Capstone Project

Cardiovascular Risk Prediction

By: Vikash Kumar

Problem Description

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.

Steps performed

1. **Data cleaning**
2. **Data visualizations**
3. **Data preprocessing**
4. **Model Implementation**
5. **Evaluation metrics**

Data Description

The dataset includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

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, are considered continuous because of the large number of possible values.)
- **Glucose:** glucose level (Continuous) Predict variable (desired target)
- **TenYearCHD:** 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) - DV

Data Cleaning

Prior to EDA, cleaning the data is essential since it will get rid of any ambiguous information that can have an impact on the results.

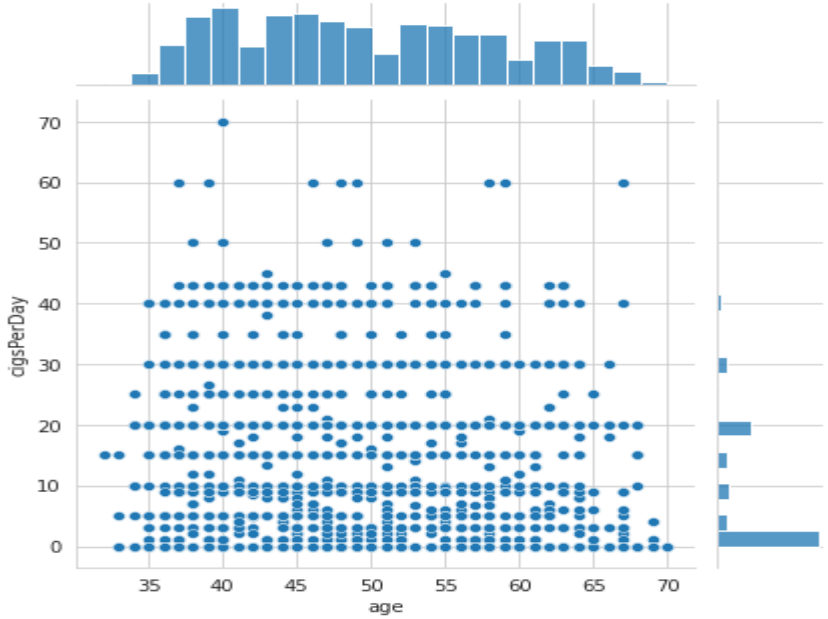
Education, cigsPerDay, BPMeds , totChol , BMI , heartRate and glucose columns have missing or null values. I have filled these columns for missing values by using KNN imputer.

For columns sex and is_smoking, I have changed the categorical value with a numerical value(like yes to 1 and no to 0, and F to 1 and M to 0).

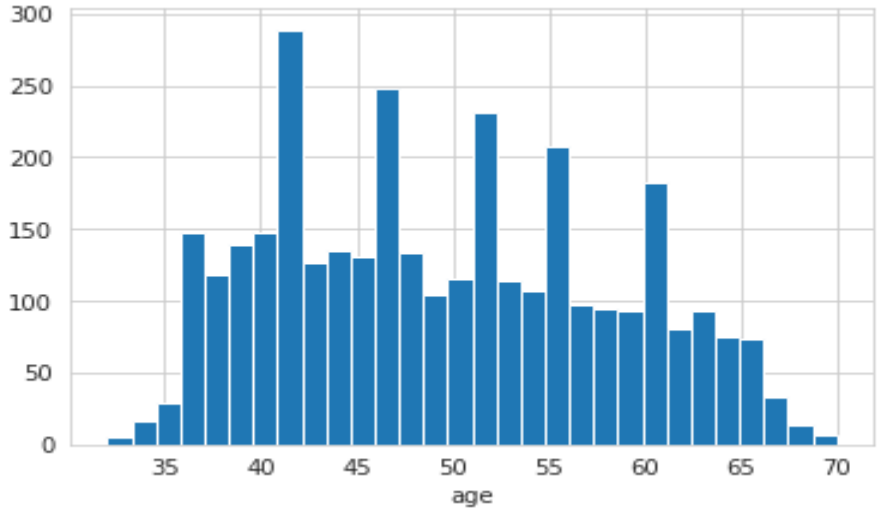
data.isna().sum()

age	0
education	87
sex	0
is_smoking	0
cigsPerDay	22
BPMeds	44
prevalentStroke	0
prevalentHyp	0
diabetes	0
totChol	38
sysBP	0
diaBP	0
BMI	14
heartRate	1
glucose	304
TenYearCHD	0

