Industrial Oriented Mini - Project Report

3-2 Summer Internship

On

**Parkinson’s Disease Classification**

BY

* + - 1. P. RUDHIR BABU-17311A12G0
      2. M. BABU RAO- 17311A12G1

Under the supervision of

**Dr.Aruna Malapati**

Assistant Professor



**Department of Information Technology**

**Sreenidhi Institute of Science and Technology**

**(an autonomous institution)**

**Yamnampet ,GhatkesarMandal,Medchal district,Hyderabad-501301**

**(June 2020**)

**SREENIDHI INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(Affiliated to Jawaharlal Nehru Technological University, Hyderabad)**

**Yamnampet (V), Ghatkesar (M), Hyderabad – 501 301**

****

**CERTIFICATE**

This is to certify that the project report entitled “**PARKINSON’S DISEASE CLASSIFICATION USING MACHINE LEARNING** ” is being submitted

by

**17311A12G0(P.RUDHIR BABU)**

**17311A12G1(M.BABU RAO)**

in partial fulfillment of the requirements for the award of **Bachelor of Technology** degree in **Information Technology** to **Sreenidhi Institute of Science& Technology** affiliated to Jawaharlal Nehru Technological University, Hyderabad (Telangana). The results embodied in the report have not been submitted to any other University or Institution for the award of any degree or diploma.

**Project Guide Project Coordinater Head Of the Department**

**Dr. Aruna Malapati Dr.K. Kranthi kumar Dr.V.V.S.S.S.Balaram**

**Assistant Professor Professor Department of IT**

**External Examiner**

Date:

**ACKNOWLEDGMENTS**

We would like to express our immense gratitude and sincere thanks to Dr.Aruna Malapati mam,assistant professor for her guidance,valuable suggestions and encouragement in completing the internship project work with in the stipulated time.

We would like to express the sincere thanks to GoalStreet ,Mr.Sheshu Kandregula sir, director of Goal Street, Mr.Durga Naveen Kandregula sir, co-founder of Goalstreet for permitting us to do our Internship project work at GoalStreet pvt ltd.Finally,we would also like thank the people who have directly or indirectly helped us and other faculty,friends and goalstreet team for their cooperation in complete in the projectwork.

M.BABURAO(17311A12G1)

P.RUDHIR BABU(17311A12G0)

**DECLARATION**

We hereby declare that this project report titled “**PARKINSON’S DISEASE CLASSIFICATION USING MACHINE LEARNING”**, is carried out on and our work submitted **to Sreenidhi Institute Of Science and Technology (SNIST)**

**(AUTONOMOUS),** is a record of an original work done by us. We here by declare that this project has not been submitted earlier to any University or Institution.

INFORMATION TECHNOLOGY (IT)

3 rd year

**17311A12G0(P.RUDHIR BABU)**

**17311A12G1(M.BABU RAO)**

**ABSTRACT**

Parkinson’s Disease is one of the most wide spread diseases in elderly people. This disease largely limits the patient’s movement and speech abilities. The patient develops a tendency to fall frequently hence, ending up hurt with various injuries. Thus, it is very important to monitor and notify either the patients or their caregivers about the severity of the disease. This work showcases a comparative study of the various datasets, algorithms and techniques available for the classification of Parkinson’s Disease. This project aims to apply multiple algorithms for classification on UCI Spiral dataset for Parkinson’s Disease.

The data set of Parkinson’s patient is available from the UCI repository and classification algorithms are applied for classifying the healthy patients from the people having Parkinson’s disease using a set of voice recording values as the attributes of the data set.

In this study, we aim to analyze and diagnose patients with Parkinson Disease (PD) by applying Machine Learning Techniques (ML) on speech datasets. In particular, we focus on applying variations of Logistic Regression, Gaussian Naïve Bayes (GNB) and Decision Tree Classifier(). The study aims to work on a previous study conducted by Istanbul University. The same datasets were used for this study and were obtained from the following UCI Database Repository

**CONTENTS**

Title page… 1

Acknowledgements… 2

Abstract……………………………….……………………………………..…….3

1. Introduction…..………………………….………………...……….….…...7

2. Data Collection..……………………………….……..…………….…...…8

3. Data Representation……………………………..……………………...9-23

3.1. Feature Engineering………………………………………………12-23

3.2. Visualization………………………………………………….......18-19

4. Models……………………………………………………………...…24-27

4.1. Logistic Regression………………………………………………24-25

4.2. Gaussian Naïve Bayes…………………………………….……..26-27

4.3. Decision Tree ………………………………………………………28

4.4.Statistical t-test analysis……………………………………….…29-30

5. Conclusion……………………………………………....…………...….31

References…………………………………………….…………………….…31

# INTRODUCTION

**Parkinson's disease** is a progressive nervous system disorder that affects movement. Symptoms start gradually, sometimes starting with a barely noticeable tremor in just one hand. Tremors are common, but the disorder also commonly causes stiffness or slowing of movement. This disease largely limits the patient’s movement and speech abilities. The patient develops a tendency to fall frequently hence, ending up hurt with various injuries.Thus, it is very important to monitor and notify either the patients or their caregivers about the severity of the disease.

We aim to build other classification models like logistic regression, Naïve Bayes, Decision Trees and others and also fine tune the parameters of the model. These models would be trained on a data set which will be engineered carefully after performing the feature engineering.

# 2.Data Collection

The PD database consists of training and test files. The training data belongs to 20 PWP (6 female, 14 male) and 20 healthy individuals (10 female, 10 male) who appealed at the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University. From all subjects, multiple types of sound recordings (26 voice samples including sustained vowels, numbers, words and short sentences) are taken. A group of 26 linear and time,frequency based features are extracted from each voice sample. UPDRS ((Unified Parkinson’s Disease Rating Scale) score of each patient which is determined by expert physician is also available in this dataset. Therefore, this dataset can also be used for regression.  
  
After collecting the training dataset which consists of multiple types of sound recordings and performing our experiments, in line with the obtained findings we continued collecting an independent test set from PWP via the same physicianâ€™s examination process under the same conditions. During the collection of this dataset, 28 PD patients are asked to say only the sustained vowels 'a' and 'o' three times respectively which makes a total of 168 recordings. The same 26 features are extracted from voice samples of this dataset. This dataset can be used as an independent test set to validate the results obtained on training set.

