CARDIVASCULAR RISK PREDICTION

PROJECT SUMMARY

Data Description

Demographic:

Sex: male or female("M" or "F")

Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioral

is_smoking: whether or not the patient is a current smoker ("YES" or "NO")

Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Medical *(history)*

BP Meds: whether or not the patient was on blood pr
essure medication (Nominal)
Prevalent Stroke: whether or not the patient had pr
eviously had a stroke (Nominal)
Prevalent Hyp: whether or not the patient was hyper
tensive (Nominal)
Diabetes: whether or not the patient had diabetes (
Nominal)
Tot Chol: total cholesterol level (Continuous)
Sys BP: systolic blood pressure (Continuous)
Dia BP: diastolic blood pressure (Continuous)
BMI: Body Mass Index (Continuous)
Heart Rate: heart rate (Continuous -
In medical research, variables such as heart rate
though in
fact discrete, yet are considered continuous becaus
e of large number of possible values.)

PROBLEM STATEMENT

The Cardivascular data consists of set of columns that speaks about the patient is suffering from heart disease or not

first we have to load the dataset and by understanding the dataset with the Meaningfull features and rows we perform the model

There are **3390** rows and **16** columns with in the dataset

FEATURES

age: Speaks about the age of the patient

education : shows the education the the patient done in four different phases 1. schooling,2.intermediate,3.UG,4.PG

****sex****:Represents whether the patient is male or female

is_smoking:Represent whether the patient is smoking or not

cigsPerDay:If smoking then how many cigars the patient can take per a day

- **BPMEDS:** is there bp in thee patient
- **PREVALENTSTROKE:** is there any previous stro ke in the patient
- **diabetes:** Is the patient suffering with dia betes

TOTCHOL :Total cholestrol

- **SYSBP** :systollic blood pressure
- **DIABP** :Diastollic blood pressure
- **BMI** :body mass index
- **HEARTRATE** :represent the heart rate of the patient
- **glucose** :glucose levels of the patient
- **TENYEARCHD:**Predicting ten year coronary hea rt disease with in the patient

Finding the co related features

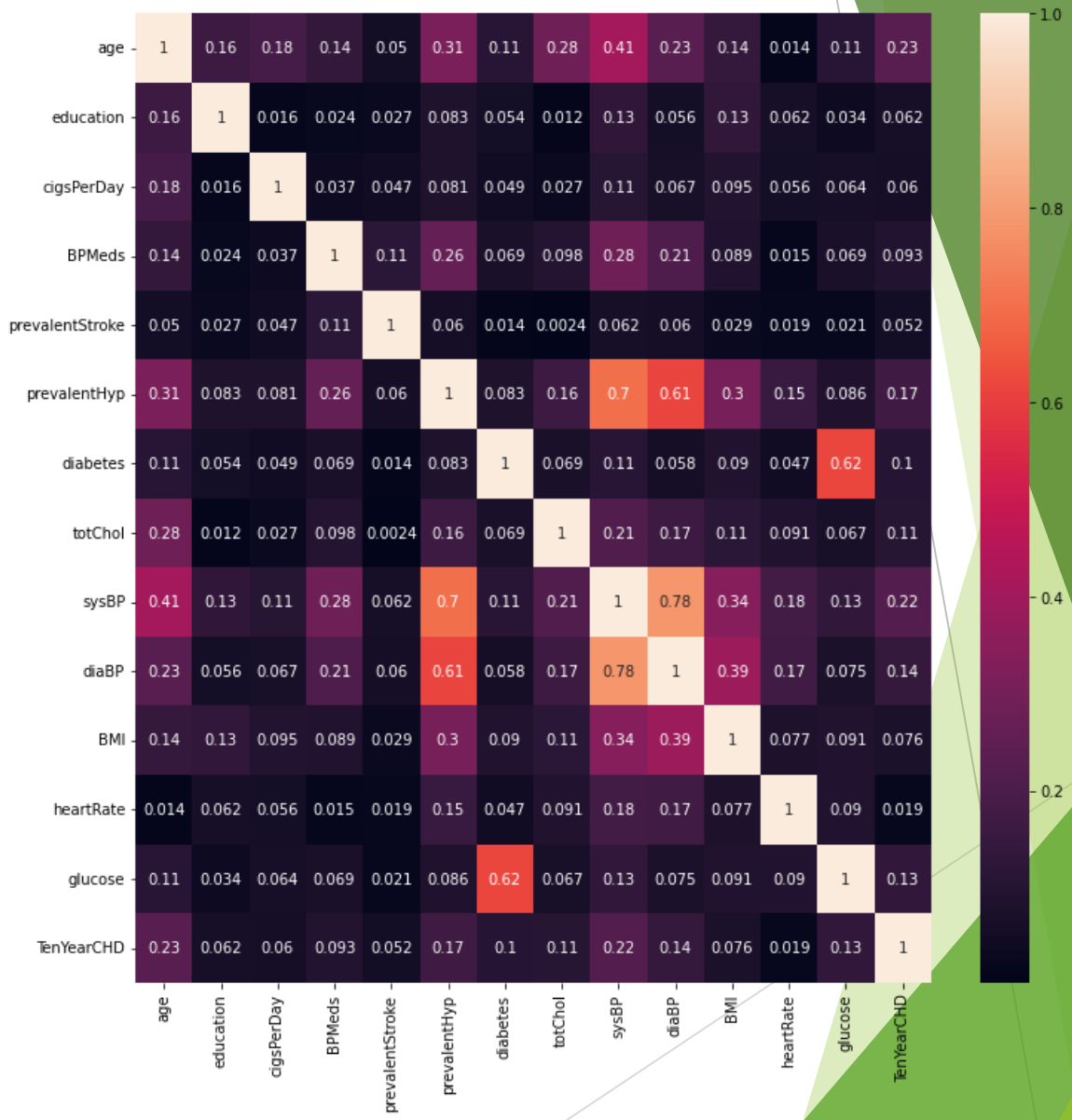
Here we can see that different featues co related with each other

