Sri Lanka Institute of Information Technology

Parkinson's Disease Detection with python

Machine Learning – IT4060
Assignment 02

Submitted by:

29/05/2022

Group Details:

Student ID	Student Name
IT19176802	Eeswar S.S
IT19244044	Samaratunga J.S
IT19114590	Gnanaseelan N.
IT19051376	Anjana W.W.M

Introduction

Description

Parkinson's disease is most likely to be inherited as a genetic disorder but it is not the case for every case. It is considered a brain disorder that causes abnormal or uncontrolled movement, such as tremors, stiffness, and loss of balance and communicating. Patients diagnosed with Parkinson often feel tightness in their muscles and they tend to loose control of their facial muscles which causes drooling. Loss of throat muscle control will cause abnormally soft voice when they speak and they tend to have trouble swallowing because of it. This could increase the danger of choking or pneumonia.

Symptoms of this disease usually start slowly but it gets worse over time. People may have difficulty walking, talking and completing their daily tasks as the disease progresses with time. They may also have psychological and behavioral changes, problems falling asleep, anxiety, memory loss and feeling of exhaustion.

Although almost anyone can be at the possibility of getting Parkinson's, some studies show that this condition can be found in more men than women. It is not yet clear why, but researches are advancing to recognize the conditions that may increase human risk. One obvious danger is age. Although most patients with Parkinson's disease start showing symptoms of the disease after age 60, about 5 to 10 percent start even before they reach 50 years. Some of the early diagnosis are inherited and forms linked to mutations in certain genes but it not the same for every case [1].

When nerve cells in the basal ganglia, which is the area of brain that controls mobility, gets disrupted or completely lose control, patents will start showing the most notable signs of Parkinson’s. Typically, it is the job of this nerve cells to produce dopamine which is a predominant chemical in human brain. With the death or paralysis of the neurons the production of dopamine decreases, causing mobility disorders associated with the Parkinson’s. The cause of the death of neurons is still a mystery to the scientists [1].

When someone is diagnosed with Parkinson's they will gradually lose sense of the nerve endings which are responsible for norepinephrine production. It is a predominant chemical messenger which keeps the sympathetic nervous system coordinated. This system regulates many bodily functions, like heart rate and Bp. This may explain some of the symptoms of Parkinson's disease which causes immobility.

1.1 Problem Addressed

Parkinson's ranked second among age-related deterioration brain disorders. It is also the most recurrent brain disorder which is related to human mobility. Estimations done by scientists reveal that it affects at least 1 percent of people who past the age of 60 worldwide.

People become more prone Parkinson's as they become old. It is now considered that the age of 60 years is when it gradually starts increasing. It is mostly found in men or people who are considered male at at birth than it is found in women or in people who are considered female at birth. Although Parkinson's disease is typically related to a person's age, it can occur in adults over the age of 20 (although this is very unlikely, and patients often have a parent or other blood relations with a similar situation) [2].

Parkinson's effects a certain part of the patient's brain causing basal ganglia to be damaged. As this area is damaged, patients lose the skills of those parts that he once kept under his own control. Scientists have discovered that Parkinson's makes significant changes in a person's brain function.

At present no blood or lab checkup can be used to to diagnose Parkinson's genetic predisposition. It is typically diagnosed by doctors with the use of patient's medical record and conducting neurological tests. It increases the possibility of Parkinson's if the start of medical procedure makes the signs of Parkinson's more frequent. Also, there are many other diseases that can cause signs such as Parkinson's disease. So it is hard and confusing to detect the Parkinson's disorder. Misjudgments are likely to be caused then diagnosing people with Parkinson's. So it is getting increasingly important to discover state-of-the-art methods to predict the risk of having Parkinson's especially when the patient has no genetic relations to the disorder. Also patents at early stages should be properly monitored for their symptoms .With the advancement of technology Parkinson's prediction can be made more accurate to reduce the risks.

As the patient lose control of throat muscle, it will cause abnormally soft voice when they speak, slur words and murmur. Such significant changes to the patient's voice are recognized as a bio marker in the early prediction of Parkinson's [3].

Dataset

This assignment's dataset was collected from the Kaggle website. Oxford Parkinson's Disease Detection Dataset was used to classify Parkinson's Disease, which was created by **Max Little of the University of Oxford.** The National Centre for Voice and Speech, Denver, Colorado collaborated to record the vocal signals. It is a CSV ASCII formatted dataset. This data set contained mainly two typed of data. There are biomedical voice measurements for Parkinson's disease and healthy people. The procedure for collecting this data is the people will be asked to speak and their speech will be recorded in various frequencies of voices. This test was conducted for 31 people, 23 with Parkinson's disease. The table of the dataset consists of 24 columns and 195-row instances. Each column represents the name, status of the disease, and features of the vocals. Of the 195 rows instances, 147 voice recording instances have Parkinson's disease, and remain 48 are for healthy people's voice recording. Oxford Parkinson's Disease Detection Dataset is well suited to make our work more successful by analyzing carefully the background of the dataset.

Description of Attributes

The following table will demonstrate the description of each attribute. One hundred ninety-five instances in 31 people with twenty-three attributes contribute the data set.

#	**Feature**	**Description**
1	name	Unique recording number and ASCII subject name
2	status	Health status of the subject 1 = Parkinson's Diseas 0 = healthy
3	MDVP:Fo(Hz)	Average vocal fundamental frequency
4	MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
5	MDVP:Flo(Hz)	Minimum vocal fundamental frequency
6	MDVP:Jitter(%)	Several measures of variation in fundamental frequency
7	MDVP:Jitter(Abs)	Several measures of variation in fundamental frequency
8	MDVP:RAP	Several measures of variation in fundamental frequency
9	MDVP:PPQ	Several measures of variation in fundamental frequency

10	Jitter:DDP	Several measures of variation in fundamental frequency
11	MDVP:Shimmer	Several measures of variation in amplitude
16	MDVP:Shimmer(dB)	Several measures of variation in amplitude
17	Shimmer:APQ3	Several measures of variation in amplitude
18	Shimmer:APQ5	Several measures of variation in amplitude
19	MDVP:APQ	Several measures of variation in amplitude
20	Shimmer:DDA	Several measures of variation in amplitude
21	NHR, HNR	Two measures of ratio of noise to tonal components in the voice
22	RPDE, D2	Two nonlinear dynamical complexity measures
23	DFA	Signal fractal scaling exponent
25	spread1, Spread2, PPE	Three nonlinear measures of fundamental frequency variation

Methodology

To predict the presence of the Parkinson's disease, a classifier is needed to reduce the over fitting. Support vector machine algorithm can be used for classification problems and regression problems. This is a binary classification problem, which is used to find the patterns in the data and understands what are the symptoms can be found in people who have Parkinson's and what are the symptoms that can found in the people who doesn't have Parkinson's.