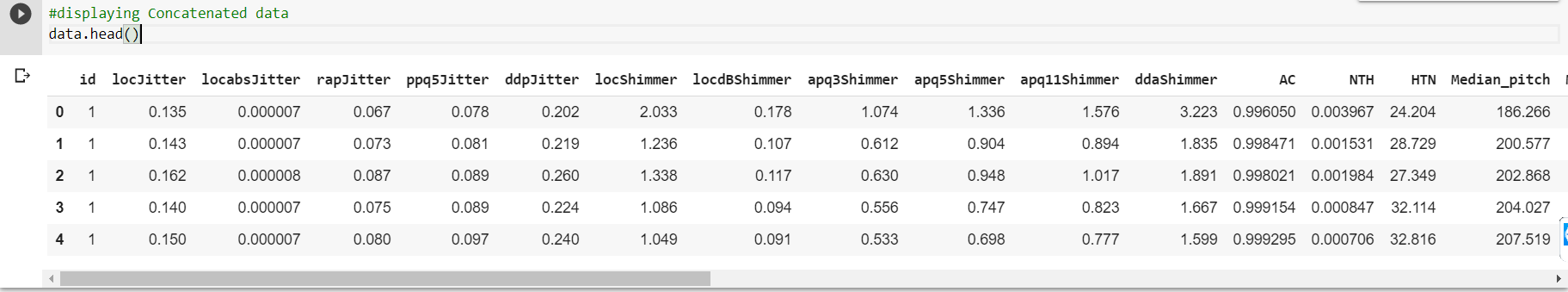
# 3.DATA REPRESENTATION

The dataset that we are using is Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Setwhich is derived from the UCI repository dataset. database contains approximately 1208 records overall.The following are the steps used to solve the problem.

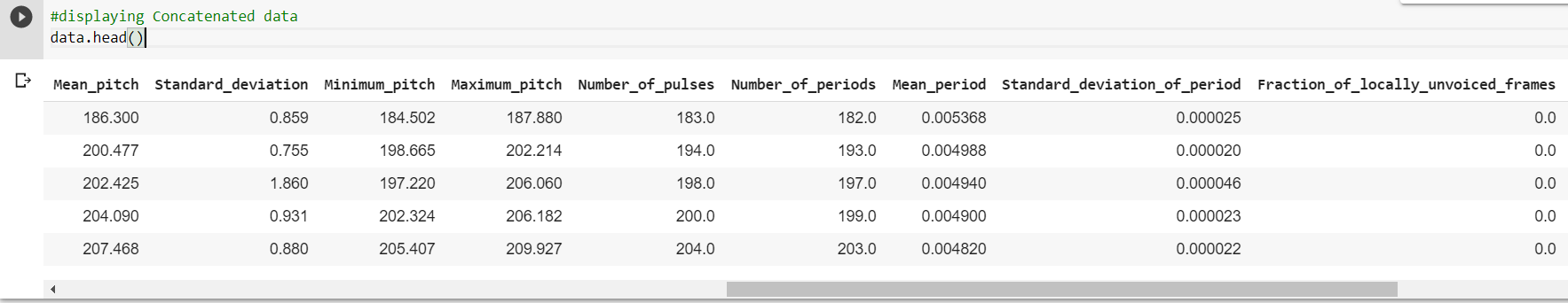
* + Read the Input dataset.
  + Perform all necessary Data Normalization, Standardization processing toprepare the transformed format of given input dataset.
  + Handle Missing values
  + Perform Exploratory Data Analysis/Visualization and bring insights of the predictor variables.
  + Apply Logistic Regression Classifier, Gaussion Naïve Bayes,Decision Tree Classifier algorithms by splitting the data into train and test sets
  + Measure and compare the performance of the models using confusion matrix and metrics like Precision and Recall

Apply statistical test to explain the goodness of the fit

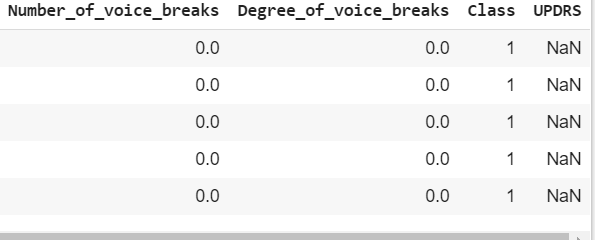
First five records of the dataset are as follows



Contd..



Contd..



There are total of 29 columns in our dataset each consisting of some specific records. Description of each column of the dataset is given below:

column1:Subject\_id  
  
colum2-27:features  
features1-5:Jitter(local),Jitter(local,absolute),Jitter(rap),Jitter(ppq5),Jitter(ddp),  
features6-11:Shimmer(local),Shimmer(local,dB),Shimmer(apq3),Shimmer(apq5),Shimmer(apq11),Shimmer(dda),  
features12-14:AC,NTH,HTN,  
features15-19:Median\_pitch,Mean\_pitch,Standard\_deviation,Minimum\_pitch,Maximum\_pitch,  
features 20-23: Number of pulses,Number of periods,Mean period,Standard deviation of period,

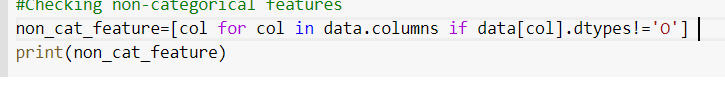
features 24-26: Fraction of locally unvoiced frames,Number of voice breaks,Degree of voice breaks

column28:UPDRS  
column 29: class information

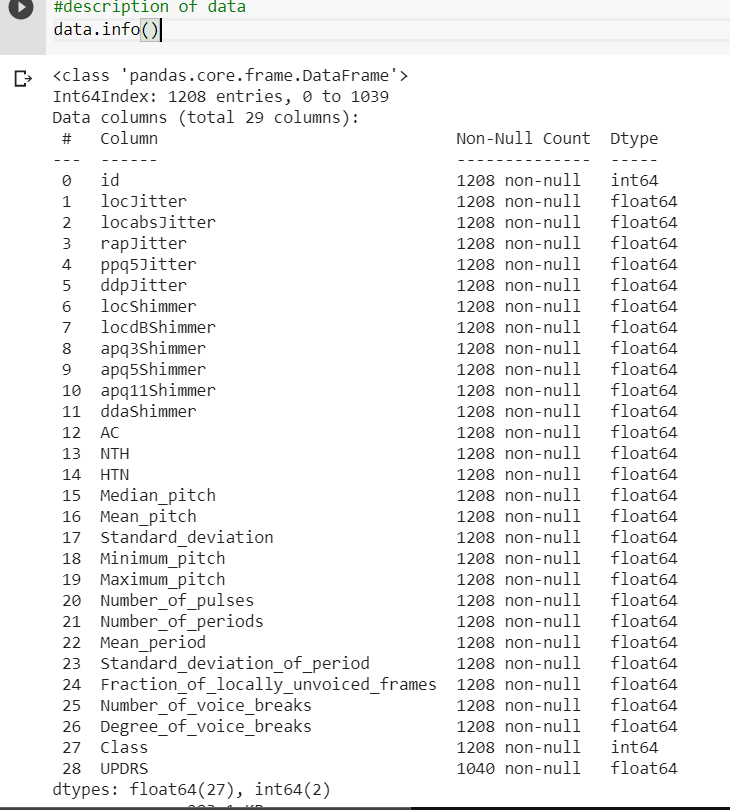
Training Data File:  
Each subject has 26 voice samples including sustained vowels, numbers, words and short  
sentences. The voice samples in the training data file are given in the  
following order:

sample# - corresponding voice samples  
1: sustained vowel (aaaâ€¦â€¦)  
2: sustained vowel (oooâ€¦...)  
3: sustained vowel (uuuâ€¦...)  
4-13: numbers from 1 to 10  
14-17: short sentences  
18-26: words  
  
Test Data File:  
28 PD patients are asked to say only the sustained vowels 'a' and 'o' three times respectively which makes a total of 168 recordings (each subject has 6 voice samples) The voice samples in the test data file are given in the following order:  
  
sample# - corresponding voice samples  
1-3: sustained vowel (aaaâ€¦â€¦)  
4-6: sustained vowel (oooâ€¦â€¦)

We have Checked categorical feature and found that there are no categorical features. So we have concluded that there are only numerical features except target Class.



