CAPSTONE PROJECT

CARDIOVASCULAR RISK PREDICTION

--Megala.A (Arunai Engineering College)

Problem Statement:

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

- The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease(CHD).

- The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Variable Each attribute is a potential risk factor. There are both demographic, behavioral

factors.

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)

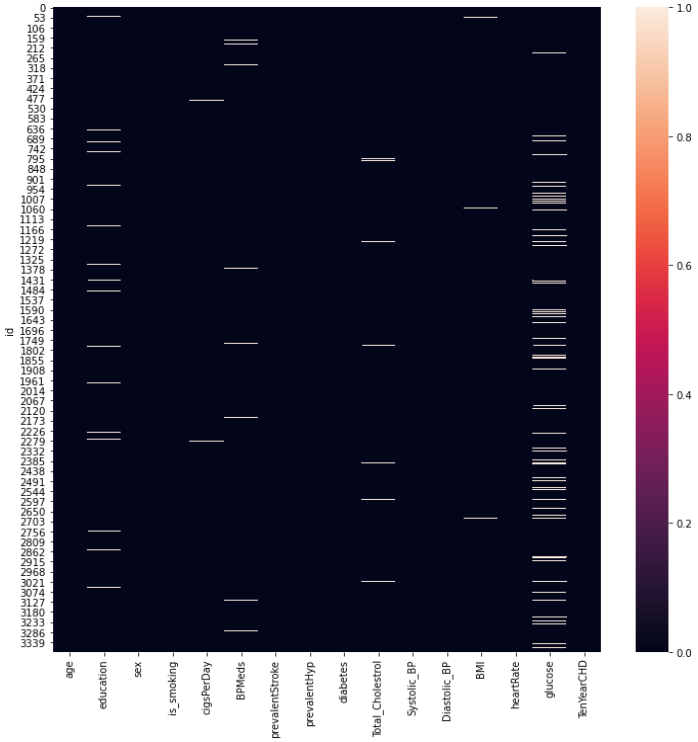
Behavioral

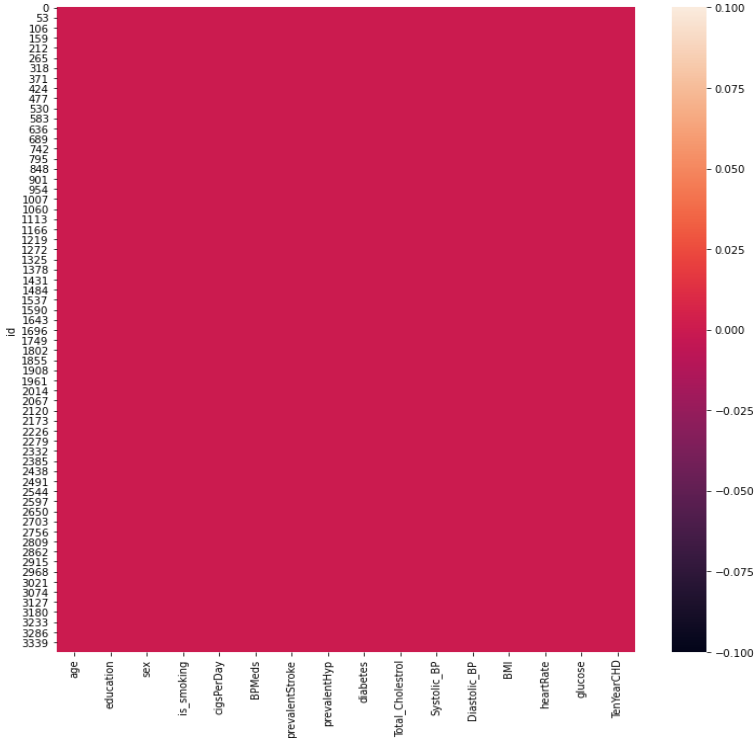
- is_smoking: whether or not the patient is a current smoker ("YES" or "NO")
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Medical(history)

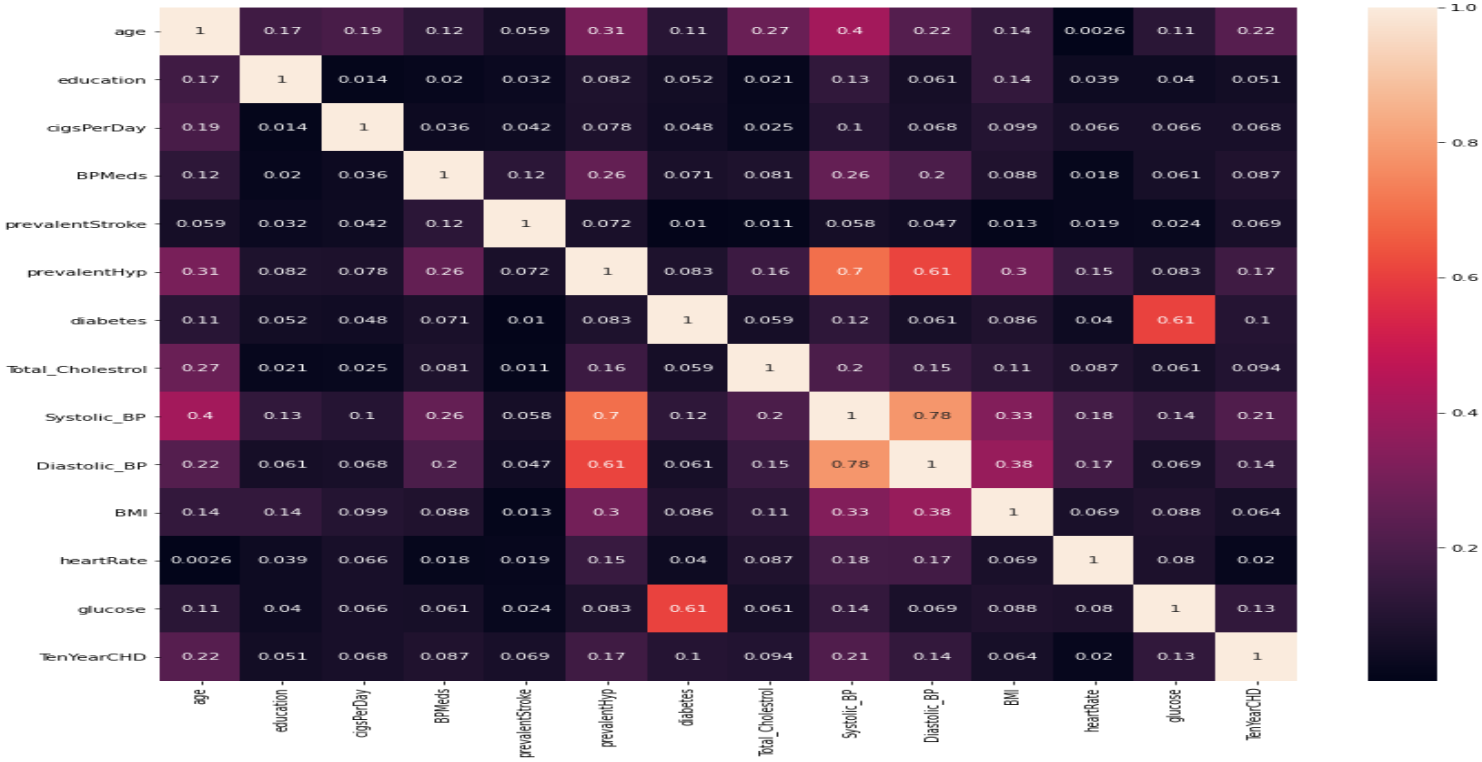
- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
- Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
- Diabetes: whether or not the patient had diabetes (Nominal) Medical(current)
- 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 because of large number of possible values.)
- Glucose: glucose level (Continuous) Predict variable (desired target)

