About Project 🡪

The given dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The number of features in Diabetics dataset are Pregnancies, Glucose, Blood Pressure, Skin Thickness, BMI, Insulin, Diabetics Pedigree Function**,** Age.

The label is outcome.

Model will predict the whether the patient has diabetics or not.

This prediction is based on the machine learning method

1-Logistic regression

2-K-Nearest Neighbour

3-Support Vector Machine

4-Decision Tree

5-Random Forest Classifiers

Tools Required 🡪

1-import pandas as pd 🡪 panel data used in creating data frames.

2- import numpy as np 🡪 Numerical python used for computational or mathematical operations. Import arrays or can do mean(), min(), max() operations.

3-data['Outcome'].value\_counts()

🡪count the values of Outcome column 0->500 and 1-->268

4-data.hist(figsize=(10,10)) 🡪Data visualisation tells the zero, max, min, mean values in featues through histogram

5-corr = data.corr() 🡪 Correlation is used to quantify the degree to which the two variable are related.correlation coefficient tells you the change in one variable when another variable changes. User can evaluate the feature variable dependency to the label or target variable.

6-split x and y in features and labels🡪

x = data.iloc[:,:-1].values 🡪 x = data.drop(columns='Outcome',axis=1)

y = data.iloc[:,-1].values 🡪 y = data['Outome']

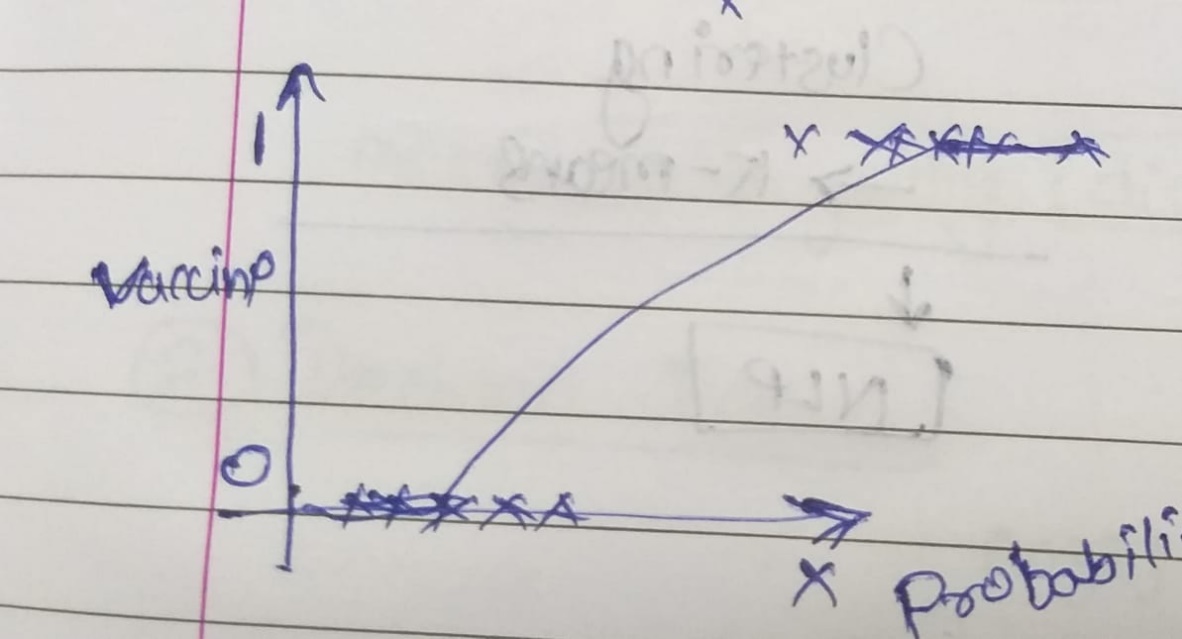
7-from sklearn.preprocessing import StandardScaler 🡪 StandardScalar transform the x value in standard form then the the x\_std.

8-from sklearn.model\_selection import train\_test\_split 🡪 train test split the x\_std and y

9-X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X,Y,test\_size=0.2, stratify=Y ,random\_state=0) 🡪stratify for removing chance of error that all diabetes can go to train and remaining to test / distribute data equally to train and test data.

