***Heart Disease Prediction Using Machine Learning***

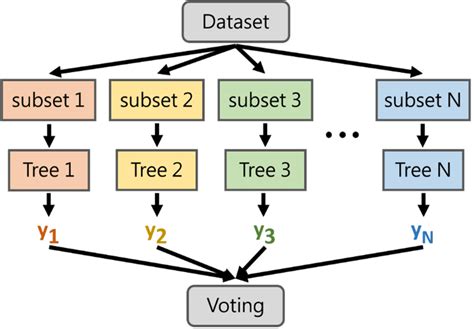
The main purpose of designing this system is to predict the risk of heart disease in a patient and trying to. we have used Various Machine Learning algorithms to evaluate and train the model, such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier and Support Vector Machine (SVM). The main objective of this project is not just predicting the risk of a heart attack but also analysing and putting some results which might help us to determine the risk factors based on the patient’s vital signs and to extract some insights which helps us to understand their working principle.

***Base Paper***

1. <https://www.researchgate.net/publication/351763446_Heart_Disease_Prediction_Using_Machine_Learning>
2. <https://www.researchgate.net/publication/346432379_Heart_Disease_Detection_Using_Machine_Learning>

***Algorithm Description***

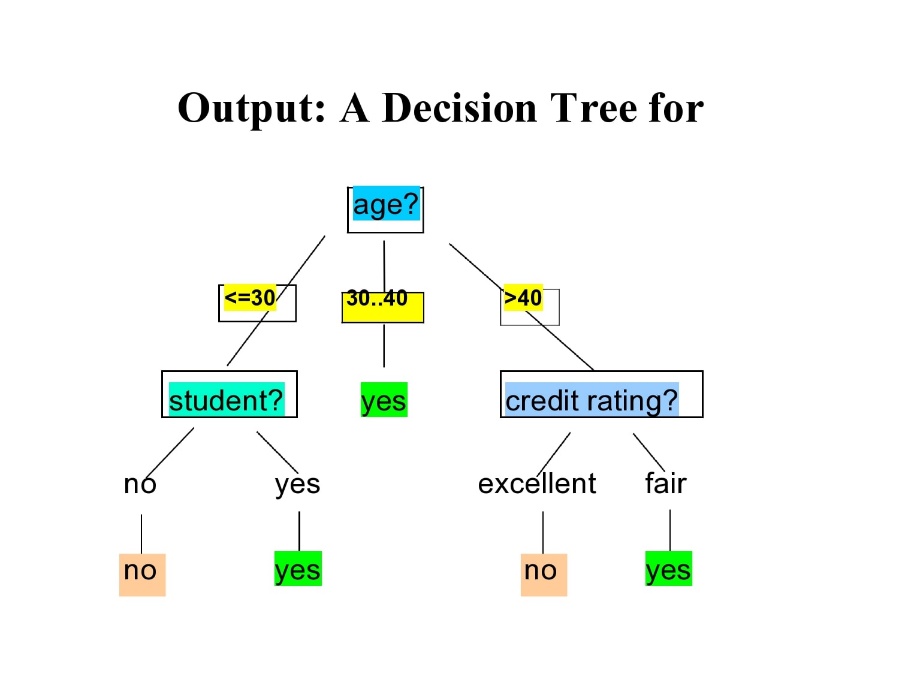
***Random Forest Classifier:*** Random Forest Classifier is an ensemble algorithm which works with multiple algorithms parallelly. This is a supervised algorithm and it can be used with both classification and regression problems. The output of the new data is estimated either by using majority voting or average voting technique. Since the algorithm works with bagging technique, multiple decision trees are used to provide the output for the specific input. This is a key difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset of those features. Random forest works best with large datasets and high dimensional.



***References:***

1. <https://www.geeksforgeeks.org/random-forest-regression-in-python/>
2. <https://www.geeksforgeeks.org/random-forest-regression-in-python/>

***Decision Tree Classifier:*** A decision tree is a tool for making decisions and the process for making decisions is in a tree like structure, decision tree is a supervised machine learning algorithm mainly used for predicting the outcome after computing all the attributes. The process flow of Decision tree goes from Root node to leave node i.e., the decision node.

