CAPSTONE PROJECT

CARDIO VASCULAR RISK PREDICTION

PRESENTED BY AMALKRISHNA N

CONTENTS:

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 3390 records and 15 attributes.
Variables

Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

DATA SUMMARY

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)

Education: The level of education of the patient (categorical values - 1,2,3,4)

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 hav 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):

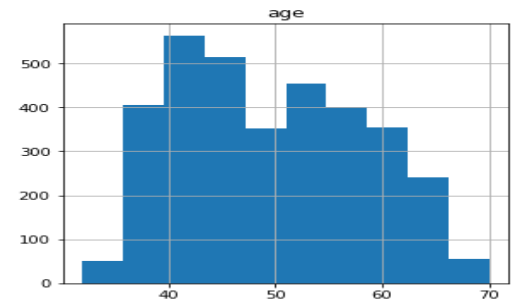
10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”)

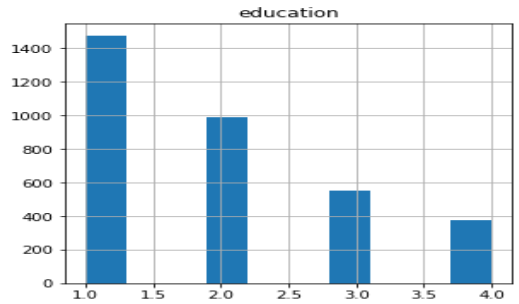
DATA CLEANING AND PROCESSING

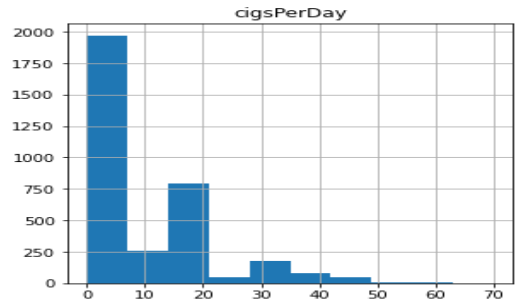
total of 7 columns are there with null values

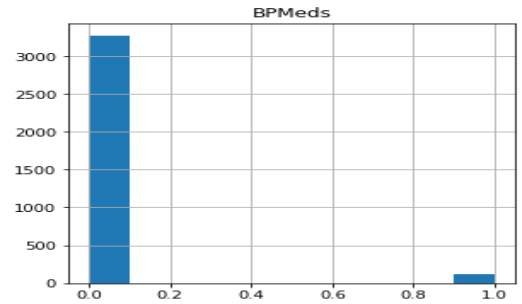
- In case of ‘cigarettes per day’ we found that the null values not conveys person is not a smoker because of that null values are filled by average value
- Null values of categorical variable column ‘education’ are filled by the mode value of the column
- Null values of continuous variable columns such as ‘total cholesterol’, ‘BMI’, ‘glucose level’ are filled with their own average values
- Null values of heart rate is filled with median value
- filling null values of BPMed with 1 if systolic bp > 140 else 0.

