GURU GHASIDAS VISHWAVIDYALAYA

BILASPUR (C.G.)



SESSION: 2021-22

A

REPORT ON

**“PARKINSON'S DISEASE DETECTION USING MACHINE LEARNING”**

**SUBMITTED TO SUBMITTED BY**

|  |  |  |
| --- | --- | --- |
| Dr. Vikas Pandey Sir | **Tushar Awasthi** |  |
| Ass. Professor |  |  |
| GGV Bilaspur |  |  |

|  |  |  |
| --- | --- | --- |
| INDEX | | |
| S. No. | **Topic** |  |
| 1 | ABSTRACT |  |
| 2 | INTRODUCTION |  |
| 3 | METHODOLOGY   * Related Research Work * Related Research Work |  |
| 4 | MATERIALS AND METHODS |  |
| 5 | DATABASE |  |
| 6 | FEATURE DESCRIPTION   * UPSIT * RBDSQ * Cerebrospinal Fluid Biomarkers * Neuroimaging markers |  |
| 7 | PREDICTION MODELS   * Multilayer Perceptron * Bayesian Network * Random Forest * Boosted Logistic Regression |  |
| 8 | WORK FLOW |  |
| 9 | DATA FLOW DIAGRAM |  |
| 10 | REQUIREMENT SPECIFICATION |  |
| 11 | SYSTEM REQUIREMENTS |  |
| 12 | SOFTWARE DESCRIPTION   * Python * NumPy * Pandas * Sckit-Learn * Jupyter Notebook |  |
| 13 | IMPLEMENTATION   * Steps for Implementation * Coding |  |
| 14 | SCREENSHOTS |  |
| 15 | TESTING   * Unit Testing * Integration Testing * Functional Testing |  |
| 16 | CONCLUSION |  |
| 17 | REFERENCE |  |

**ABSTRACT**

Diagnosis of the Parkinson disease through machine learning approaches provides better understanding from PD dataset in the present decade. Parkinson’s disease (PD) is one of the major public health problems in the world. It is a well-known fact that around one million people suffer from Parkinson’s disease in the United States whereas the number of people suffering from Parkinson’s disease worldwide is around 5 millions. Thus, it is important to predict Parkinson’s disease in early stages so that early plan for the necessary treatment can be made. People are mostly familiar with the motor symptoms of Parkinson’s disease; however an increasing amount of research is being done to predict the Parkinson’s disease from non-motor symptoms that precede the motor ones. If early and reliable prediction is possible then a patient can get a proper treatment at the right time. No motor symptoms considered is Rapid Eye Movement (REM) sleeping Behaviour Disorder (RBD) and olfactory loss. Developing machine learning models that can help us in predicting the disease can play a vital role in early prediction. In this paper we extend a work which used the non-motor features such as RBD and olfactory loss. Along with this the extended work also uses important biomarkers. In this paper we try to model this classifier using different machine learning models that have not been used before. Thus, it is concluded that these models can be used for early prediction of Parkinson’s disease.

**INTRODUCTION**

Parkinson’s disease is a progressive disorder that affects movement leading to shaking, stiffness and difficulty with walking balance and coordination. This is the one of the important projects in machine learning because it comes under healthcare domain and healthcare is one of those fields where machine learning can contribute a lot. So, Parkinson’s symptoms usually begin gradually and get worse over time. Parkinson affects a mostly people who are more than 50 years of age so there are also cases when people below that age also get affected with it. This is one of the critical diseases in healthcare and let’s sees how we can build a machine learning system to diagnose this particular disease.

The main cause of Parkinson’s disease is actually unknown. However, it has been researched that the combination of environmental and genetic factors plays an important role in causing Parkinson’s disease. For general understanding the Parkinson’s disease is treated as disorder of the central nervous system which is the result of loss of cells from various parts of the brain. These cells also include substantia nigra cells that produce dopamine. Dopamine plays a vital role in the coordination of movement. It acts as a chemical messenger for transmitting signals within the brain. Due to the loss of these cells, patients suffer from movement disorder.

The symptoms of PD can be classified into two types i.e. **non-motor** and **motor** symptoms. Many people are aware of the motor symptoms as they can be visually perceived by human beings. These symptoms are also called as **cardinal symptoms**. These include resting tremor, slowness of movement (bradykinesia), postural instability (balance problems) and rigidity. It is now established that there exists a time-span in which the non-motor symptoms can be observed. These symptoms are called as dopamine-non-responsive symptoms. These symptoms include cognitive impairment, sleep difficulties, loss of sense of smell, constipation, speech and swallowing problems, unexplained pains, drooling, constipation and low blood pressure when standing. It must be noted that none of these non-motor symptoms are decisive, however when these features are used along with other biomarkers from Cerebrospinal Fluid measurement (CSF) and dopamine transporter imaging, they may help us to predict the PD.

**METHODOLOGY**

Different researchers have used different features and data to predict Parkinson’s disease. We have used biomedical voice of human as the main feature. The voice dataset for Parkinson disease has been retrieved from UCI Machine learning repository from Centre for Machine Learning and Intelligent Systems.

**Related Research Work:** 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 recording 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.

The following are the Attribute Information:

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, MDVP:APQ, Shimmer:DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, PPE

|  |  |  |
| --- | --- | --- |
| Attribute Information: | | |
| Matrix column entries (attributes): | | |
| name: | ASCII subject name and recording number | |
| MDVP:Fo(Hz) | Average vocal fundamental frequency | |
| MDVP:Fhi(Hz) | Several measures of variation in fundamental frequency | |
| MDVP:Flo(Hz) |
| MDVP:Jitter(%) |
| MDVP:Jitter(Abs) |
| MDVP:RAP |
| MDVP:PPQ |
| Jitter:DDP |
| MDVP:Shimmer | Several measures of variation in amplitude | |
| MDVP:Shimmer(dB) |
| Shimmer:APQ3 |
| Shimmer:APQ5 |
| MDVP:APQ |
| Shimmer:DDA |
| NHR, HNR | Two 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 | Three nonlinear measures of fundamental frequency variation | |
| spread2 |
| PPE |

The data contains 5875 number of instances and 26 attributes.

**MATERIALS AND METHODS**

A flowchart of the proposed analysis is shown down. The data was first collected and the required non-motor and biomarker features are then extracted. Then different machine learning algorithms are employed for the classification task. Finally, a comparative analysis is made based on the accuracy provided by different machine learning models.