['id', 'locJitter', 'locabsJitter', 'rapJitter', 'ppq5Jitter', 'ddpJitter', 'locShimmer', 'locdBShimmer', 'apq3Shimmer', 'apq5Shimmer', 'apq11Shimmer', 'ddaShimmer', 'AC', 'NTH', 'HTN', 'Median\_pitch', 'Mean\_pitch', 'Standard\_deviation', 'Minimum\_pitch', 'Maximum\_pitch', 'Number\_of\_pulses', 'Number\_of\_periods', 'Mean\_period', 'Standard\_deviation\_of\_period', 'Fraction\_of\_locally\_unvoiced\_frames', 'Number\_of\_voice\_breaks', 'Degree\_of\_voice\_breaks', 'Class', 'UPDRS']



In our dataset target column is ‘Class’,it has two categorical values encoded into 1 and 0.Which indicates whether the patient has disease as 1 and not as 0.

**3.1 Feature Engineering**

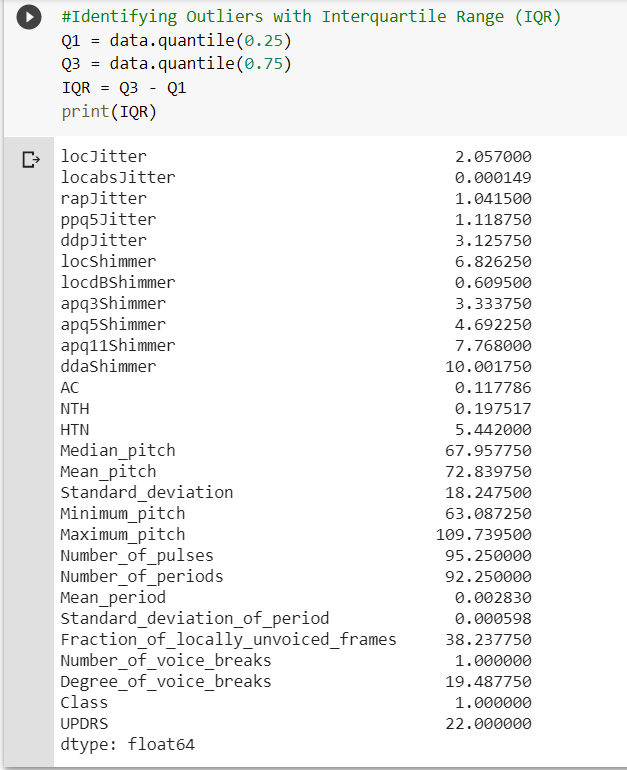
 Algorithms require features with some specific characteristic to work properly. Here, the need for **feature engineering** arises. I think feature engineering efforts mainly have two goals:

* Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
* Improving the performance of machine learning models.

**Handling outliers**

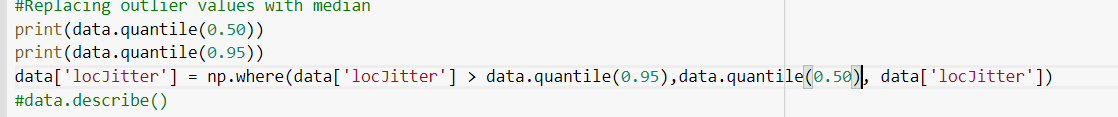
Using Inter Quantile range to remove outliers and to make a lower and upper boundary for each and every feature in the dataset.

In descriptive statistics, the interquartile range, also called the midspread, middle 50%, or H‑spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q₃ − Q₁.



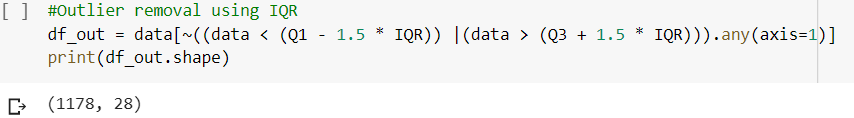
**Replacing Outliers with median values**

In this technique, we replace the extreme values with median values. It is advised to not use mean values as they are affected by outliers. The first line of code below prints the 50th percentile value, or the median, which comes out to be 140. The second line prints the 95th percentile value, which comes out to be around 326. The third line of code below replaces all those values in the 'Loan\_amount' variable, which are greater than the 95th percentile, with the median value. Finally, the fourth line prints summary statistics after all these techniques have been employed for outlier treatment.

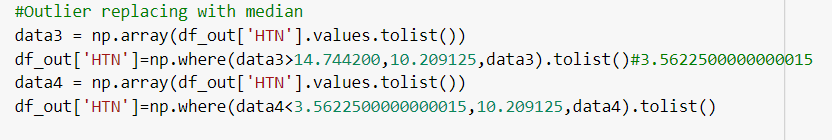
****

****

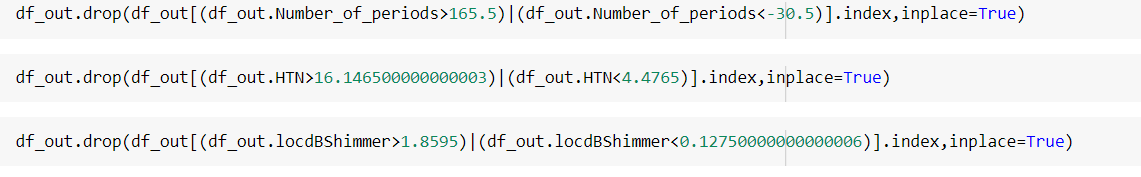
Therefore we have replaced all the outlier values with median values.And again we are performing the Interquantile range method to sort out the data without outliers and storing it into a new dataframe named ‘df\_out’.



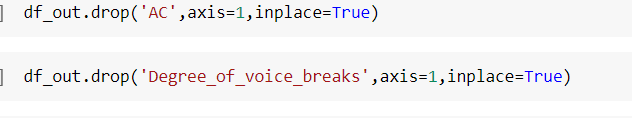
And again after applying Interquantile range for individual attributes rather than collectively performing for all the attributes at a time.



