - PD detection
- Can be diagnosed at early stage

FEELS

- Loss of interest
- Anxious
- Unhappy, frightened and panicked
- Stiff limbs during the day and night
- Severe headaches

3.2 Ideation & Brainstorming

Why Parkinson's Disease Detection?

Globally, disability and death due to PD are increasing faster than for any other neurological disorder. The number of patients with PD in India is estimated to be 7 million. The rise in PD prevalence estimates calls attention to the increasing individual and societal burden and the pressing need for measures to address and impact this challenging disease. Early in the disease it affects the voice and rigidity, slowness of movement, and difficulty with walking. Depression and anxiety and other symptoms include sensory, sleep, and emotional problems.

Why Voice Dataset?

Early symptoms of Parkinson's affect the voice of the patient that includes symptoms like weakening of voice, jitter, shimmering and shaking of voice while talking. Hence our project detects the disease using speech recordings as a dataset there by we can detect the disease as early as possible. This detects whether a person is affected by Parkinson's disease or not.

Need for classification:

A classification algorithm is the mostly preferred predictive algorithm for detecting disease. Classification is a function that weighs the input features (voice datasets) so that the output separates one class into positive values (1- PD positive) and the other into negative values (0-PD negative).

Why Support Vector Machine algorithm (SVM)?

There are many algorithms used for classification in machine learning but SVM is better than most of the other algorithms used as it has a better accuracy in result. Basically, SVM finds a hyper-plane that creates a boundary between the types of data. In SVM, we plot each data item in the dataset in an N- dimensional space, where N is the number of features/attributes in the data and finds the optimal hyperplane to separate the data.

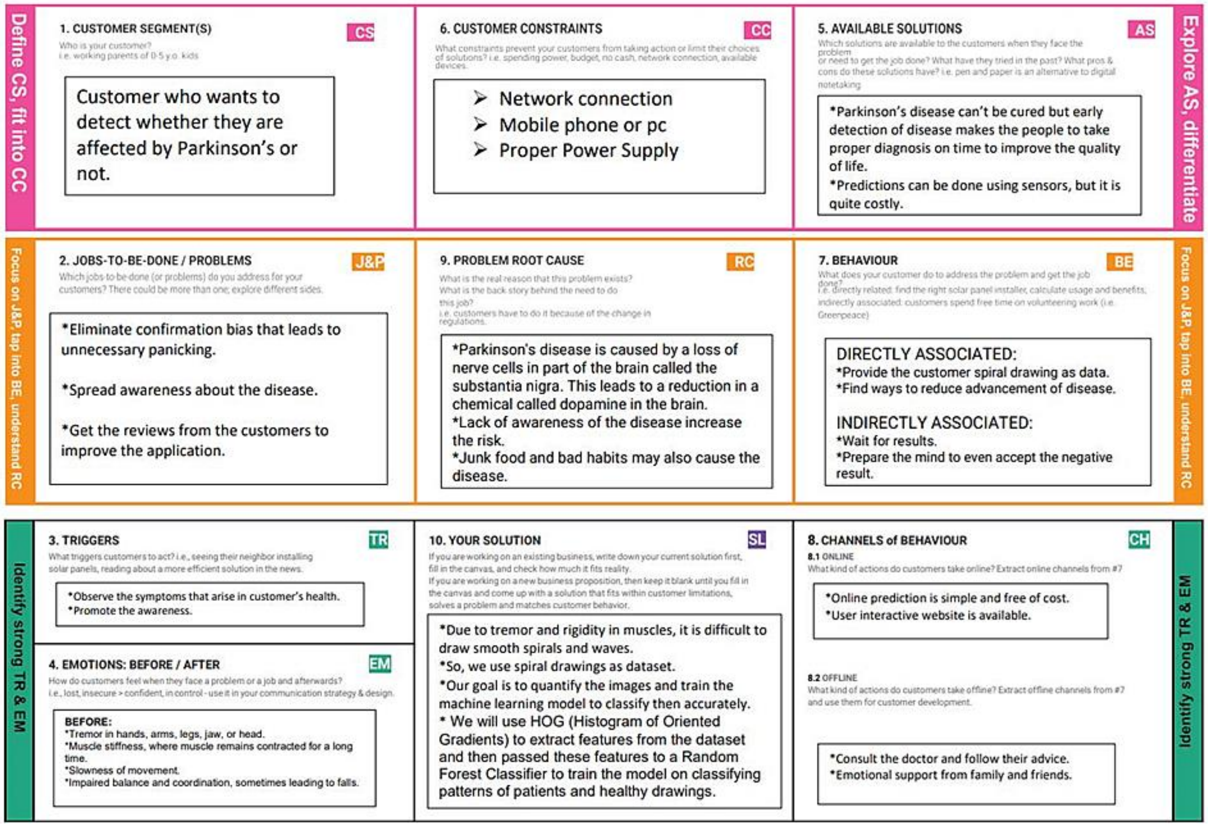
3.3 Proposed Solution:

Support vector machines classifier (SVM) in machine learning,is a set of related supervised learning methods widely used in pattern recognition, classification, voice activity detection and regression analysis .

A proposed system for classifying subtypes of Parkinson’s disease is strongly linked to disease duration and severity.

3.4 PROBLEM SOLUTION FIT:

The Problem-Solution Fit simply means that you have found a problem with your customer and that the solution you have realized for it actually solves the customer’s problem.



Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Who is your customer?
i.e. working parents of 0-5 y.o. kids

Customer who wants to detect whether they are affected by Parkinson’s or not.

6. CUSTOMER CONSTRAINTS CC

What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices.

- Network connection
- Mobile phone or pc
- Proper Power Supply

5. AVAILABLE SOLUTIONS AS

Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital notetaking

*Parkinson’s disease can’t be cured but early detection of disease makes the people to take proper diagnosis on time to improve the quality of life.
*Predictions can be done using sensors, but it is quite costly.

Explore AS, differentiate

Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS J&P

Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.

*Eliminate confirmation bias that leads to unnecessary panicking.

*Spread awareness about the disease.

*Get the reviews from the customers to improve the application.

9. PROBLEM ROOT CAUSE RC

What is the real reason that this problem exists? What is the back story behind the need to do this job?
i.e. customers have to do it because of the change in regulations.

*Parkinson's disease is caused by a loss of nerve cells in part of the brain called the substantia nigra. This leads to a reduction in a chemical called dopamine in the brain.
*Lack of awareness of the disease increase the risk.
*Junk food and bad habits may also cause the disease.

7. BEHAVIOUR BE

What does your customer do to address the problem and get the job done?
i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace)

DIRECTLY ASSOCIATED:
*Provide the customer spiral drawing as data.
*Find ways to reduce advancement of disease.

INDIRECTLY ASSOCIATED:
*Wait for results.
*Prepare the mind to even accept the negative result.

Focus on J&P, tap into BE, understand RC

Identify strong TR & EM

3. TRIGGERS TR

What triggers customers to act? i.e., seeing their neighbor installing solar panels, reading about a more efficient solution in the news.

*Observe the symptoms that arise in customer’s health.
*Promote the awareness.

4. EMOTIONS: BEFORE / AFTER EM

How do customers feel when they face a problem or a job and afterwards?
i.e., lost, insecure > confident, in control - use it in your communication strategy & design.

BEFORE:
*Tremor in hands, arms, legs, jaw, or head.
*Muscle stiffness, where muscle remains contracted for a long time.
*Slowness of movement.
*Impaired balance and coordination, sometimes leading to falls.

10. YOUR SOLUTION SL

If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality.
If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behavior.

*Due to tremor and rigidity in muscles, it is difficult to draw smooth spirals and waves.
*So, we use spiral drawings as dataset.
*Our goal is to quantify the images and train the machine learning model to classify then accurately.
* We will use HOG (Histogram of Oriented Gradients) to extract features from the dataset and then passed these features to a Random Forest Classifier to train the model on classifying patterns of patients and healthy drawings.

8. CHANNELS of BEHAVIOUR CH

8.1 ONLINE
What kind of actions do customers take online? Extract online channels from #7

*Online prediction is simple and free of cost.
*User interactive website is available.

8.2 OFFLINE
What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.

*Consult the doctor and follow their advice.
*Emotional support from family and friends.