**DATABASE**

In this study the data from Parkinson’s Progression Markers Initiative (PPMI) database was obtained. PPMI is an observational, multicentre study that collects clinical and imaging data and biologic samples from various cohorts that can be used by researchers to establish markers of disease progression in PD. PPMI has established a comprehensive, standardized, longitudinal PD data and biological sample repository that can play a vital role in the development of tools which assist in prediction of PD to obtain the recent information, the official website of PPMI (www.ppmi-info.org) can be visited. This dataset is similar to the one used in the database on 8th August 2016. On this date the data of 184 normal patients and 402 early PD subjects were collected. It is noted that PPMI has observations from each of the patients at different time intervals. Thus the data of each patient at different periods like screening or baseline, first visit, second visit and so on are available. In the present investigation the data at baseline observation are considered in, the authors have used features from University of Pennsylvania Smell Identification Test, RBD screening questionnaire, CSF Markers of Aβ1-42,α- syn, P-tau181, Ttau, T-tau/Aβ1-42, P-tau181/Aβ1-42 and P-tau181/T tau, and SPECT measurements of striatal binding ratio (SBR) data. In this study these features have been used because we felt that they are a good combination of non-motor features and biomarkers. The details of these features are given in next section.

**FEATURE DESCRIPTION**

1. **University of Pennsylvania Smell Identification Test (UPSIT):** Olfactory dysfunction is an important marker of Parkinson’s disease. It acts as sensitive and early marker for Parkinson’s disease. It is a fact that most of the people who suffer from PD have olfactory loss however it doesn’t mean that all the people with olfactory loss are suffering from PD. Olfactory dysfunction are in various forms for instance it may be impairment in odour detection or odour differentiation. A study by Posen showed that about 10% of the subjects who were suffering from odour dysfunction were at the risk of PD. For quantifying this odour loss, the data of University of Pennsylvania Smell Identification Test is used. This test is commercially available and is also one of the most reliable tests. The procedure of the test is as follows. A subject is provided with 4 different 10-page booklets. Each of these pages has a different odour. A subject has to scratch the page and smell it. For each of this page, there exists a question with four options. Depending on the odour the subject selects one of the options. This procedure is repeated for all the pages in all the booklets. Once the test is completed the UPSIT score is calculated. The maximum score can be 40 when the subject identifies each of the odours correctly. One main advantage of this is that the test takes only a few minutes. For the present analysis the UPSIT score at baseline check-up from PPMI has been taken.
2. **REM sleep Behaviour Disorder Screening Questionnaire (RBDSQ):** RBD is another non-motor symptom that plays an important role in early prediction of Parkinson’s disease. People suffering from RBD have disturbances in sleep. These disturbances include vivid, aggressive or action packed dreams. Similar to olfactory loss, studies have shown that disorder in sleep behaviour increases the risk of being affected with Parkinson’s disease. For quantifying this non-motor symptom, the REM Sleep Behaviour Disorder Screening Questionnaire is used. The RBDSQ is a 10-item patient self-rating instrument. The test contains ten short questions with answers as yes or no. A yes is equivalent to 1 and a no is equivalent to 0. The ten questions are divided such that each of the group of the questions provides the observations about a particular behaviour. Some of the examples of the questions from are “I sometimes have vivid dreams”, “The dream content mostly matches my nocturnal behaviour”, and “My sleep is frequently disturbed”, etc. As some of the subjects may have a bed partner, they can also be used in this test. Each of the answers is provided as either one or zero. In the present study the feature for sleep disorder is obtained by summing up all the answers. This sum can be a maximum of 12 if we take the first nine questions. It is observed here that a higher score in this case means a higher risk of PD in contrast to that of UPSIT score. This RBDSQ score is taken from PPMI.
3. **Cerebrospinal Fluid Biomarkers:** Biomarkers play a pivotal role in this analysis. Without the aid of biomarkers, the prediction of PD is less accurate. The biomarkers are the significant factors in increasing the accuracy of the model. Biomarkers need to be sensitive, reproducible and must be closely associated with the disease. Cerebrospinal fluid is a clear, colourless body fluid found in the brain. It has more physical contact with the brain as compared to any other fluid. Due to the close proximity with the brain, any protein or peptide which is related to the brain specific functionalities or disease are diffused into CSF. Hence, the CSF can act as an important biomarker for brain related diseases which in the present case is Parkinson’s disease.

The CSF samples are collected from PPMI. In PPMI, for each of subjects the CSF samples are obtained and certain measurements are made. These measurements include Aβ1- 42(amyloid beta (1-42), T-tau (total tau) and P-tau181 (tau phosphorylated at threonine). According to PPMI Research Laboratory these three are the important biomarkers that can be extracted from the CSF fluid. Along with this the concentration of α-Syn was also collected from PPMI database. Kang et al have mentioned that ratios like Ttau/Aβ1-42, P-tau181/Aβ1-42 and P-tau181/T-tau also play a significant role in early detection of Parkinson’s disease. In the present investigation the measurements of Aβ1- 42, T-tau and P-tau181 and also the ratios T- tau/Aβ1-42, Ptau181/Aβ1-42 and P-tau181/T-tau are taken.

1. **Neuroimaging markers:** Single-photon emission computed tomography (SPECT) is a neuroimaging technique that uses gamma rays. The SPECT is a common routine for helping a doctor to decide whether a subject is suffering from neurodegenerative diseases. According to, the SPECT imaging can detect the dopaminergic transporter loss during the early stages of PD. When a subject has an abnormal scanning then the person has more probability of being affected with Parkinson’s disease or other neuro degenerative disease. However, a normal scan denotes that the subject is suffering from other type of diseases. DatScan SPECT imaging obtained from PPMI imaging centres are used in this study. At PPMI the striatal binding ratios were calculated. The DatScan SPECT images are collected according to the PPMI imaging protocol. These raw images are then reconstructed so as to ensure consistency among different imaging centres. After this attenuation correction is performed on these images. After this the Gaussian filter is applied and it is followed by normalization. Finally the required part is extracted from the images and then the striatal binding ratio for left and right caudate, the left and right putamen are calculated. In this paper, these four striatal binding values are used as neuroimaging biomarkers.