Dropping the rows of attributes which has outliers.

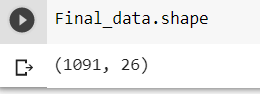


Dropping the attributes which has more than 50 percent outliers.



Storing the resulting data of dataframe ‘df\_out’ into new dataframe ‘Final\_data’.

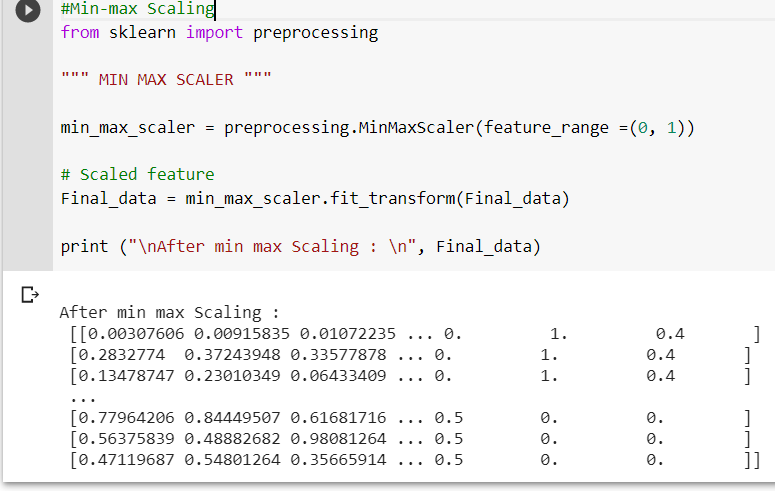
Final\_data=df\_out



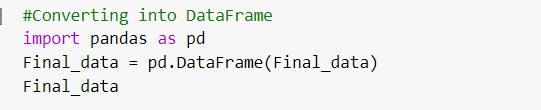
The shape of the data after removing outliers.

**Min-max Feature Scaling**

**Min**-**max normalization** is one of the most common ways to **normalize** data. For every feature, the minimum value of that feature gets transformed into a 0, the **maximum** value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1.

****

After feature scaling the ‘Final\_data’ ndarray needs to be converted into the dataframe.



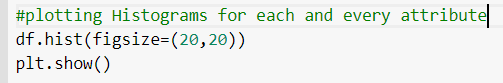
Once again we need to add column names to our ‘Final\_data’ dataframe by using below code:

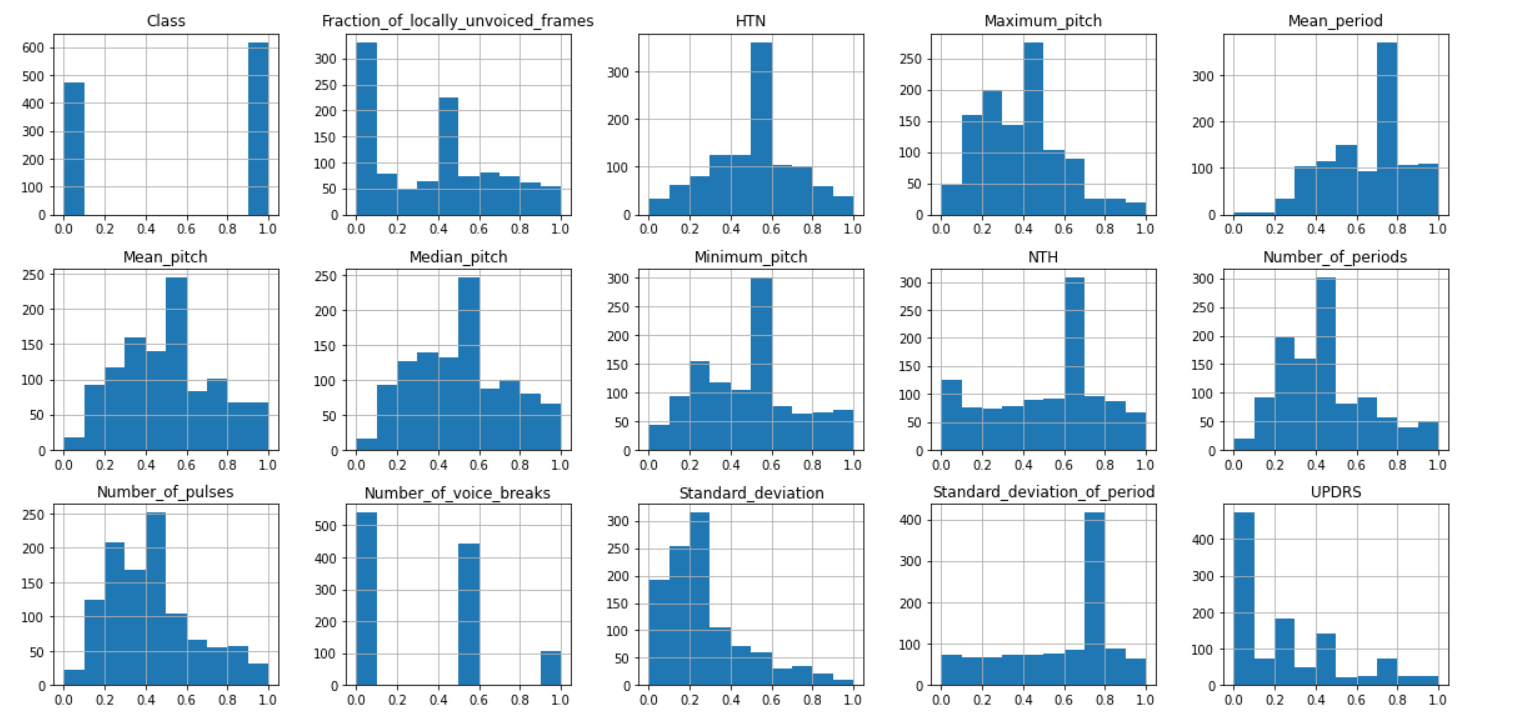
Final\_data.columns=['locJitter','locabsJitter','rapJitter','ppq5Jitter','ddpJitter','locShimmer','locdBShimmer','apq3Shimmer','apq5Shimmer','apq11Shimmer','ddaShimmer','NTH','HTN','Median\_pitch','Mean\_pitch','Standard\_deviation','Minimum\_pitch','Maximum\_pitch','Number\_of\_pulses','Number\_of\_periods','Mean\_period','Standard\_deviation\_of\_period','Fraction\_of\_locally\_unvoiced\_frames','Number\_of\_voice\_breaks','Class','UPDRS']