Missing Values & After filling NaN Values

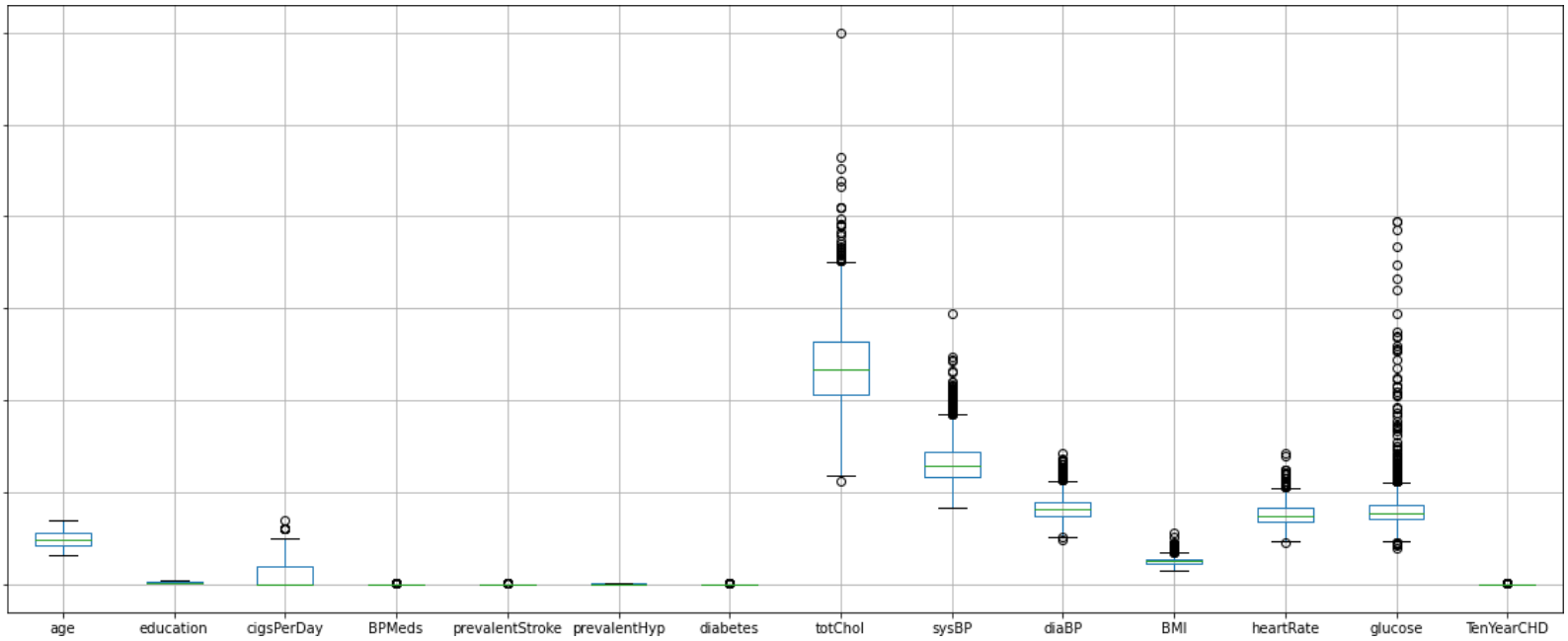




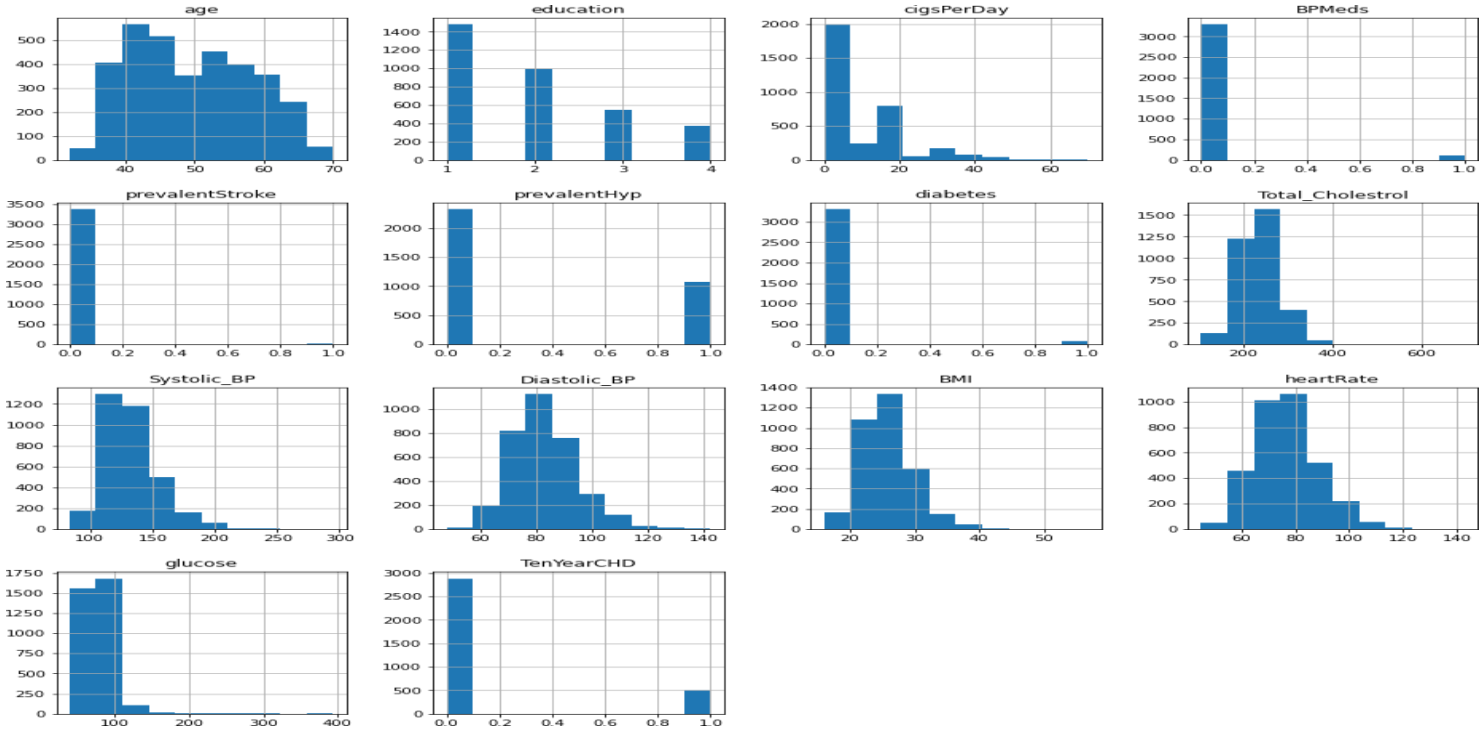
CORRELATION BETWEEN FEATURES



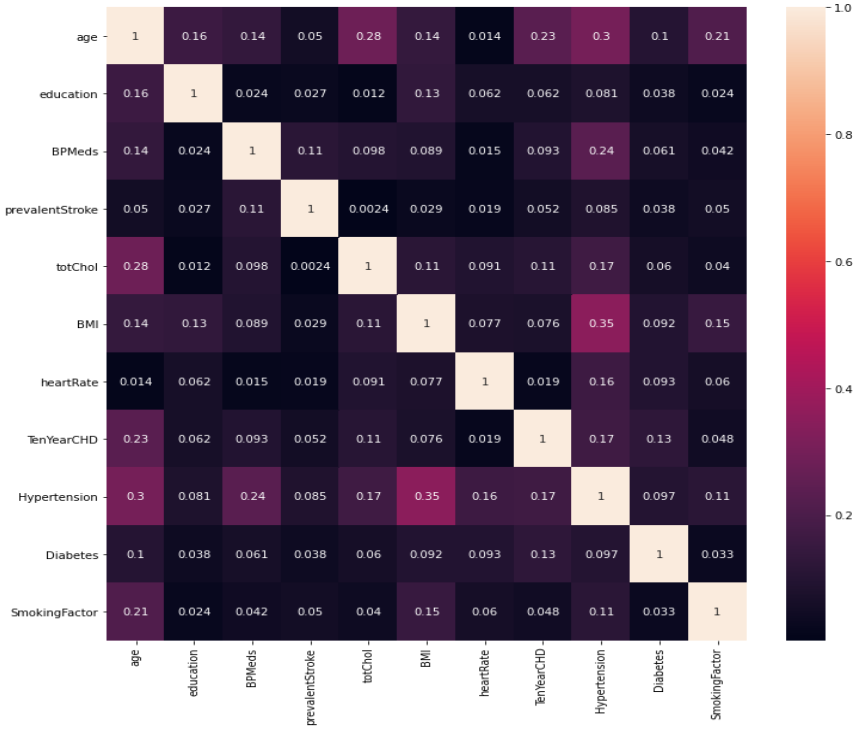
OUTLIER DETECTION



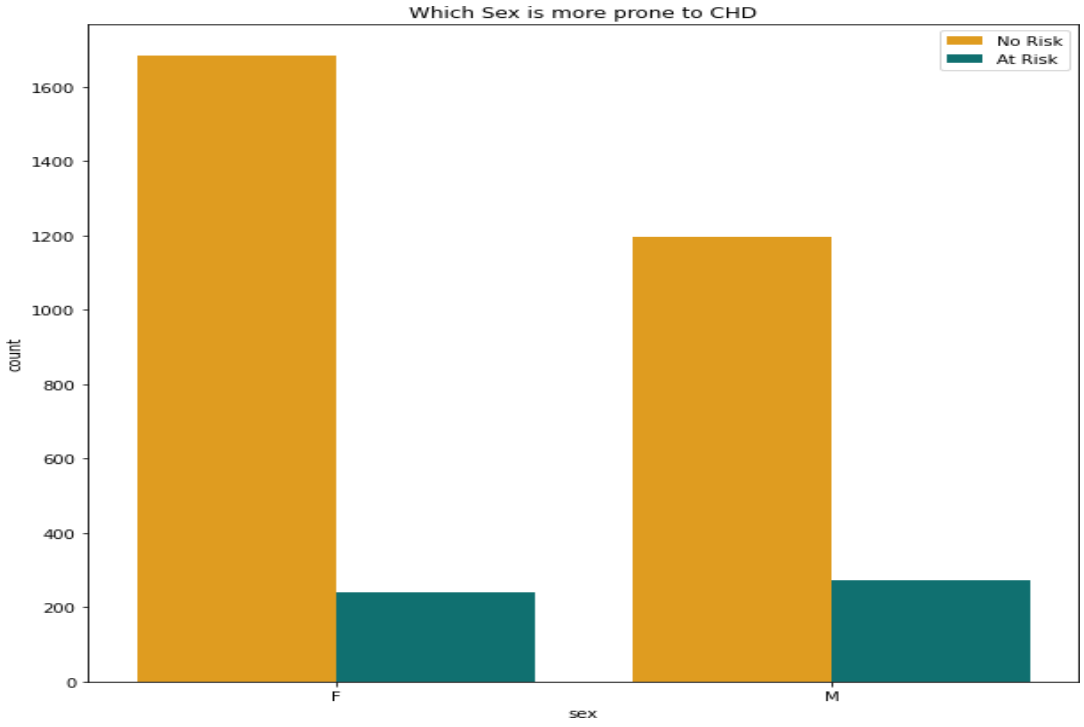