Identify strong TR & EM

4. REQUIREMENT ANALYSIS

4.1 Functional requirement:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)

FR-1	Analyzing Symptoms	1. Stiffness in muscles 2. Rigidity andslowness in bodymovements 3. Breaking of voice and shivering in tone 4. Difficulty with walking 5. Emotional and behavioíal changes 6. Dementia and depression
FR-2	Collecting voicedataset	1. Speech and voice recordings of the patient iscollected. 2. Various voiceparameters are measured.
FR-3	Working on dataset	1. Voice recording is measured against theparameters. 2. Data is preprocessed and dependent variablesare found. 3. Data is split intotrain and testdata. 4. Training and testingis done and the modelisevaluated.

FR-4	Applying SVM algorithm	1. SVM finds a hyper-plane that creates a boundary between the types of data. 2. We plot each data item in the dataset in an N-dimensional space. 3. The algorithm tries to find the optimal hyperplane which can be used to classifydataset into healthy person or person suffering from Parkinson.
FR-5	Providing insights of dataset	1. Raw datacollection and sharing of data and systems are essential factors in hospital management. 2. According to thesedata appropriate measures can be taken. 3. Providing dataset without error. 4. Providing treatment for the patients who aresuffering from Parkinson.

4.2 Non-Functional requirements

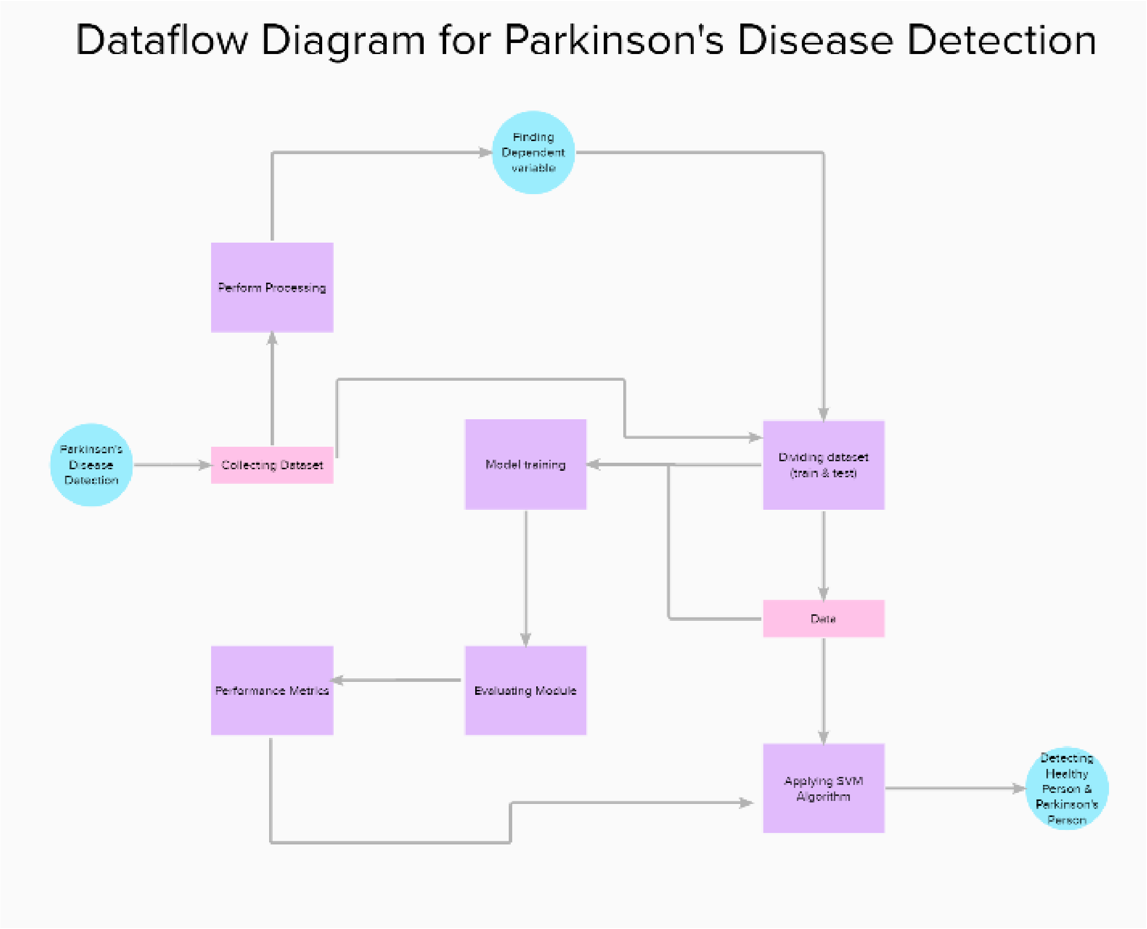
FR No.	**Non-Functional Requirement**	**Description**