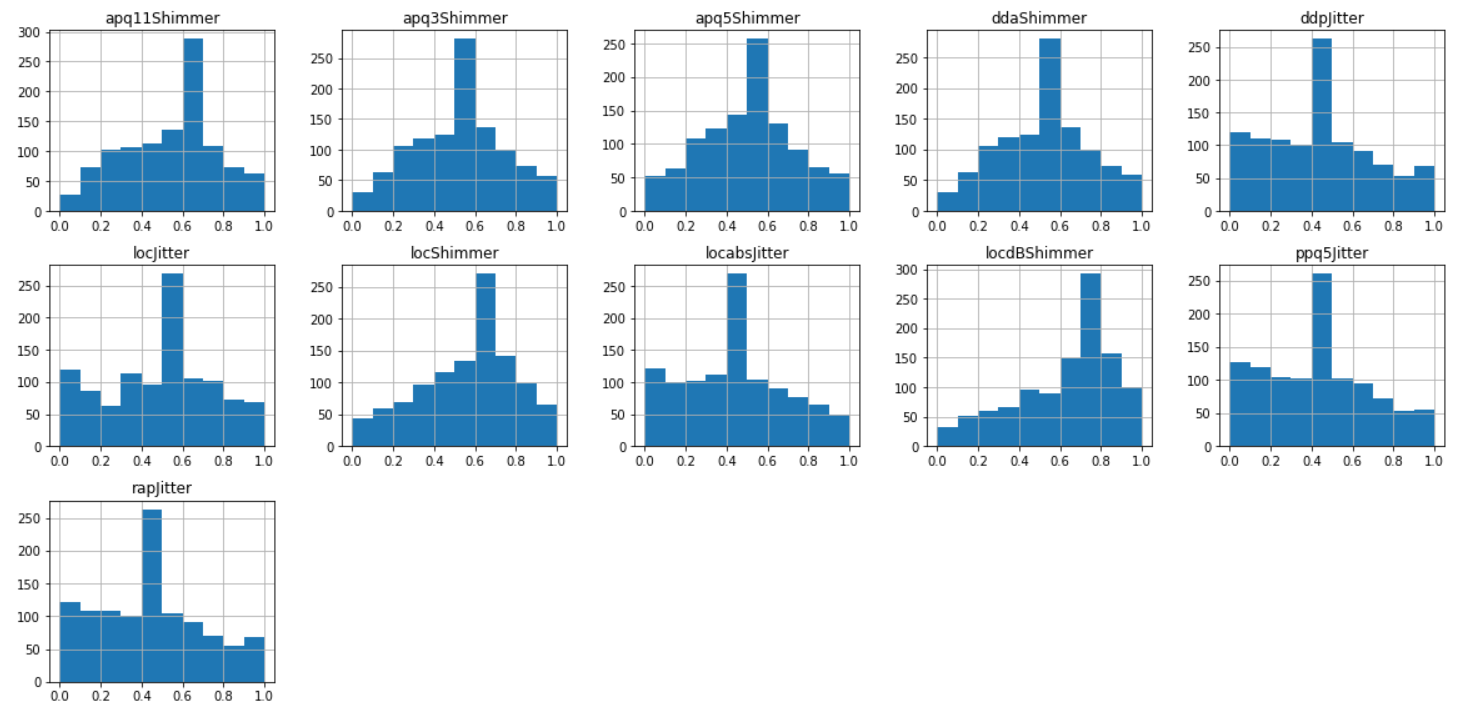
Converting ‘Class’ attribute of float data type to int.And storing Final\_data in ‘df’



**3.2Visualization**

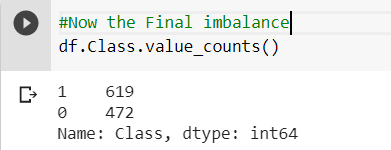




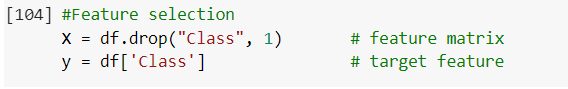


Here the purpose of Visualization is to observe whether the data is distributed normally or not and to observe and remove the outliers. If data in each attribute is not normally distributed we have to apply some normalization techniques.But we have data distributed normally after removing outliers and normalization.And we can gain some meaningful insights from the observations above histograms.

Since our final dataset is ‘df’,the imbalance in the target class is:



Now we are seperating all the feature and target label values into two different Dataframes.



**Feature Selection**

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

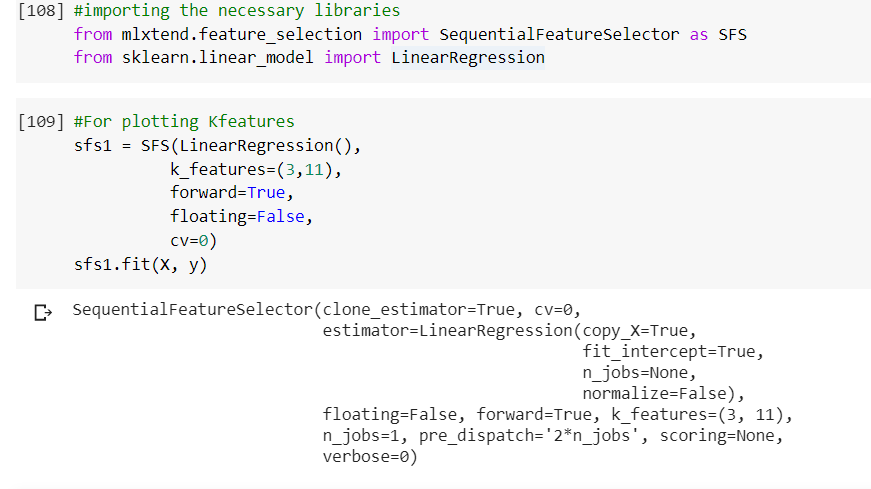
**i) Backward Elimination:** In backward elimination, we start with the full model (including all the independent variables) and then remove the insignificant feature with highest **p-value(> significance level).** This process repeats again and again until we have the final set of**significant**features.

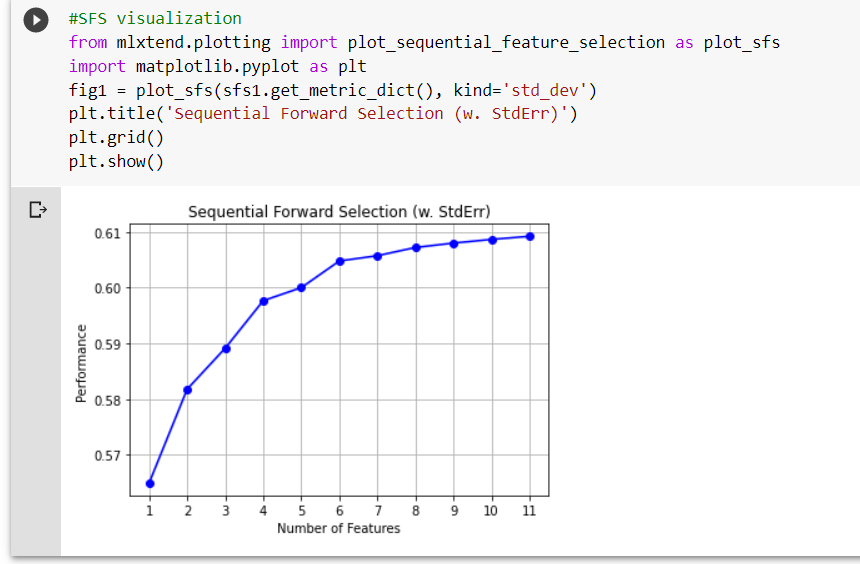
****

The above Backward elimination method will result in the important features which will contribute much to the ‘Class’ Label.