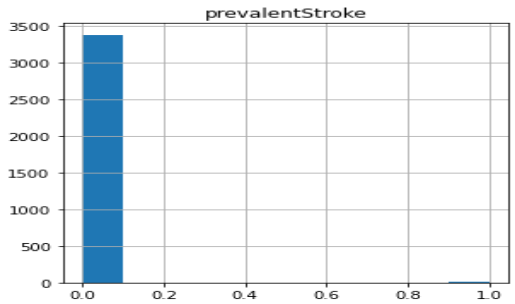
EXPLORATORY DATA ANALYSIS

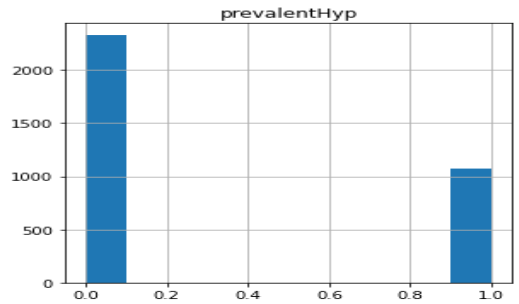

age

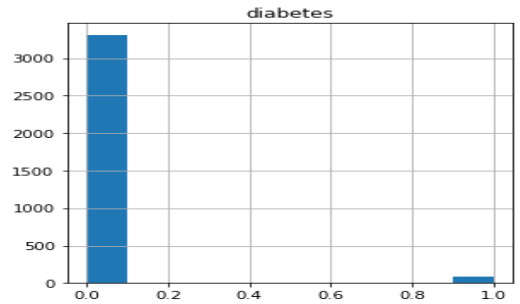

education

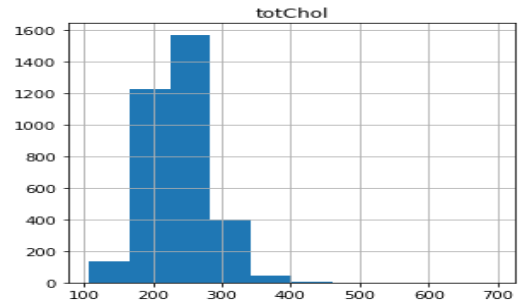

cigsPerDay

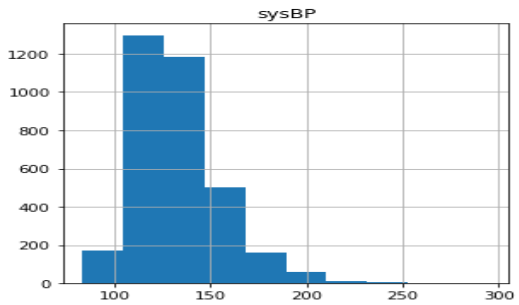

BPMeds

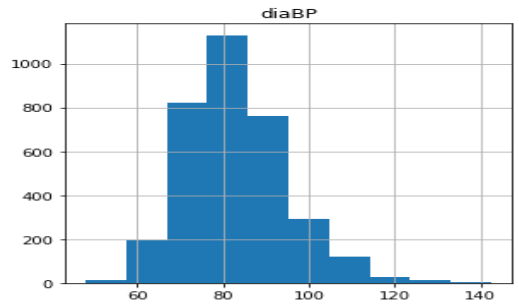

prevalentStroke

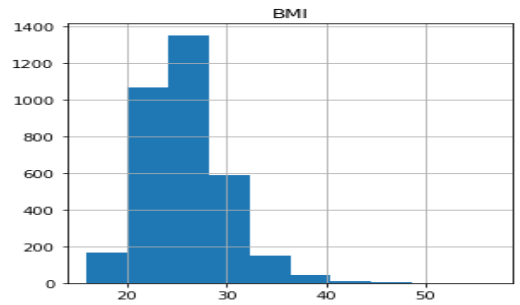

prevalentHyp

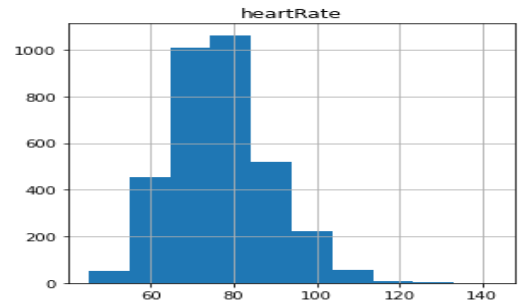

diabetes

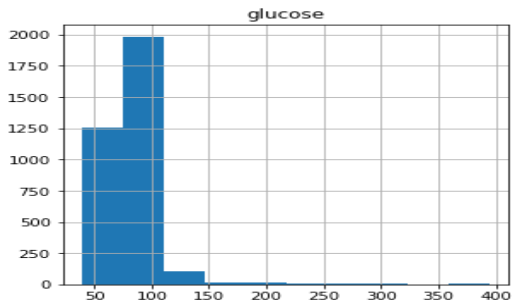

totChol

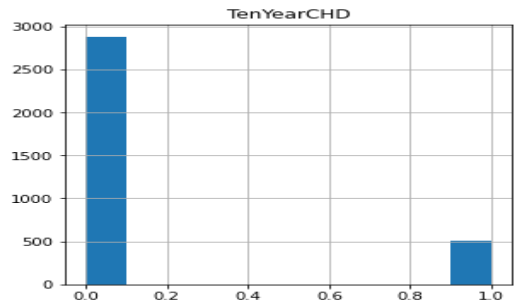

sysBP


diaBP


BMI

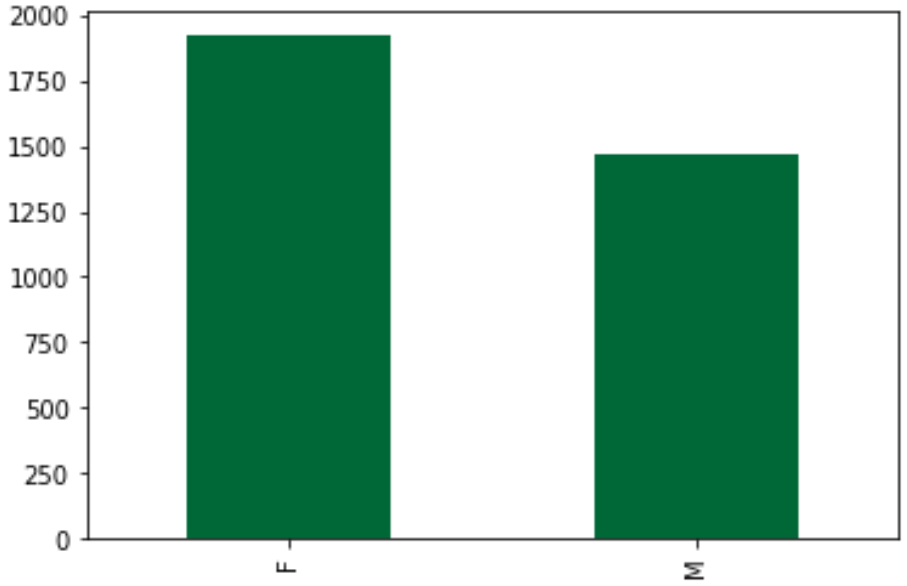

heartRate


glucose


TenYearCHD

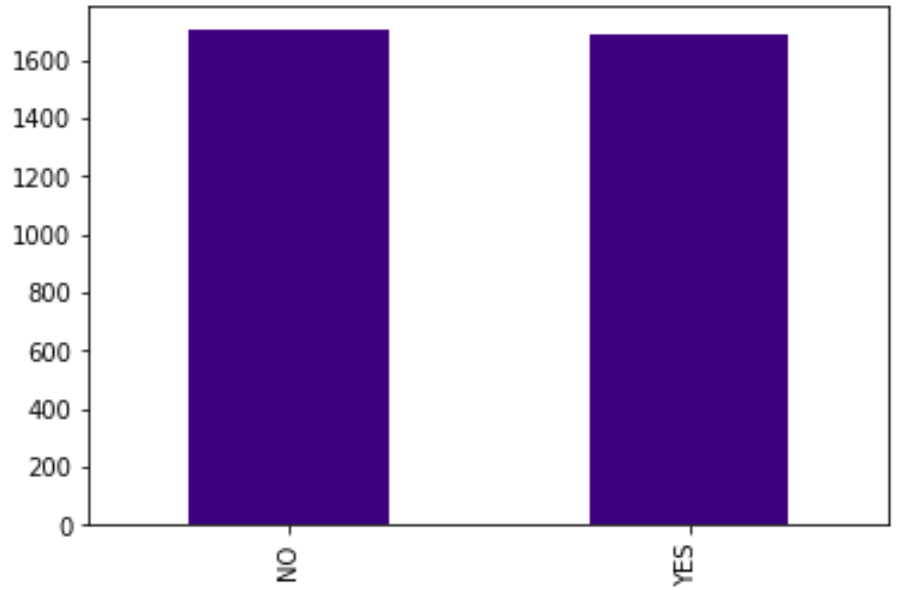
- age column is almost showing a good distribution near to normal distribution
- more peoples are belongs to education class 1 then it uniformly decreasing towards class 4
- In the overall data set average cigarettes per day is in 1-10 range.(Includes smokers and non-smokers)
- most of the peoples are not medicating for blood pressure issue. BP medicating peoples are under 200 in numbers
- The patient with previously suffered with a stroke is very less in number numerically it is 22
- Around 1000 patients are previously hypertensive
- Diabetic patients are very less in number somewhere around 100
- The total cholesterol level of patients is distributed well with a little skew. cholesterol level of almost 1600 peoples are in the range of 225-275
- most of the patient's systolic BP in a range of 100-150 and diastolic BP in a range of 75-85
- Body Mass Index is highly concentrated around 25 range
- most of the patient's heart rate are around 75-80
- glucose level highly concentrated around 100
- Ten year CHD risk peoples lesser in number and it shows a class imbalance of dependent feature

Gender distribution

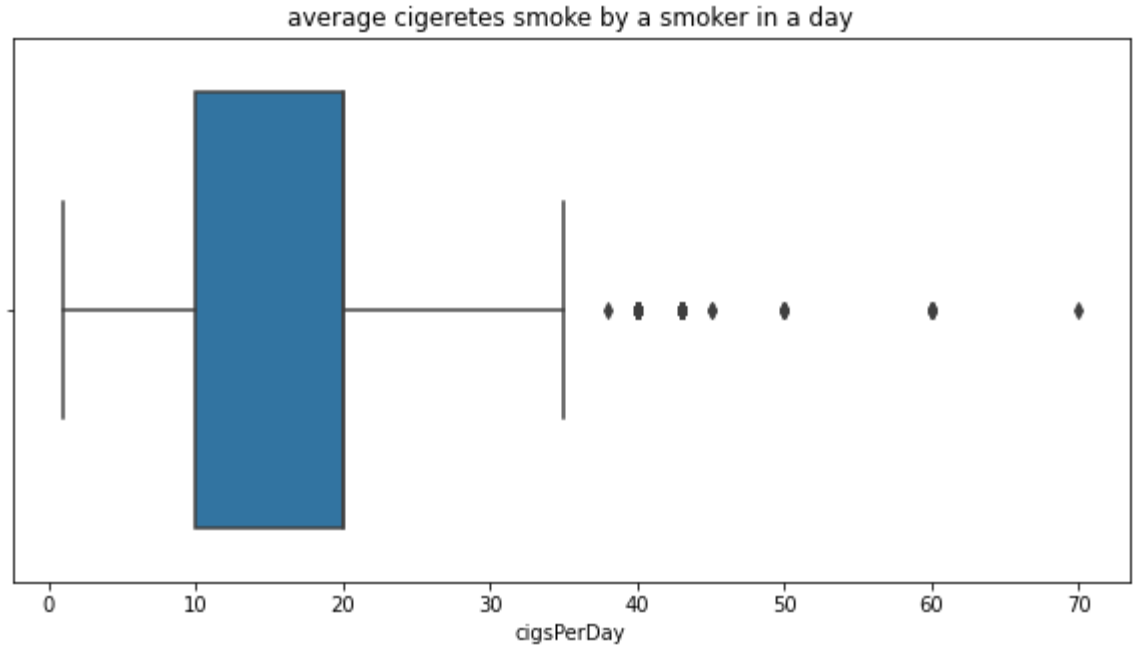


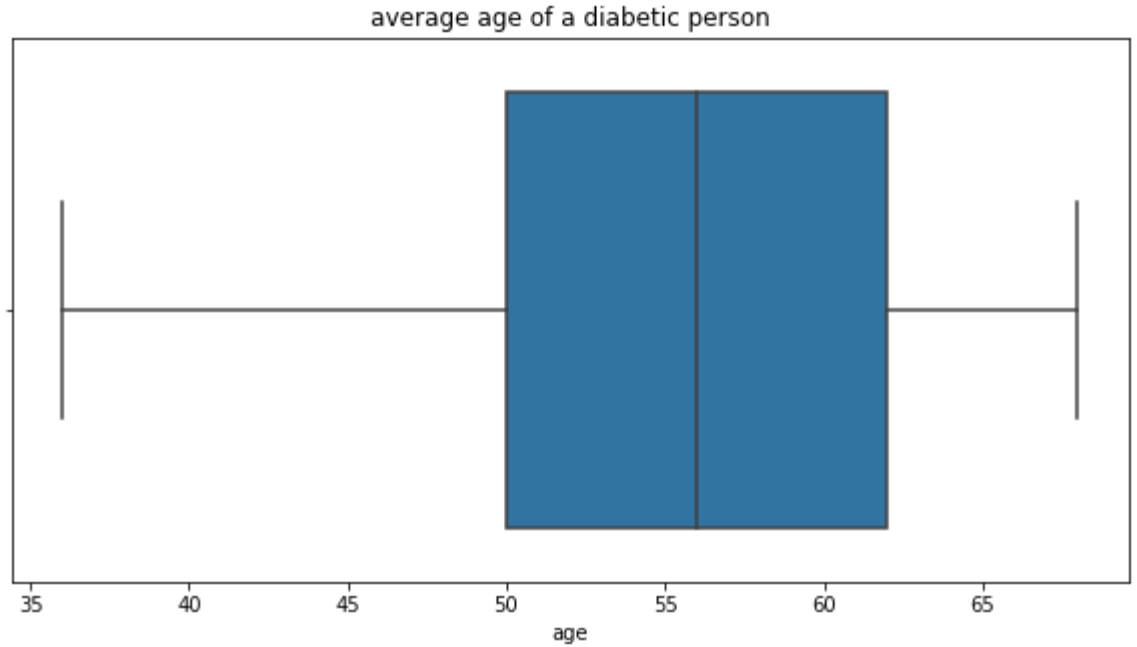
- female peoples are more in number

smokers and non smokers



- smokers and non-smokers are almost the same in number.
- non-smokers are just 16 more than smokers