NFR-1	**Usability**	1. Usable systems are straightforward to use by as many peopleas possible, bothin case of either end users or administrators to view the hospital records when needed.
NFR-2	**Security**	**Patient identification:** 1. To recognize andanalyze the patient perfectly.
NFR-3	**Reliability**	1. Understanding the current trend and working on to it to solvethe problem in an efficient manner. 2. Being software as a service, HMS is highlyresilient to any technology disruptions, downtime, or crashes experienced by othertechnology systems.
NFR-4	**Performance**	**Response time:** 1. Providing acknowledgment in minimaltimeabout the patient information. **Comfortability:** 2. To ensure that the guidelines andaccessibilities are followed.
NFR-5	**Availability**	1. Better coordination with the hospital management to provide all its resourcesaccessible when needed. 2. Accessibility of all medical facilities.

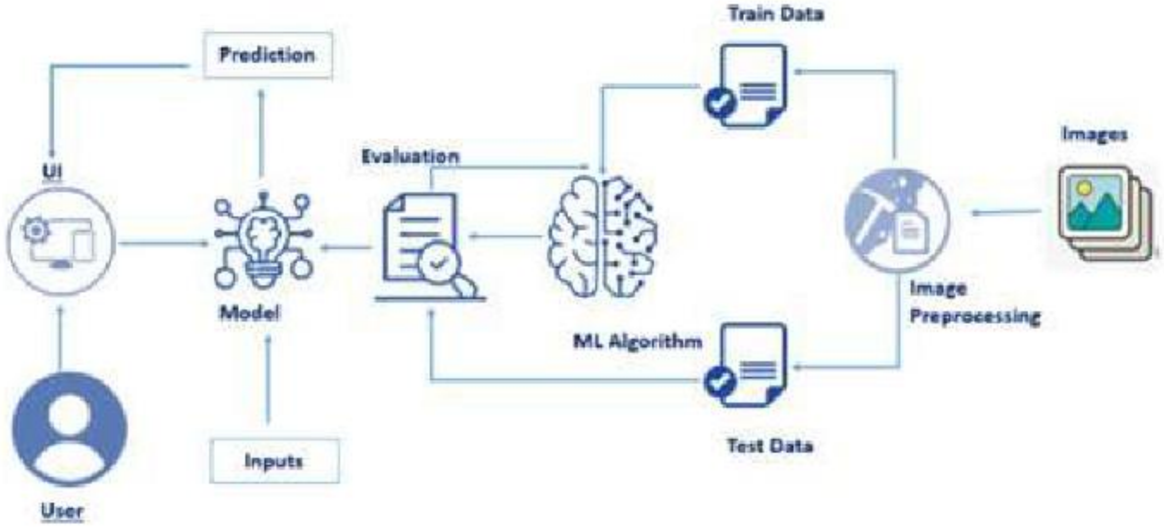
NFR-6	**Scalability**	1. Make sure thatthe work is done in more efficient way with the appropriate resources. 2. Make complex decisions understandablewith proper data.

5. PROJECT DESIGN

5.1 Data Flow Diagrams


Dataflow Diagram for Parkinson's Disease Detection
Finding Dependent variable
Perform Processing
Parkinson's Disease Detection
Collecting Dataset
Model training
Dividing dataset (train & test)
Date
Performance Metrics
Evaluating Module
Applying SVM Algorithm
Detecting Healthy Person & Parkinson's Person

5.2 Solution & Technical Architecture



5.3 User Stories:

User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web user)	Registration	USN-1	As a user, I can register for the application and I will receive	I can access my account / dashboard	Low	Sprint-1

			confirmation email.			
	Data Collection	USN-2	Once logged in, voice data is collected and perform pre-processing.	I can access the data and preform data preprocessing.	Medium	Sprint-2
	Implementation	USN-3	Application splits the dataset for training and testing in 80:20 ratio.	I can split the data into train and test data, and perform training and testing.	Medium	Sprint-3
	Implementation	USN-4	Application uses to find the hyperplane.	I can find the hyperplane by finding the maximum margin.	High	Sprint-3
	Deployment	USN-5	Predict the results.	I can find if the person has parkinson's disease or not.	High	Sprint-4