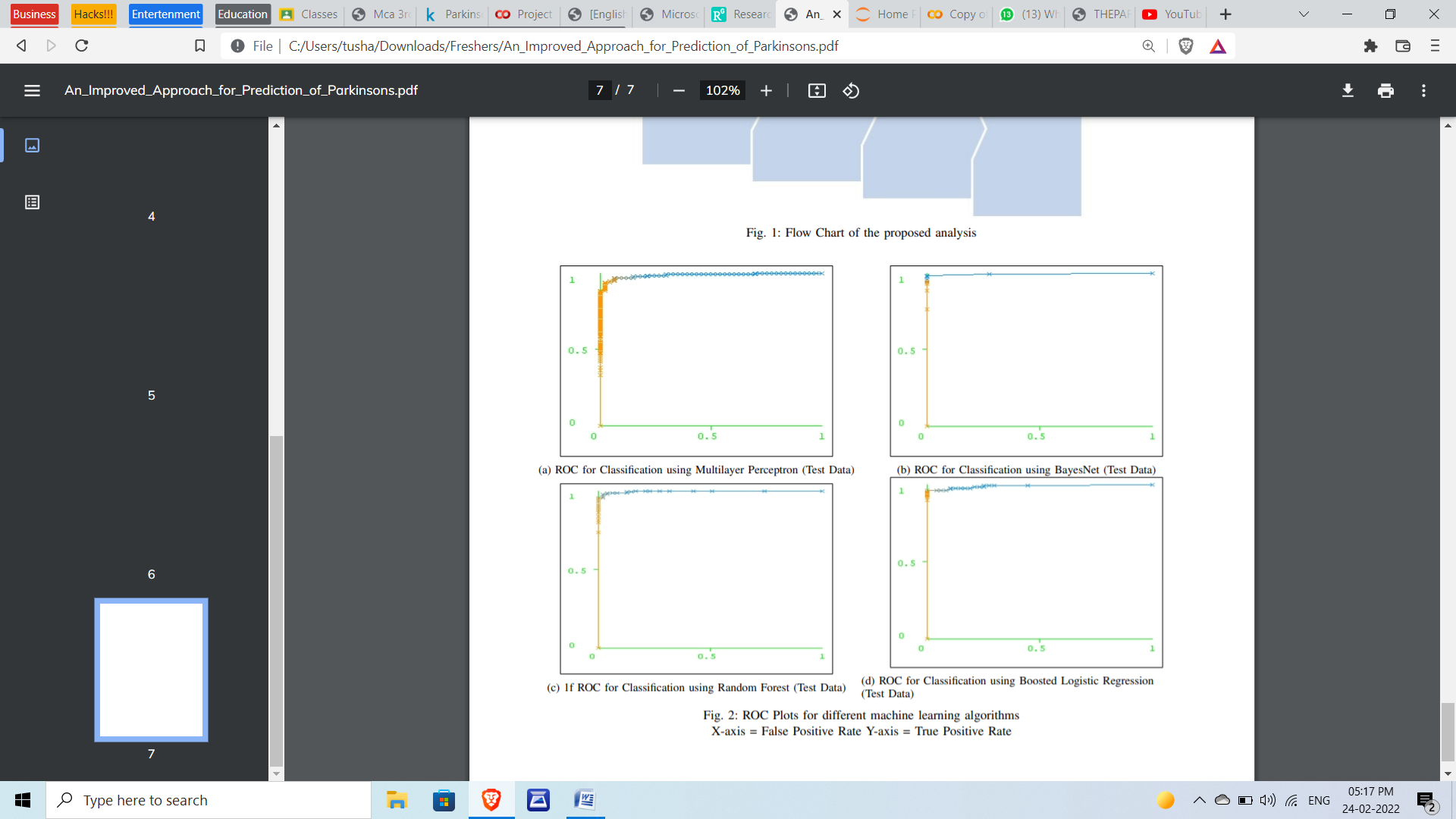
**PREDICTION MODELS**

### Support vector machine (SVM): For this particular project we use Support vector machine.  SVMs are supervised [machine learning algorithms](https://www.sciencedirect.com/topics/computer-science/machine-learning-algorithm), and they are used for classification and regression analysis. The SVM performs both linear classification and nonlinear classification. The nonlinear classification is performed using the Kernel function. In nonlinear classification, the kernels are homogenous polynomial, complex polynomial, Gaussian [radial basis function](https://www.sciencedirect.com/topics/computer-science/radial-basis-function), and hyperbolic tangent function. The SVM is performed excellently by the proper selection of kernel, the parameters of the kernel, notably the [Gaussian kernel](https://www.sciencedirect.com/topics/engineering/gaussian-kernel) is preferred as it has a single parameter. The SVM method outperforms the other methods, and hence it is the most commonly used [machine learning technique](https://www.sciencedirect.com/topics/engineering/machine-learning-technique), especially in industrial applications. The SVM method is considered the best method for diagnosing coronary diseases. The SVM method has also suffered from potential setbacks such as high memory consumption when it processes large volumes of data. It is not easy to interpret the parameters of the solved SVM method. The SVM method requires all the input data to be correctly labeled before the process. The advantages of the SVM method are the better accuracy in classification and the best performance in the analysis.

**For distinguishing early PD and healthy normal subjects:** In this study, four more different machine learning classifiers are chosen for classification task. A brief description of each of them is provided in this section. WEKA is used for classification using Multilayer Perceptron, Bayesian Network, Random Forest, and Boosted Logistic Regression. The main motive is to find an algorithm that can improve the already reported accuracy as well as to see how various models are performing. Firstly, the dataset is normalized using the Normalize filter in WEKA. Then the dataset is divided in such a way that 70% is used for training and the rest 30% is used for testing. While partitioning the dataset the same class proportion in both the test and train data is maintained. For example, if the proportion of healthy people in the complete data is 40% then both in training and testing the proportion of healthy people to PD subjects is maintained at 40%. This type of partitioning is known as stratified partitioning. The accuracy, recall, precision and fmeasure for each these algorithms are computed and the ROC of each of the classifiers are plotted. Finally the performance measures of different classifiers used in this paper as well as in are compared.