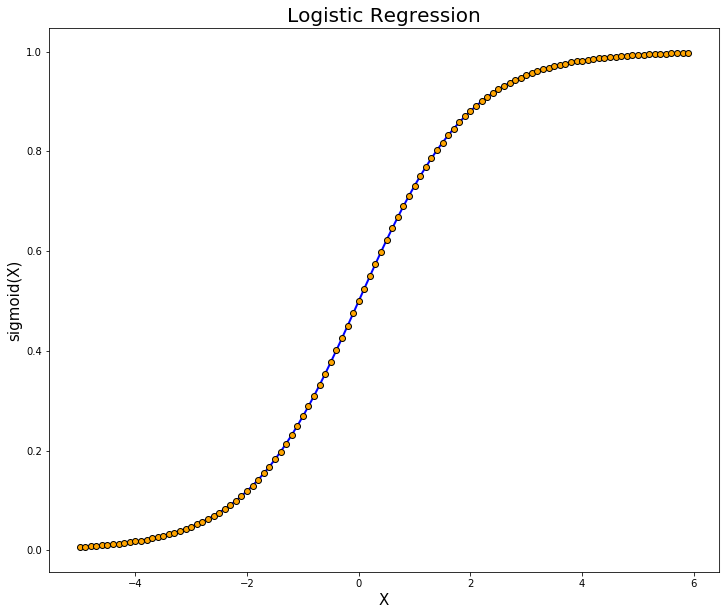


***References:***

1. <https://www.ibm.com/topics/decision-trees>
2. <https://www.mastersindatascience.org/learning/introduction-to-machine-learning-algorithms/decision-tree/>

***Logistic Regression:*** Logistic Regression is a Supervised algorithm which mostly works in the case of binary classification problems. Logistic regression is a sophisticated algorithm where the data to be trained using this algorithm should be properly presented i.e., Normalized/Scaled, Columns should be Converted to numerical and data should be neat and clean. The output is presented in the form of logit score, where this helps us to predict the likelihood of an event occurring of a given problem. The main reason of getting a **S** curve in the below chart is that the sigmoid function does the trick of converting the given number in the range between 0 and 1.

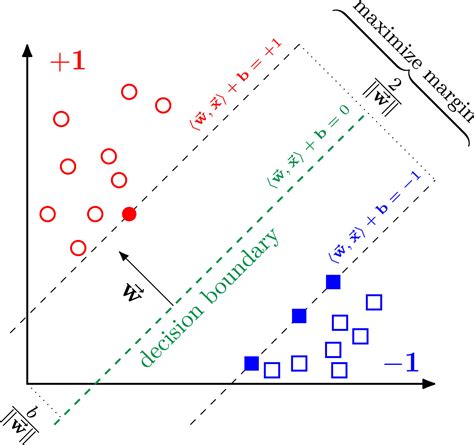
Sigmoid(x) = ***Y = 1 / 1+e -z***



***References:***

1. <https://www.geeksforgeeks.org/understanding-logistic-regression/>
2. <https://www.ibm.com/topics/logistic-regression>

***Support Vector Machine:*** Support vector machines are basically a supervised learning algorithm which classifies the data points by drawing a linear curve and a non-linear curve depending on the data it is dealing with. The boundary that separated the 2 or more classes is called as a hyperplane, though there is a possibility of having some million hyperplanes for our data, but we need to find the hyperplane with maximum margin from all the training points, which makes the algorithm more efficient while predicting on new dataset, it can easily classify on which side the new data belongs to.



***References:***

1. <https://www.geeksforgeeks.org/support-vector-machine-algorithm/>
2. <https://scikit-learn.org/stable/modules/svm.html>