DISTRIBUTION OF DATA



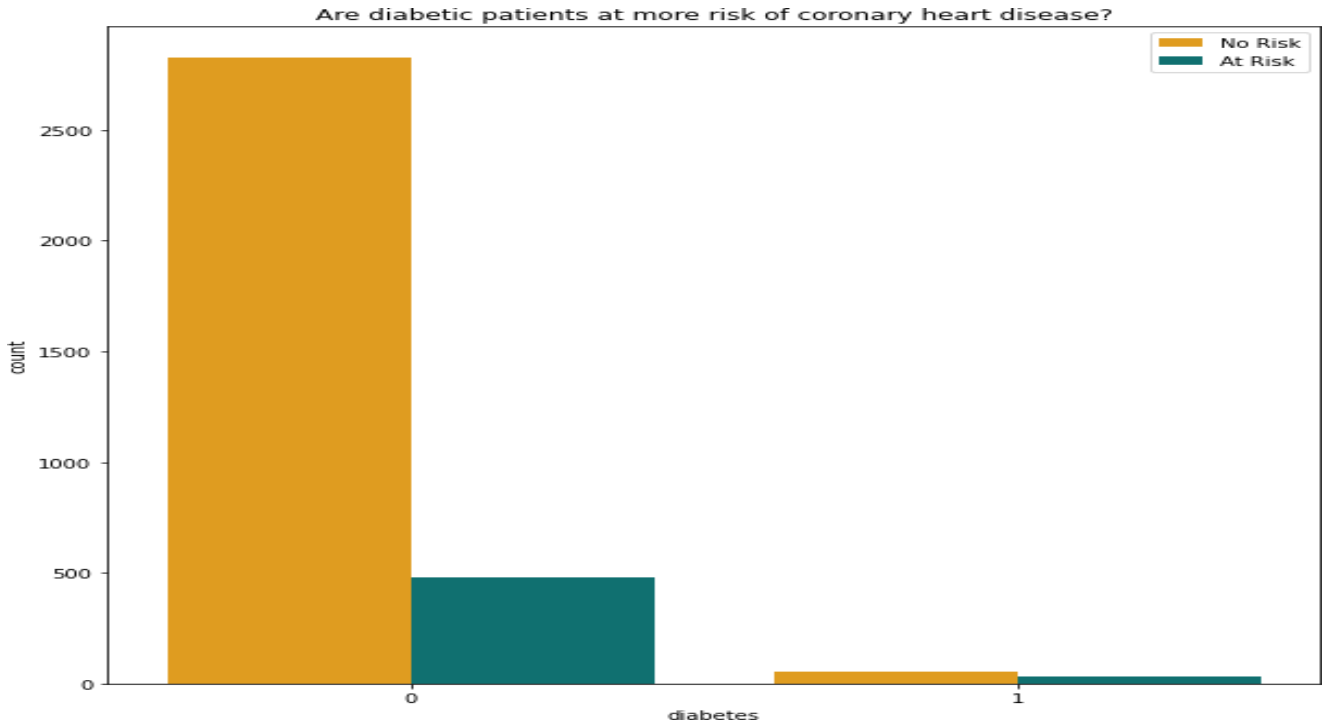
CORRELATION BETWEEN FEATURES AFTER FEATURE ENGINEERING



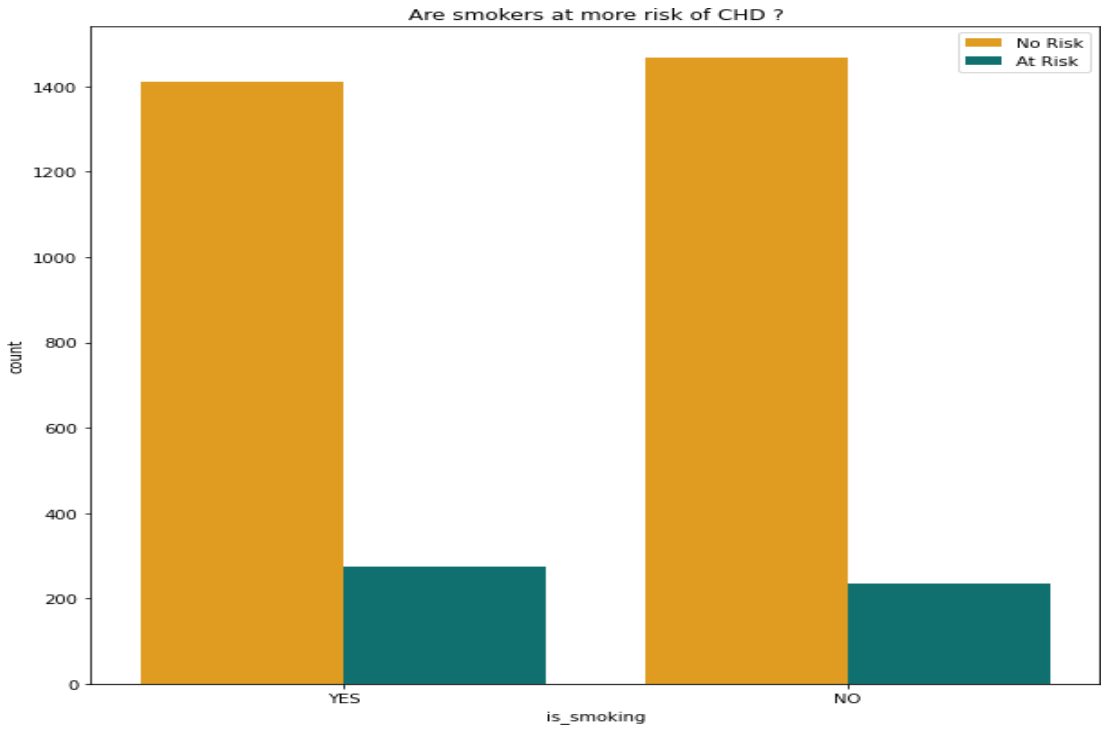
WHICH SEX IS MORE PRONE TO CHD ?



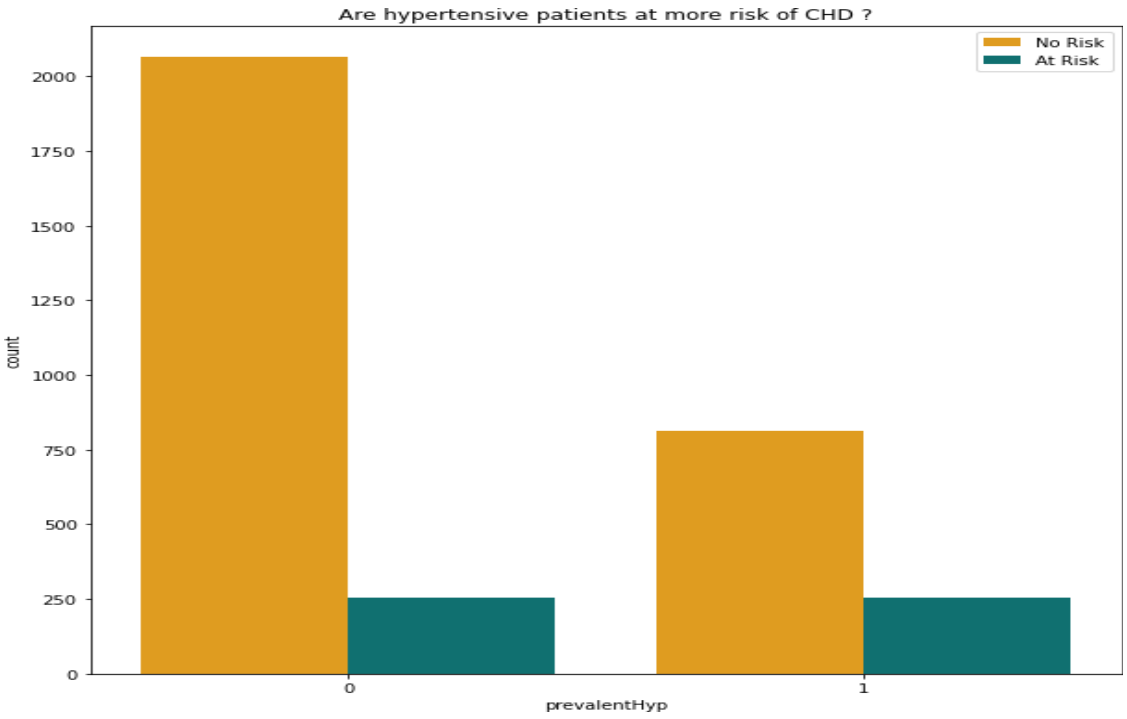
ARE DIABETIC PATIENTS AT MORE RISK OF CHD ?