As it is resulting only the 8 features we also have to verify the curse of dimensionality point, by plotting features one at a time on X-axis vs ‘Class’ target label on the Y-axis.

For performing this operation we have to import ‘*SequentialFeatureSelector*’ from *mlxtend.featureselection* and we are applying the prediction plot based on *Linear regression* classifier with *kfeatures*.



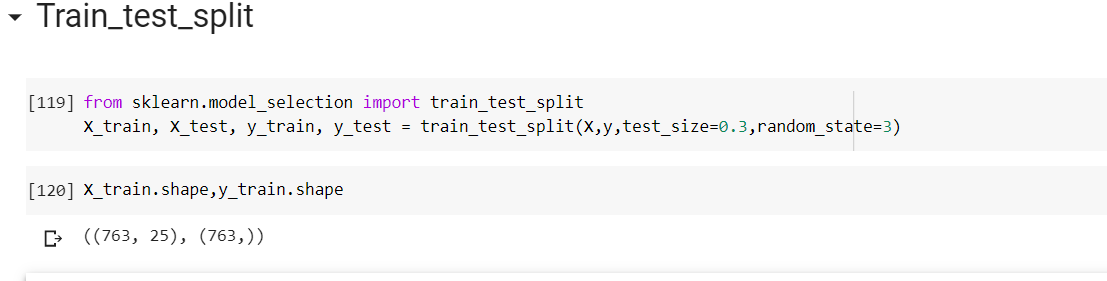


Therefore there are 11 best features out of 25.

Since often we’ve been obtaining the accuracy and precision for best 11 features.

We are considering all the 25 features, for train\_test\_split for implementing Classification models.As of now we have finished feature engineering part and we are going to apply Classification models.

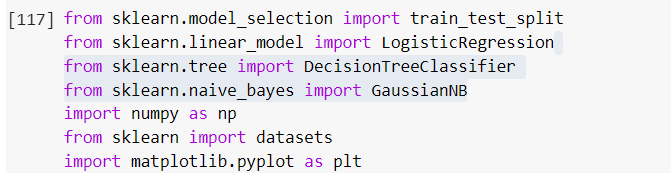
Separated features and target label for training purpose.

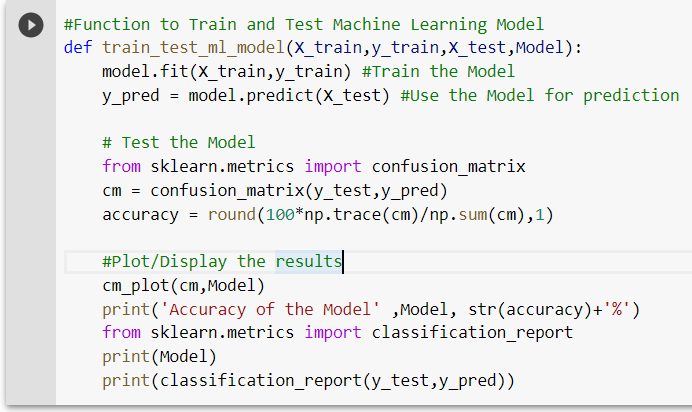


**Function declaration**

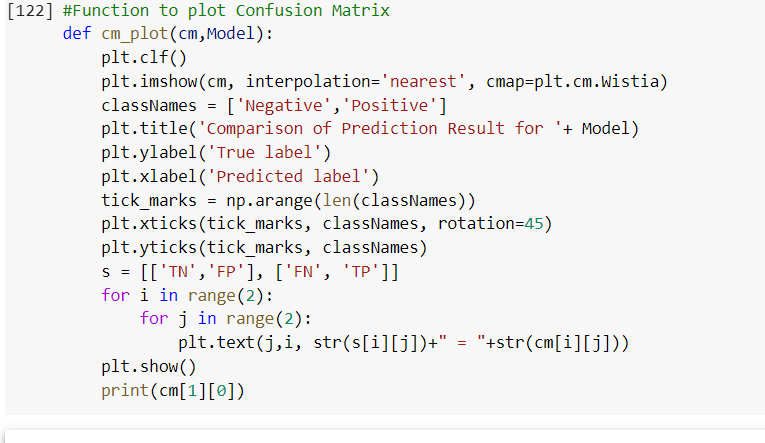
Function to Train and Test Machine Learning Model

Importing the necessary packages





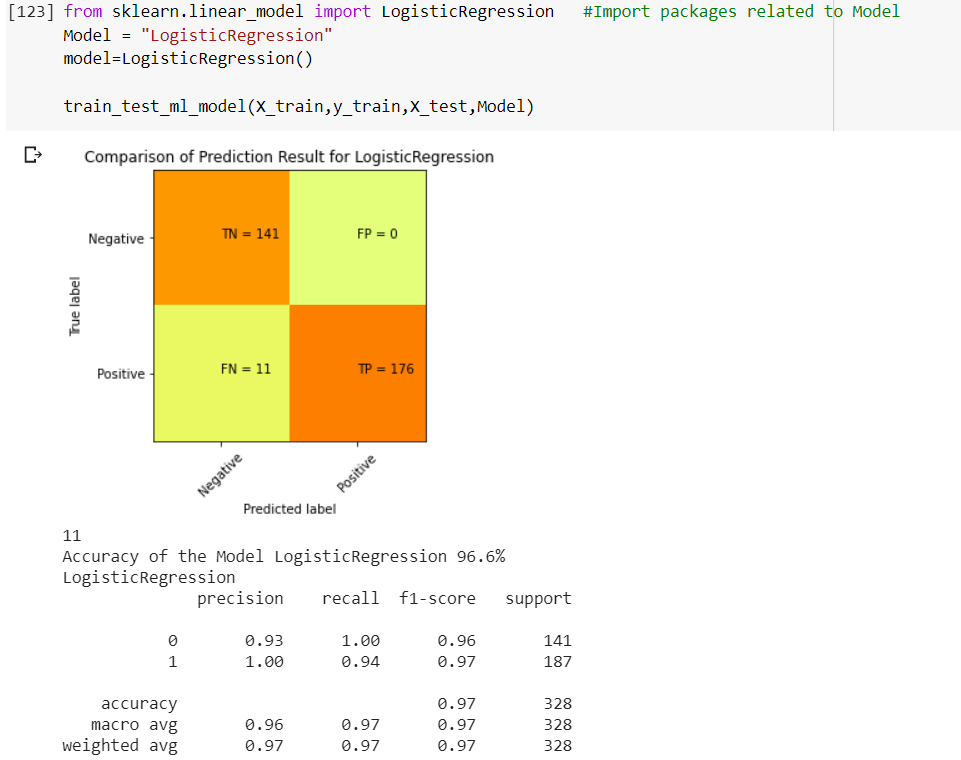
The above defined function ‘train\_test\_ml\_model(X\_train,y\_train,X\_test,y\_test)’ will fit the corresponding classification Model given in the function calling, thus fit(X\_train,y\_train) function will train the machine with percentage of training sample data once it is invoked with ‘model’ .we are plotting confusion matrix and calculating accuracy