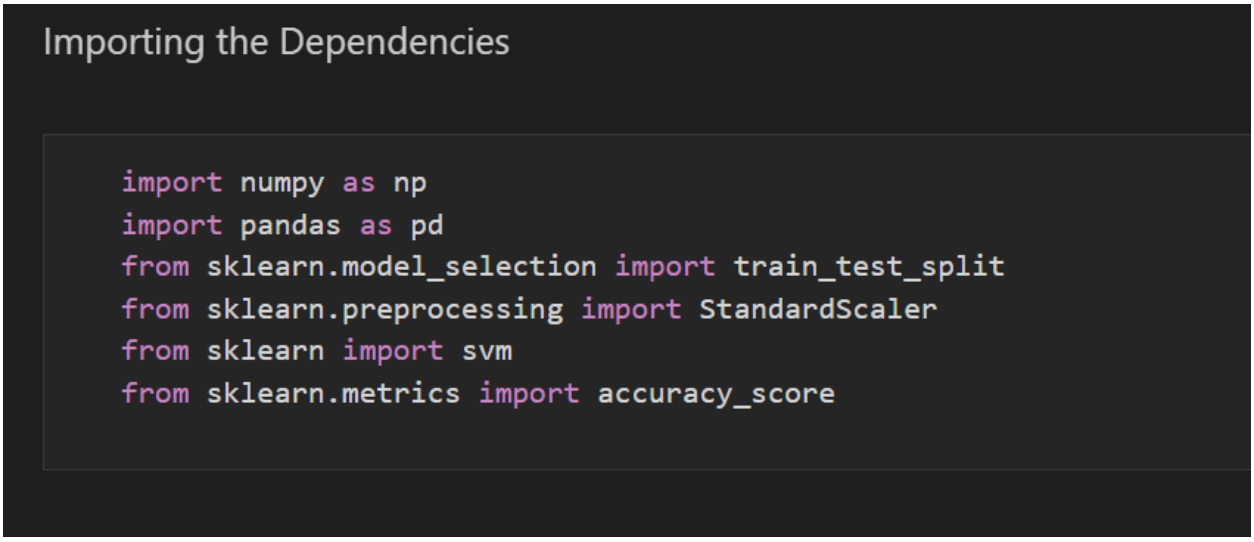
SVM

SVM is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, primarily it is used in classification problems.

The objective of SVM algorithm is to find the best decision boundary that can separate n dimensional space into classes (n means number of features). So that we will be able to put the new data point in the suitable category in future. This best decision boundary is called as Hyper plane.

Implementation

1) Importing the dependencies



Dependencies are the libraries or functions that we need for our project.

- The numpy library is useful for making arrays.
- This pandas library is useful for creating pandas data frame. Data frames are structured data. So if we have the data in a structured table we can analyze that data easily and process the data easily.

- We have imported the train_test_split from sklearn library. sklearn library is one of the most important library's when it comes to machine learning. This contains several machine learning models and other functions that we need.
- To split the data training and testing data set we need to use train_test_split.
- From the sklearn.preprocessing we have imported the standartScaler. We have to process the data before feeding into our machine learning model. And for this purpose we use standard scalar function. And this is used to standardize our data in common ring.
- SVM represents support vector machine model which we have used to predict the Parkinsons.
- Finally we have accuracy score which we have imported from sklearn metrics. This is used to evaluate the model. This is used to give the accuracy score and tell us how good the model is.

Reading the dataset

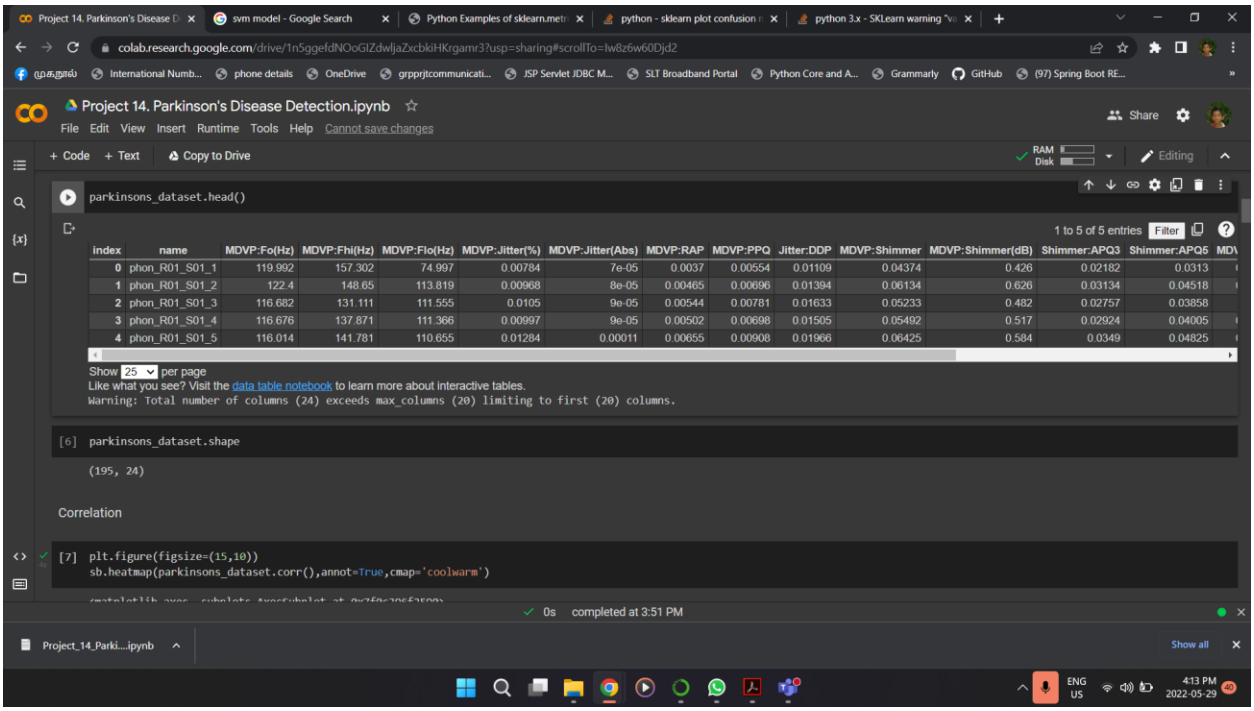
Pandas library read_csv() method used to read the data set. The dataset then assigned to the variable called Parkinson's data.