- when we are taking the average exclusively of smokers(excluding non-smokers data), it will be the more accurate average of cigarettes per day

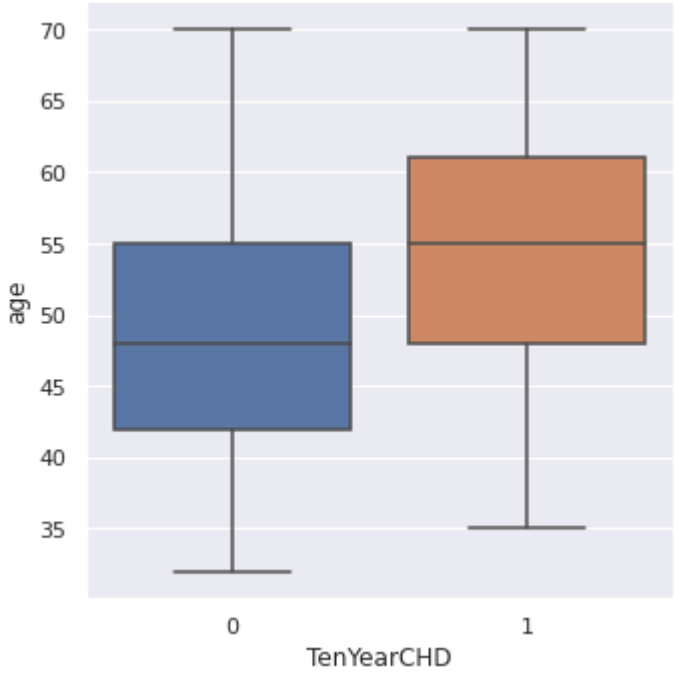
- it shows that a smoker will smoke 10-20 cigarettes per day on an average

- Plotted box plot with age and diabetic data

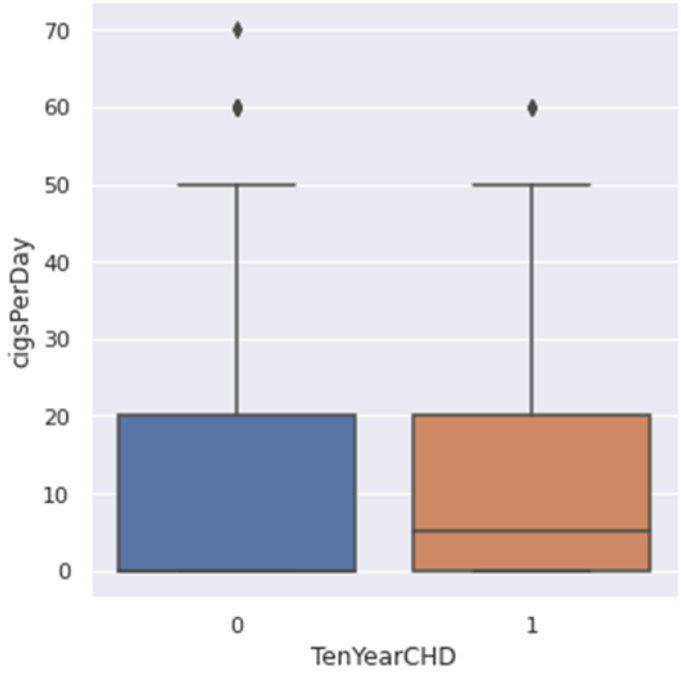
- age of diabetic person mostly come under range of 50-62

- average age of a diabetic person is 56

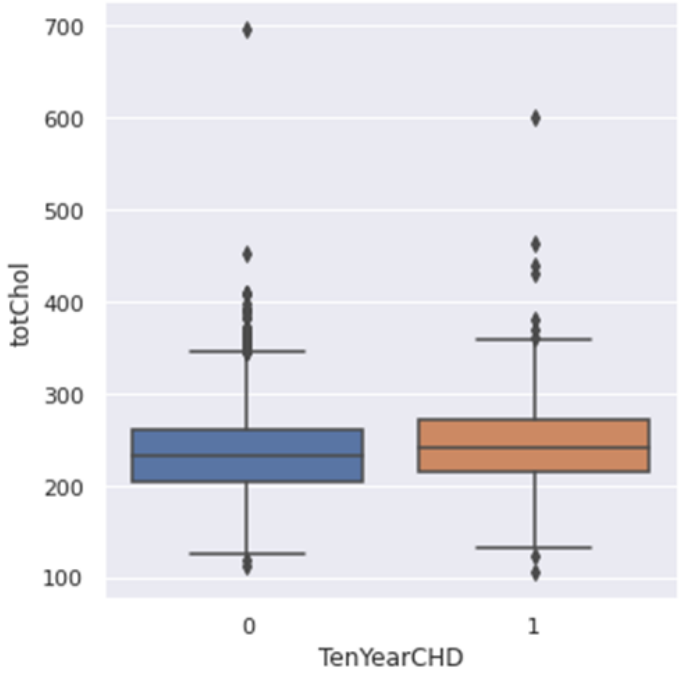
Distribution of age over target variable



Cigarettes Per Day distribution of patients under risk and non-risk



Total Cholesterol level distribution



• Age of patients without risk of CHD for 10 years is more in a range of 42-55

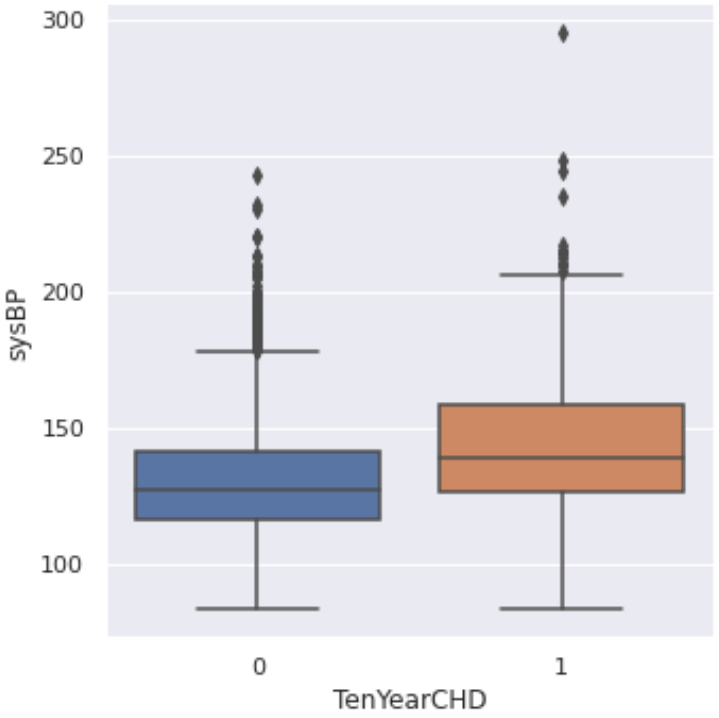
•age of patients under risk is from 48 to 60

•average cigarettes per day of all kind of patients under 10 year CHD risk and not distributed more in 0-20 range