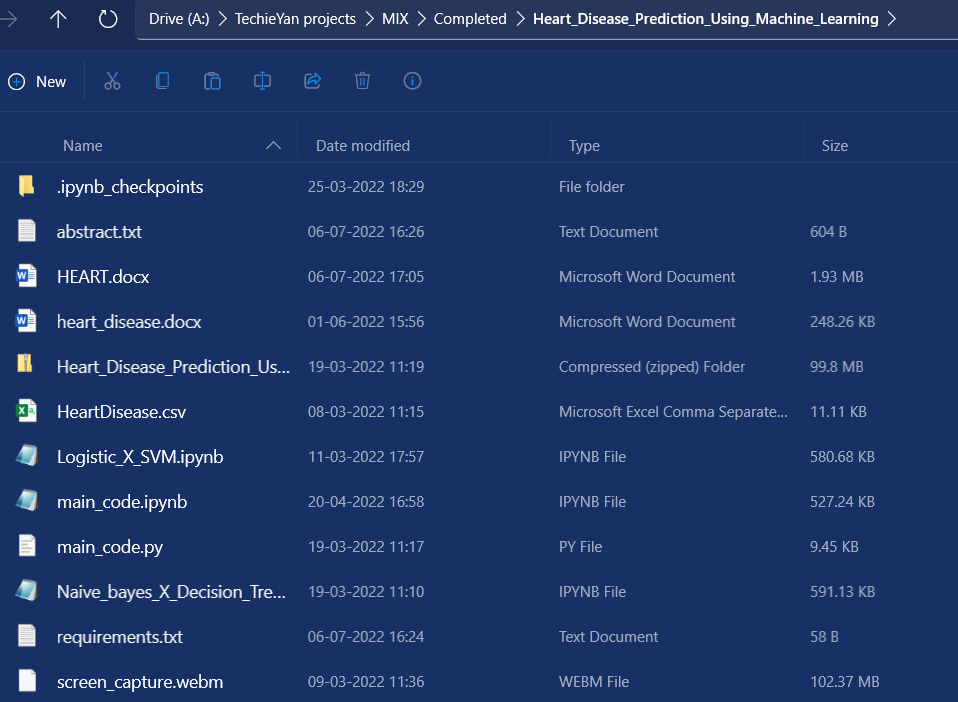
***Steps to Execute the Code!***

**Note:** Make sure you have added path while installing the software’s.

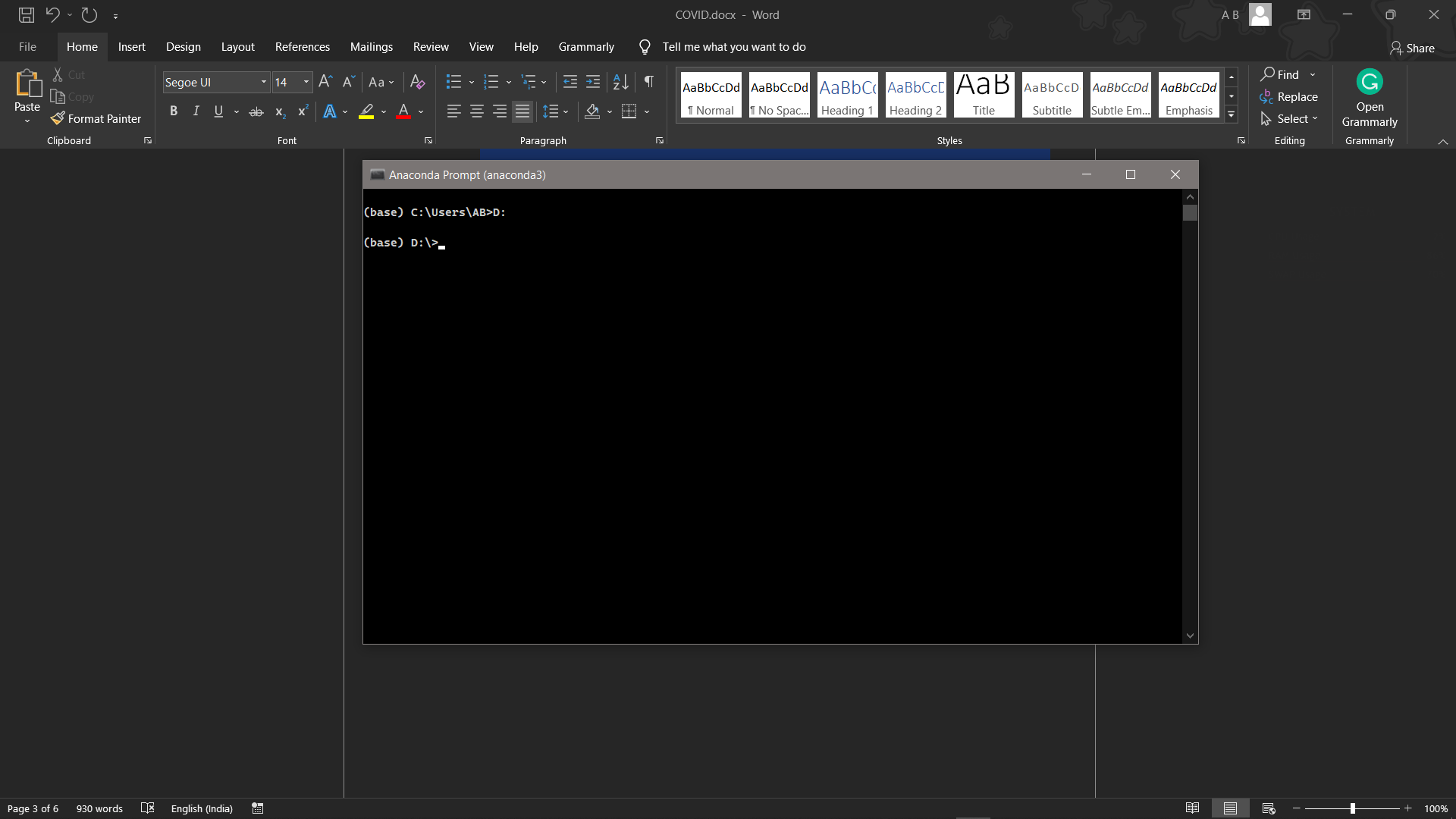
<https://techieyantechnologies.com/2022/07/how-to-install-anaconda/>