6.PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	10	High
Sprint-1		USN-2	As a user, I will receive confirmation emailonceI haveregistered for the application	8	High
Sprint-2		USN-3	As a user, I can register for the application through Facebook	2	Low
Sprint-1		USN-4	As a user, I can register for the	6	Medium

			application through Gmail.		
Sprint-1	Login	USN-5	As a user, I can log into the applicatio n by entering email & password.	10	High
Sprint-2	Data Collection	USN-6	Once loggedin, voice data is collected andperform pre-processing.	4	Medium
Sprint-3	Implementation	USN-7	As an admin, split the dataset for training and testing in 80:20 ratio.	6	Medium

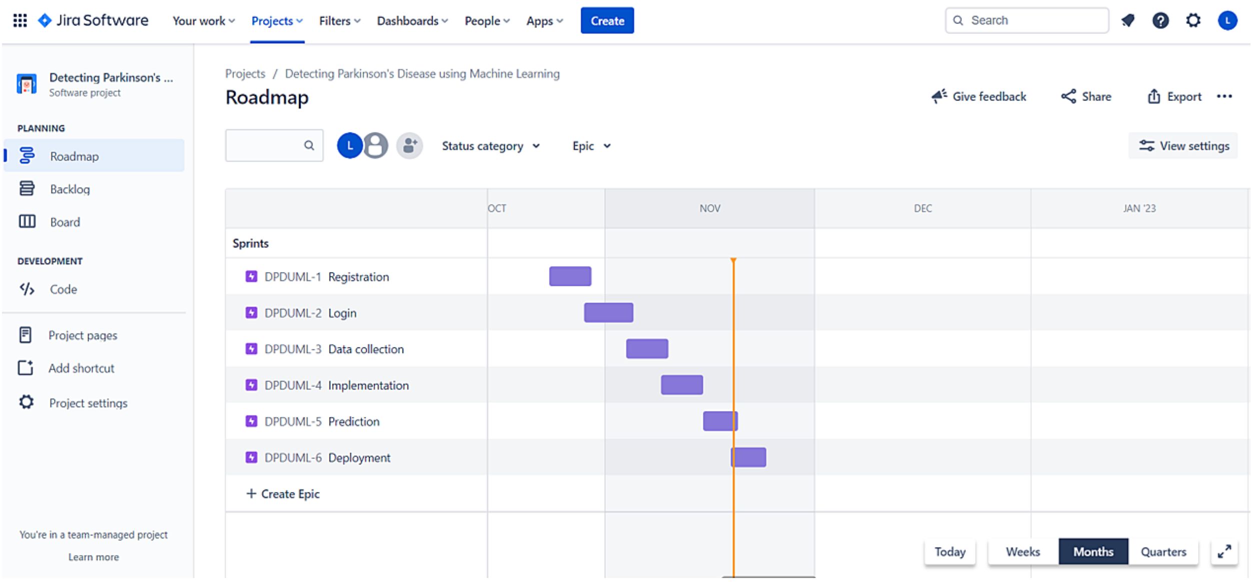
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority
Sprint-3	Implementation	USN-8	Application uses to find the hyperplane.	10	High
Sprint-4	Deployment	USN-9	Predict the results.	10	High
Sprint-4		USN-10	Deploy the model on IBMcloud.	6	Medium

6.2 Sprint Delivery Schedule:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	10	6 Days	24 Oct 2022	29 Oct 2022	10	29 Oct 2022
Sprint-2	10	6 Days	31 Oct 2022	05 Nov 2022	10	05 Nov 2022
Sprint-3	10	6 Days	07 Nov 2022	12 Nov 2022	10	12 Nov 2022

Sprint-4	10	6 Days	14 Nov 2022	19 Nov 2022	10	19 Nov 2022

6.3 Reports from JIRA:



Projects / Detecting Parkinson's Disease using Machine Learning

Roadmap

Sprints

DPDUML-1 Registration

DPDUML-2 Login

DPDUML-3 Data collection

DPDUML-4 Implementation

DPDUML-5 Prediction

DPDUML-6 Deployment

+ Create Epic

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1

Importing the Dependencies:

- NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays.