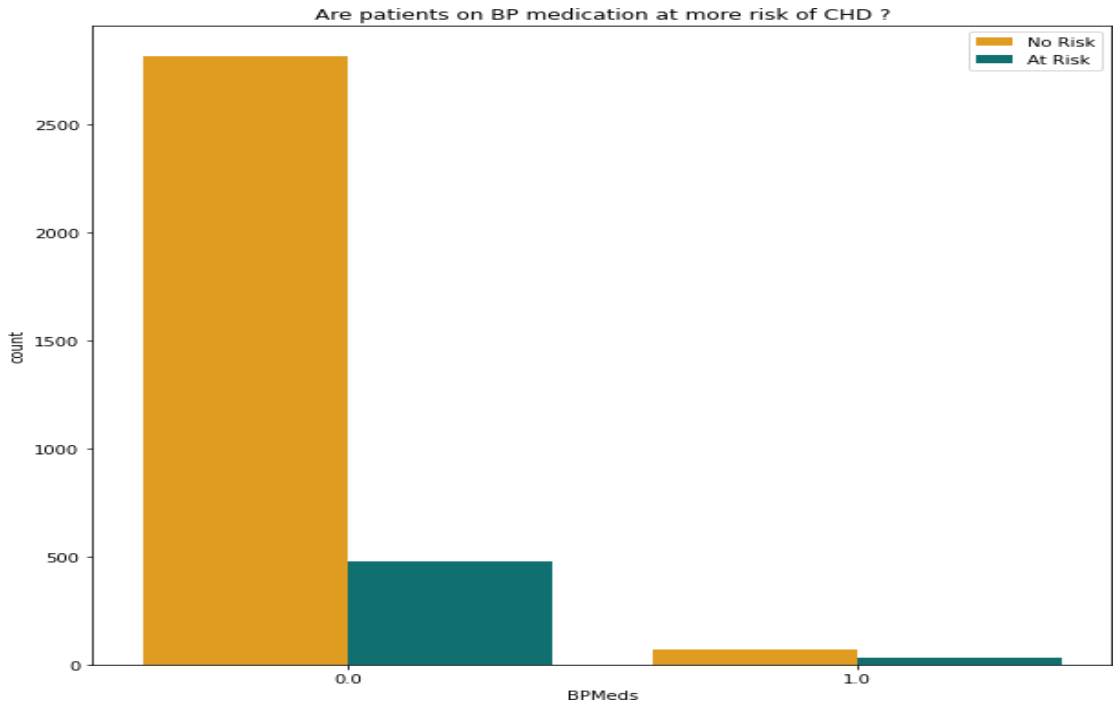
ARE SMOKERS AT MORE RISK OF CHD ?



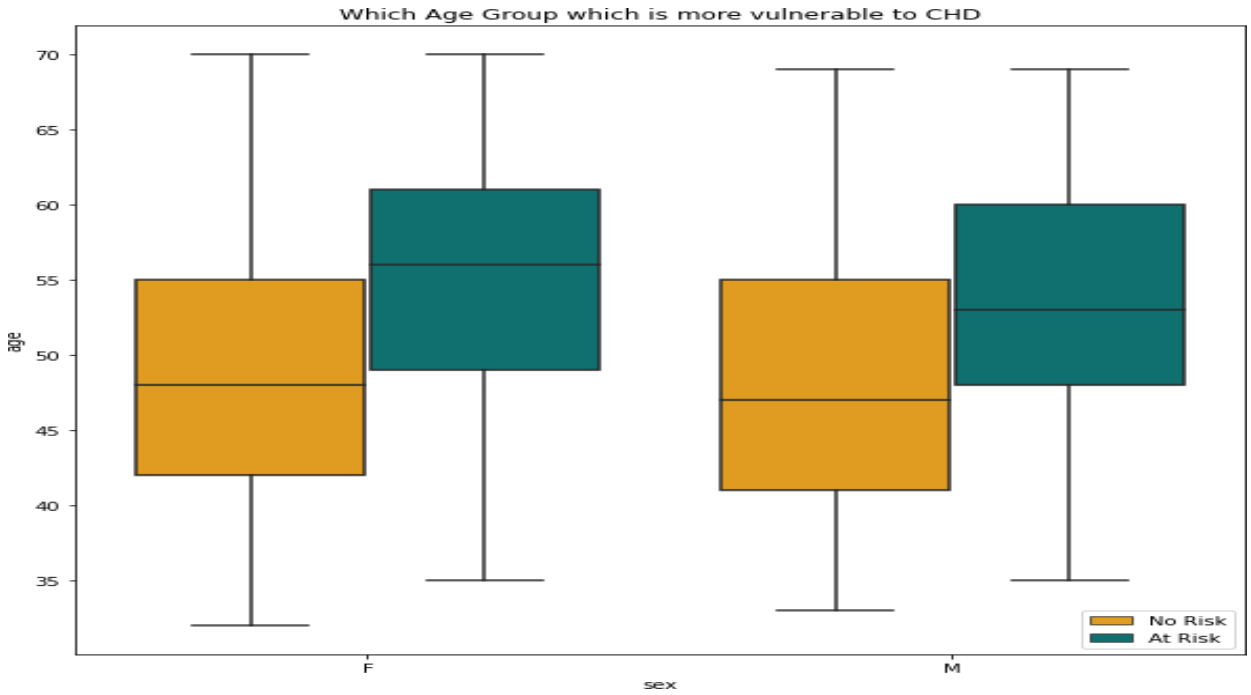
ARE HYPERTENSIVE PATIENTS AT MORE RISK OF CHD ?



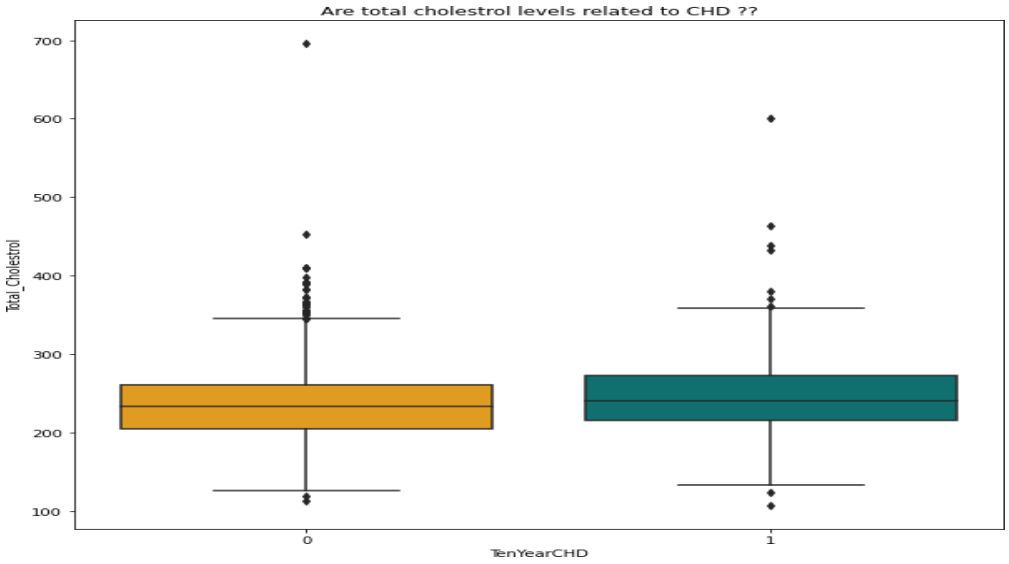
Are patients on BP medication at more risk of CHD ?



WHICH AGE GROUP IS MORE VULNERABLE TO CHD ?