<https://techieyantechnologies.com/2022/06/get-started-with-creating-new-environment-in-anaconda-configuring-jupyter-notebook-and-installing-libraries-using-requirements-txt-2/>

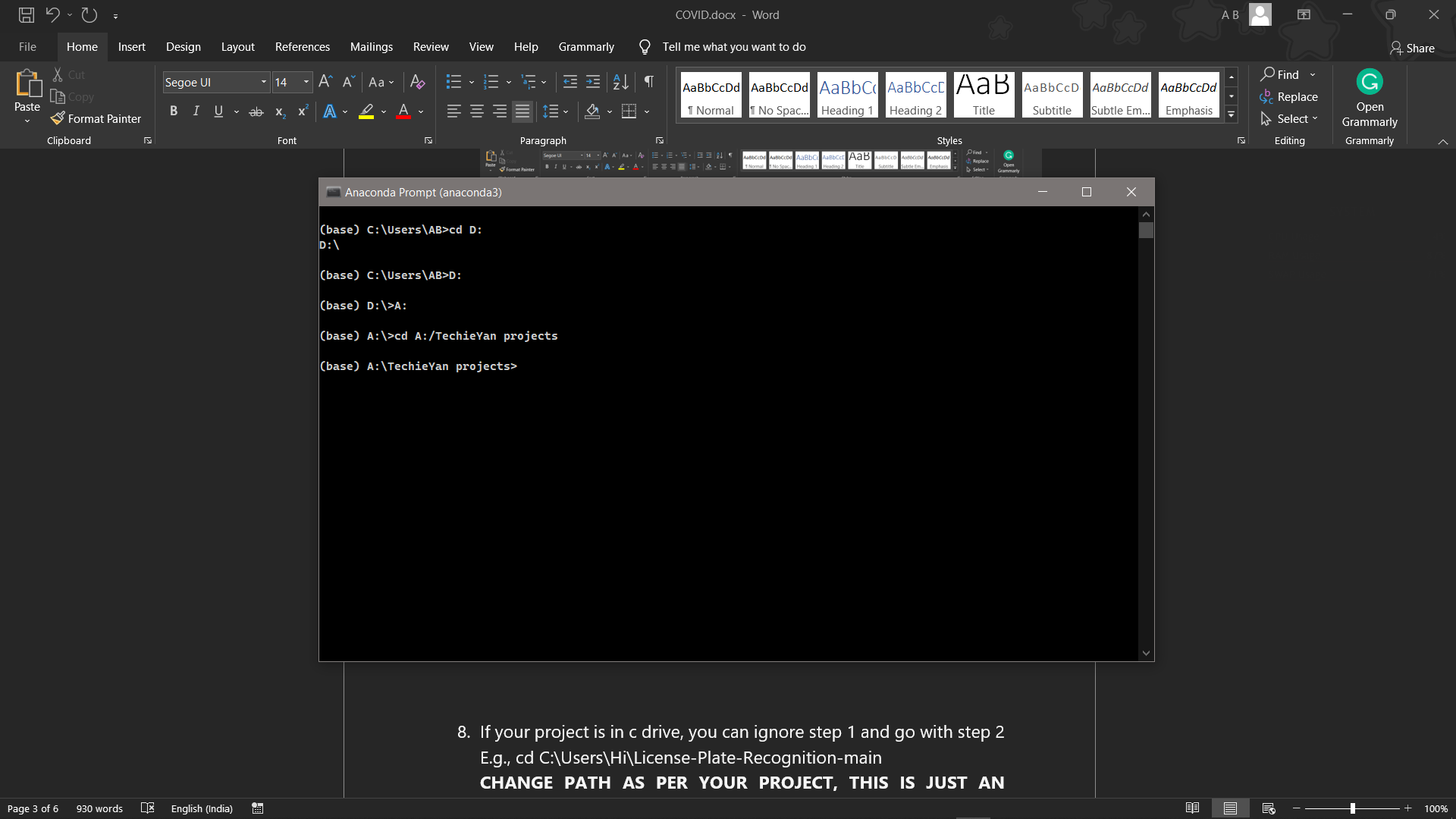
1. Install the prerequisites/software’s required to execute the code from reading the above blog which is provided in the link above.
2. Press windows key and type in anaconda prompt a terminal opens up.
3. Go to the directory where your requirement.txt file is present, not just requirement.txt, if you want to execute any .py or .ipynb files, you need to go to that specific folder or path, where they are saved.