1. **Multilayer Perceptron:** Multilayer perceptron is a feedforward artificial neural network. The basic principle of multilayer perceptron is that it takes the input and maps it to a nonlinear space, and then it tries to predict the corresponding outputs. MLP architecture is viewed as a multiple layers of nodes, with each layer being fully connected with the next layer. Each node in the MLP is interpreted as a neuron that has an activation function which is non-linear. The back-propagation algorithm which is a supervised learning technique is used for training the model. The number of hidden layers in the MLP have a significant impact on the performance of the classifier
2. **Bayesian Network:** The Bayesian network is one of the probabilistic graphical models used in machine learning. The Bayes Net corresponds to graphical model structures which are known as directed acyclic graph (DAG). This graphical model is understood in the following manner. The nodes in the graph represent the random variables and the edge between node x and node y denotes the probabilistic dependencies among random variables corresponding to the respective nodes. Hence the nodes that are not connected in the Bayesian network are the random variables which are independent to each other. Different computational and statistical methods are used to estimate the conditional dependencies. Bayes Network learning uses various search algorithms and quality measures. In the present model K2 learning algorithm for searching is used.
3. **Random Forest:** Random forest is part of ensemble learning method that is used for classification, regression and other tasks. In Random forest, there are many decision trees. For a given input, each of the decision trees classify it as yes/no (in case of binary classification). Then once each of the trees have classified as yes/no, the value which has the majority among them is taken as output. The advantages are that this algorithm runs effectively on large inputs and it also helps in estimating which of the features are important.
4. **Boosted Logistic Regression:** Logistic regression was developed by statistician David Cox in 1958. A logistic model is used to predict the binary class using one or more features. Logit- the natural algorithm for an odds ratio is the central mathematical concept behind logistic regression. Logistic regression is well suited in case when one wants to establish relationship between a categorical outcome variable and one or more categorical or continuous predictor variables.

Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning. It belongs to the family of machine learning algorithms which convert weak learners to strong ones. AdaBoost is used for boosting different classifiers.

****

**WORK FLOW**

Now let’s understand the work flow which we are going to follow:

The first thing that we need in machine learning is the data which we use to train our machine learning model. So for this we need the Parkinson’s data, the data which contains certain details about the patients who has Parkinson’s and those patients who doesn’t have Parkinson’s. So what happen, our machine learning model can find the patterns in this data and it can understand what are the symptoms can be found in people who have Parkinson’s and what are the symptoms are common in people who doesn’t have Parkinson’s so using this data our machine learning model can detect this particular disease so we need this data.

So once we have this data we need to process this data so we cannot feed the raw data to out machine learning model so we need to do some processing so once we do the processing we need to split this data to training data and test data so this is the general procedure which we follow in every machine learning projects so we split this data into training data and testing data and we use this training data to train our machine learning model so this is where our machine learning model will understand about the data. okhy so for this particular projects we are going to use a support vector machine model or a support vector machine classifier. So once we train this support vector machine model with the training data, we will then evaluate this model to find how this model is working, what is the performance of this model using the test data. So this is the reason why we splitting the data into training data and testing data, so the training data is used for the training the model and the test data is used for evaluating our model so once we evaluate the model we will have a trained support vector machine model so when you give new data this particular model can detect whether a person has Parkinson or not, whether that person is affected by Parkinson or ears on healthy person.

**So this is the DFD (Data Flow Diagram) which we will follow:**

The data flow diagram (DFD) is one of the most important tools used by system analysis. Data flow diagrams are made up of number of symbols, which represents system components. Most data flow modeling methods use four kinds of symbols: Processes, Data stores, Data flows and external entities.

**REQUIREMENT SPECIFICATION**

A Software Requirements Specification (SRS) is a description of particular software product, program or set of programs that performs a set of functions in a target environment (IEEE Std. 830-1993).

* 1. **Purpose:** The purpose of software requirements specification specifies the intentions and intended audience of the SRS
  2. **Scope:** The scope of the SRS identifies the software product to be produced, the capabilities, application, relevant objects etc. We are proposed to implement Passive Aggressive Algorithm which takes the test and trained data set from the cancer data set.
  3. **Definitions, Acronyms and Abbreviations Software Requirements Specification:** It’s a description of a particular software product, program or set of programs that performs a set of function in target environment.
  4. **References:** IEEE Std. 830-1993, IEEE Recommended Practice for Software Requirements Specifications thy Sierra and Bert Bates.
  5. **Overview:** The SRS contains the details of process, DFD’s, functions of the product, user characteristics. The non-functional requirements if any are also specified.
  6. **Overall description:** The main functions associated with the product are described in this section of SRS. The characteristics of a user of this product are indicated. The assumptions in this section result from interaction with the project stakeholders.

**SYSTEM REQUIREMENTS**

* Hardware Requirements:
  + Processor : above 500 MHz
  + Ram : 4 GB
  + Disk : 4 GB
  + Input device : Standard Keyboard and Mouse.
  + Output device : VGA and High Resolution Monitor.
* Software Requirements:
  + Operating System : Windows 7 or higher
  + Programming : Python 3.6 and related   
     libraries
  + Software : Google Colab
  + Another S/w : Anaconda Navigator and   
     Jupyter Notebook.

**SOFTWARE DESCRIPTION**

1. **Python:** Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. C Python, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of its variant implementations. C Python is managed by the non-profit Python Software Foundation.
2. **NumPy:** NumPy is a general-purpose array-processing package. It provides a highperformance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:
   1. A powerful N-dimensional array object • Sophisticated (broadcasting) functions
   2. Tools for integrating C/C++ and Fortran code
   3. Useful linear algebra, Fourier transform, and random number capabilities 24
   4. Besides its obvious scientific uses, NumPy can also be used as an efficient multidimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.
3. **Pandas:** Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data. In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data. Prior to Pandas, Python was majorly used for data mining and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.
4. **Sckit-Learn:**
   1. Simple and efficient tools for data mining and data analysis
   2. Accessible to everybody, and reusable in various contexts
   3. Built on NumPy, SciPy, and matplotlib 15
   4. Open source, commercially usable - BSD license.
5. **Jupyter Notebook:**
   1. The Jupyter Notebook is an incredibly powerful tool for interactively developing and presenting data science projects.
   2. A notebook integrates code and its output into a single document that combines visualizations, narrative text, mathematical equations, and other rich media.
   3. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
   4. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.
   5. The Notebook has support for over 40 programming languages, including Python,R, Julia, and Scala.
   6. Notebooks can be shared with others using email, Drop box, Git Hub and the Jupyter Notebook.
   7. Your code can produce rich, interactive output: HTML, images, videos, LATEX, and custom MIME types.
   8. Leverage big data tools, such as Apache Spark, from Python, R and Scala. Explore that same data with pandas, scikit-learn, ggplot2, Tensor Flow.

**IMPLEMENTATION**

**Steps for Implementation:**

1. Install the required packages for building the ‘Passive Aggressive Classifier’.
2. Load the libraries into the workspace from the packages.
3. Read the input data set.
4. Normalize the given input dataset.
5. Divide this normalized data into two parts: a. Train data b. Test data (Note: 80% of Normalized data is used as Train data, 20% of the Normalized data is used as Test data.)