- Pandas is defined as an open-source library that provides high-performance data manipulation in Python.Pandas is used to analyze data.
- Sklearn (or Scikit-learn) is a Python library that offers various features for data processing that can be used for classification, clustering, and model selection.
- Model_selection is a method for setting a blueprint to analyze data and then using it to measure new data.
- Seaborn - data visual(themes) Matplotlib - uses graph
- Pylab - namespace

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn import svm
from sklearn.metrics import accuracy_score
import seaborn as sns
import matplotlib.pyplot as plt
import pylab as pl
```

Data Collection & Analysis:

- loading the data from csv file to a Pandas DataFrame

```
parkinsons_data = pd.read_csv('/content/parkinsons.csv')
```

- printing the first 5 rows of the dataframe

```
parkinsons_data.head()
```

- number of rows and columns in the dataframe

```
parkinsons_data.shape
```

- getting more information about the dataset

```
parkinsons_data.info()
```

- Jitter - adding a small amount of variability(horizontal or vertical) to the data to ensure all data points are visible. Multidimensional Voice Program (MDVP) analysis. Neural Human Renderer (NHR) Harmonics-to-noise ratio. perceived phonatory effort (PPE)
- getting some statistical measures about the data

```
parkinsons_data.describe()
```

- distribution of target Variable

```
parkinsons_data['status'].value_counts()
```

- Count Plot

```
sns.countplot(parkinsons_data['status'].values)
plt.xlabel('Status Value')
plt.ylabel('Status Counts')
plt.title("Parkinson's Counts in Dataset")
plt.show()
```

```
X=parkinsons_data.iloc[:,
[1,2,3,4,5,6,7,8,9,10,11,12,12,14,15,16,18,19,20,21,22,23]].values
y=parkinsons_data.iloc[:,17].values
print(X)
print(y)
```

7.2 Feature 2

- Dataset Splitting emerges as a necessity to eliminate bias to training data in ML algorithms. Modifying parameters of a ML algorithm to best fit the training data commonly results in an overfit algorithm that performs poorly on actual test data. For this reason, we split the dataset into multiple, discrete subsets on which we train different parameters.
- Splitting the dataset

```
X_train,X_test,y_train,y_test=train_test_split(X, y, test_size=0.2,
random_state=1)
```

- Feature Scaling is a method to scale numeric features in the same scale or range (like:-1 to 1, 0 to 1).
- This is the last step involved in Data Preprocessing and before ML model training.
- It is also called as data normalization.

- We apply Feature Scaling on independent variables.
- We fit feature scaling with train data and transform on train and test data.
- Feature scaling

```
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

- PCA Principal component analysis
- Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space. The input data is centered but not scaled for each feature before applying the SVD.
- The fit(data) method is used to compute the mean and std dev for a given feature so that it can be used further for scaling.The transform(data) method is used to perform scaling using mean and std dev calculated using the .fit() method.The fit_transform() method does both fit and transform.

- Applying PCA

```
from sklearn.decomposition import PCA
pca = PCA(n_components = None)
X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)
```

```
variance = pca.explained_variance_ratio_
variance.tolist()
```

- Support Vector Machine Model

```
model = svm.SVC(kernel='linear')
```

- training the SVM model with training data

```
model.fit(X_train, Y_train)
```

- accuracy score on training data

```
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(Y_train, X_train_prediction)
print('Accuracy score of test data : ', test_data_accuracy)
```

- Building a Predictive System

```
input_data =
(197.07600,206.89600,192.05500,0.00289,0.00001,0.00166,0.00168,0.00498,0.01098,0.09700,0.00563,0.00680,0.00802,0.01689,0.00339,26.77500,0.422229,0.741367,-7.348300,0.177551,1.743867,0.085569)
```

- changing input data to a numpy array

```
input_data_as_numpy_array = np.asarray(input_data)
```

- reshape the numpy array

```
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)
```

- standardize the data

```
std_data = scaler.transform(input_data_reshaped)
prediction = model.predict(std_data)
print(prediction)
```

8.TESTING

8.1 Test Cases

Section	Total Cases	Not Tested	Fail	Pass
Print Engine	7	0	0	7
Client Application	195	0	0	195
Security	0	0	0	0
Outsource Shipping	0	0	0	0
Exception Reporting	9	0	0	9