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

Data analyzing and preprocessing

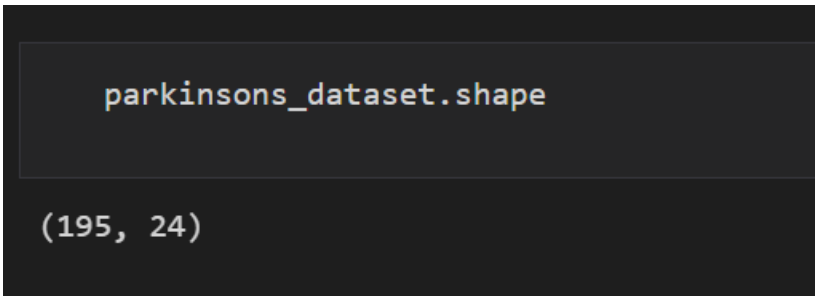
The head() function used to get first 5 rows in dataset for quickly testing if right type of data in our dataset.

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



index	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer	MDVP:Shimmer(dB)	Shimmer:APQ3	Shimmer:APQ5
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	7e-05	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313
1	phon_R01_S01_2	122.4	148.65	113.819	0.00968	8e-05	0.00465	0.00696	0.01394	0.06134	0.626	0.03134	0.04518
2	phon_R01_S01_3	116.682	131.111	111.555	0.0105	9e-05	0.00544	0.00781	0.01633	0.05233	0.482	0.02757	0.03858
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	9e-05	0.00502	0.00698	0.01505	0.05492	0.517	0.02924	0.04005
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	0.584	0.0349	0.04825

The shape function will give the number of rows and columns that present in the dataset. So we are having 195 rows and 24 columns in our data set.



Model

Parkinson's_Disease_Detection (1).ipynb

File Edit View Insert Runtime Tools Help Last saved at 5:02 PM

Comment Share M

+ Code + Text Connect Editing

```
[ ]  1 svm_model = svm.SVC(kernel='linear')
```

```
[ ]  1 modelhistroy = svm_model.fit(Xtrain, Ytrain)
```

Model Evaluation

Accuracy Score

```
1 Xtrain_predict = svm_model.predict(Xtrain)
2 train_accuracy = accuracy_score(Ytrain, Xtrain_predict)
```

```
[ ]  1 print('Accuracy score of training data : ', train_accuracy)
```

```
Accuracy score of training data :  0.8846153846153846
```

```
[ ]  1 Xtest_predict = svm_model.predict(Xtest)
     2 test_accuracy = accuracy_score(Ytest, Xtest_predict)
```

```
[ ]  1 print('Accuracy score of test data : ', test_accuracy)
```

```
Accuracy score of test data :  0.8717948717948718
```

```
[ ]  1 sb.set()
     2 get_ipython().run_line_magic('matplotlib','inline')
     3 confusionmat =confusion_matrix(Ytest, Xtest_predict)
     4 sb.heatmap(confusionmat.T,square=True,fmt='d',cbar=False)
     5 plt.xlabel('True class')
```

In order to get more information about the dataset we can use info function. Which will give some information about the dataset. It tells how many rows, columns and non-null values. Null values is missing values.

Project 14. Parkinson's Disease Detection.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

Share

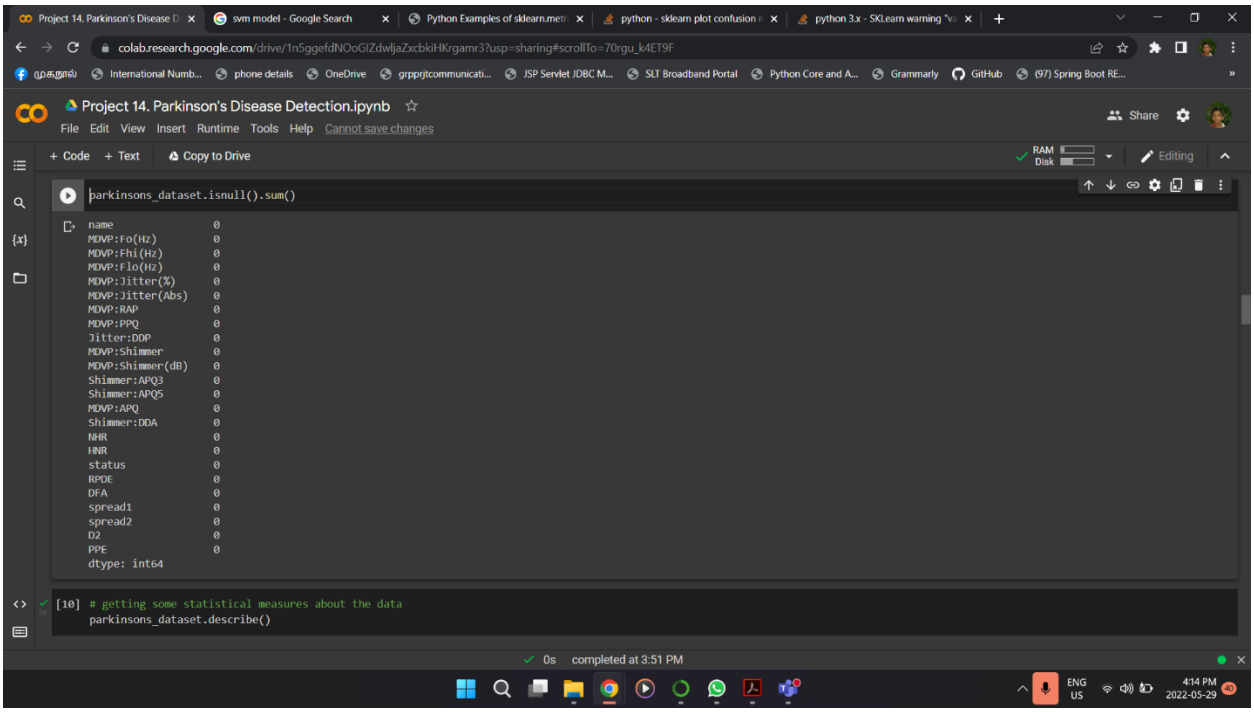
+ Code + Text Copy to Drive RAM Disk Editing