1.Sysbp and diabp features are highly co related with each other with .78 percentage among others

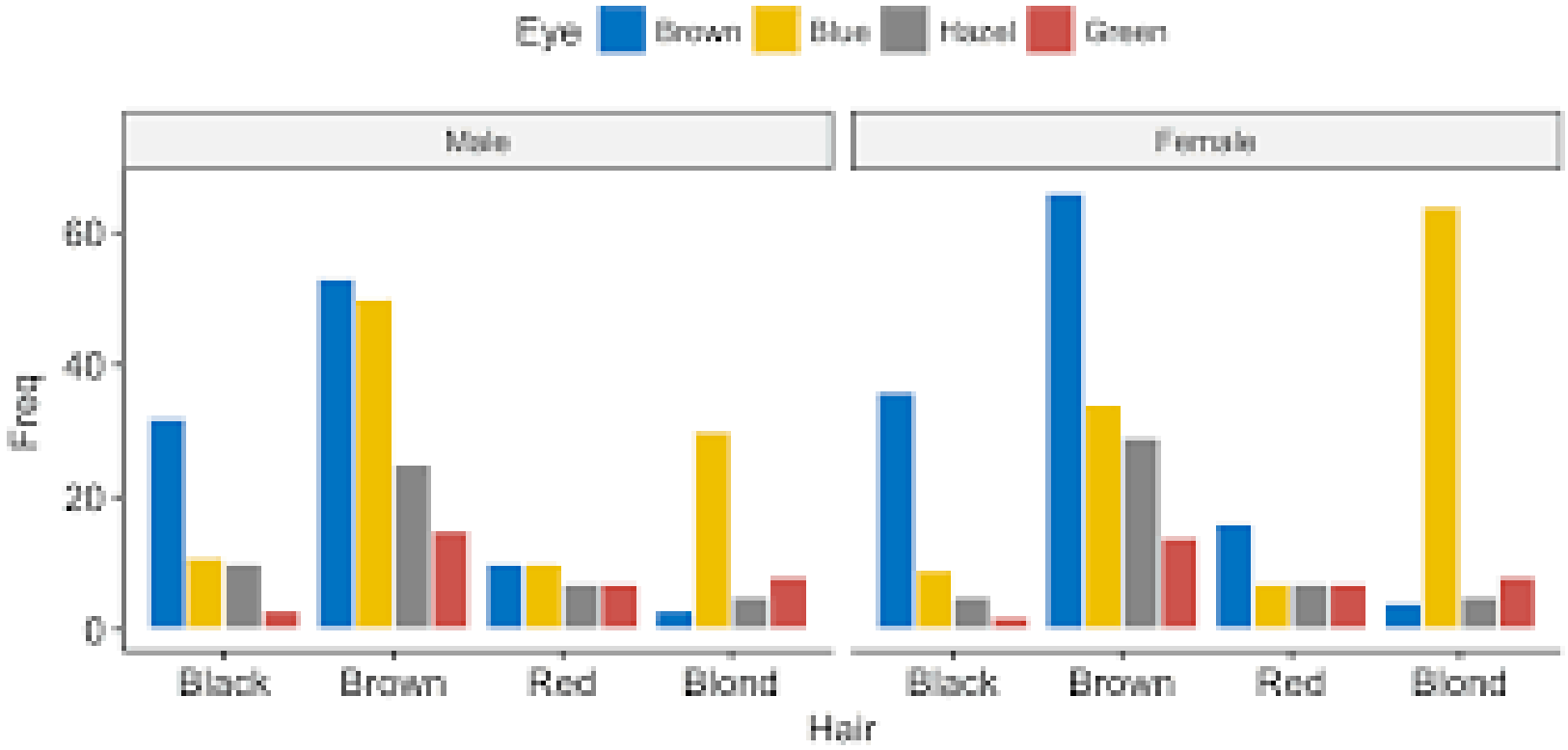
2.*There is no significant correlation between independ ent and dependent varibles*

3.but there is nhigh correlation between independent v ariables

4. Also some of the other features tht are correlatd with eeach oter with some percentage



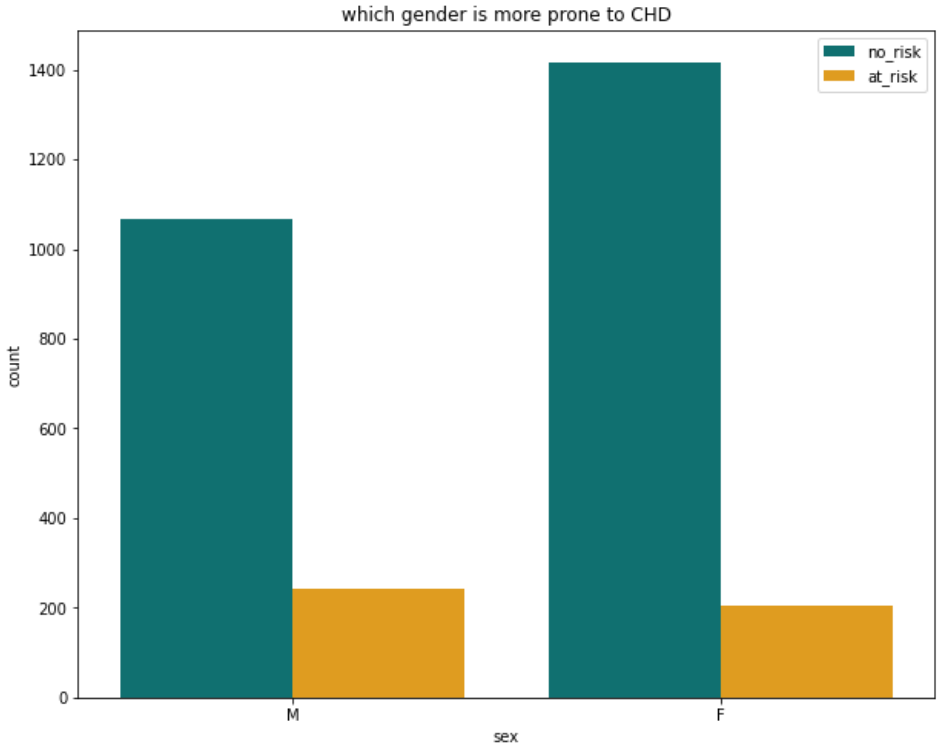
HERE WE HAVE DONE SOME CATEGORICAL VISUALISATION ON THE DATASET



Gender prone too heart disease

By the graph we intend that females are less effe ctive of disease when compared to males