•but in case of class under risk average of distribution come to 5

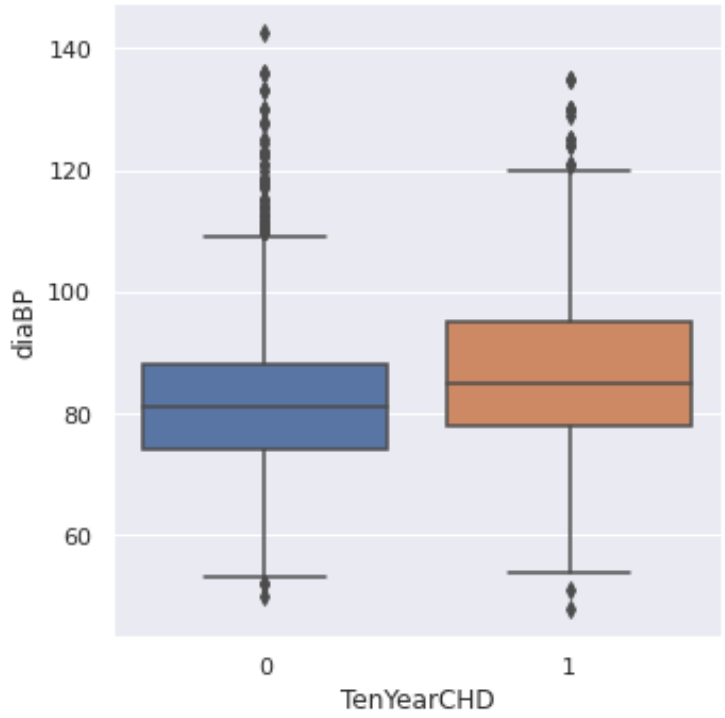
cholesterol level of patients under risk is higher

Systolic blood pressure



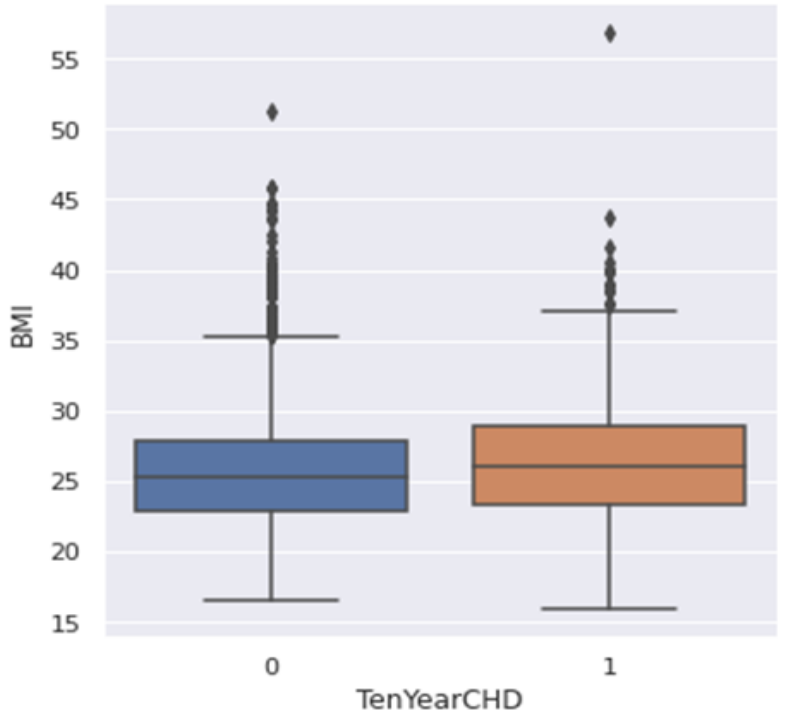
- systolic BP of patients under risk is higher

Diastolic blood pressure



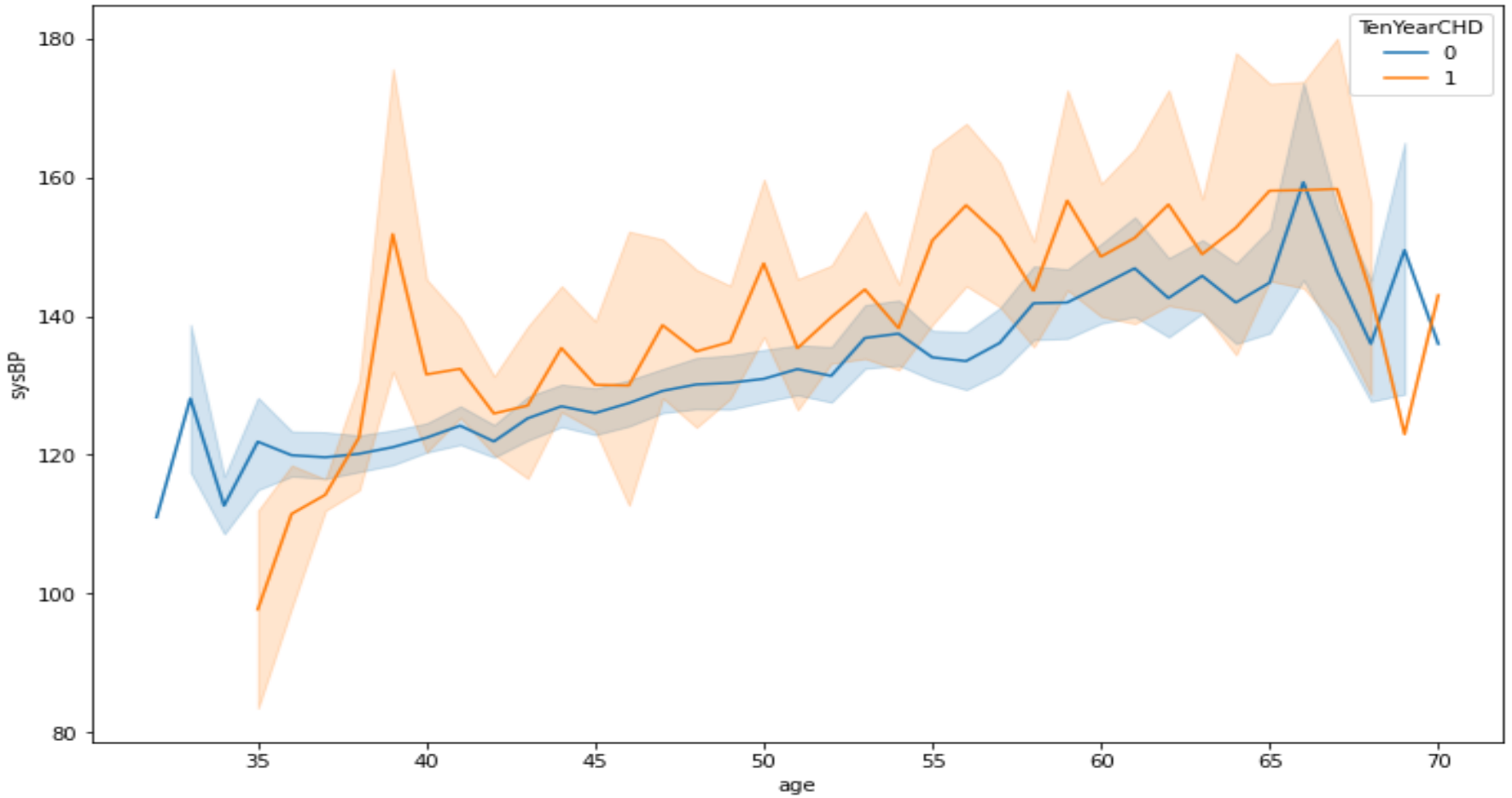
- Diastolic BP of patients under risk is higher

Body Mass Index



- BMI distribution patients under risk is slightly higher
- shows chances of obesity or overweight of risk patients

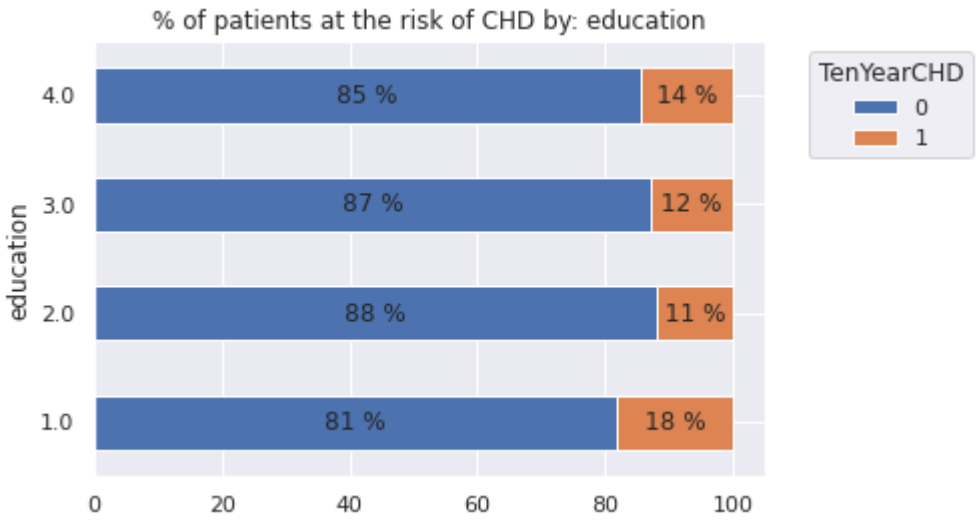
systolic BP variation over age for both target class