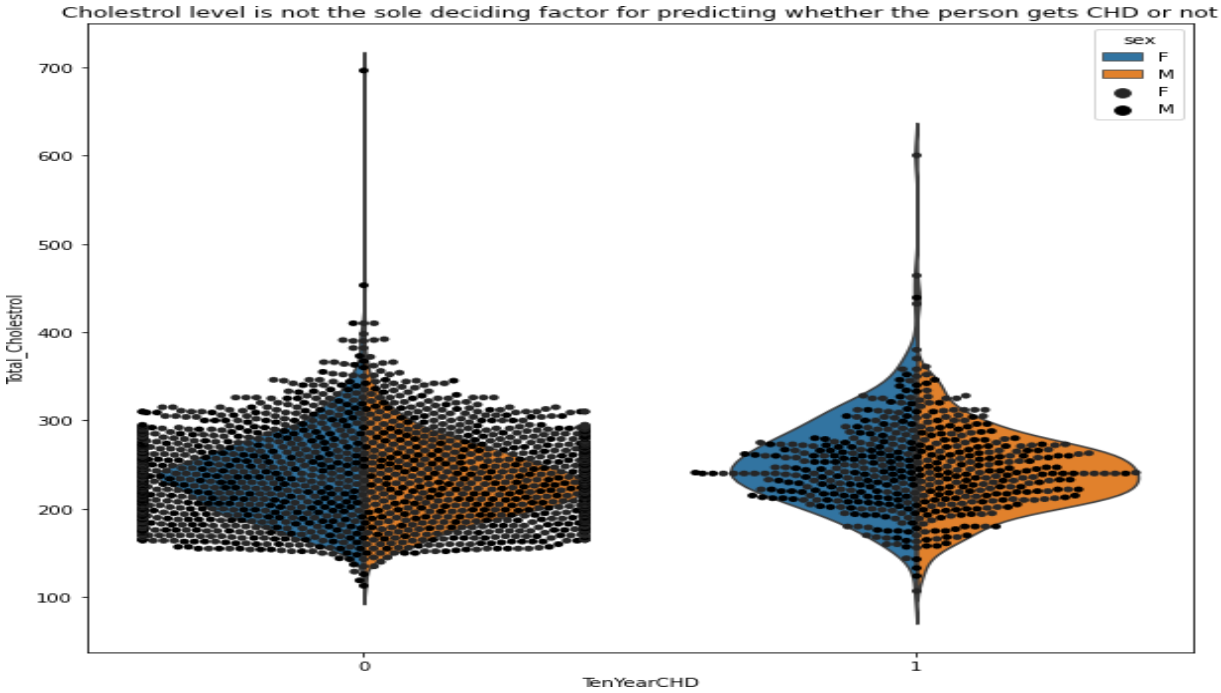


ARE TOTAL CHOLESTEROL LEVELS RELATED TO CHD ?

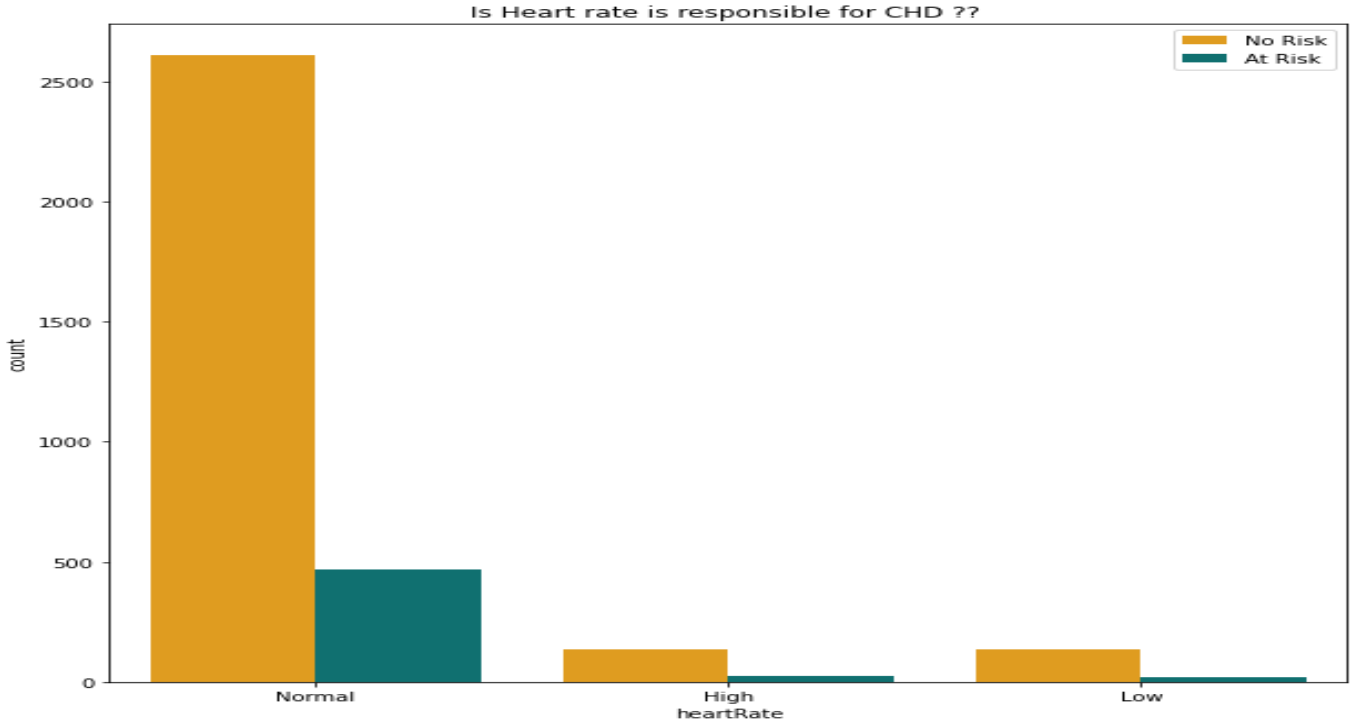


This indicates that cholestrol level is not the sole deciding factor for predicting whether the person gets coron ary heart disease or not.
People with similar levels of cholestrol have got coronary heart disease as well as are free from coronary hea rt disease.
Clearly, there is no direct correlation of coronary heart disease with the cholestrol level.

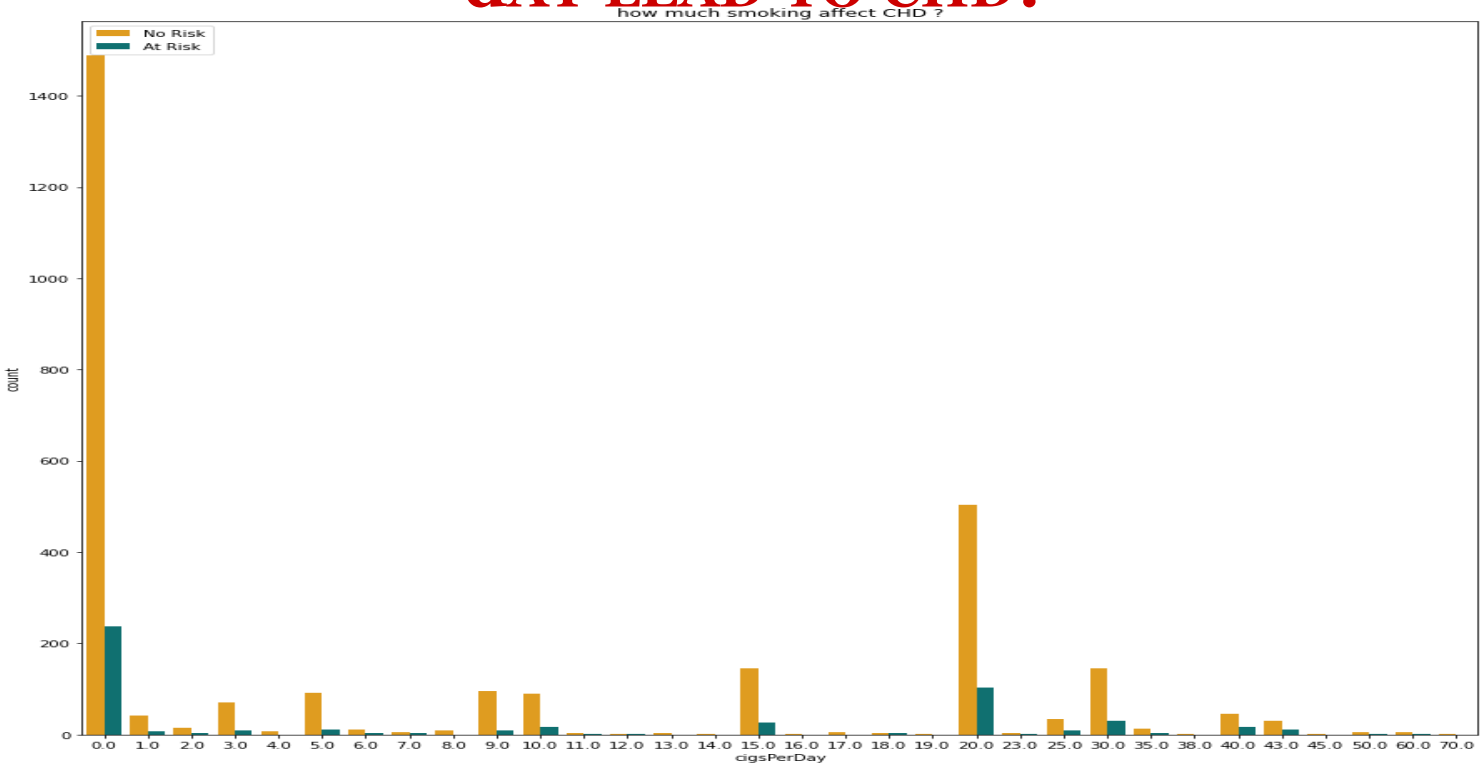
Cholestrol level is not the sole deciding factor for CHD