**CODING**

**Importing the Dependencies**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn import svm

from sklearn.metrics import accuracy\_score

**Data Collection & Analysis**

# loading the data from csv file to a Pandas DataFrame

parkinsons\_data = pd.read\_csv('/content/parkinsons.csv')

# loading the data from csv file to a Pandas DataFrame

parkinsons\_data = pd.read\_csv('/content/parkinsons.csv')

# number of rows and columns in the dataframe

parkinsons\_data.shape

# getting more information about the dataset

parkinsons\_data.info()

# checking for missing values in each column

parkinsons\_data.isnull().sum()

# getting some statistical measures about the data

parkinsons\_data.describe()

# distribution of target Variable

parkinsons\_data['status'].value\_counts()

1 --> Parkinson's Positive

0 --> Healthy

# grouping the data bas3ed on the target variable

parkinsons\_data.groupby('status').mean()

**Data Pre-Processing**

**Separating the features & Target**

X = parkinsons\_data.drop(columns=['name','status'], axis=1)

Y = parkinsons\_data['status']

print(X)

print(Y)

**Splitting the data to training data & Test data**

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)

**Data Standardization**

scaler = StandardScaler()

scaler.fit(X\_train)

X\_train = scaler.transform(X\_train)

X\_test = scaler.transform(X\_test)

print(X\_train)

**Model Training**

**Support Vector Machine Model**

model = svm.SVC(kernel='linear')

# training the SVM model with training data

model.fit(X\_train, Y\_train)

**Model Evaluation**

**Accuracy Score**

# 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)

**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)

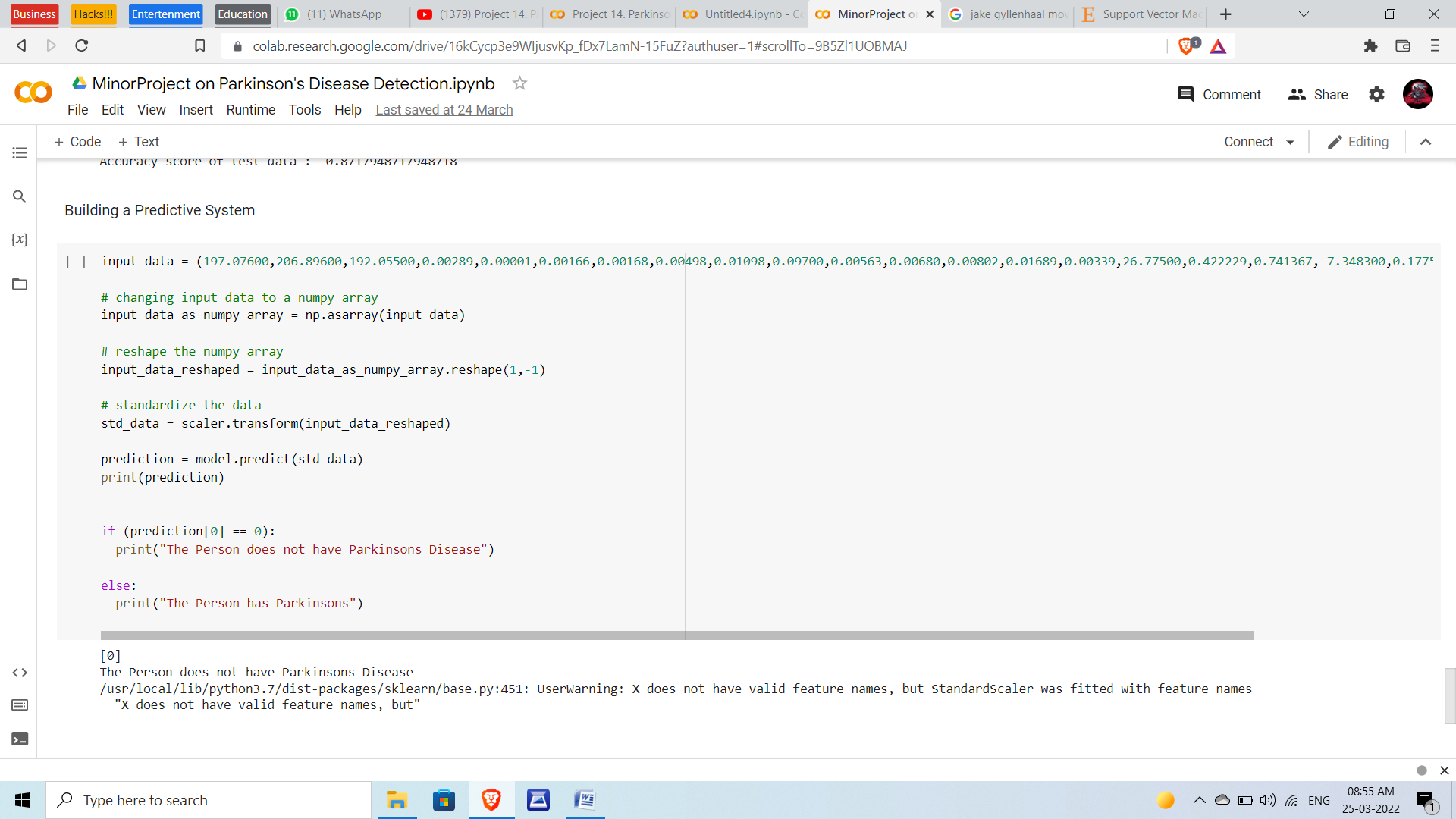
if (prediction[0] == 0):

  print("The Person does not have Parkinsons Disease")