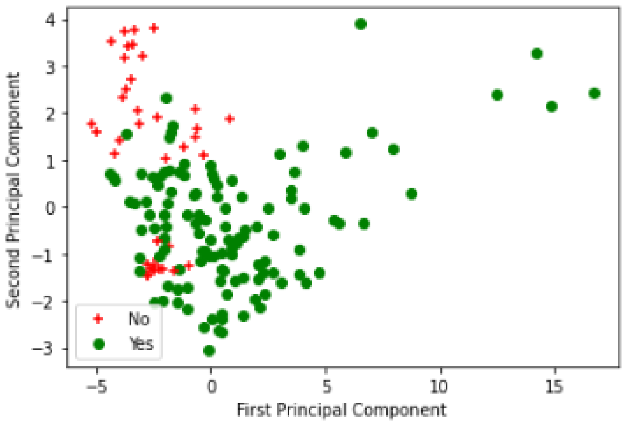
Final Report Output	4	0	0	4
Version Control	2	0	0	2

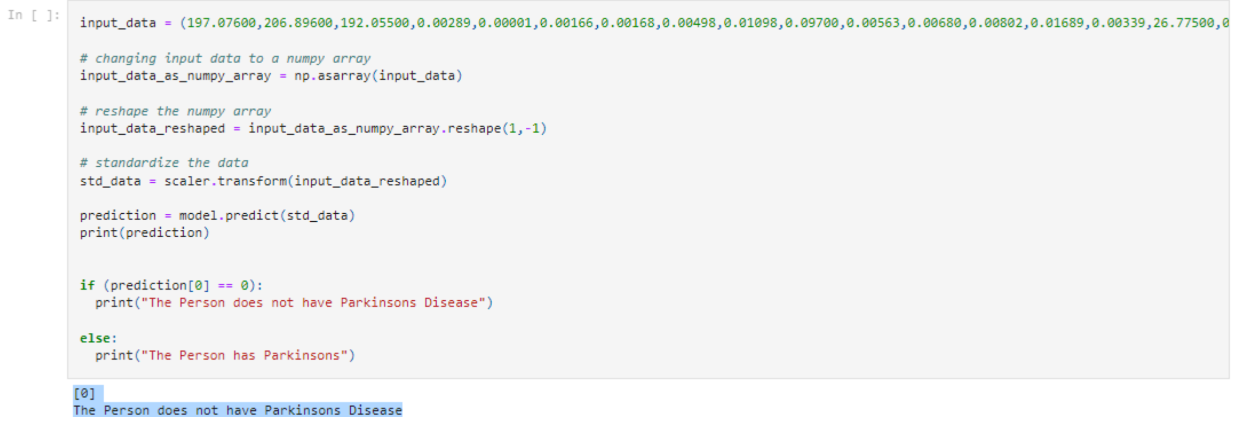
8.2 User Acceptance Testing:

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	5	3	2	5	15
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	0	0	0
Skipped	0	0	0	0	0
Won't Fix	0	5	2	0	7
Totals	19	13	11	25	69

9. RESULTS

9.1 Performance Metrics





```
In [ ]:
input_data = (197.07600,206.89600,192.05500,0.00289,0.00001,0.00166,0.00168,0.00498,0.01098,0.09700,0.00563,0.00680,0.00802,0.01689,0.00339,26.77500,0

# changing input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the numpy array
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the data
std_data = scaler.transform(input_data_reshaped)

prediction = model.predict(std_data)
print(prediction)


if (prediction[0] == 0):
  print("The Person does not have Parkinsons Disease")

else:
  print("The Person has Parkinsons")
```

```
[0]
The Person does not have Parkinsons Disease
```

10.ADVANTAGES & DISADVANTAGES:

SVM classifiers perform well in high-dimensional space and have excellent accuracy. SVM classifiers require less memory because they only use a portion of the training data.SVM performs reasonably well when there is a large gap between classes.High-dimensional spaces are better suited for SVM.When the number of dimensions exceeds the number of samples, SVM is useful.SVM uses memory effectively.

11. CONCLUSION:

Disease analysis and classification play an important role in the health care industry using Machine Learning. Detection and analysis of disease from the various difficult dataset is the most important issue in the research trend, of the lack of data samples and sufficient features to detect Parkinson's disease. These processes create difficulties in the accurate result gain. However, there are several types of disease diagnosis techniques available from the sensor to image analysis, every technique needs a data set and an approach to handle such huge datasets. There are several different methods for the diagnosis and classification of Parkinson’s disease. This survey offered different machine learning techniques to solve the Parkinson’s disease analysis and classification problem. From the analysis, effective feature selection and appropriate classifier will give better result.