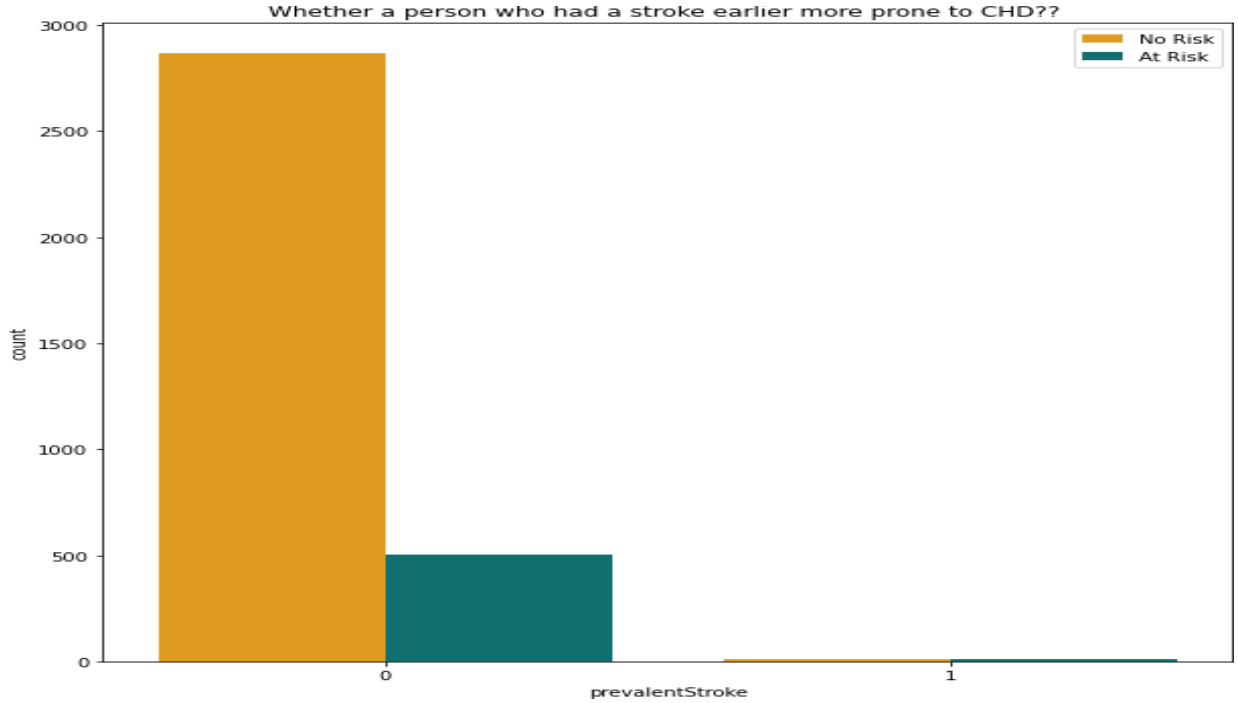
Can Heart rate possibly define the risk of CHD ?



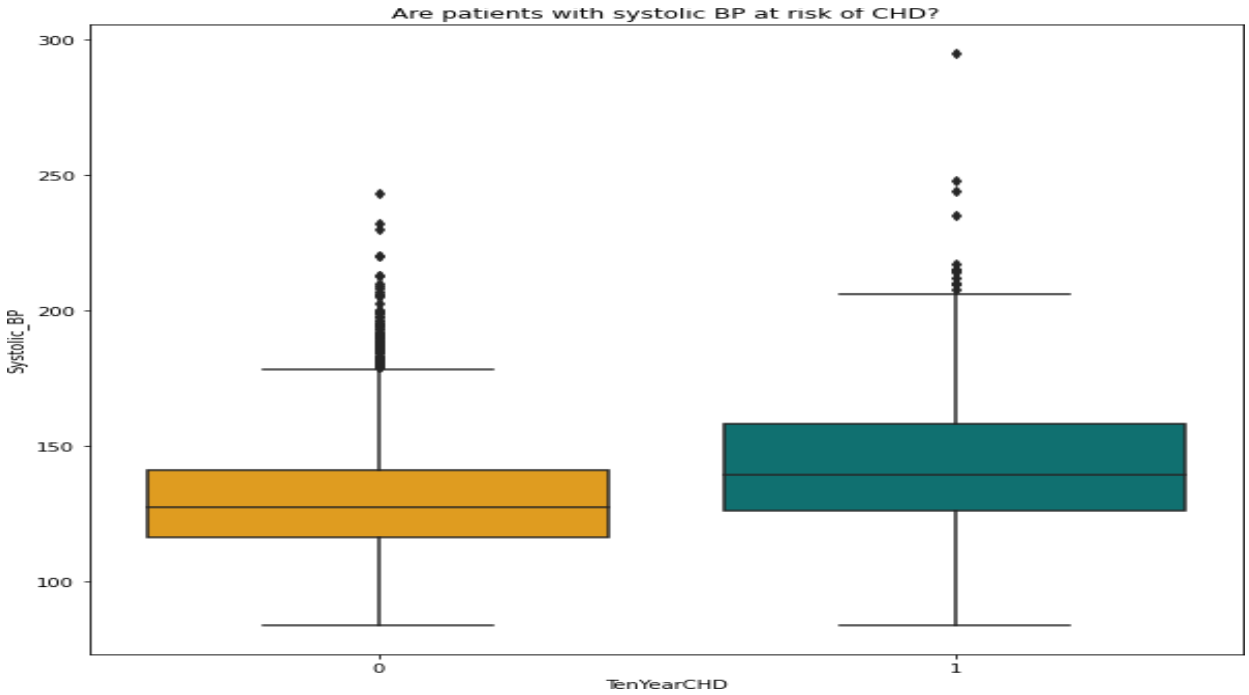
Can smoking number of cigarettes per day lead to CHD?



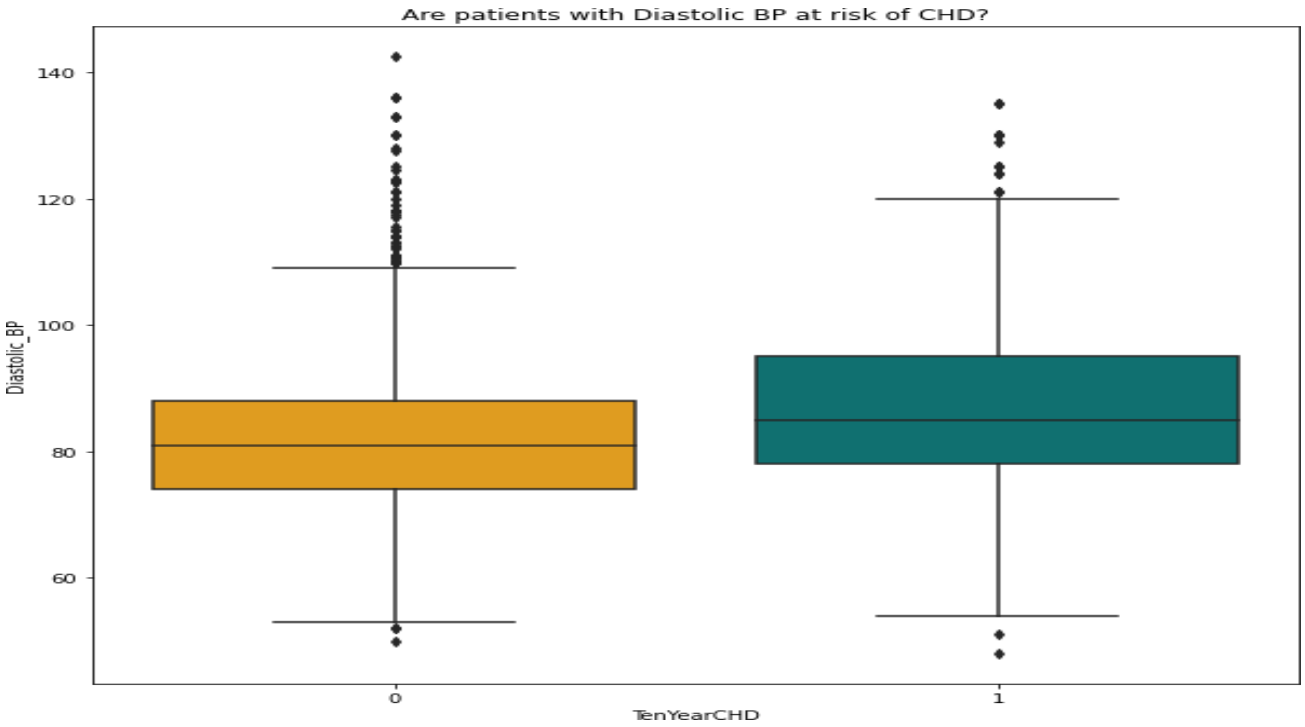
ONE WHO HAD A STROKE EARLIER MORE PRONE TO CHD ?