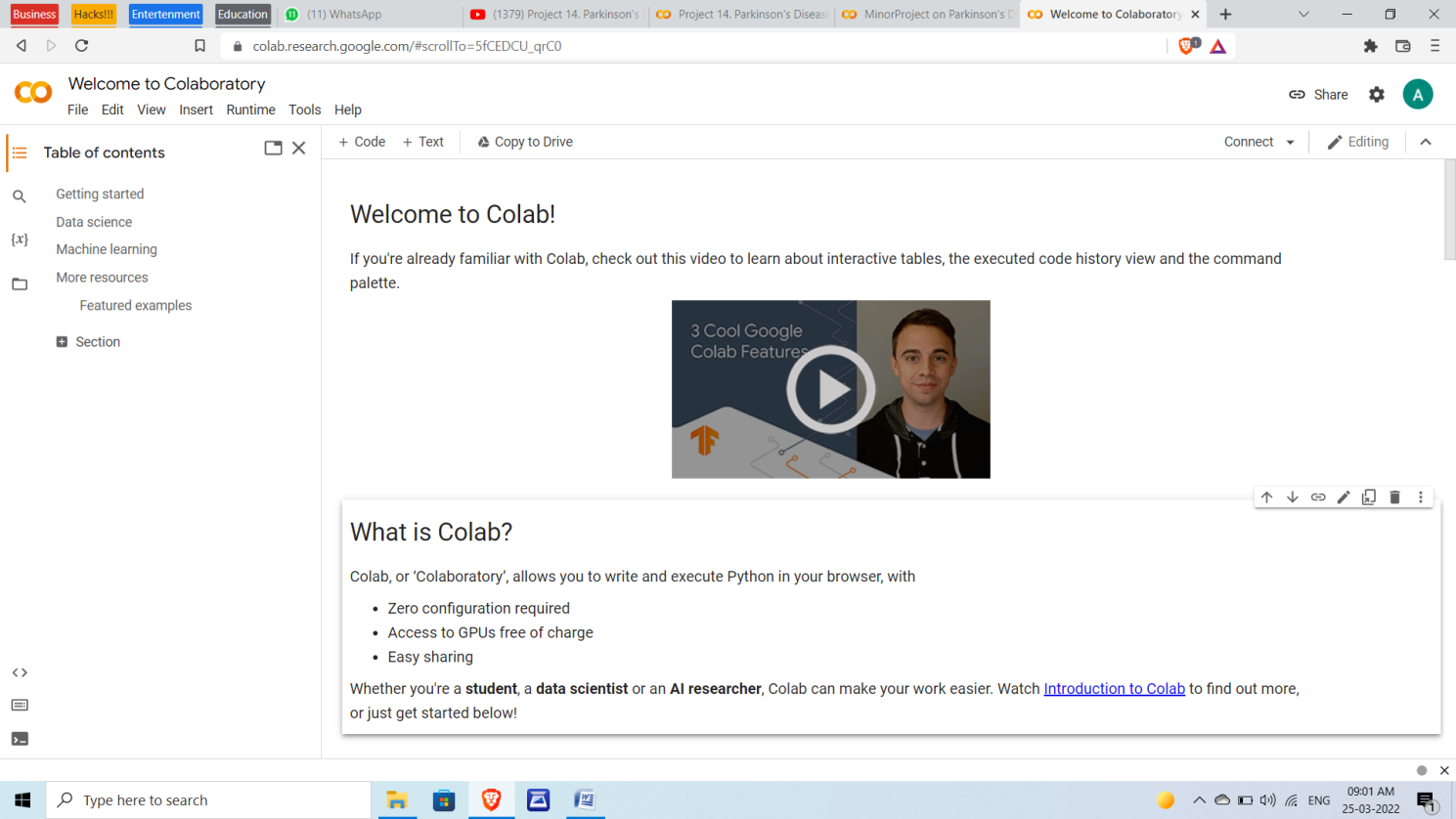
else:

  print("The Person has Parkinsons")