- As the green represents the no risk of disease and yellow represents at risk

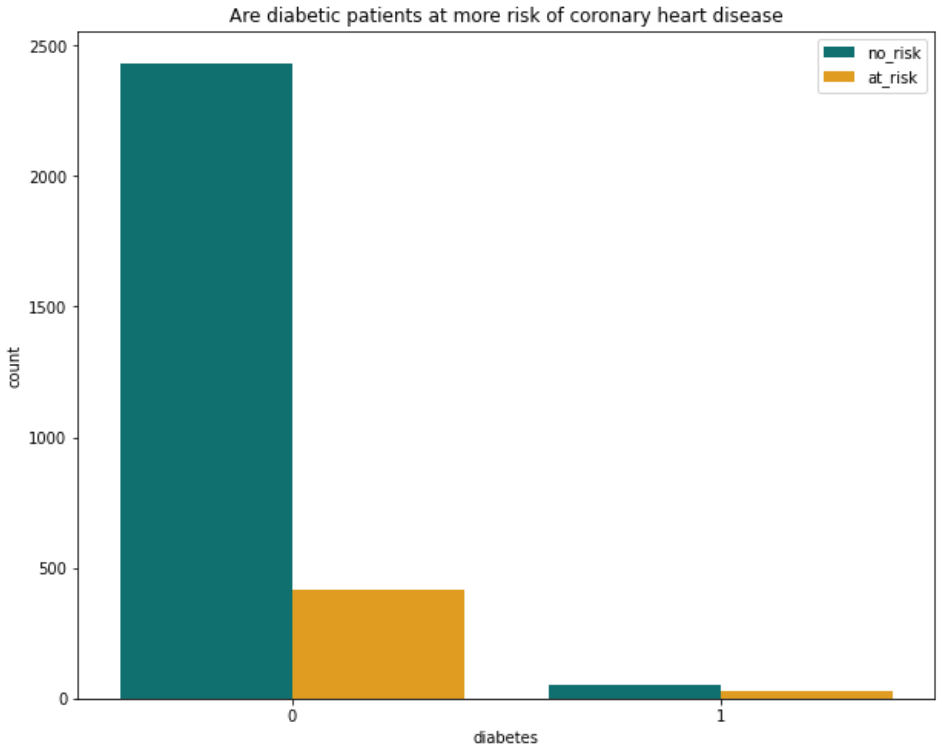


Are diabetic patients at more risk of coronary heart disease

we representts the diabetic patients at more risk o f coronary heart disease

0 indicates tenyearCHD occurs and 1 indicates not

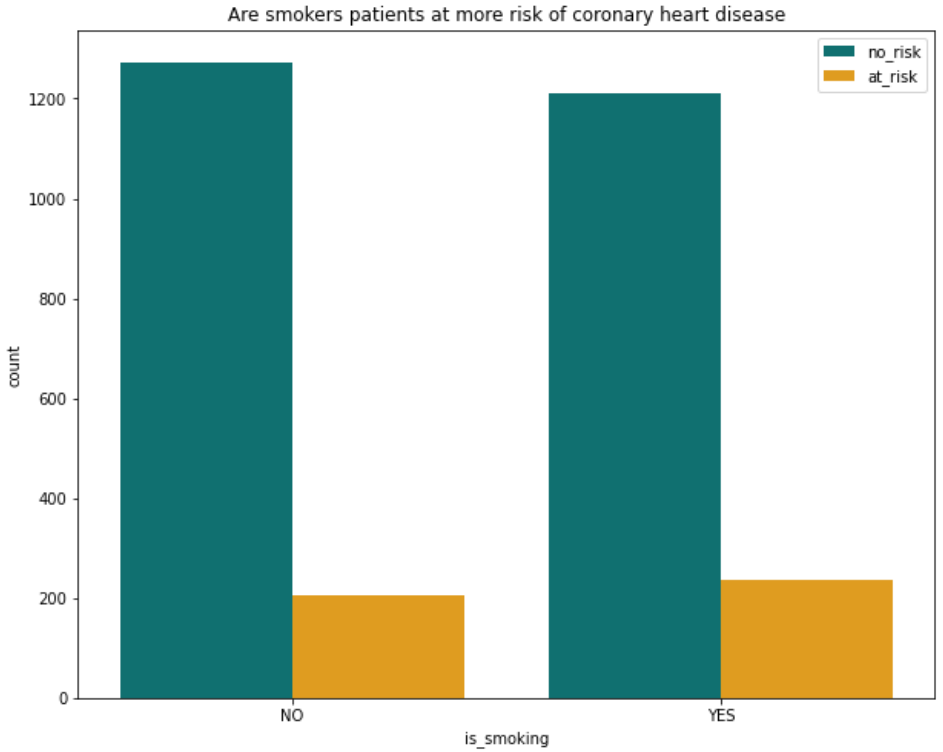
here we ca see that diabetes patients are more ser ioiusly occurs with CHD