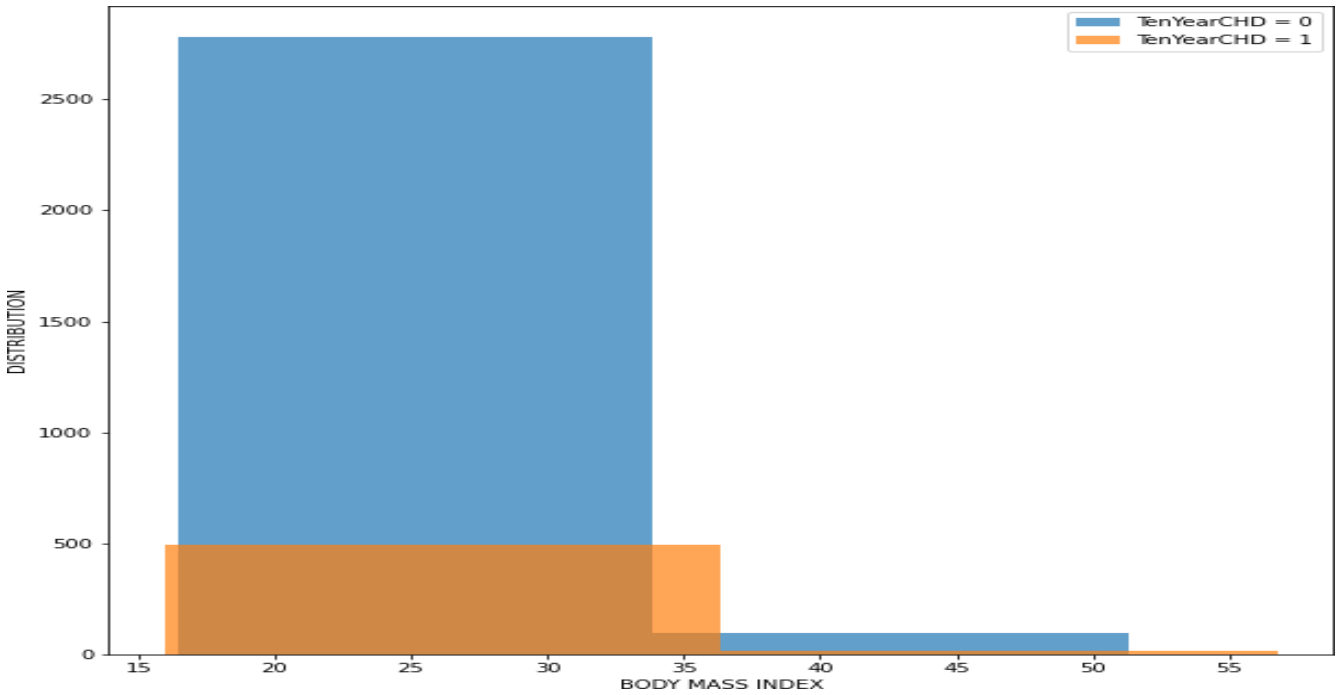
ARE PATIENTS WITH SYSTOLIC BP AT RISK OF CHD ?



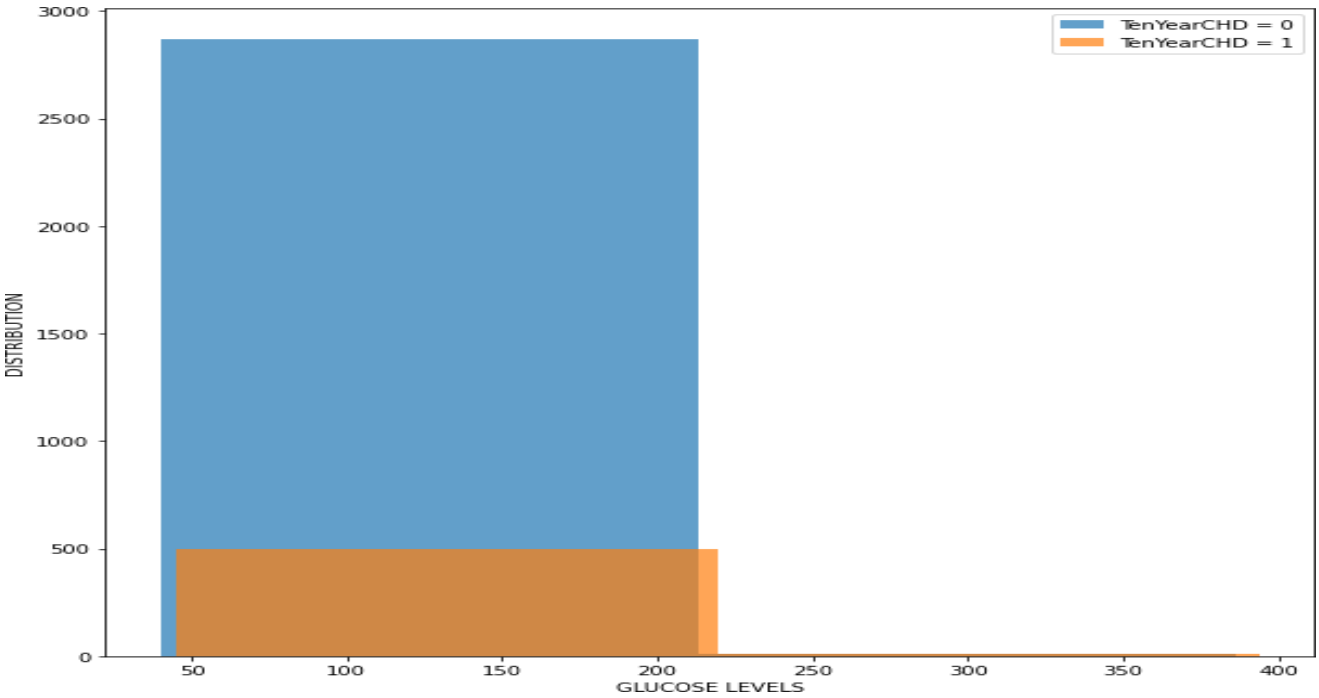
ARE PATIENTS WITH DIASTOLIC BP AT RISK OF CHD ?



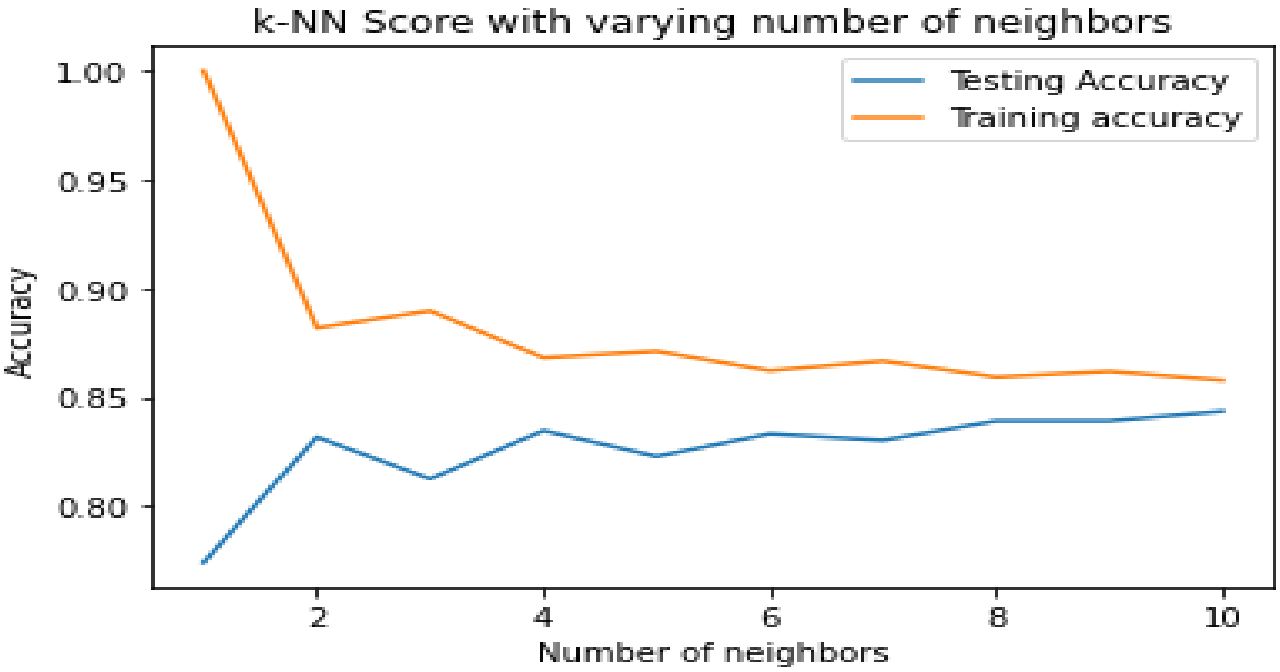
IS PATIENTS BMI IMPORTANT TO SHOW THE RISK OF CHD ?