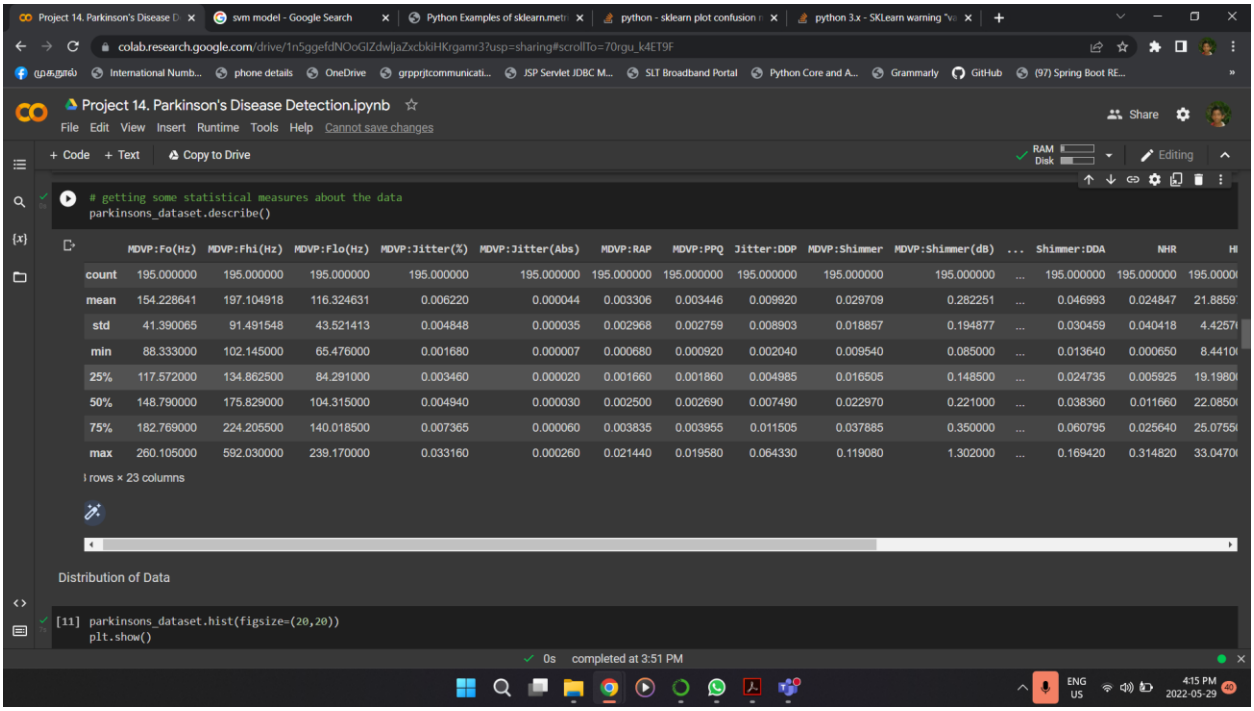
```
parkinsons_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   name              195 non-null    object
 1   MDVP:Fo(Hz)       195 non-null    float64
 2   MDVP:Fhi(Hz)      195 non-null    float64
 3   MDVP:Flo(Hz)      195 non-null    float64
 4   MDVP:Jitter(%)    195 non-null    float64
 5   MDVP:Jitter(Abs)  195 non-null    float64
 6   MDVP:RAP          195 non-null    float64
 7   MDVP:PPQ          195 non-null    float64
 8   Jitter:DDP        195 non-null    float64
 9   MDVP:Shimmer      195 non-null    float64
 10  MDVP:Shimmer(dB)  195 non-null    float64
 11  Shimmer:APQ3      195 non-null    float64
 12  Shimmer:APQ5      195 non-null    float64
 13  MDVP:APQ          195 non-null    float64
 14  Shimmer:DDA       195 non-null    float64
 15  NHR               195 non-null    float64
 16  HNR               195 non-null    float64
 17  status            195 non-null    int64
 18  RPDE              195 non-null    float64
 19  DFA               195 non-null    float64
 20  spread1           195 non-null    float64
 21  spread2           195 non-null    float64
 22  D2                195 non-null    float64
 23  PPE               195 non-null    float64
dtypes: float64(22), int64(1), object(1)
memory usage: 36.7+ KB
```

0s completed at 3:51 PM

Data preprocessing is a very important step. Most of the datasets contains missing values, may be in an unusable format. So here we check for missing values and there are no null values in our dataset.

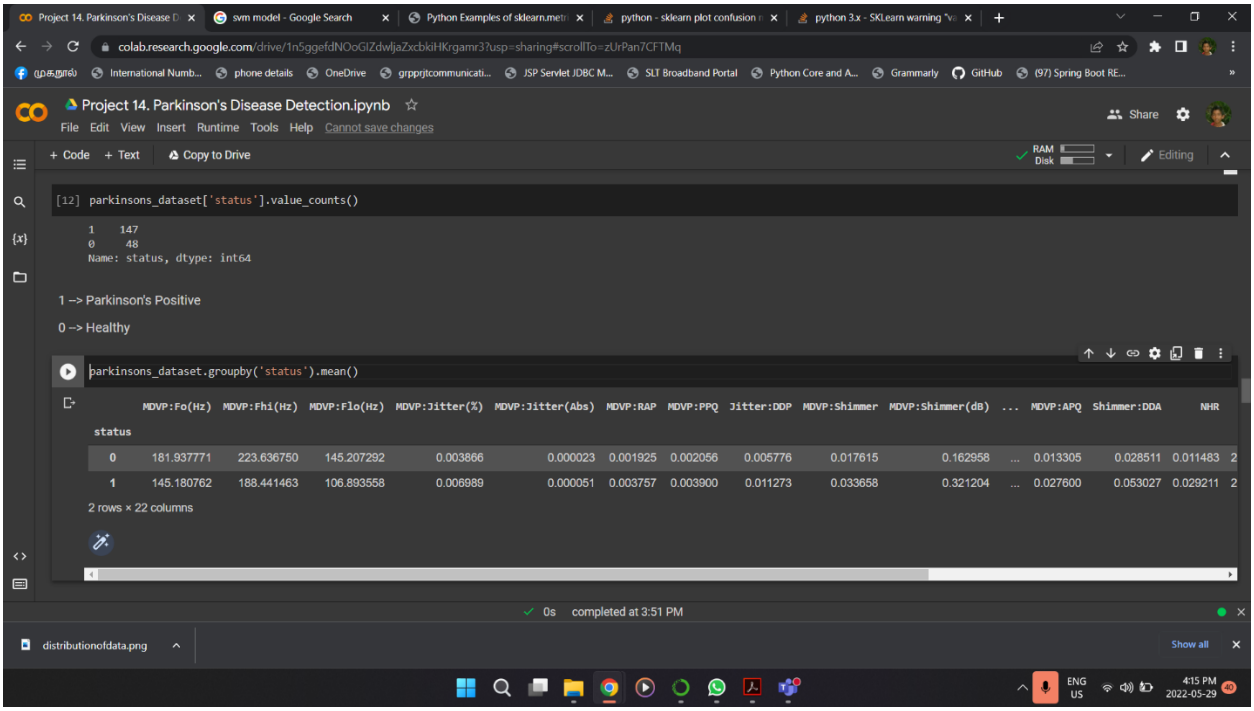




Feature selection

First we will group the data based on the target variable (status column). So we will be using group by function here and we are going to get mean value for each column.