1. <<directory of your file:>>. E.g., If my file is in d drive, then
2. Type d:



1. cd d:\License-Plate-Recognition-main #CHANGE PATH AS PER YOUR PROJECT, THIS IS JUST AN EXAMPLE

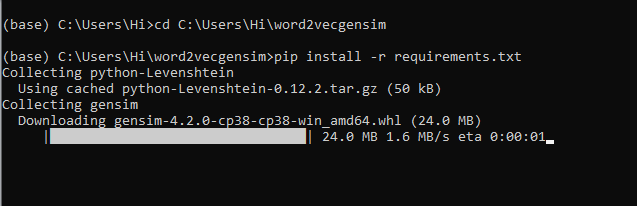


1. If your project is in c drive, you can ignore step 4 and go with step 5.

E.g., cd C:\Users\Hi\License-Plate-Recognition-main

**CHANGE PATH AS PER YOUR PROJECT, THIS IS JUST AN EXAMPLE**

1. Run pip install -r requirements.txt or conda install requirements.txt (Requirements.txt is a text file consisting of all the necessary libraries required for executing this python file. If it gives any error while installing libraries, you might need to install them individually.), example: pip install “module\_name” i.e., pip install pandas



1. If you would like to run .ipynb file, Please [follow the link to setup and open jupyter notebook](https://techieyantechnologies.com/2022/06/get-started-with-creating-new-environment-in-anaconda-configuring-jupyter-notebook-and-installing-libraries-using-requirements-txt-2/), You will be redirected to the local server there you can select which ever .ipynb file you’d like to run and click on it and execute each cell one by one by pressing shift+enter.

**Make sure to change the path of the dataset in the code**

***Dataset Description***

The Dataset is collected form Kaggle Repository which contains 303 Instances with 14 features. Some of the important features which contribute in developing the best model are Blood sugar, cholesterol, chest pain, age, thalach etc. This study's goal is to predict the presence of heart disease in patients where this presence is valued from no presence to likely presence. The researchers accelerating their research works to develop software with the help of machine learning algorithms which can help doctors to decide both prediction and diagnosing of heart disease. The main objective of this research project is to predict the heart disease of a patient using machine learning algorithms.



***Issues Faced***

1. We might face an issue while installing specific libraries, in this case, you might need to install the libraires manually. Example: pip install “module\_name/library” i.e., pip install pandas

2. Make sure you have the latest or specific version of python, since sometimes it might cause version mismatch.

3. Adding path to environment variables in order to run python files and anaconda environment in code editor, specifically in any code editor.