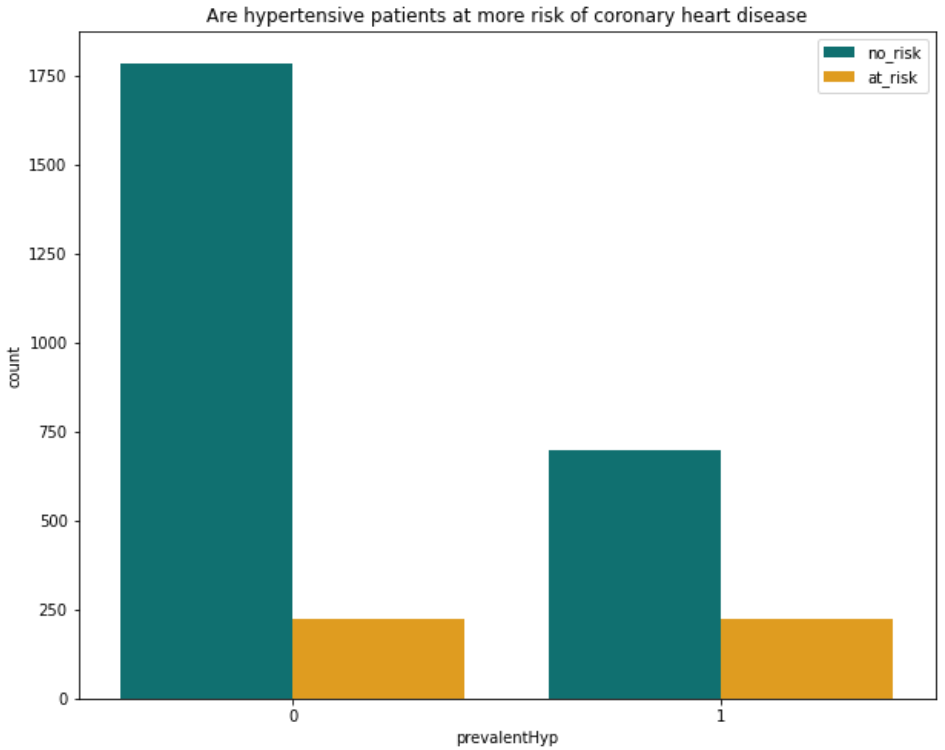
Are smokers patients at more risk of coronary heart diseas

WE compared that smoking effects the CHD then we noptices that

there is no relation of smoiking causes tha CHD by this graph



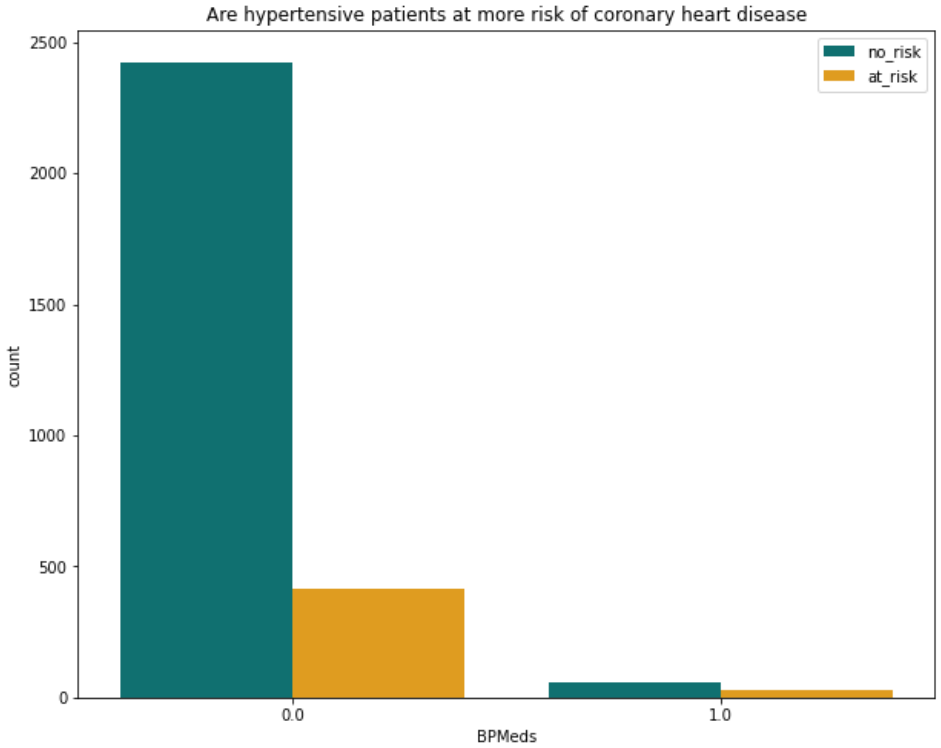
Are hypertensive patients at more risk of coronary heart disease



by this hypertensive patients at not at all more risk of coronary heart disease by prevlenthyp

it doesn't occurs difference among them

Are patients with blood pressure on mediacation at more risk of coronary heart disease