0- Represents healthy people
1- Represents people affected by Parkinson disease



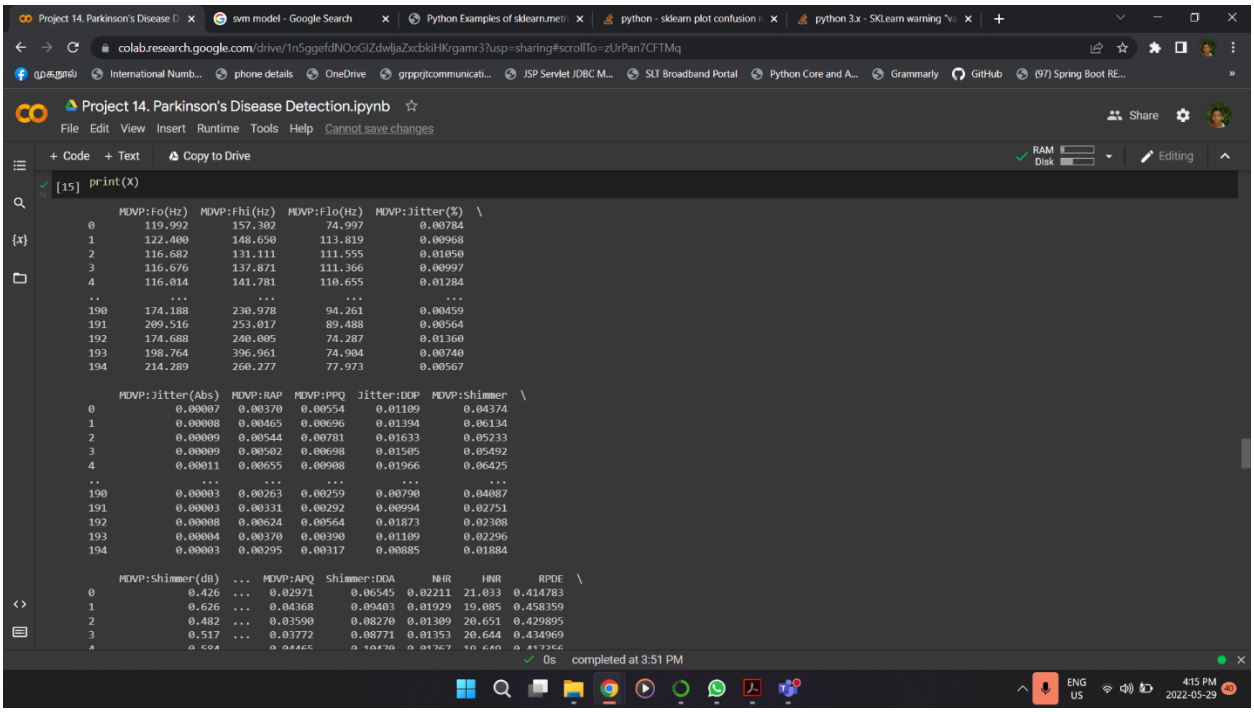
So in the output we have found the mean for each cases
So first case is the people who doesn't have Parkinson's and we took all the mean values for people who doesn't have Parkinson's. And also we have taken all the mean values for people who does have Parkinson's as well. Parkinson's people have a very less frequency in the first column where as the healthy people have voice frequency high. So we can see a difference in values right in each of the columns. So there is a clear distinction between the healthy people and the people who affected by

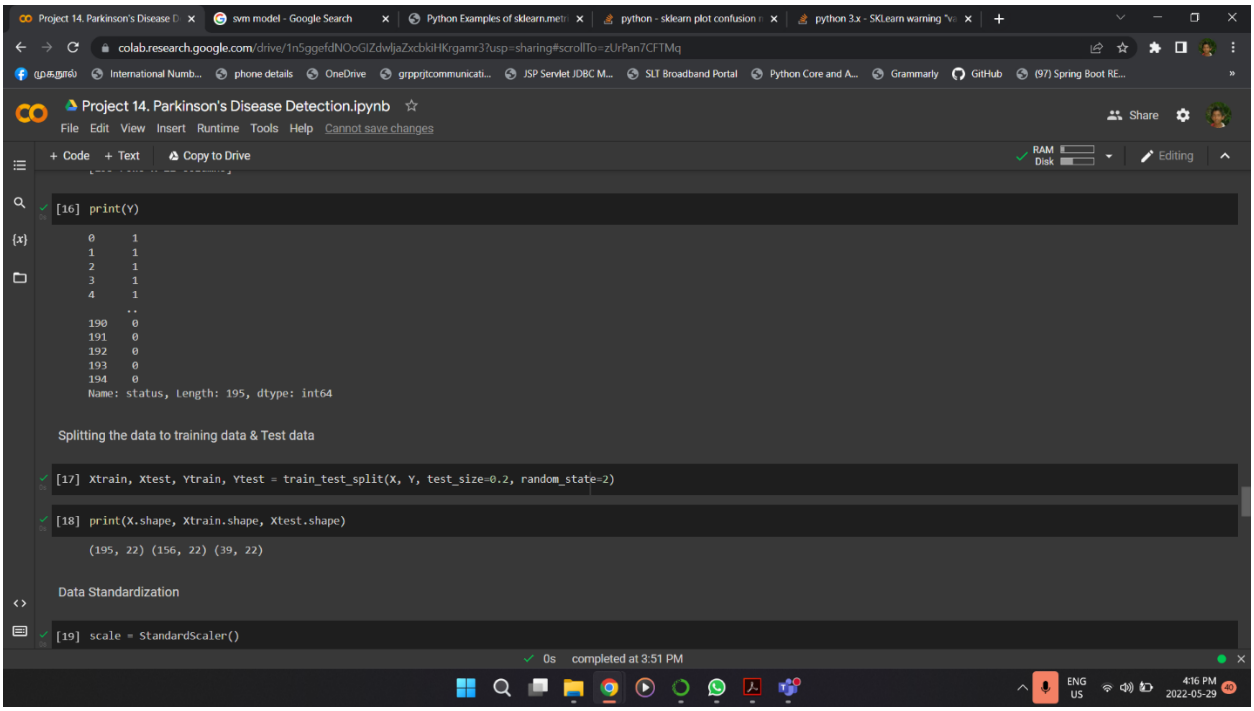
Parkinson's. So this difference will be understood by our machine learning model and it will be used for the future prediction.