Algorithms and Accuracy🡪

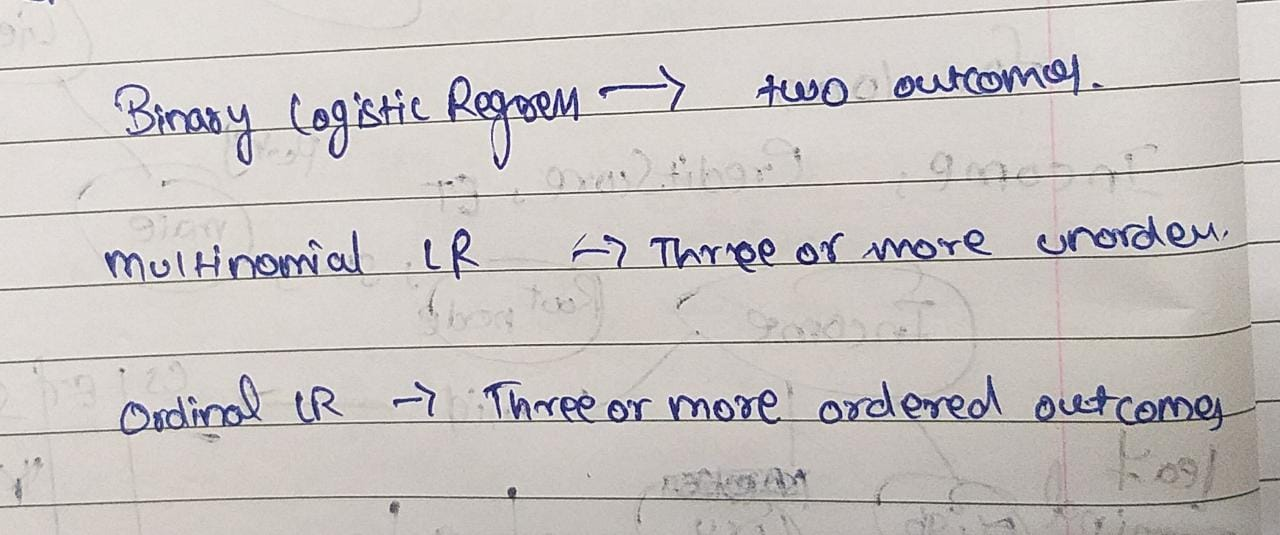
1-Logistic Regression



Uses the sigmoid function for getting the log( probability of diabetic ) 🡪

Log(p/1-p) and here we consider the threshold value as classification

That tell or gives output in o and 1.

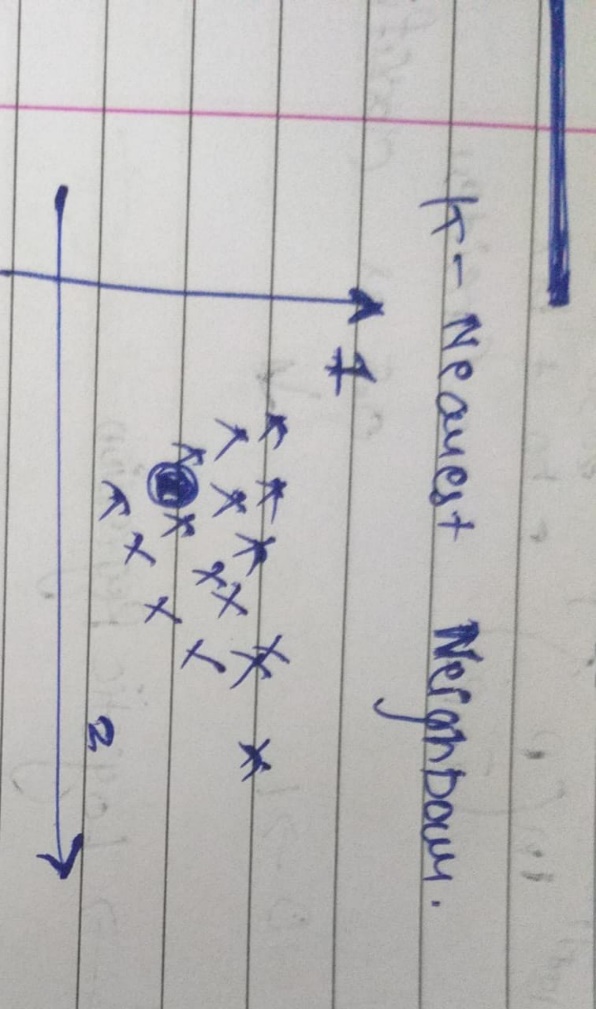


It’s accuracy is

Training accuracy of the Logistic Regression Model: 0.7785016286644951

Accuracy of predict value and accurate Y\_test value: 0.7792207792207793

2-k-Nearest Neighbour🡪



Tells you the nearest class by the instance by calculating the distance by 🡪

1-Ecludian

2-Manhattan

3-Minkowski

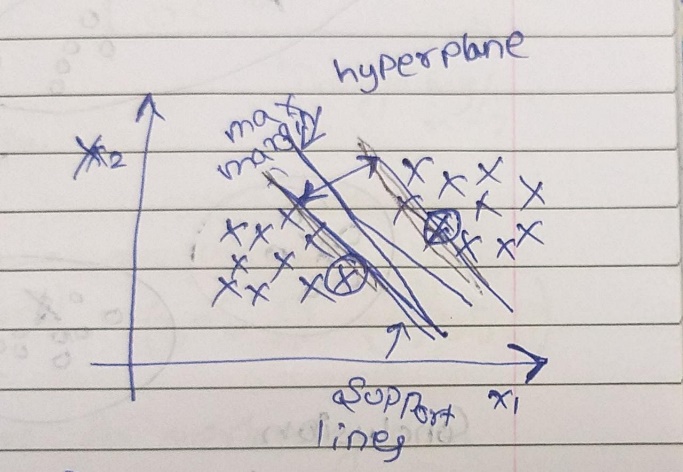
Attributes are metric 🡪 distance preference,

n\_neighbours = 5(by default)