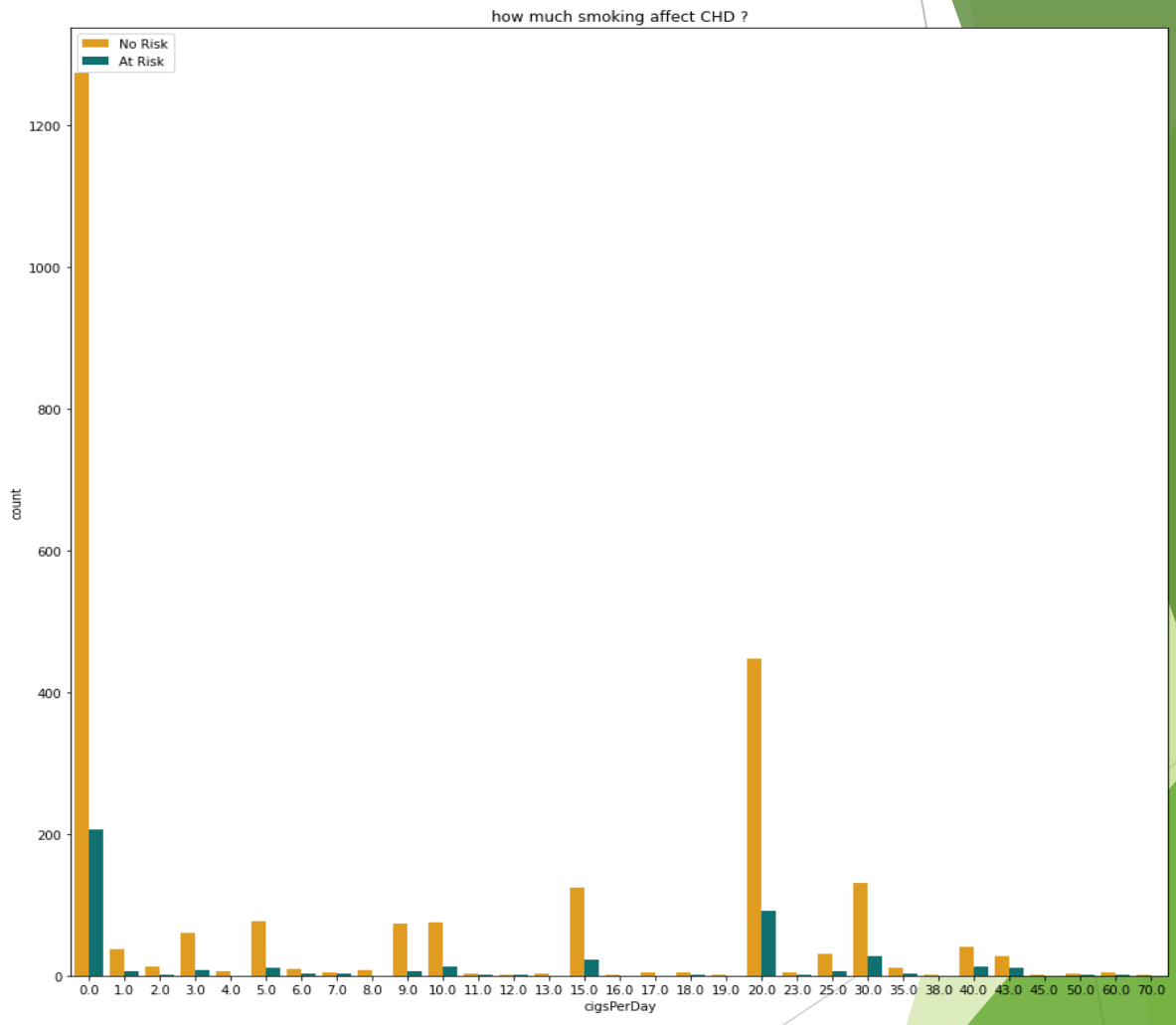


Patience with bloodpressure can indicatsthe symlp toms of occuring the CHD

theres almost equal chancewsn oif getiing the CH D with BP

How much smoking affect CHD

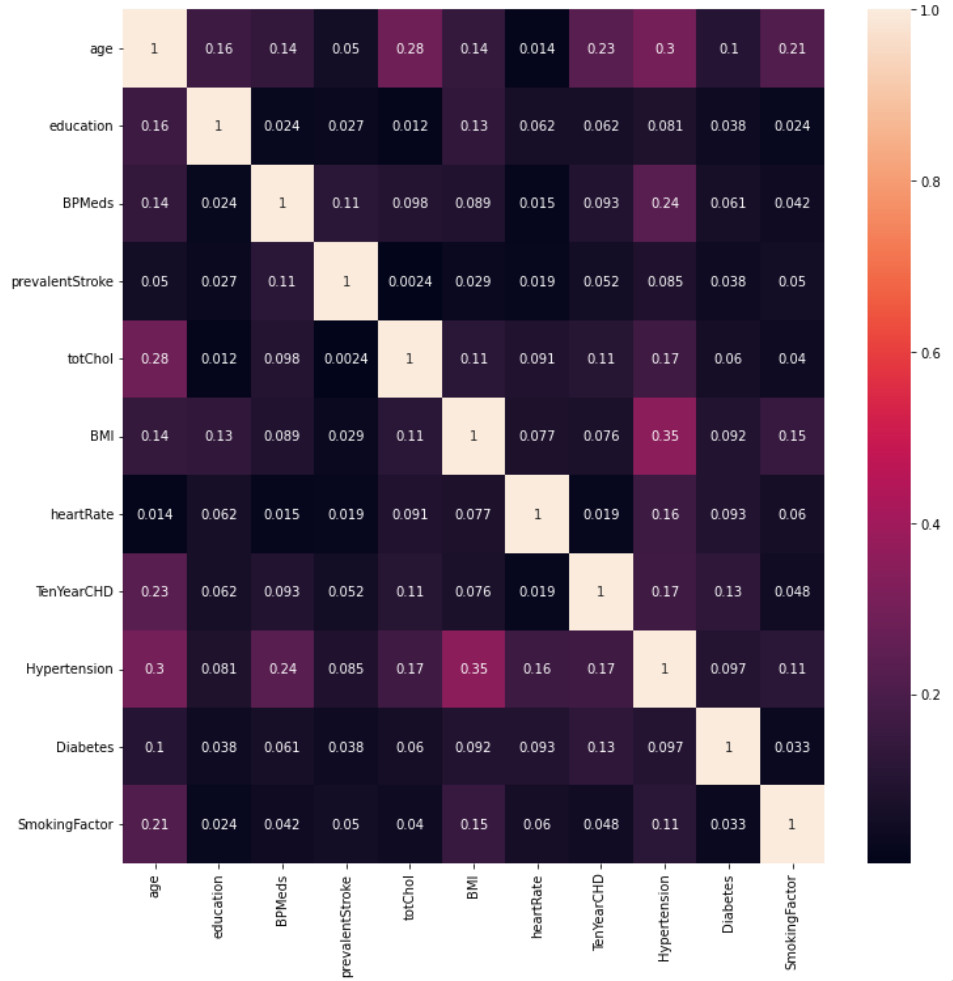
By understaning the graph we may know that smoking doesn’t put any efforts to get effected by the heart disease