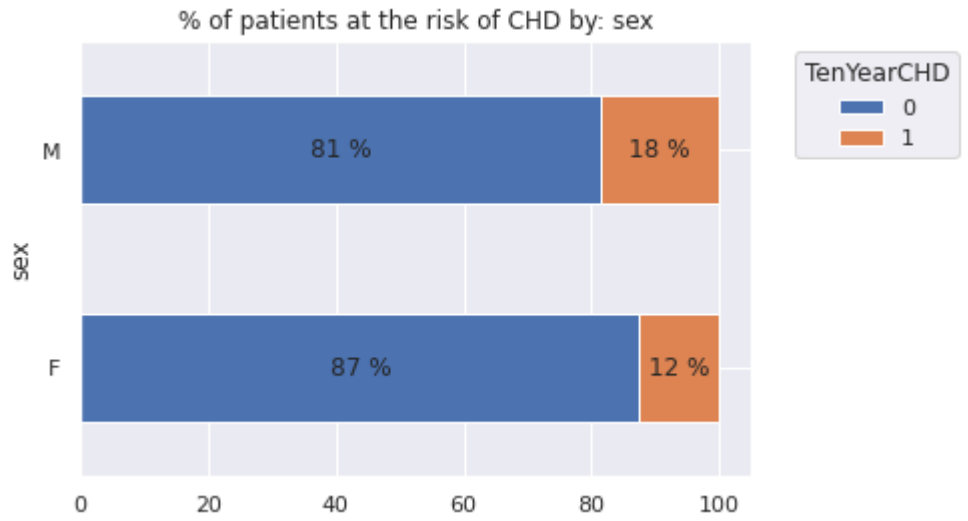


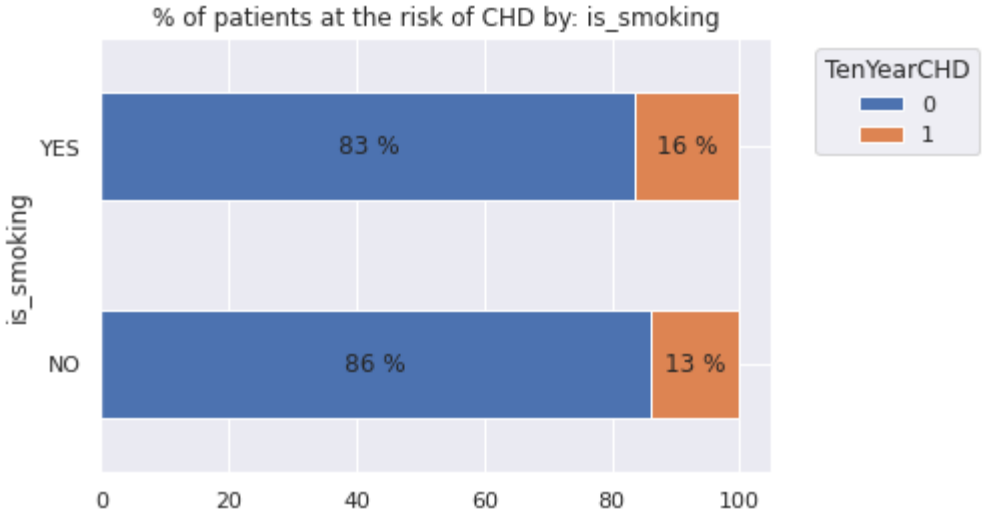
- The systolic BP of patients under both target variable class is increasing over age

- As per blood pressure there will be a same trend in diastolic BP

stacked bar chart of percentage of patients under risk or not over all categorical features







<u>% of patients at the risk of CHD by: education</u>

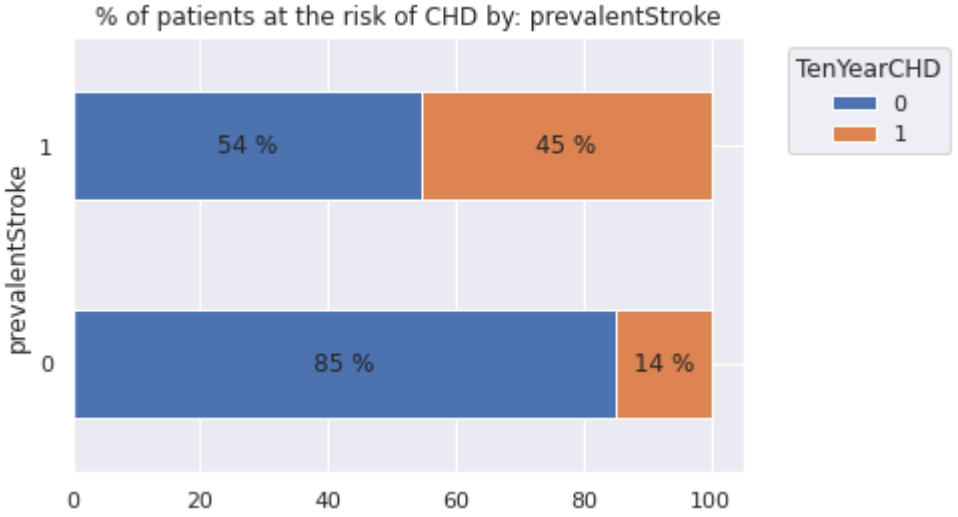
- in case of educational categories risk percentage of CHD is almost equal
- In level 1 category 18% of CHD risk

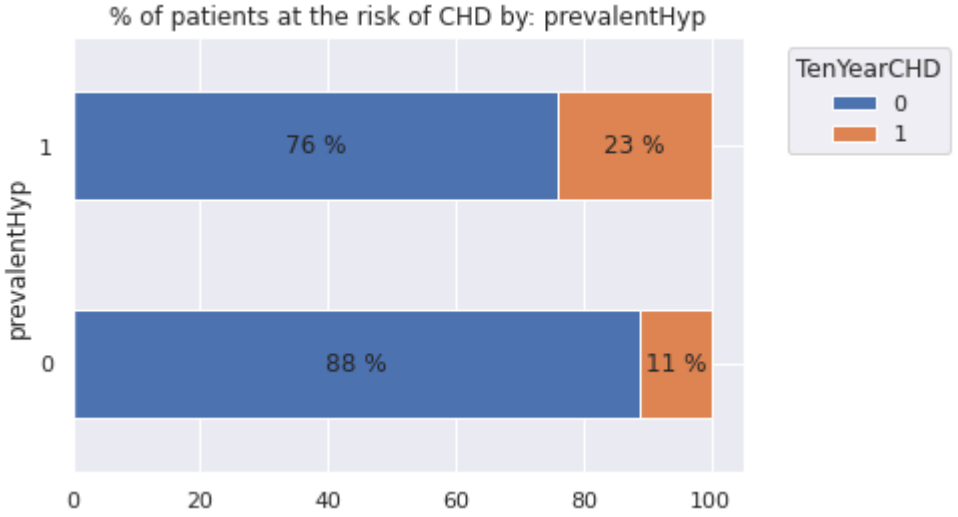
<u>% of patients at the risk of CHD by: sex</u>