12. FUTURE SCOPE:

Identify constraints that limit our project's options for developing a solution.Our future work will be to increase the size of the dataset and increase the accuracy of the model.

Hospitals want to automate the detecting the disease persons from eligibility process (real time) based on the account detail. To automate this process by show the prediction result in web application or desktop application. To optimize the work to implement in Artificial Intelligence environment.

13. APPENDIX:

Source Code

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn import svm
from sklearn.metrics import accuracy_score
import seaborn as sns
import matplotlib.pyplot as plt
import pylab as pl

# loading the data from csv file to a Pandas DataFrame
parkinsons_data = pd.read_csv('/content/parkinsons.csv')

# printing the first 5 rows of the dataframe
parkinsons_data.head()

# number of rows and columns in the dataframe
parkinsons_data.shape

# getting more information about the dataset
parkinsons_data.info()



# getting some statistical measures about the data
parkinsons_data.describe()

# distribution of target Variable
parkinsons_data['status'].value_counts()

#Count Plot
sns.countplot(parkinsons_data['status'].values)
```

```
plt.xlabel('Status Value')
plt.ylabel('Status Counts')
plt.title("Parkinson's Counts in Dataset")
plt.show()

X=parkinsons_data.iloc[:,
[1,2,3,4,5,6,7,8,9,10,11,12,12,14,15,16,18,19,20,21,22,23]].values
y=parkinsons_data.iloc[:,17].values
print(X)
print(y)

#Splitting the dataset
X_train,X_test,y_train,y_test=train_test_split(X, y, test_size=0.2,
random_state=1)

#Feature scaling
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)

#Applying PCA
from sklearn.decomposition import PCA
pca = PCA(n_components = None)
```

```
X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)
variance = pca.explained_variance_ratio_
variance.tolist()

#N components = 2
pca = PCA(n_components = 2)
X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)

#Shape after applying PCA
X_train.shape

#Data after PCA
for i in range(0, X_train.shape[0]):
    if y_train[i] == 0:
        c1 = pl.scatter(X_train[i,0],X_train[i,1],c='r', marker='+')
    elif y_train[i] == 1:
        c2 = pl.scatter(X_train[i,0],X_train[i,1],c='g', marker='o')
pl.xlabel("First Principal Component")
pl.ylabel("Second Principal Component")
pl.legend([c1, c2], ['No', 'Yes'])
pl.show()
```

```
# grouping the data based on the target variable
parkinsons_data.groupby('status').mean()

X = parkinsons_data.drop(columns=['name','status'], axis=1)
Y = parkinsons_data['status']

print(X)
print(Y)

X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
test_size=0.2, random_state=2)

print(X.shape, X_train.shape, X_test.shape)

scaler = StandardScaler()

scaler.fit(X_train)

X_train = scaler.transform(X_train)

X_test = scaler.transform(X_test)
```

```
print(X_train)

model = svm.SVC(kernel='linear')

# training the SVM model with training data
model.fit(X_train, Y_train)

# accuracy score on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(Y_train,
X_train_prediction)
print('Accuracy score of training data : ', training_data_accuracy)

# accuracy score on training data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(Y_test, X_test_prediction)

print('Accuracy score of test data : ', test_data_accuracy)

input_data =
(197.07600,206.89600,192.05500,0.00289,0.00001,0.00166,0.00168
,0.00498,0.01098,0.09700,0.00563,0.00680,0.00802,0.01689,0.0033
9,26.77500,0.422229,0.741367,-
7.348300,0.177551,1.743867,0.085569)

# changing input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the numpy array
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the data
std_data = scaler.transform(input_data_reshaped)

prediction = model.predict(std_data)
print(prediction)

if (prediction[0] == 0):
  print("The Person does not have Parkinsons Disease")
else:
  print("The Person has Parkinsons")
```

GitHub & Project Demo Link

GITHUB LINK: https://github.com/IBM-EPBL/IBM-Project-28340-1660110830

DEMO VIDEO LINK: https://www.youtube.com/embed/1oMdecrN41o