Plotting the correlation matrix using heatmap

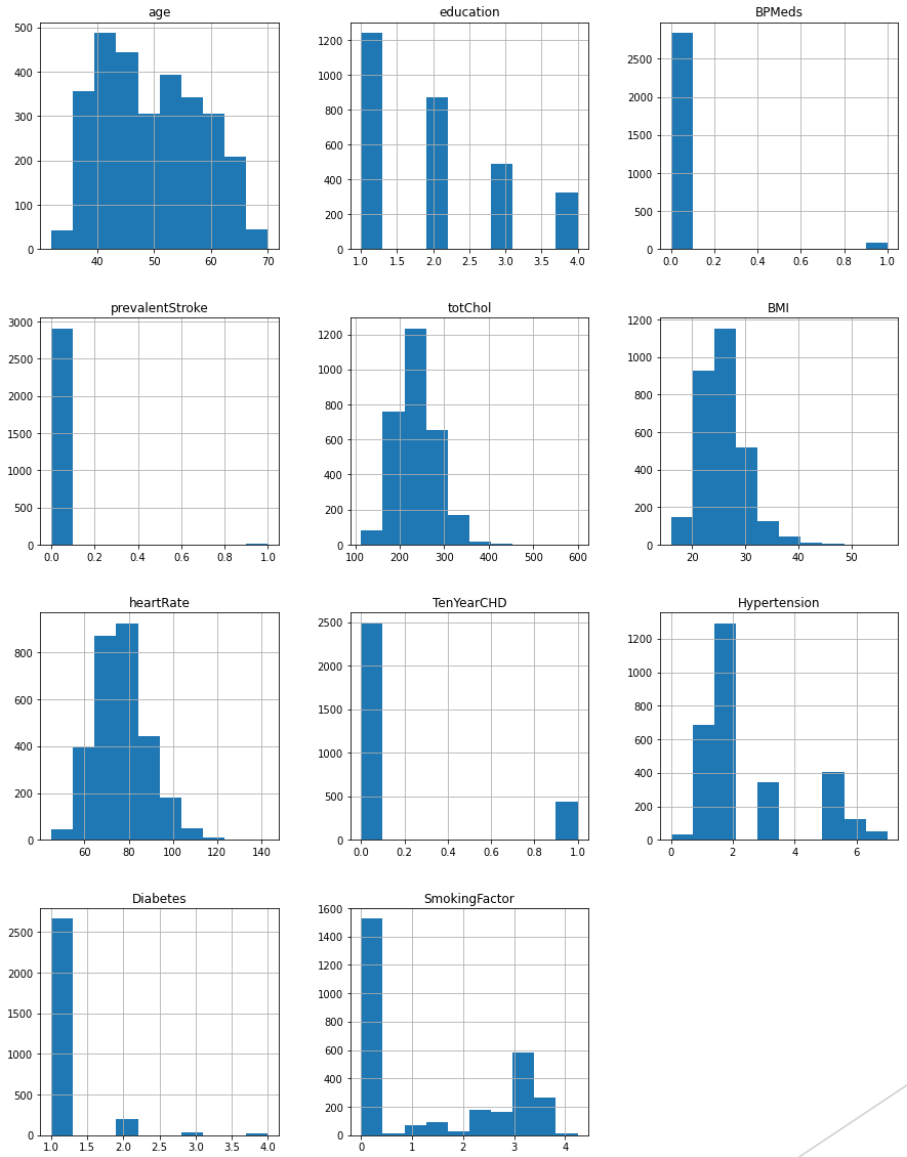
Here we can see that the multicollinearity doesn't exists between the columns

Every feature consists of its own correlation



plot histogram to see the distribution of the dat

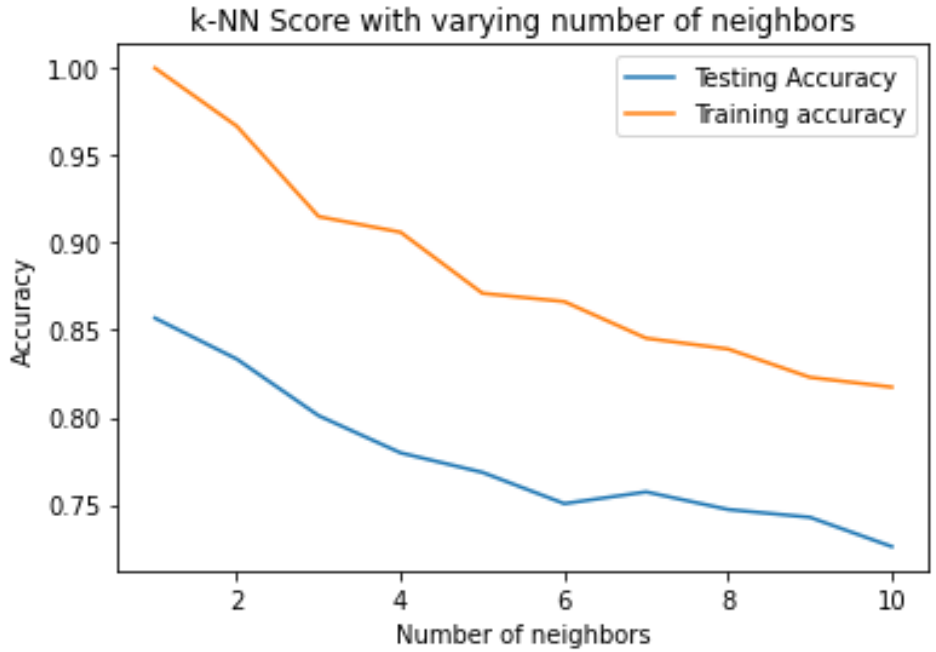
BPMeds and prevalentStroke does not help in explaining variance so we can remove those columns