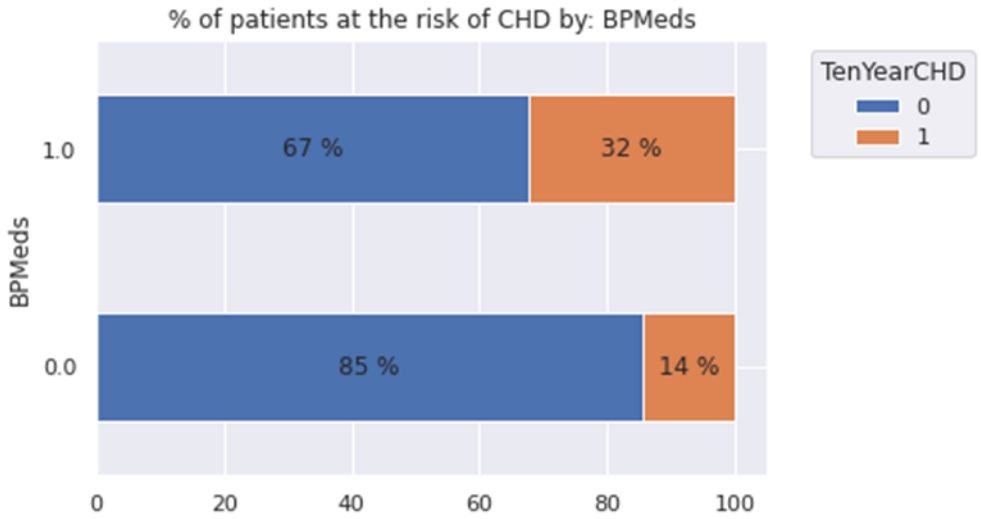
- 18% of male category peoples are under risk of CHD
- Only 12% of female peoples are under risk

<u>% of patients at the risk of CHD by: whether smoking or not</u>

- 16% of smoking peoples are come under risk of CHD
- Only 13% of non-smoking peoples are come under risk







<u>% of patients at the risk of CHD by: prevalent stroke</u>

- In case of prevalently stroke happened patients, 45% are under risk
- As compared to patients without prevalent stroke it is much higher

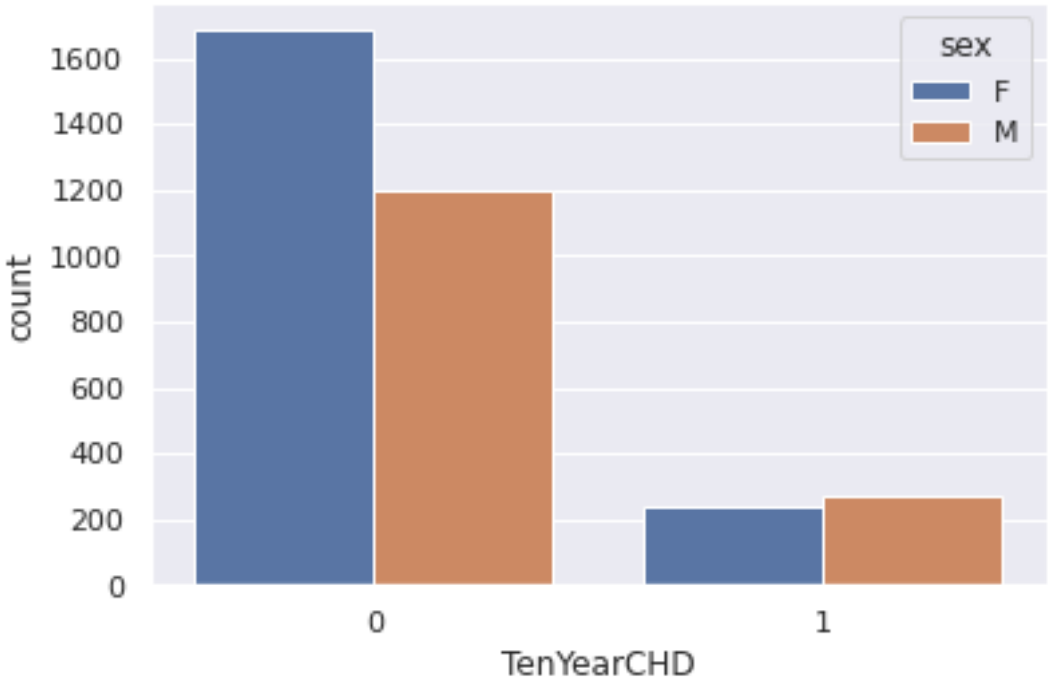
<u>% of patients at the risk of CHD by: prevalent hyper tension</u>

- Patients with prevalent hyper tension 23% are come under the risk of CHD

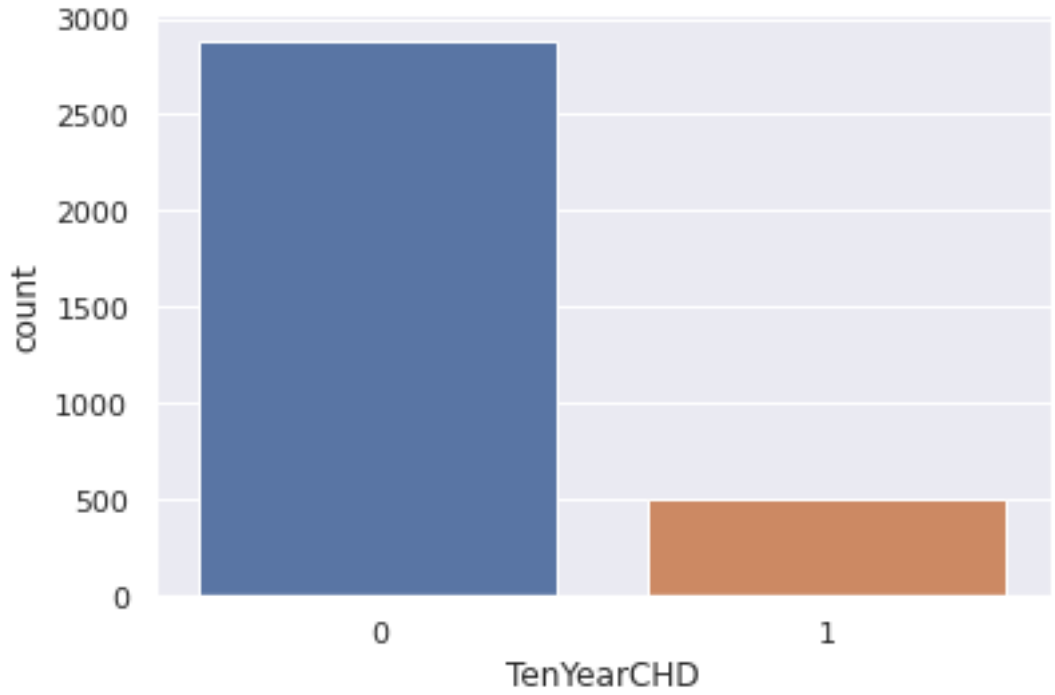
<u>% of patients at the risk of CHD by: whether medicating for BP or not</u>