Separating the feature and targe

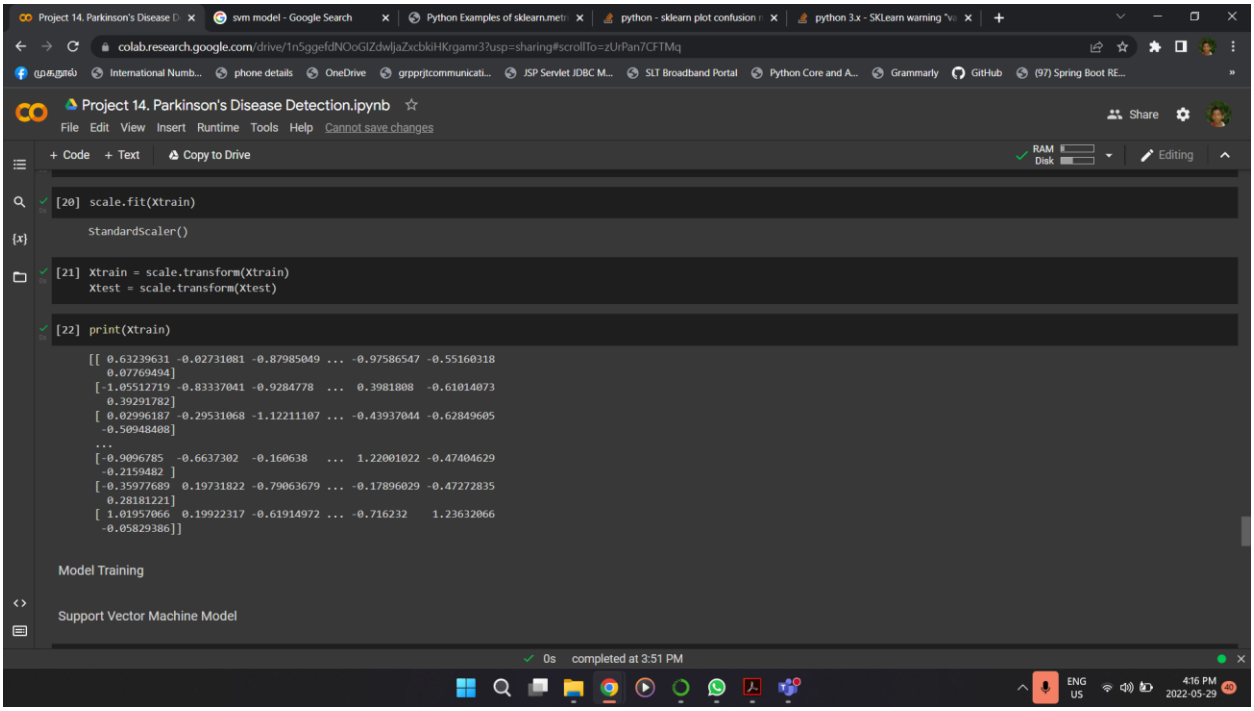
Target represents status column
Feature represents other than target column

So here we are taking all the features in x column and status column in y
This drop function drops the name column and status column and store it in x
And we will store all the status value in y.