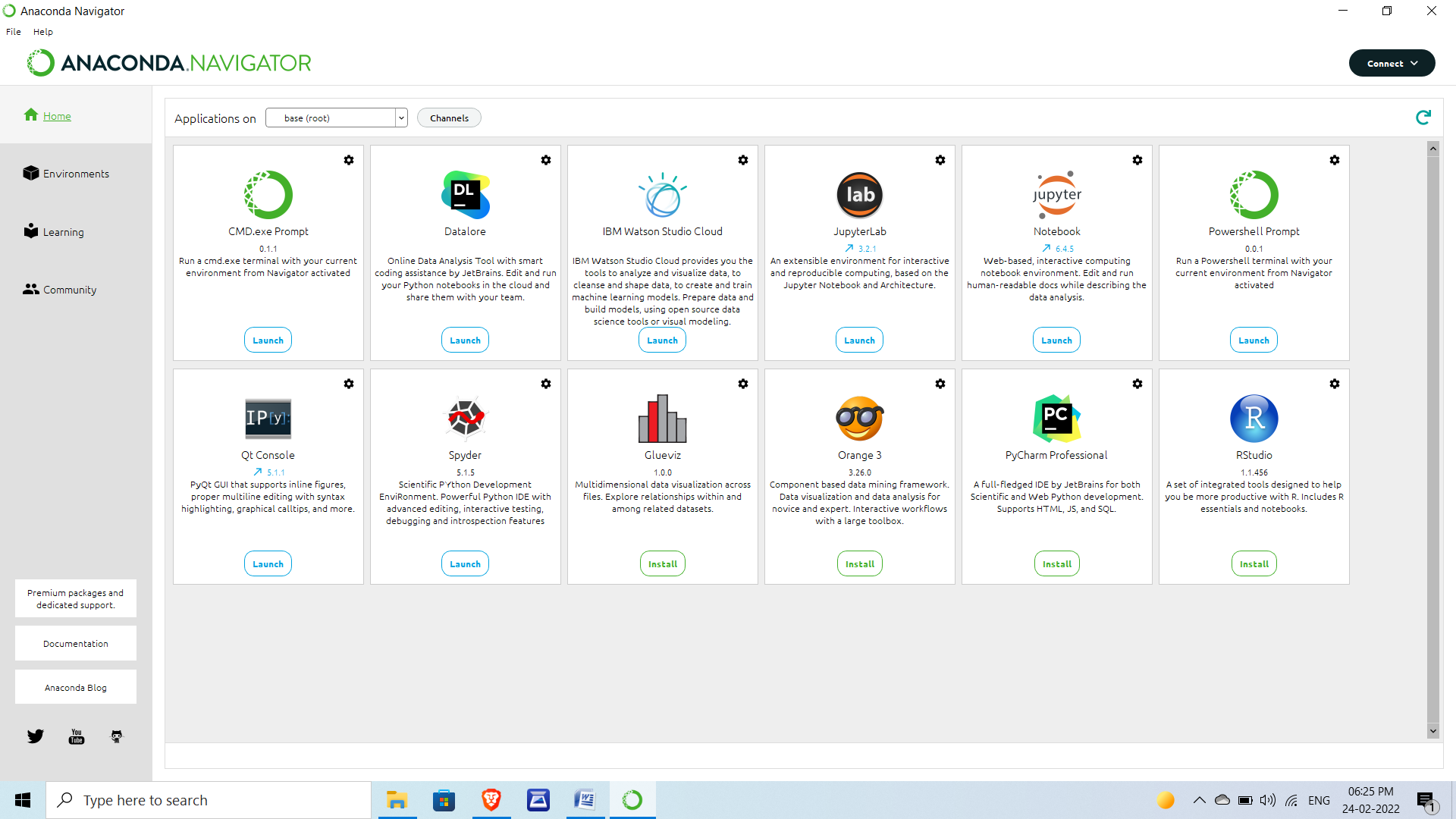
**SCREENSHOTS**



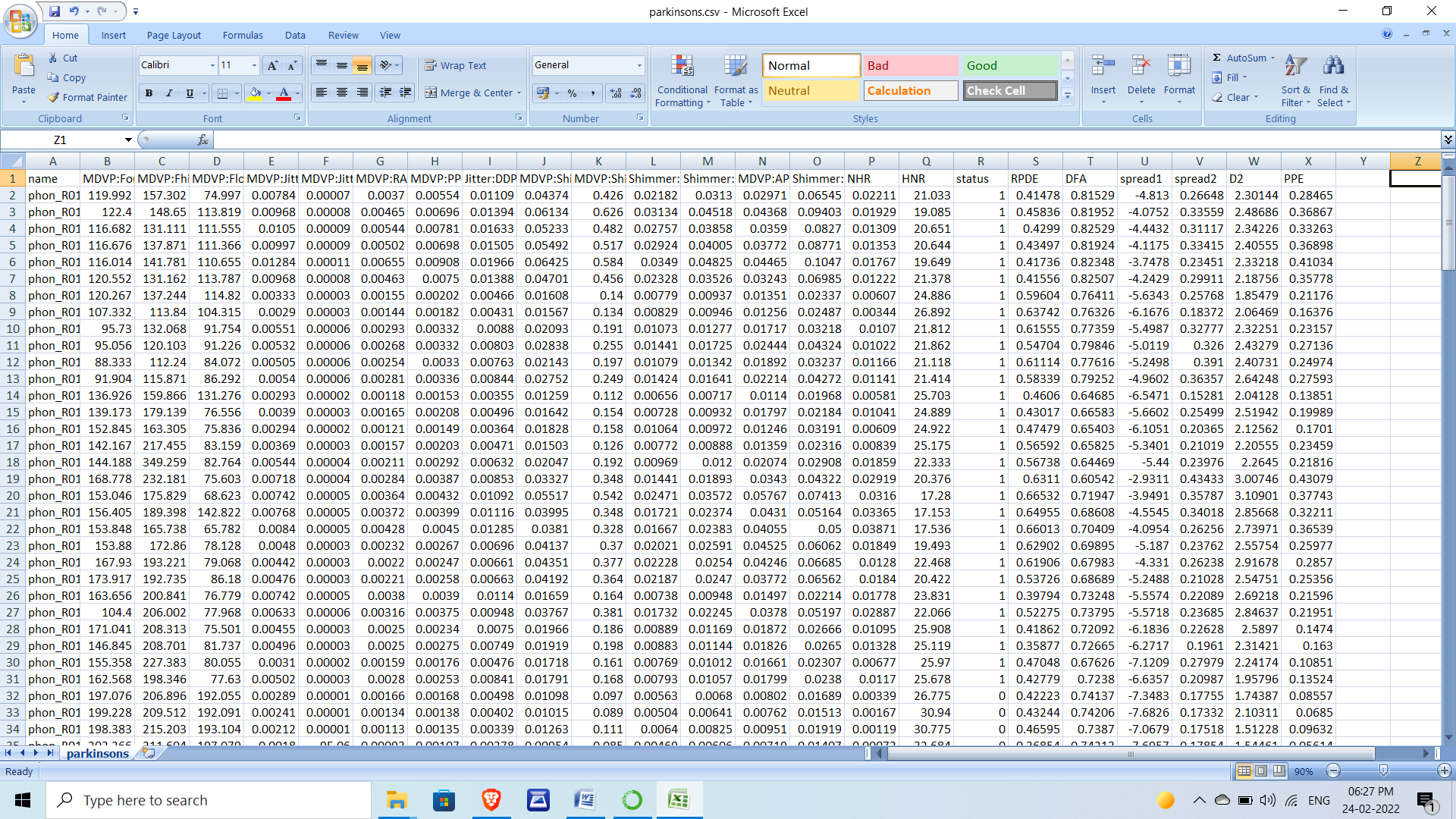
Output or Predictive System



Google Colab lobby



Another s/w Anaconda Navigator-Jupyter

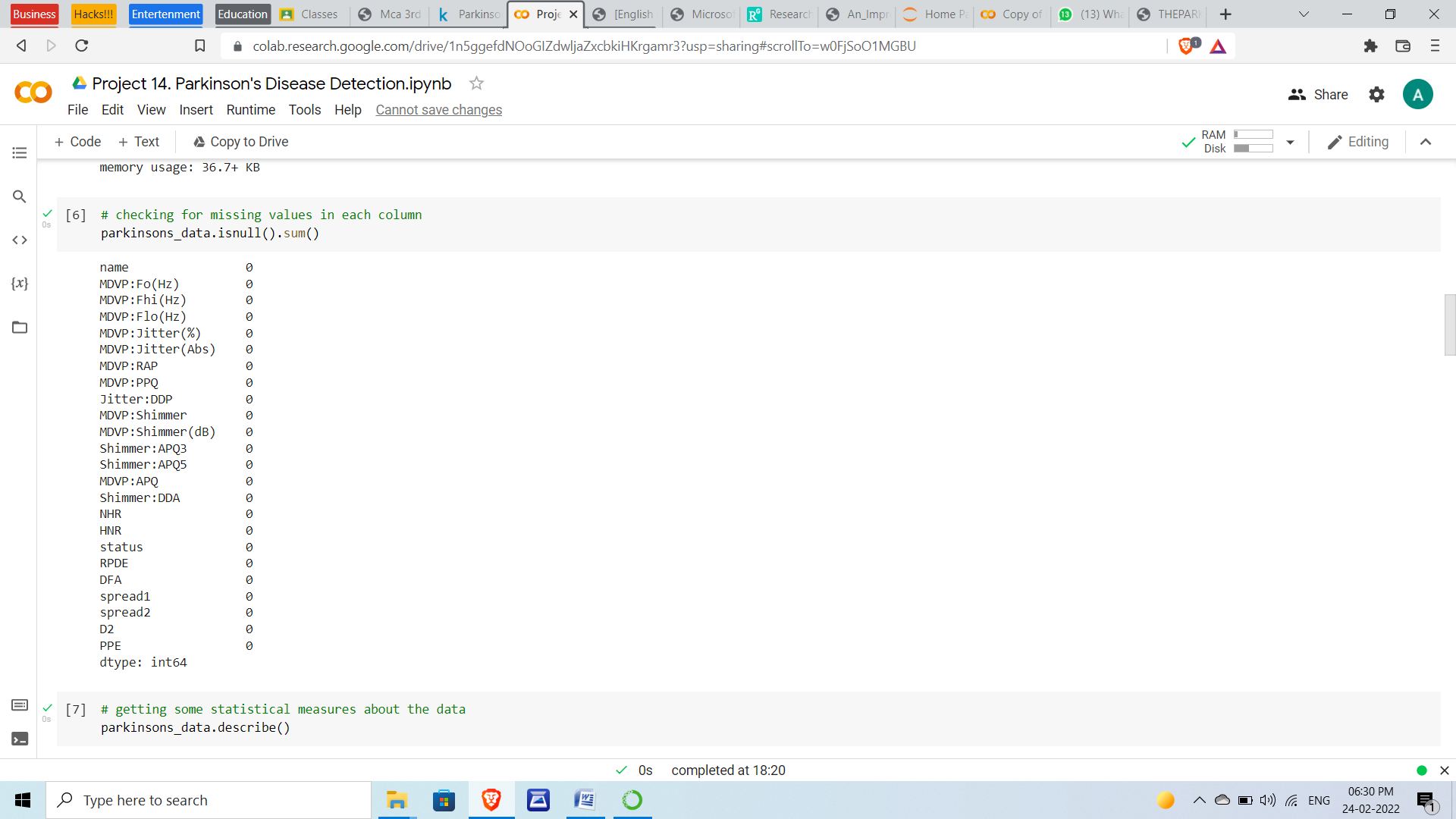


Data Set of Parkinson’s Project

**TESTING**

**UNIT TESTING**: It is the level of software testing where individual units and the components are tested. In the proposed project the data of an individual person is taken and tested. The accuracy is high 100% when tested with a single person data.

**INTEGRATION TESTING:** It may be level of software testing where individual units are combined and it tested as a gaggle. In the proposed project all the data is combined and tested. The accuracy level is 94.87%. This testing will test whole project at a time. It reduces the time complexity in integration testing.



Checking Missing values

**FUNCTIONAL TESTING:** Functional testing may be a sort of software testing that validates the software against the functional requirements/specifications. This testing is detecting Parkinson’s will based on machine learning algorithm. ML algorithm will boost up the speed.

Typically, functional testing involves the following steps:

* Identifying the functions of that the software is expected to perform.
* Create input-data based on the function's specifications.
* It determines the output based up on the function's specifications.
* Execute the test case. •Compare the actual and expected outputs.

**CONCLUSION**

The diagnosis of Parkinson’s Disease is not direct which means that one particular test like blood test or ECG cannot determine whether a person is suffering from PD or not. Doctors go through the medical history of a patient, followed by a thorough neurological examination. They find out at least two cardinal symptoms among the subjects and then predict whether the subject is suffering from PD. The misdiagnosis rate of PD is significant due to a no definitive test. In such case it will be helpful for us to aid the doctor by providing a machine learning model. The prediction models are developed using machine learning techniques of boosted logistic regression, classification trees, Bayes Net and multilayer perceptron based on these significant features. It is observed that the performance is better. It is demonstrated that Boosted Logistic Regression produce superior results. These results encourage