- Patients who where medicating for BP 32% of them are come under CHD risk

Distribution of gender over both target class

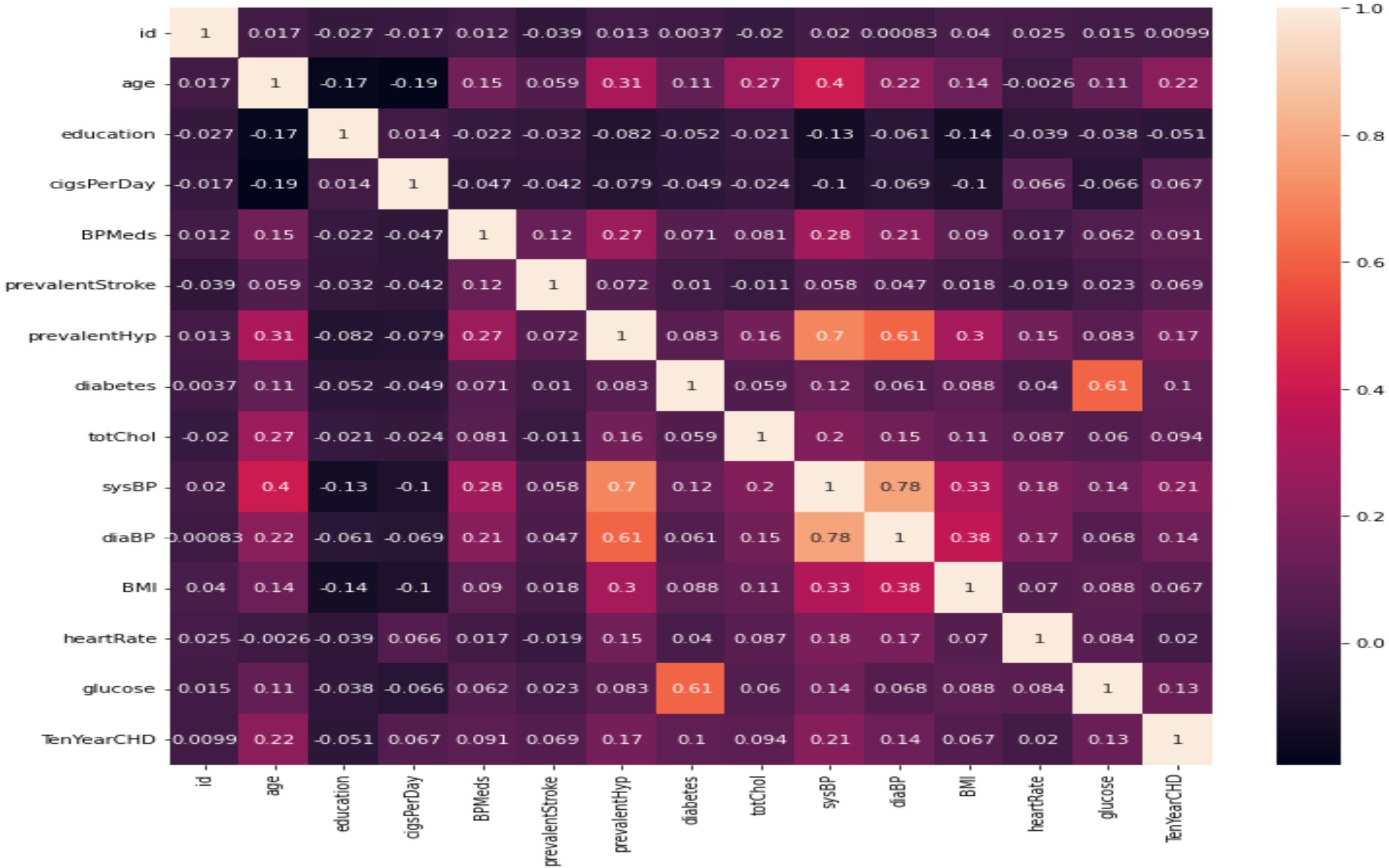


Target variable



- under non-risk category females are more
- under risk category males are more

- showing class imbalance in target variable


id 1 0.017 -0.027 -0.017 0.012 -0.039 0.013 0.0037 -0.02 0.02 0.00083 0.04 0.025 0.015 0.0099
age 0.017 1 -0.17 -0.19 0.15 0.059 0.31 0.11 0.27 0.4 0.22 0.14 -0.0026 0.11 0.22
education -0.027 -0.17 1 0.014 -0.022 -0.032 -0.082 -0.052 -0.021 -0.13 -0.061 -0.14 -0.039 -0.038 -0.051
cigsPerDay -0.017 -0.19 0.014 1 -0.047 -0.042 -0.079 -0.049 -0.024 -0.1 -0.069 -0.1 0.066 -0.066 0.067
BPMeds 0.012 0.15 -0.022 -0.047 1 0.12 0.27 0.071 0.081 0.28 0.21 0.09 0.017 0.062 0.091
prevalentStroke -0.039 0.059 -0.032 -0.042 0.12 1 0.072 0.01 -0.011 0.058 0.047 0.018 -0.019 0.023 0.069
prevalentHyp 0.013 0.31 -0.082 -0.079 0.27 0.072 1 0.083 0.16 0.7 0.61 0.3 0.15 0.083 0.17
diabetes 0.0037 0.11 -0.052 -0.049 0.071 0.01 0.083 1 0.059 0.12 0.061 0.088 0.04 0.61 0.1
totChol -0.02 0.27 -0.021 -0.024 0.081 -0.011 0.16 0.059 1 0.2 0.15 0.11 0.087 0.06 0.094
sysBP 0.02 0.4 -0.13 -0.1 0.28 0.058 0.7 0.12 0.2 1 0.78 0.33 0.18 0.14 0.21
diaBP 0.00083 0.22 -0.061 -0.069 0.21 0.047 0.61 0.061 0.15 0.78 1 0.38 0.17 0.068 0.14
BMI 0.04 0.14 -0.14 -0.1 0.09 0.018 0.3 0.088 0.11 0.33 0.38 1 0.07 0.088 0.067
heartRate 0.025 -0.0026 -0.039 0.066 0.017 -0.019 0.15 0.04 0.087 0.18 0.17 0.07 1 0.084 0.02
glucose 0.015 0.11 -0.038 -0.066 0.062 0.023 0.083 0.61 0.06 0.14 0.068 0.088 0.084 1 0.13
TenYearCHD 0.0099 0.22 -0.051 0.067 0.091 0.069 0.17 0.1 0.094 0.21 0.14 0.067 0.02 0.13 1
id
age
education
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
diaBP
BMI
heartRate
glucose
TenYearCHD
1.0
0.8
0.6
0.4
0.2
0.0

FEATURE ENGINEERING FOR MODELING

LABEL ENCODING

In data set ‘sex’ and ‘is_smoking’ columns are in a labeled form with only two variable, it converted to numeric form

ONE HOT ENCODING

Only ‘education’ column have to be changed by one hot encoding in the data set.

SMOTE

In the dataset the target variable has 2 classes which is considerably in an imbalanced condition with very less number of datapoints with positive class. It will affect prediction of the model when we train the model with a highly imbalanced target class. By SMOTE in the dataset for training it was resolved.

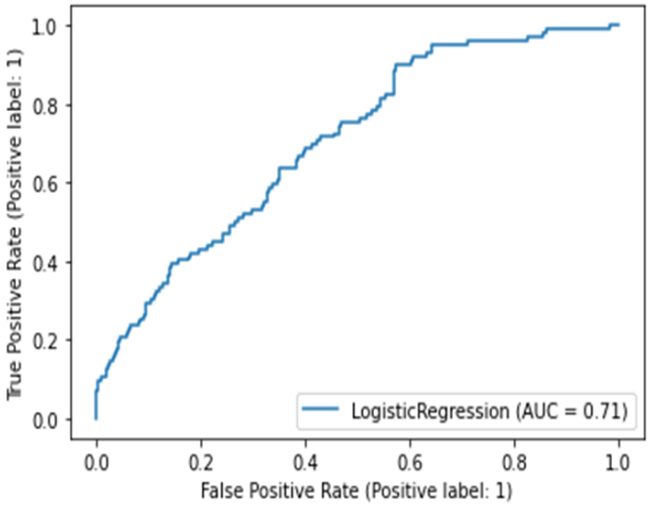
SCALING

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. Generally after scaling all independent features comes to a 0 - 1 range. It was performed as data pre-processing, to handle highly varying magnitudes. Here we used it only for KNN algorithm