4. Make sure to change the **paths in the code** accordingly where your dataset/model is saved.

**Refer to the Below links to get more details on installing python and anaconda and how to configure it.**

<https://techieyantechnologies.com/2022/07/how-to-install-anaconda/>

<https://techieyantechnologies.com/2022/06/get-started-with-creating-new-environment-in-anaconda-configuring-jupyter-notebook-and-installing-libraries-using-requirements-txt-2/>

***Note:***

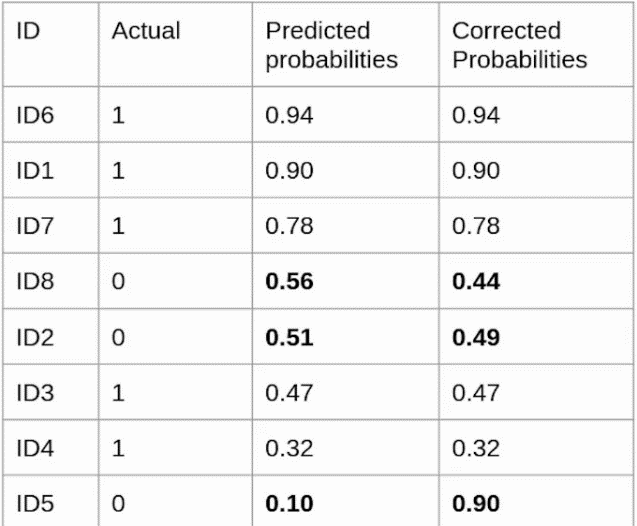
**All the required data has been provided over here. Please feel free to contact me for model weights and if you face any issues.**

<https://www.linkedin.com/in/abhinay-lingala-5a3ab7205/>

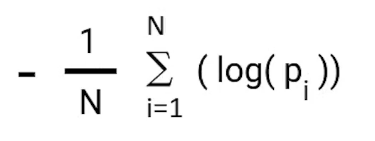
***Yes, you now have more knowledge than yesterday, Keep Going.***

***Evaluation Metrics***

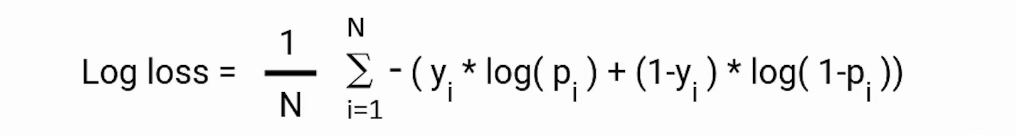
Evaluation metrics are considered as one of the most important steps in any machine learning and deep learning projects, where it will allow us to evaluate how good our model is performing on the new data or on unseen data. There are a lot of evaluation metrics which can be used in order to assess how good our model is performing, in our case, since we are dealing with binary classification and neural network, we are going to sue binary\_cross\_entropy/log loss, which basically compares the actual class with the predicted probabilities and then it calculates a corrected probability by subtracting it with the probability of a datapoint belonging to class1 with the predicted probability, i.e. for the case of ID8 it is actually class 0, but the probability is of class 1 is 0.56, so we subtract (1 – 0.56), we get 0.44 that is our corrected probability. Then Log loss is calculated by applying log transformation on each of the calculated\_probablities. The the average of the negative corrected\_probablities are taken which will gives us the log loss/binary\_cross\_entropy, the lower the value the better our model is performing.