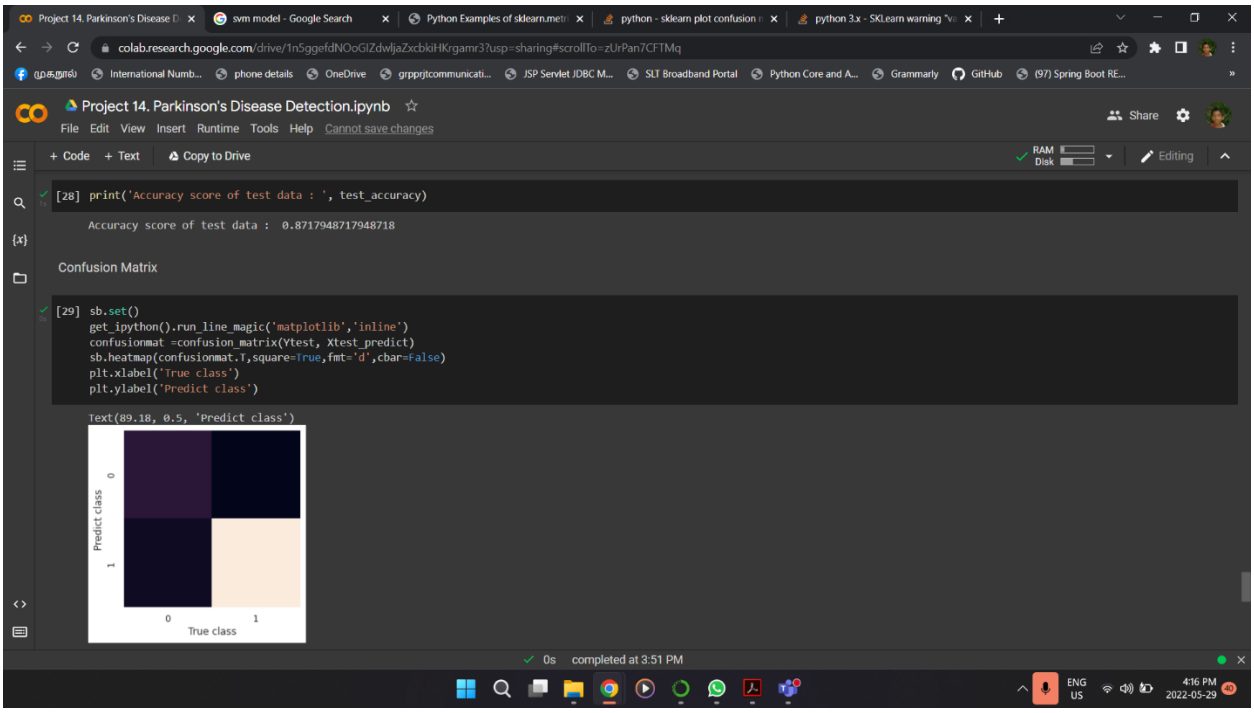




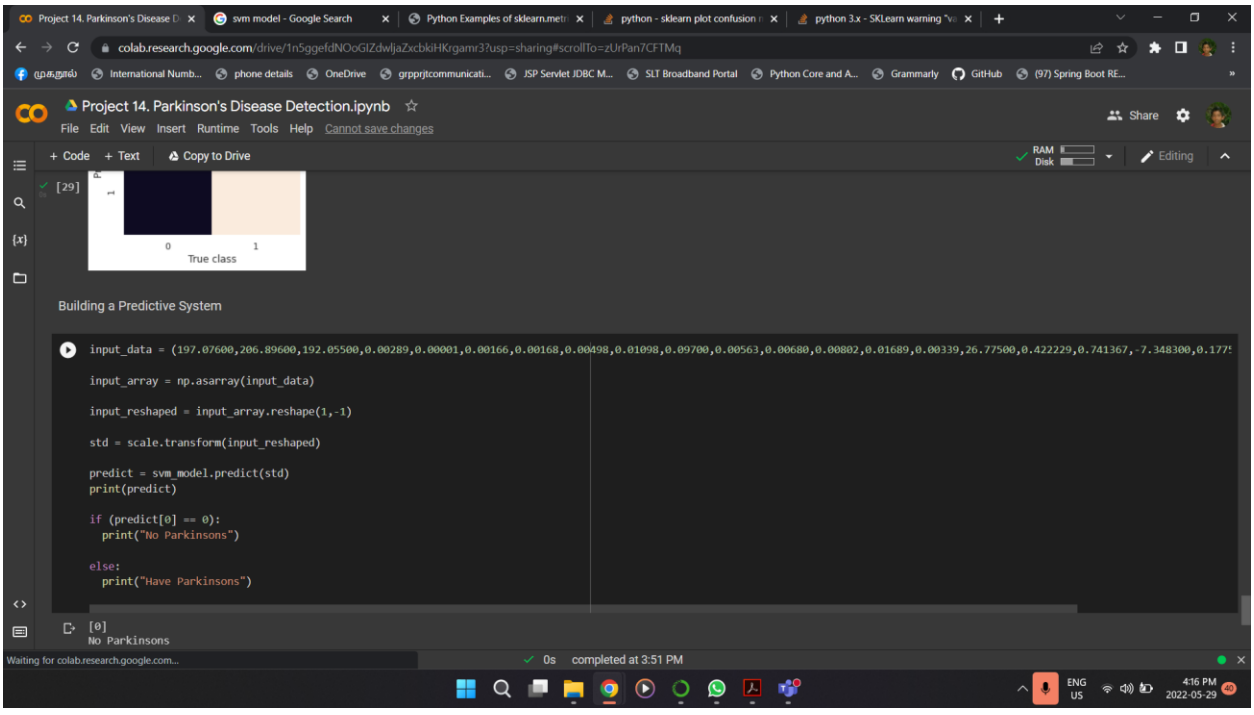
Split the data into training and testing sets.