Training accuracy of the KNN Model: 0.8175895765472313

Accuracy of predict value and accurate Y\_test value: 0.7792207792207793

3-Support Vector Machine🡪



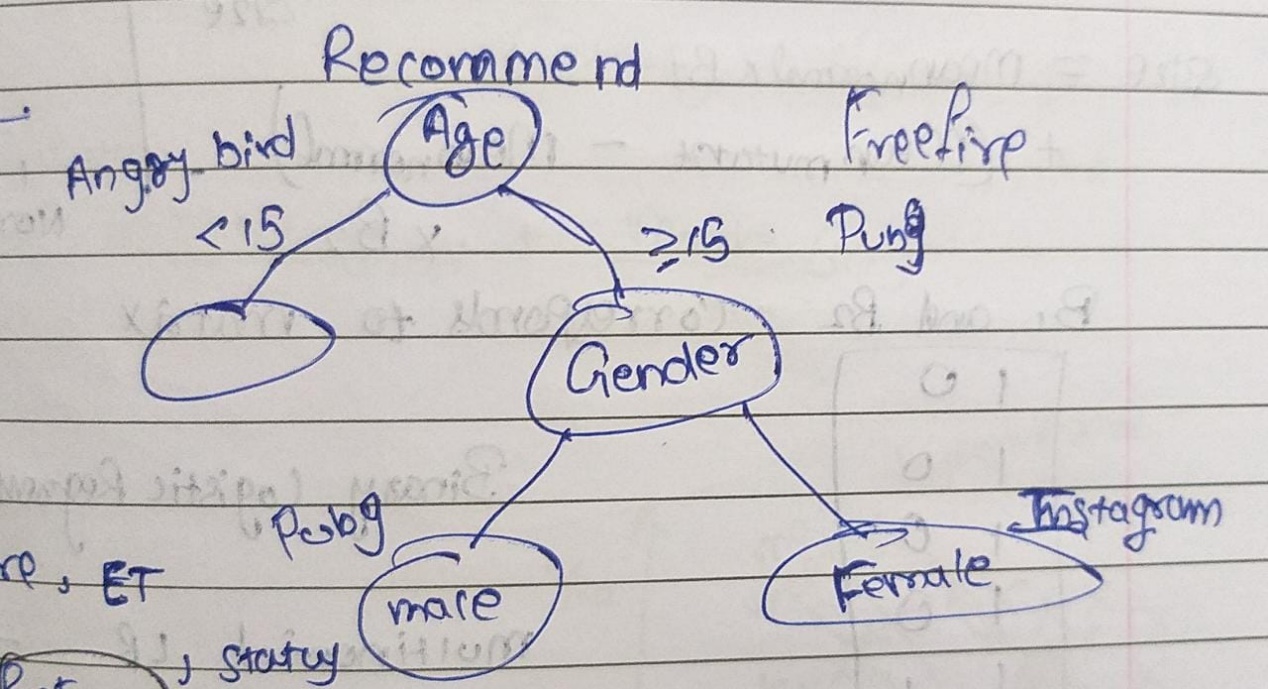
Here x 🡪 support vectors, Best margin to distinguish the area

Mostly used algorithm 🡪 Best accuracy, but most risky, easy for calculation due to hyperplane in between classes that defines positive and negative margin.

Training accuracy of the SVM Model: 0.8127035830618893

Accuracy of predict value and accurate Y\_test value: 0.7792207792207793

4-Desicion Tree🡪



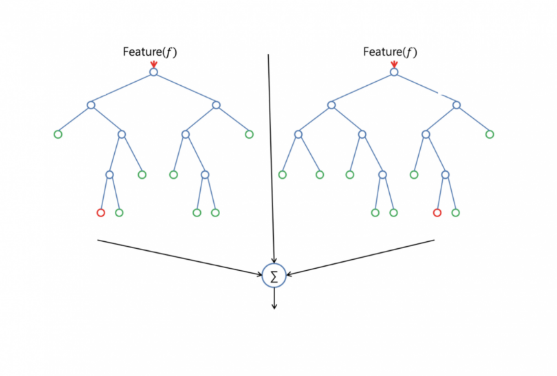
That splits in more defined ways.

The accuracy of the Model by Decision tree Method 🡪

Training accuracy of the Decision Tree Model 1.0

Accuracy of predict value and accurate Y\_test value 0.7142857142857143

5-Random Forest 🡪



Training accuracy of the Random Forest Model 0.988599348534202

Accuracy of predict value and accurate Y\_test value 0.7402597402597403

Conclusion🡪

Best fit algorithm is K-Nearest Neighbour because its accuracy is 77.922 and its training accuracy is 81.750.

To find accuracy use confusion matric 🡪

Accuracy = (TP+TN)/ (TP+TN+FP+FN)

Second best fit model is Support Vector machine because its accuracy is 77.922 and its training accuracy is 81.270.

-------------X----------X--------------X-----------X-----------------

Name 🡪 Devansh Aheer

Major Project

Machine Learning

June to July