ML model implementation KNN

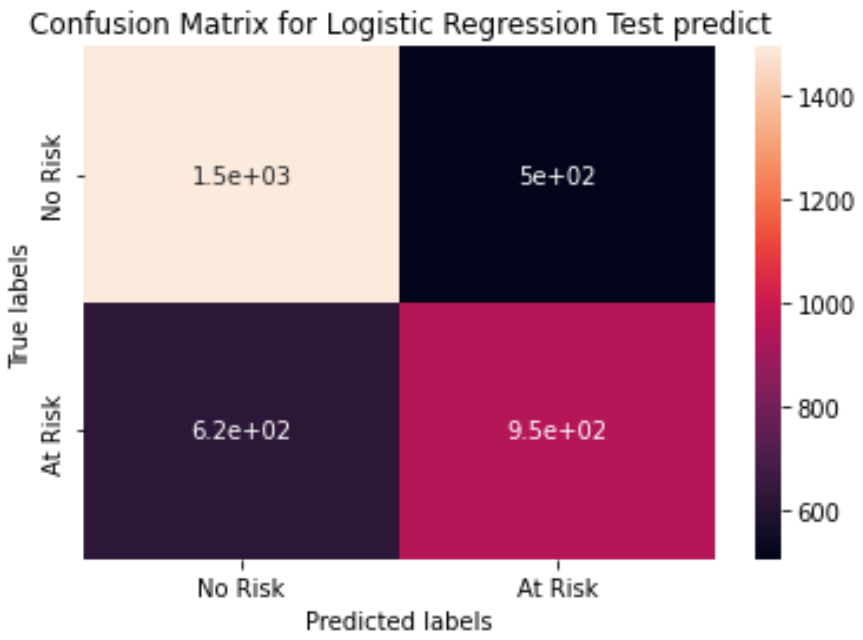
Webcan see that test score is incr easing as number of neighbours inc reases.

Let's try to find best parameter f or knn.