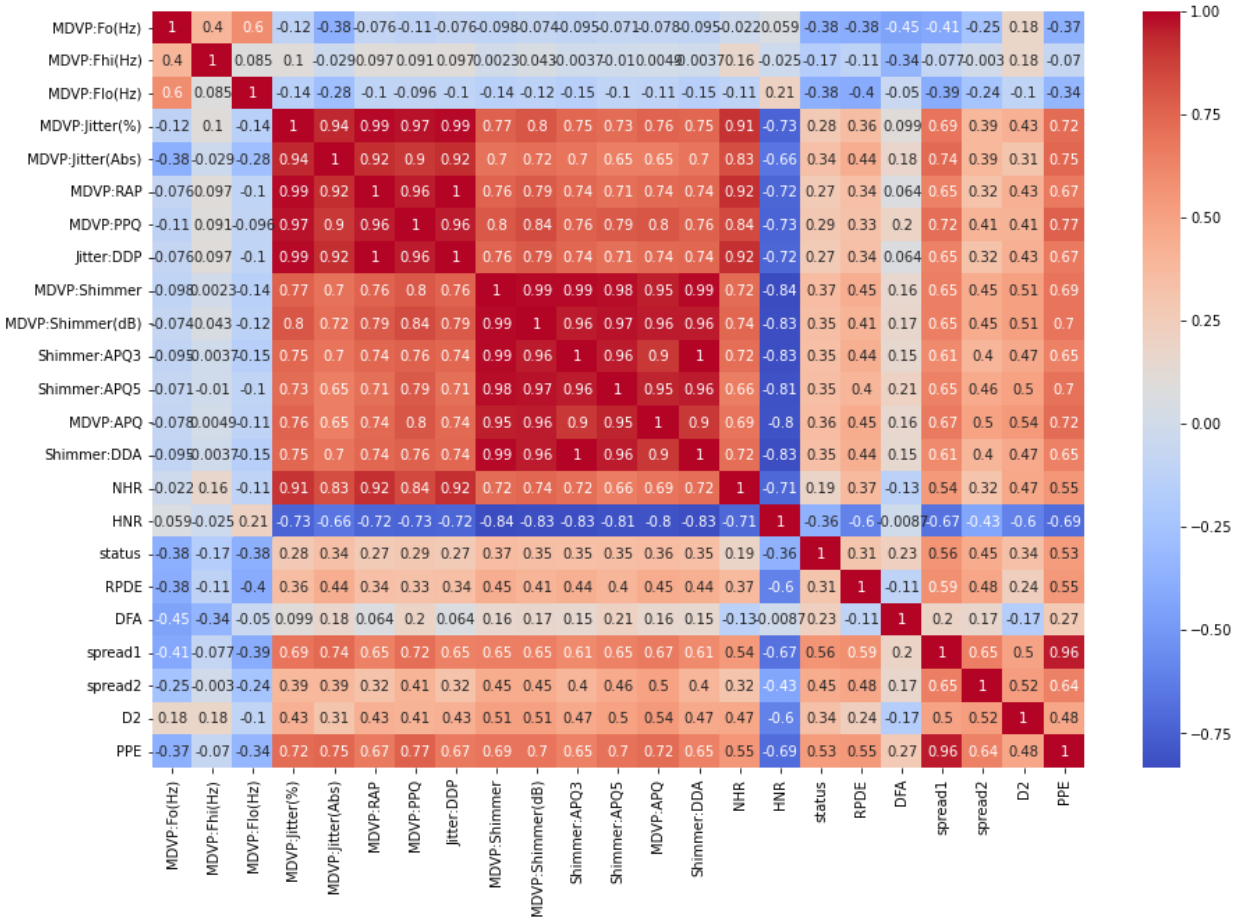
Accuracy score and Confusion matrix



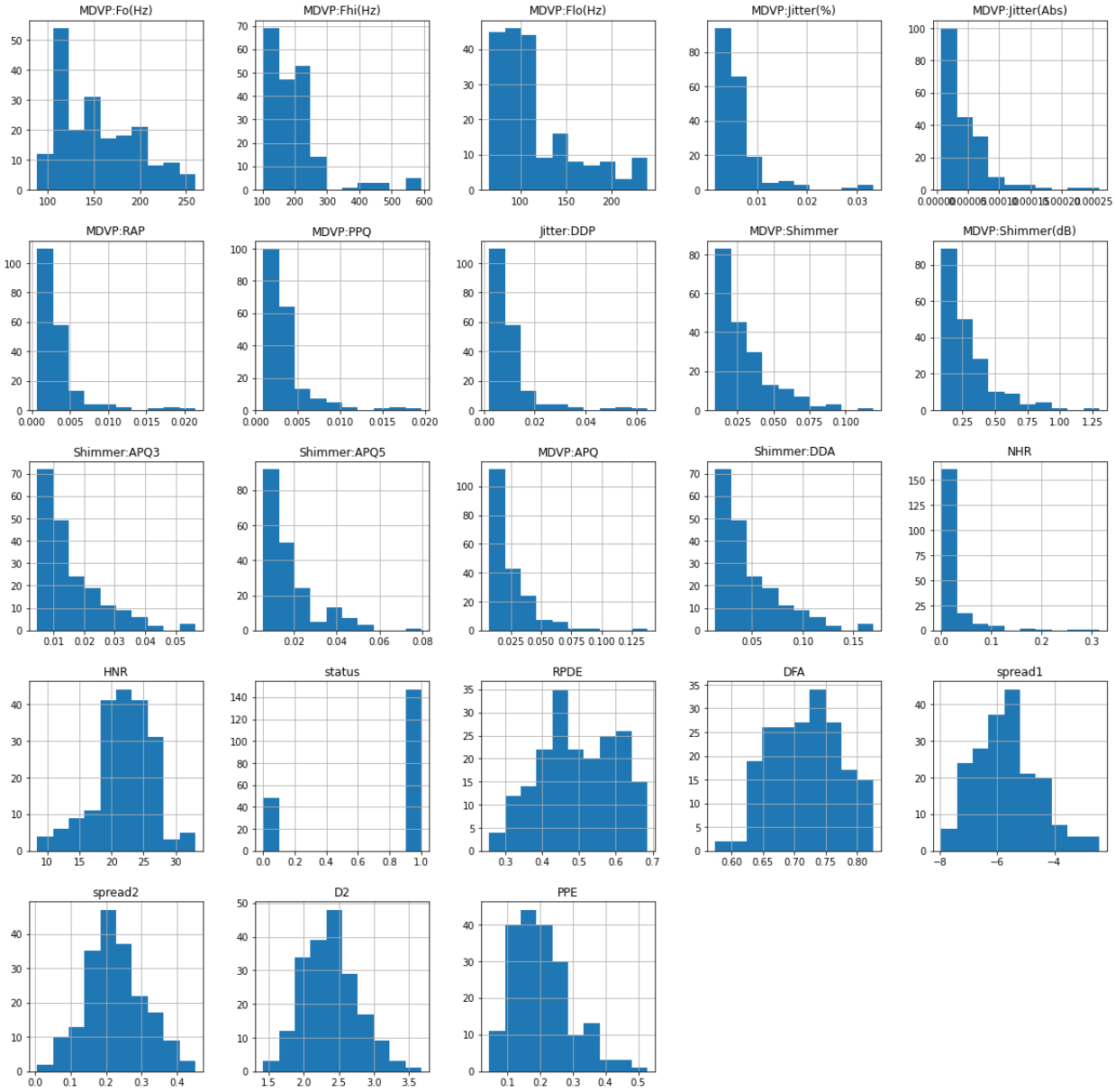
Prediction result for new patient data



Correlation



Data Distribution



Contribution

Easwar

- Model evaluation
- Report writing
- Data set searching

Nivethika

- Report writing
- Separating the features and target
- Distribution of data

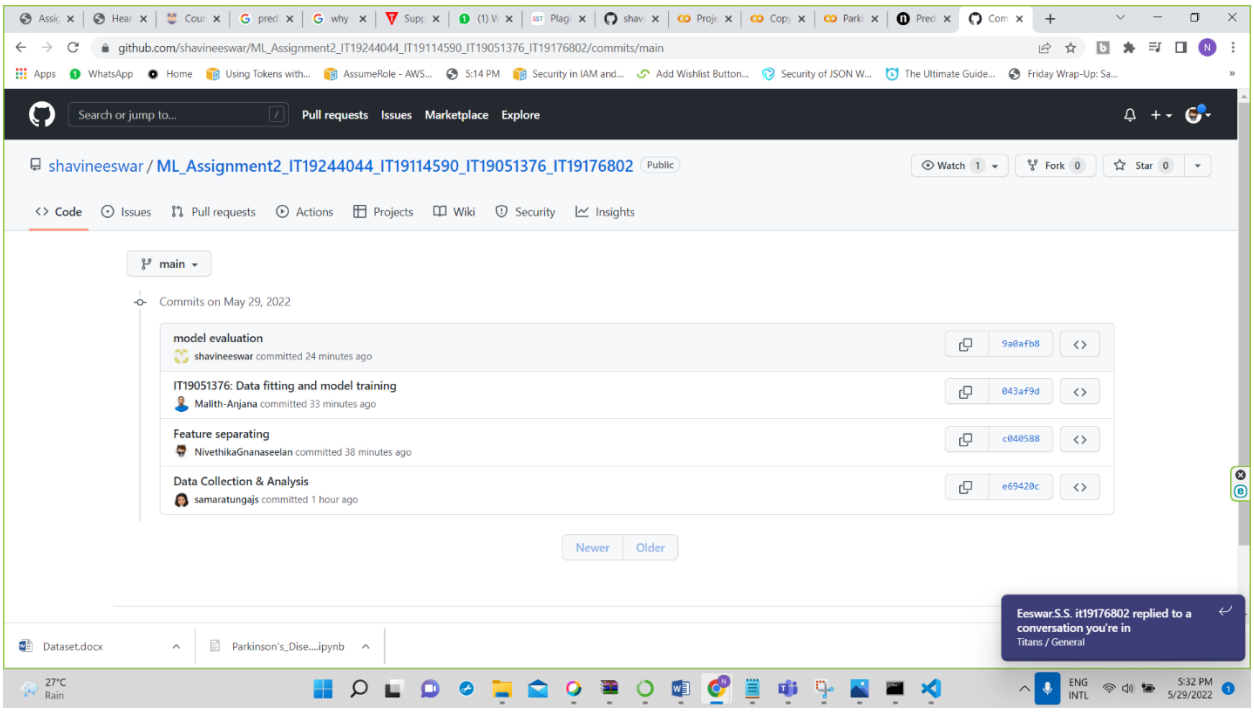
Malith

- Report writing
- Training the model
- Data set searching

Jaymini

- Video making
- Data collection and analysis
- Report writing

Commits



References

Celik, E., & Omurca, S. I. (2019). Improving Parkinson's disease diagnosis with machine learning methods. 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, EBBT 2019. https://doi.org/10.1109/EBBT.2019.8742057

Karapinar Senturk, Z. (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. Medical Hypotheses, 138, 109603. https://doi.org/10.1016/J.MEHY.2020.109603

Mathur, R., Pathak, V., & Bandil, D. (2019). Parkinson Disease Prediction Using Machine Learning Algorithm. Advances in Intelligent Systems and Computing, 841, 357–363. https://doi.org/10.1007/978-981-13-2285-3_42

Appendix

Github link:
https://github.com/shavineeswar/ML_Assignment2_IT19244044_IT19114590_IT19051376_IT19176802.git

Video Link:

https://drive.google.com/file/d/1M7sq-F60AhCsQhhRw_s2gphGIb0M5d1W/view?usp=sharing

https://mysliit-my.sharepoint.com/:v:/g/personal/it19244044_my_sliit_lk/EQNXLt9ENNtHl_QxHfm-n_oBcSn0vL054sPb8_Dr_9zJzA?e=aCaLCs