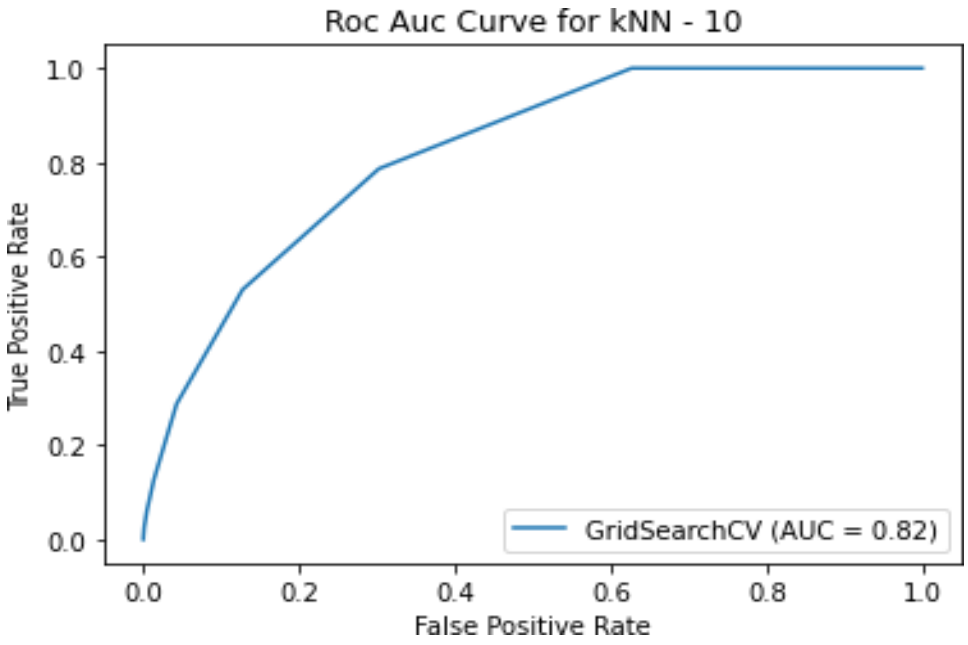
CAN PATIENTS GLUCOSE LEVELS SHOW THE RISK OF CHD ?



K-NN SCORE WITH VARYING NUMBER OF NEIGHBORS



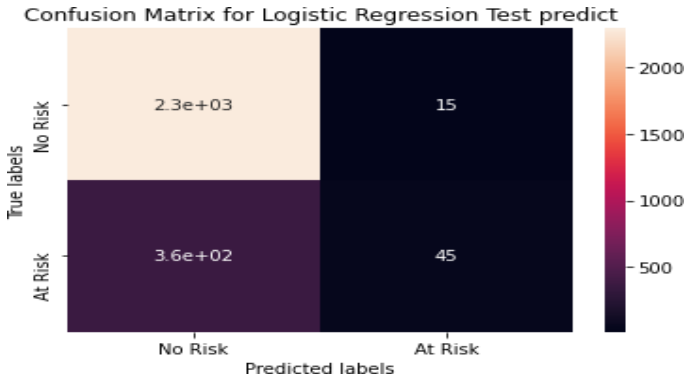
ROC AUC CURVE FOR KNN

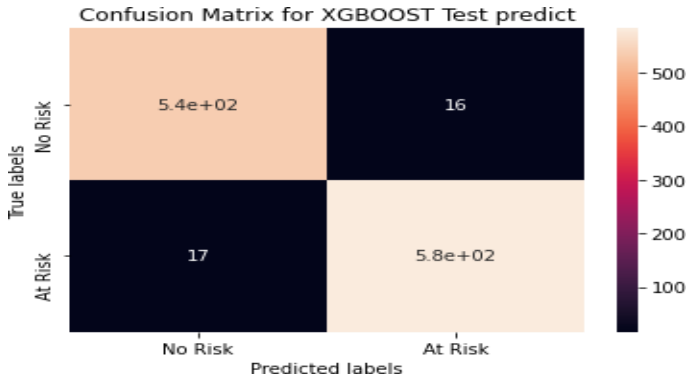


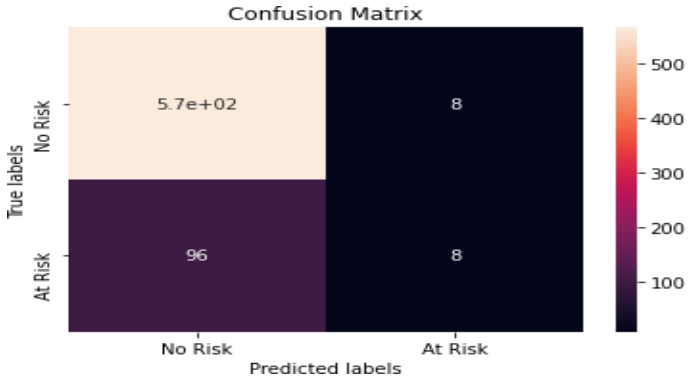
THE BEST FITTING MODEL:-

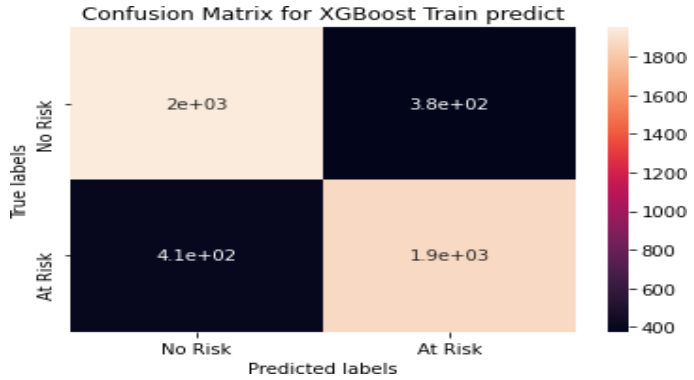
Sr.no	ML Model	Test Accuracy Score	Train Accuracy Score
1	Naive Bayes Classifier	81	83
2	KNN	84	86
3	Logistic Regression	84	86
4	Decision Tree	75	76
5	Random Forest	89	99.8
6	Gradient Boost	87	90
7	XGBoost	97	83

Confusion Matrix

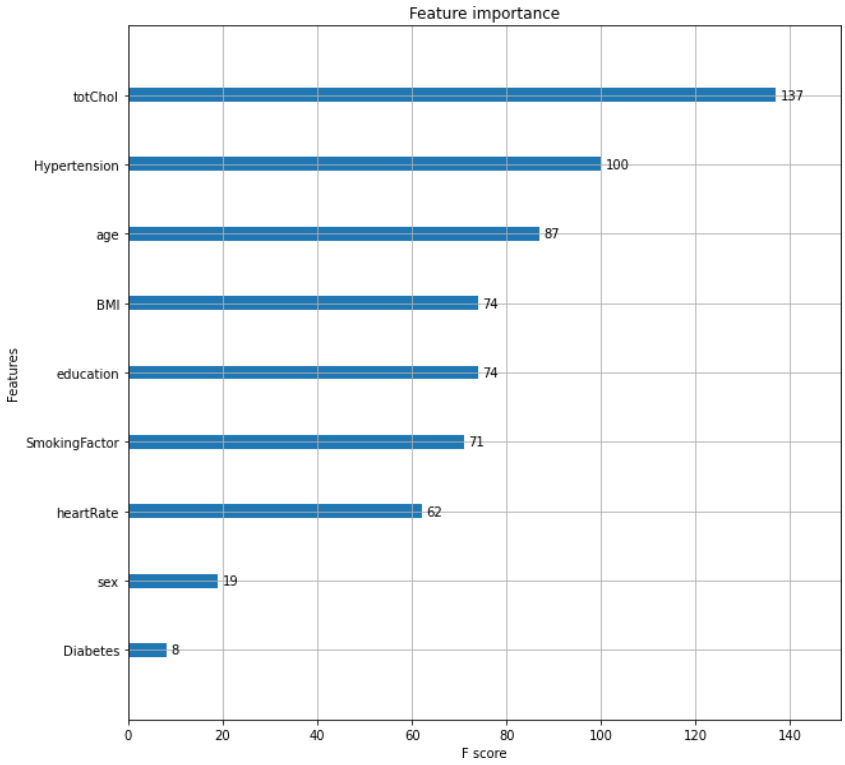








The Feature importance



Precision and Recall:-

- The precision is the proportion of relevant results in the list of all returned search results.
- The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned.
- In this project we are giving more importance to recall because predicting that the person doesn't have a disease when he have one can risk that persons life.

CHALLENGES:

- Less amount of data available made it difficult to predict properly.
- Missing relevant/Important features in our dataset like Chest pain location, chest pain type, Family history of coronary artery, Exercise, etc.
- The dataset was imbalanced and hence we were not able to apply some models properly.

Conclusion:

A cardiovascular disease detection model has been built using no of ML classification modelling techniques.
This project once deployed can possibly help predict the patients for cardiovascular disease based to their past medical history Blood pressure, Body mass index, Sugar levels etc.

The algorithms used in building the model are Logistic regression, Decision trees, KNN, Random forest classifier, Naive bayes classifier, Gradient boost and XGboost.
The top three models with best accuracy are C
forest & XGboost with accuracy of 87%, 89%,

CONCLUSION:

And to conclude we started with loading the data. So far we have done EDA, null values treatment, encoding of categorical columns, feature selection and then model building.

In all of these models our accuracy revolves in the range of 75 to 97%. And there is no such improvement in accuracy score even after hyperparameter tuning.

Also it is concluded that accuracy of XGboost is highest as compared to all the algorithms used i.e. 97%.

This performance could be due to various reasons like: No proper pattern of data, lack of data, not enough rele

with enough data we can train our model eve