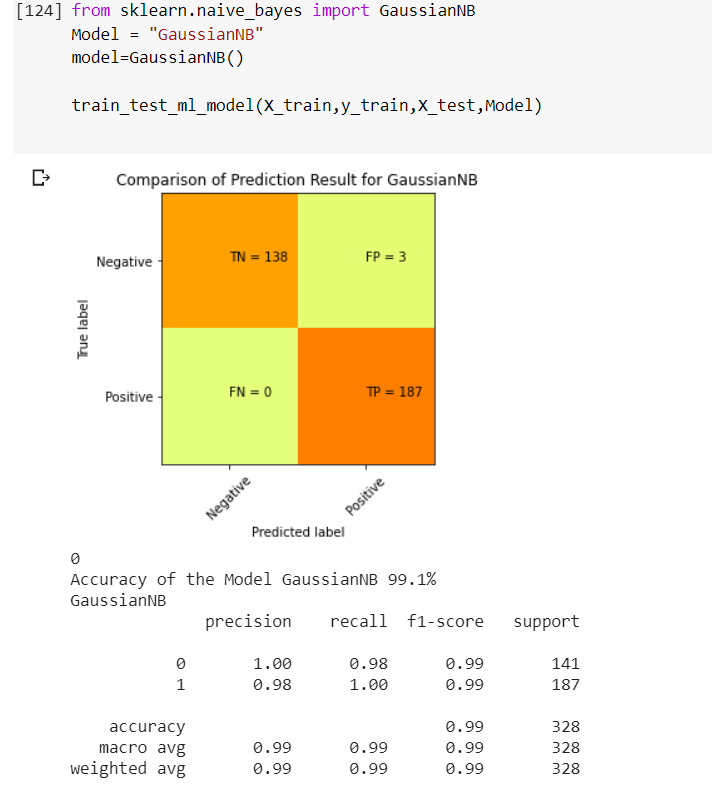
**4.Applying Classification models**

4.1 LogisticRegressionClassifier

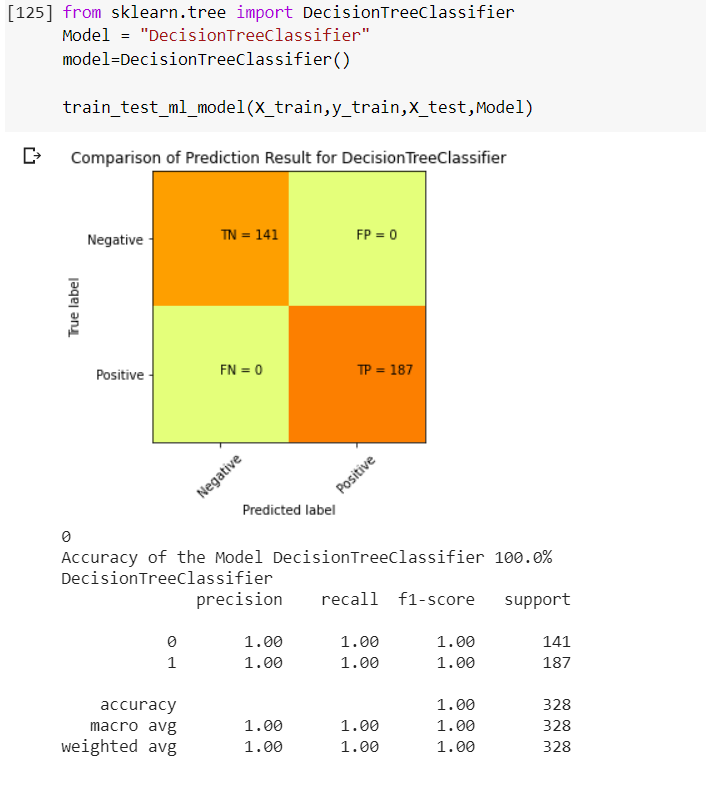
**Logistic regression** is a **classification** algorithm used to assign observations to a discrete set of classes. ... **Logistic regression** transforms its output using the **logistic** sigmoid function to return a probability value.



4.2.**Gaussian Naïve bayes Classifier**



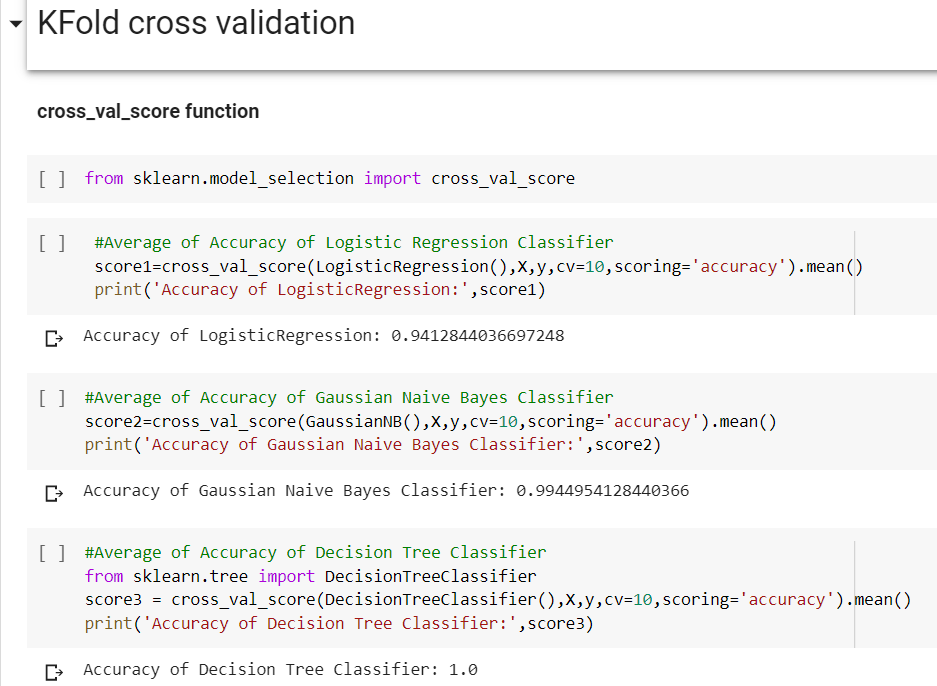
**4.3.Decision Tree Classifier**

****

**Kfold-Cross-validation:**

**k**-**Fold Cross**-**Validation**. **Cross**-**validation is** a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called **k** that refers to the number of groups that a given data sample **is** to be split into.