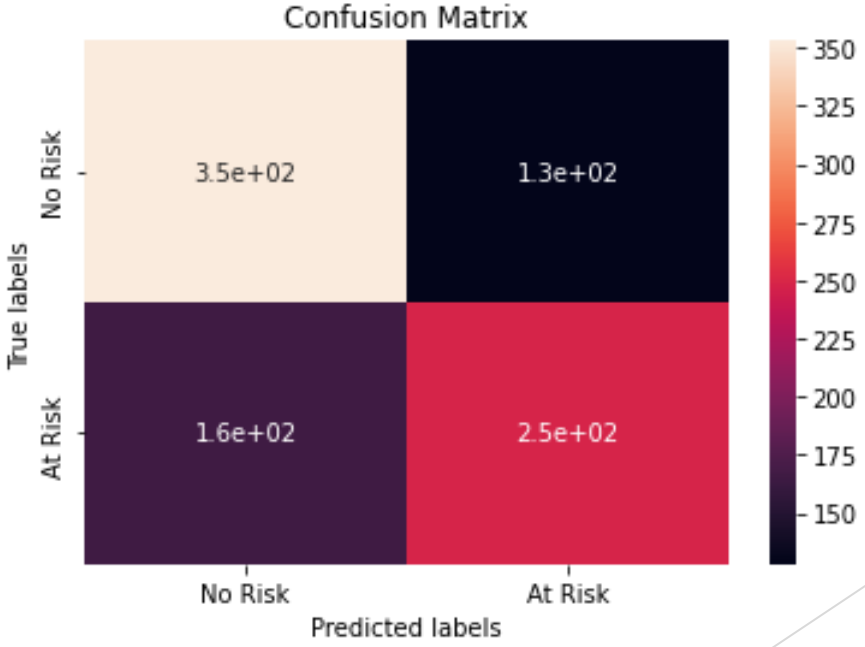


LOGISTIC REGRESSION

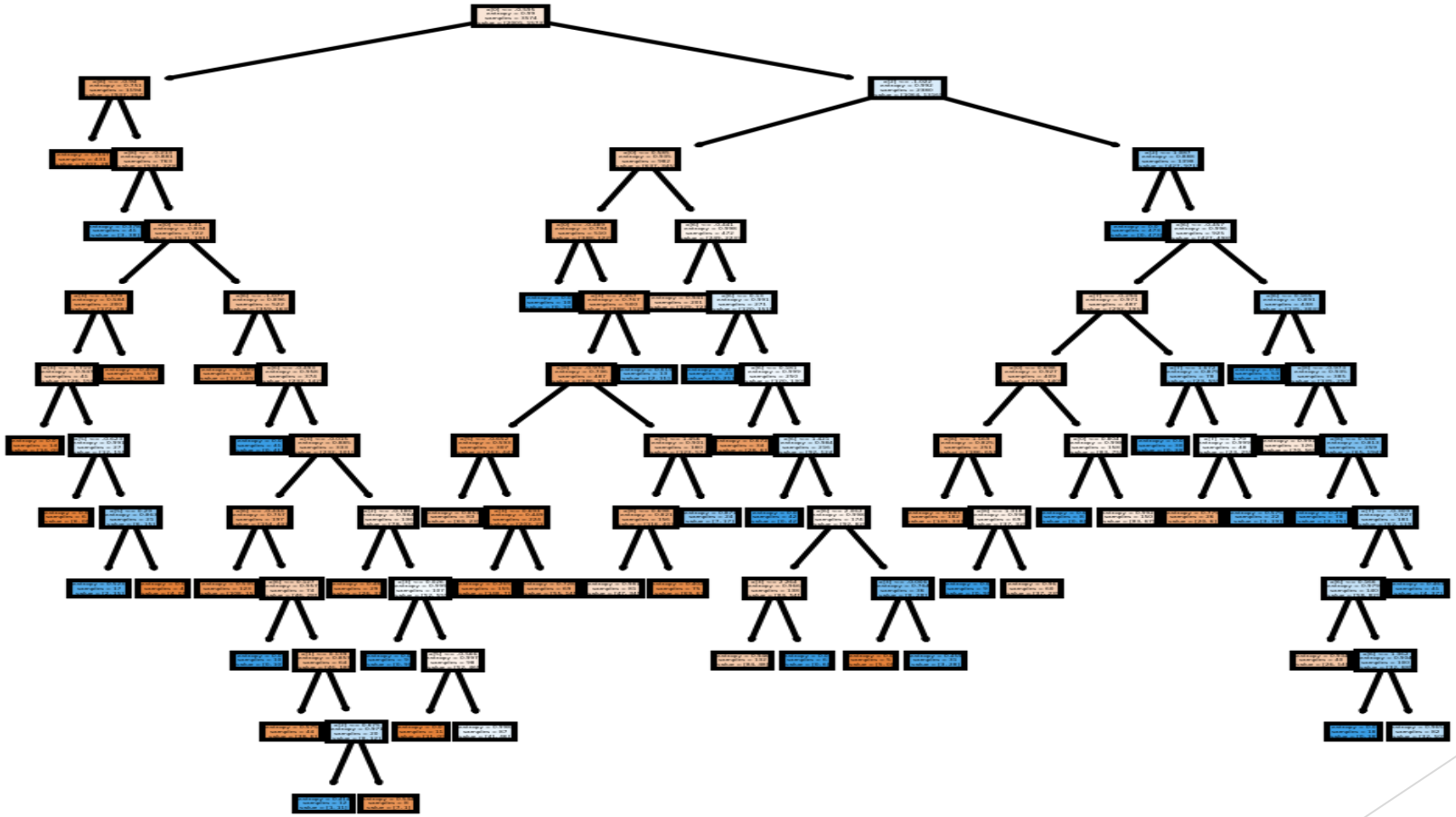
confusion matrix for train data



confusion matrix for test data

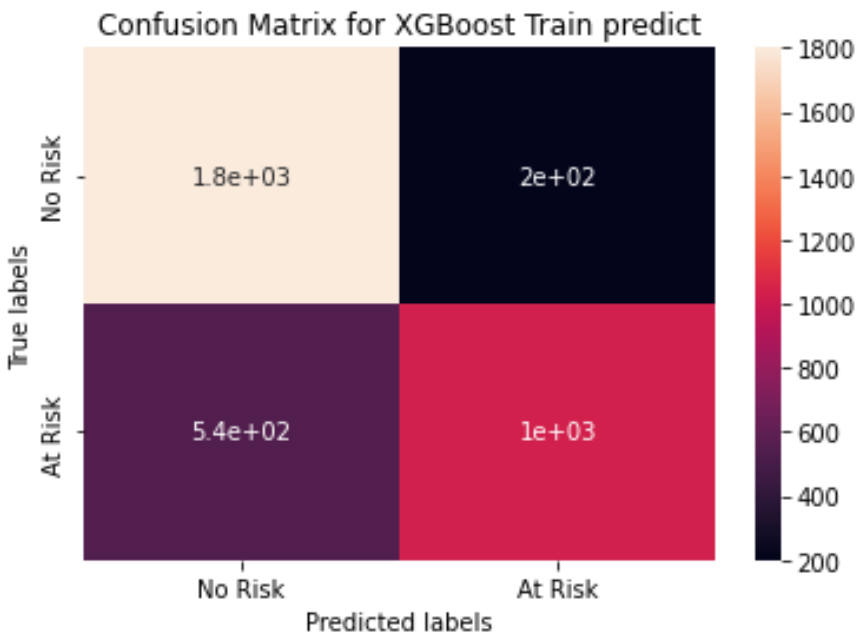


DECISION TREE CLASSIFIER

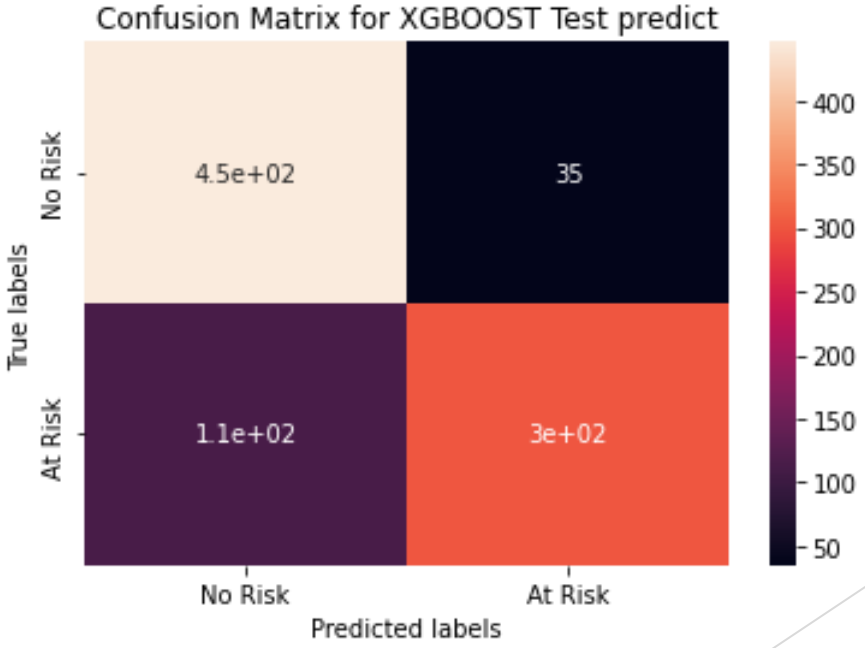


XGBOOST

CONFUSION matrix for Xgboost(train)



Confusion matrix for xgboost (test)



By selecting alla the accuracies of the model gradientboosting classifier appears to be best accuracy

The accuracy of KNN : 0.86

The accuracy of Logistic regression : 0.68

the acciracy of decision tree : 0.79

The accuracy of randomforestclassifier : 0.
88

the accuracy of XGBoost : 0.83

The accuiracy of gradientboostingclassifier
 : 0.89

THANK YOU