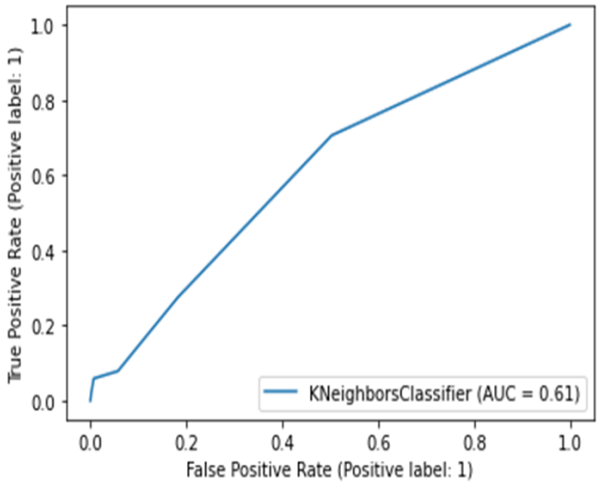
MODELING

- LOGISTIC REGRESSION
- KNN CLASSIFIER
- XGBOOST CLASSIFIER
- RANDOM FOREST CLASSIFIER

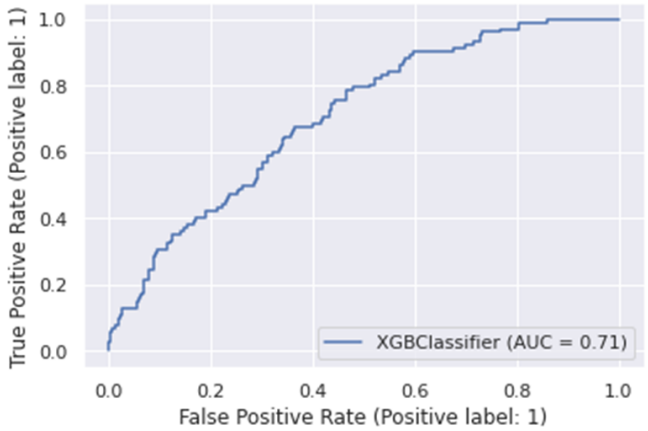
LOGISTIC REGRESSION



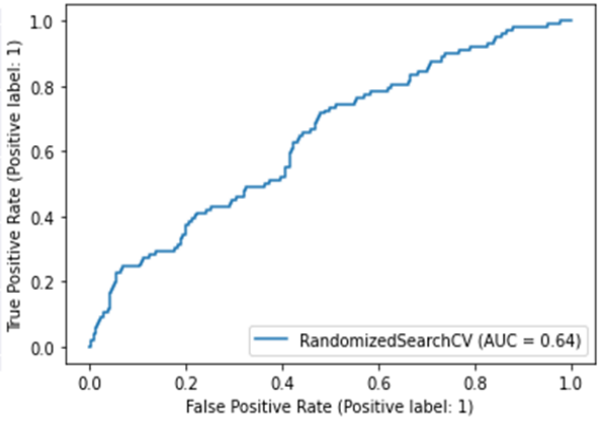
KNN



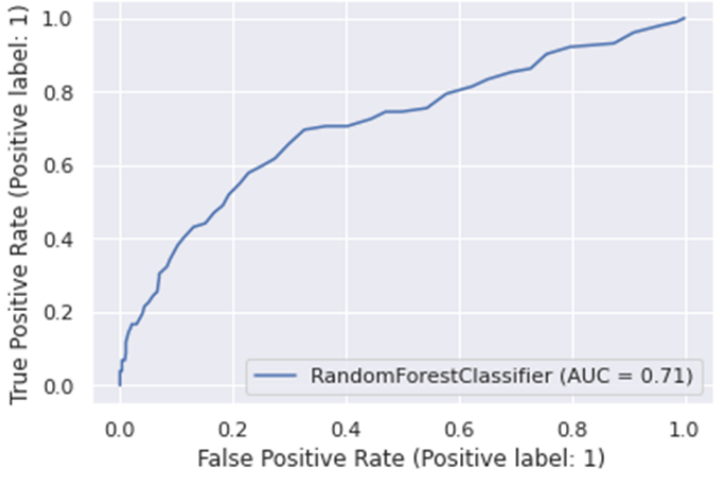
XGB CLASSIFIER



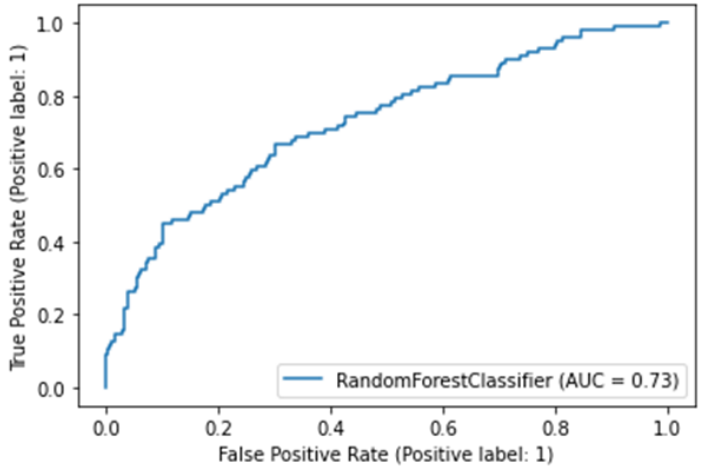
XGB WITH RANDOMIZED SEARCH CV



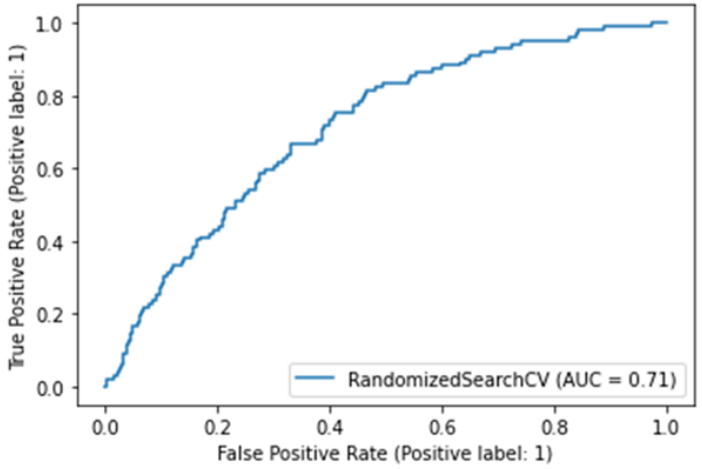
RANDOMFOREST CLASSIFIER



RANDOM FOREST WITH GRID SEARCH CV



RANDOM FOREST WITH RANDOMIZED SEARCH CV



LOGISTIC REGRESSION

	precision	recall	f1-score	support
0	0.87	0.91	0.89	576
1	0.35	0.25	0.29	102
accuracy			0.82	678
macro avg	0.61	0.58	0.59	678
weighted avg	0.79	0.82	0.80	678

KNN

	precision	recall	f1-score	support
0	0.88	0.90	0.89	576
1	0.34	0.27	0.30	102
accuracy			0.81	678
macro avg	0.61	0.59	0.60	678
weighted avg	0.79	0.81	0.80	678

XGB CLASSIFIER

	precision	recall	f1-score	support
0	0.86	0.95	0.90	576
1	0.31	0.12	0.17	102
accuracy			0.83	678
macro avg	0.58	0.54	0.54	678
weighted avg	0.78	0.83	0.79	678

XGB WITH RANDOMIZED SEARCH CV

	precision	recall	f1-score	support
0	0.87	0.94	0.91	576
1	0.41	0.23	0.29	102
accuracy			0.83	678
macro avg	0.64	0.58	0.60	678
weighted avg	0.80	0.83	0.81	678

RANDOMFOREST CLASSIFIER

	precision	recall	f1-score	support
0	0.86	0.99	0.92	576
1	0.70	0.07	0.13	102
accuracy			0.86	678
macro avg	0.78	0.53	0.52	678
weighted avg	0.83	0.86	0.80	678

RANDOM FOREST WITH GRID SEARCH CV

	precision	recall	f1-score	support
0	0.87	0.90	0.89	576
1	0.32	0.26	0.29	102
accuracy			0.81	678
macro avg	0.60	0.58	0.59	678
weighted avg	0.79	0.81	0.80	678

RANDOM FOREST WITH RANDOMIZED SEARCH CV

	precision	recall	f1-score	support
0	0.88	0.90	0.89	576
1	0.33	0.28	0.31	102
accuracy			0.81	678
macro avg	0.60	0.59	0.60	678
weighted avg	0.79	0.81	0.80	678

CONCLUSION

Mainly used 4 types of different classification algorithms:

Logistic regression, k-nearest neighbors' classification, XG booster and Random Forest classification. After all, totally verified 7 classification model predictions.

Evaluated each model output with more than five evaluation parameters.

Out of all evaluations Random Forest with randomized search cv gave maximum good results for this classification prediction problem with better predictions.

- Here we got a test score accuracy of 81%
- area under the Receiver operating characteristic curve (AUC-ROC) of the model is 71%.
- f1 score for class 1 is 31% and class 0 is 89%
- precision of class 0 prediction is 88% and class 1 prediction is 33%
- recall of class 0 is 90% and class 1 is 28%

Conclusions from EDA

- An average smoker smokes 10 to 20 cigarettes per day. It is a very danger condition, moreover CHD it is very near to lung cancer risk
- Average age of a CHD risk person is 55age
- Systolic and diastolic blood pressures have a very significant role in CHD risk of patients. With higher blood pressure peoples are more come under risk category
- Blood pressure of all peoples increasing with respect to age
- BMI of risk patients are also at a slightly higher level. We can conclude from that the obesity condition peoples make them under risk category
- Male peoples are more under risk of CHD (18% of male peoples under risk but females are 12%)
- Smoking also induces risk of CHD (16% of smoking peoples are under risk)
- Prevalent stroke causes high risk of CHD (45% of peoples with prevalent stroke were under risk)
- Peoples with prevalent hypertension and medicating peoples for BP are more prone to risk of CHD in upcoming 10 years
- Age, systolic BP, diastolic BP, diabetes is much positively correlated to ten-year risk of CHD

THANK YOU