## Parameter tuning using k fold cross validation

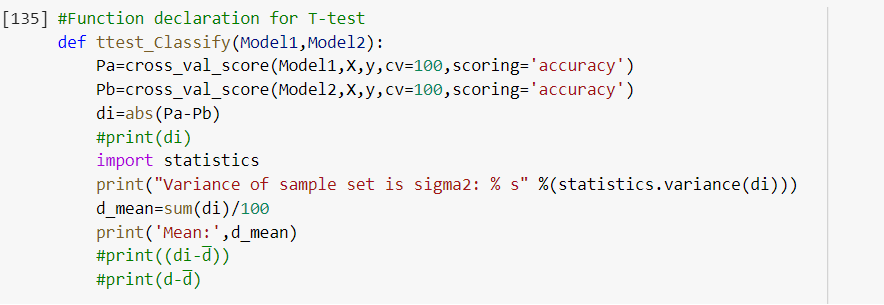


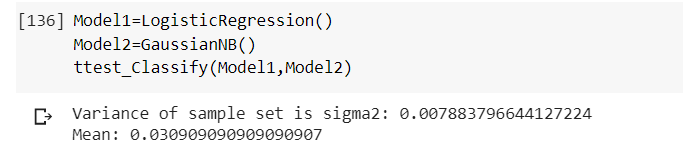
Over here the accuracy of Gaussian Naïve bayes Classifier is best since accuracy of decision tree Classifier indicating that itself it is overfitting the model. The worst model according to accuracy is LogisticRegression compared to GaussianNB in terms of accuracy.

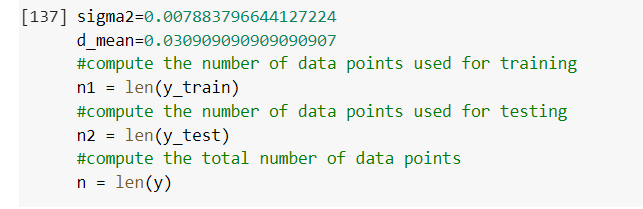
**4.4.Paired students’ t-test analysis**

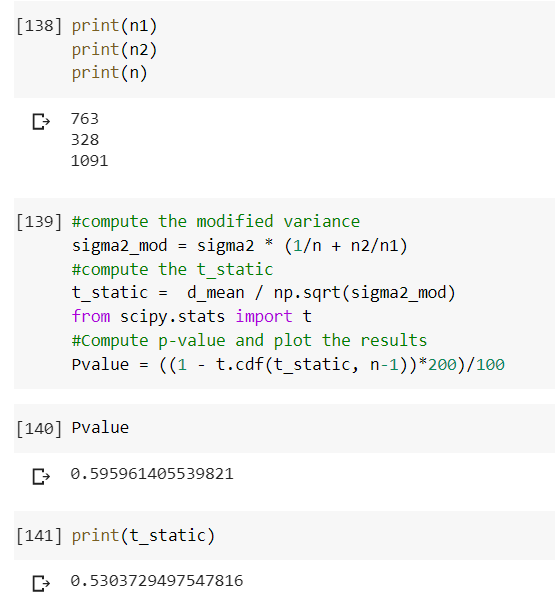
Comparing the performance of machine learning (ML) methods for a given task and selecting a final method is a common operation in applied ML.

The purpose of this post is to, initially, demonstrate why we need to use statistical methods for choosing the final model. Then, it explains why one the frequently used statistical hypothesis tests (i.e., paired Student’s t-test) is inadequate for comparing the performance of ML models. Finally, this post demonstrates how the corrected version of the paired Student’s t-test can be implemented for examining the performance of ML models.





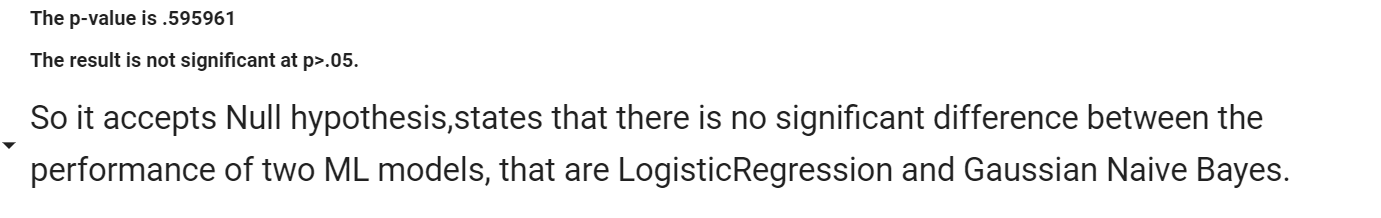


****

Since the null hypothesis is that there is nothing going on or there is no difference between the performance of two ML models, a P-value smaller than the considered significance level rejects the null hypothesis in favor of the alternative hypothesis, which assumes the ML models perform differently. In addition, a P-value greater than the significance levels shows that we fail to reject the null hypothesis.

**5.Conclusion**

In our classification model



**References:**

Code link: <https://colab.research.google.com/drive/1j0nnJLZUOtBgx5RcvaYT67n4afK68nzf#scrollTo=XV9omkefVtB5>

<https://www.google.com/search?rlz=1C1RLNS_enIN878IN878&sxsrf=ALeKk02bRZ4jBRjP8e-MzepkvkGCwI5kew%3A1593461753885&ei=-Uv6Xv7BNY_c9QPphJr4CA&q=logistic+regression+classification&oq=logistic+regression+&gs_lcp=CgZwc3ktYWIQARgAMgQIIxAnMgQIIxAnMgQIIxAnMgQIABBDMgQIABBDMgQIABBDMgQIABBDMgIIADIECAAQQzIECAAQQzoECAAQR1D6dFj6dGDSjgFoAHABeACAAaUCiAGlApIBAzItMZgBAKABAaoBB2d3cy13aXo&sclient=psy-ab>

<https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114>

<https://github.com/codebasics/py/blob/master/ML/12_KFold_Cross_Validation/12_k_fold.ipynb>

[https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings#](https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings)

<https://colab.research.google.com/drive/1inSJ3kvLpiGOHAj4f_cmIfpDu7ckd6ry#scrollTo=MRVw_9vGE_RF>

<https://www.pluralsight.com/guides/cleaning-up-data-from-outliers>