***Log loss calculation for corrected\_probablities***



***Log Loss formula without calculating corrected\_probablities***

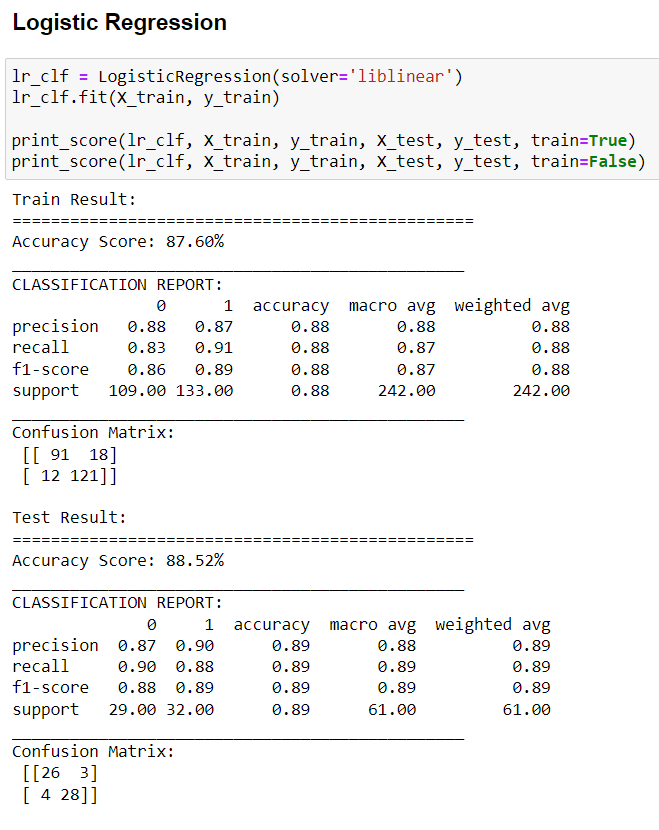


***Reference:***

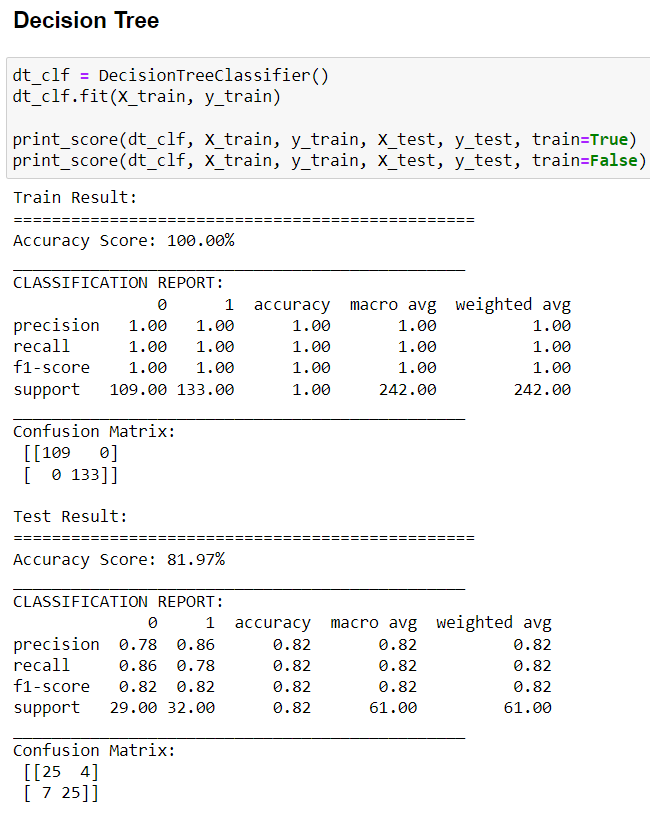
<https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/>

***Results***

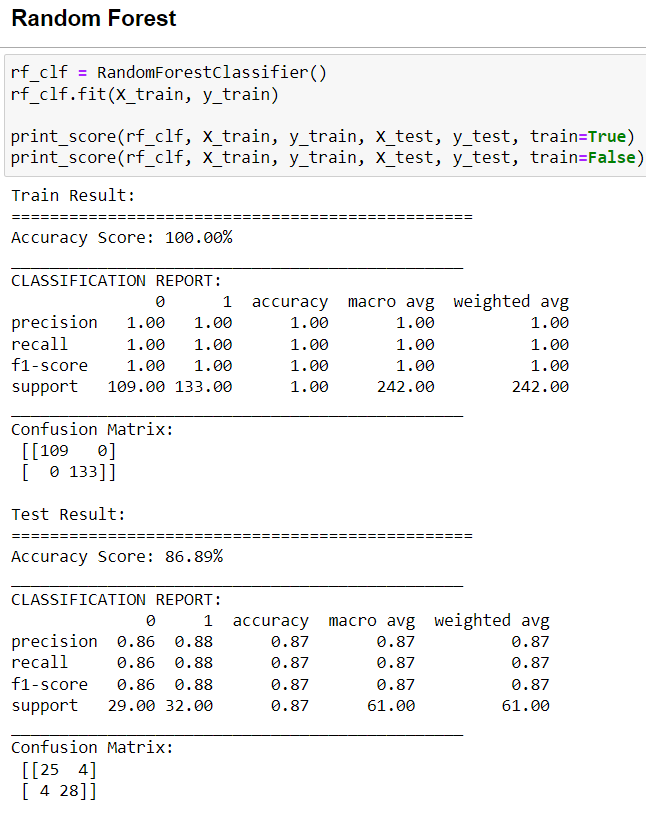
1. ***Logistic Regression***

****

1. ***Decision Tree***



1. ***Random Forest Classifier***



1. ***Support Vector Machine***