us to try other ensemble learning techniques. The present work employs different machine learning algorithms which are not used in. This study plays an important role in having a comparative analysis of various machine learning algorithms. In conclusion, this model can provide the nuclear experts an assistance that can aid them in better and accurate decision making and clinical diagnosis. It is also found that the proposed method is fully automated and provides improved performance and hence can be recommended for real life applications.

**REFERENCES**

1. High Accuracy Detection of Early Parkinson’s Disease through Multimodal Features and Machine Learning. R Prashant, Sumantra Dutta Roy, Pravat K. Mandal, Shantanu Ghosh.
2. Amit S., Ashutosh M., A. Bhattacharya, F. Revilla, (2014, March).Understanding Postural Response of Parkinson’s Subjects Using Nonlinear Dynamics and Support Vector Machines. Austin J, Biomed Eng 1(1): id1005.
3. Parkinson’s disease detection using olfactory loss and REM sleep disorder features. R. Prashanth-EMBS Member, S. Dutta Roy-IEEE Member, and P. K. Mandal, S. Ghosh
4. K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, W. Poewe, B. Mollenhauer, P.-E. Klinik, T. Sherer, M. Frasier, C. Meunier, A. Rudolph, C. Casaceli, J. Seibyl, S. Mendick, N. Schuff, Y. Zhang, A. Toga, K. Crawford, A. Ansbach, P. De Blasio, M. Piovella, J. Trojanowski, L. Shaw, A. Singleton, K. Hawkins, J. Eberling, D. Brooks, D. Russell, L. Leary, S. Factor, B. Sommerfeld, P. Hogarth, E. Pighetti, K. Williams, D. Standaert, S. Guthrie, R. Hauser, H. Delgado, J. Jankovic, C. Hunter, M. Stern, B. Tran, J. Leverenz, M. Baca, S. Frank, C.-A. Thomas, I. Richard, C. Deeley, L. Rees, F. Sprenger, E. Lang, H. Shill, S. Obradov, H. Fernandez, A. Winters, D. Berg, K. Gauss, D. Galasko, D. Fontaine, Z. Mari, M. Gerstenhaber, D. Brooks, S. Malloy, P. Barone, K. Longo, T. Comery, B. Ravina, I. Grachev, K. Gallagher, M. Collins, K.L. Widnell, S. Ostrowizki, P. Fontoura, T. Ho, J. Luthman, M.v.d. Brug, A.D. Reith, P. Taylor, The Parkinson Progression Marker Initiative (PPMI), Prog Neurobiol, 95 (2011) 629-635.
5. <https://www.michaeljfox.org/understanding-parkinsons/living-withpd/topic.php?smell-loss>
6. <https://www.youtube.com/>
7. <https://www.google.com/>
8. <https://www.kaggle.com/nidaguler/parkinsons-data-set>
9. <https://www.slideshare.net/>
10. <https://www.researchgate.net/>