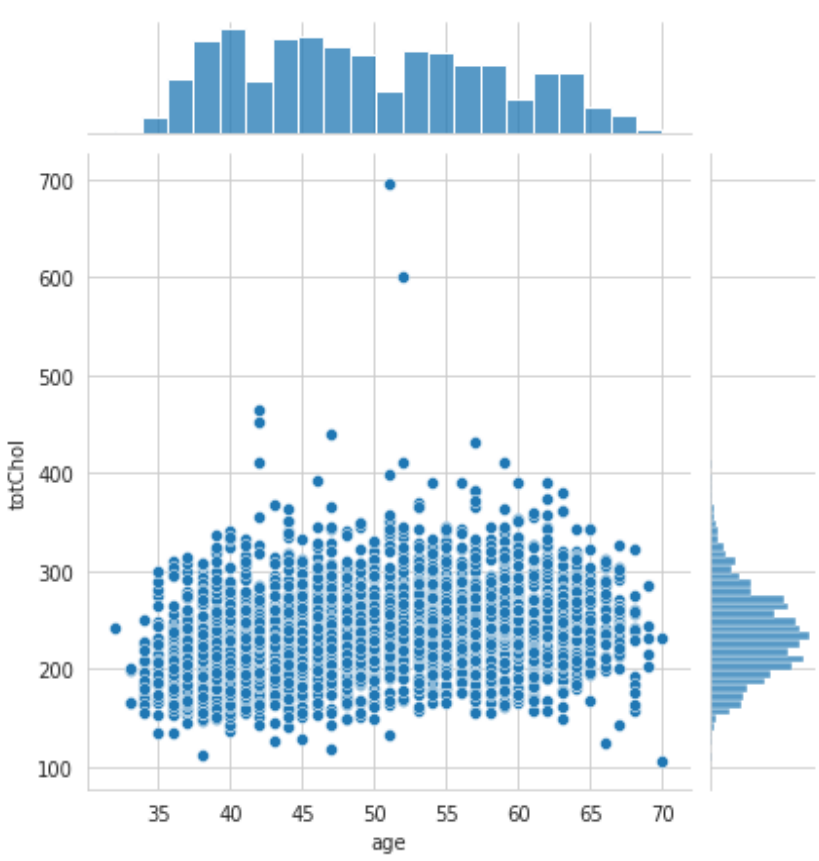
Exploratory Data Analysis

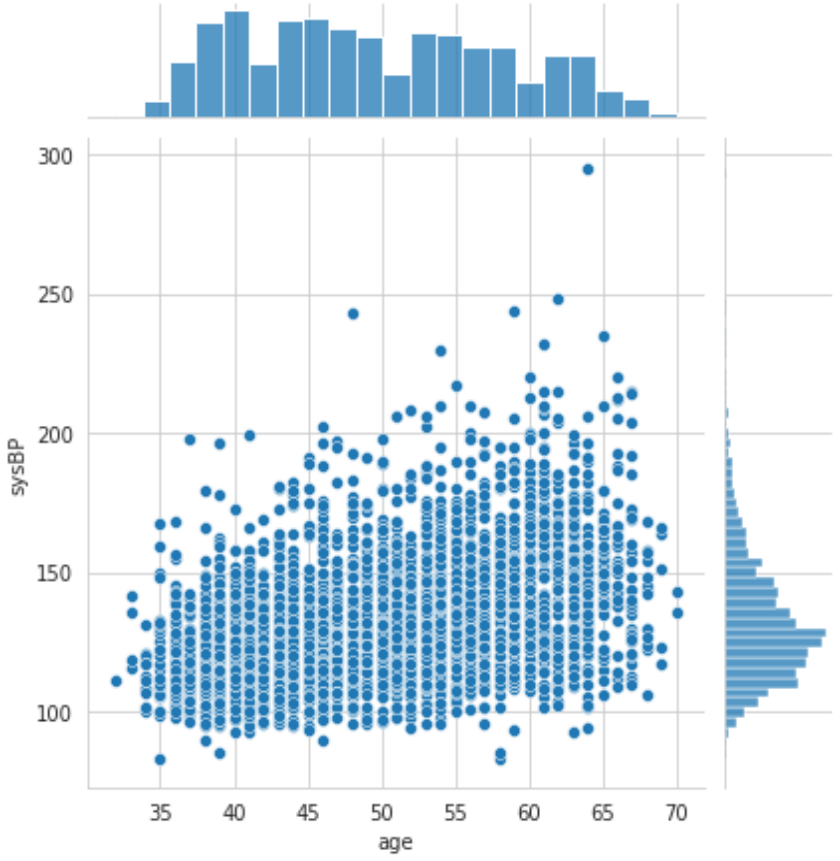


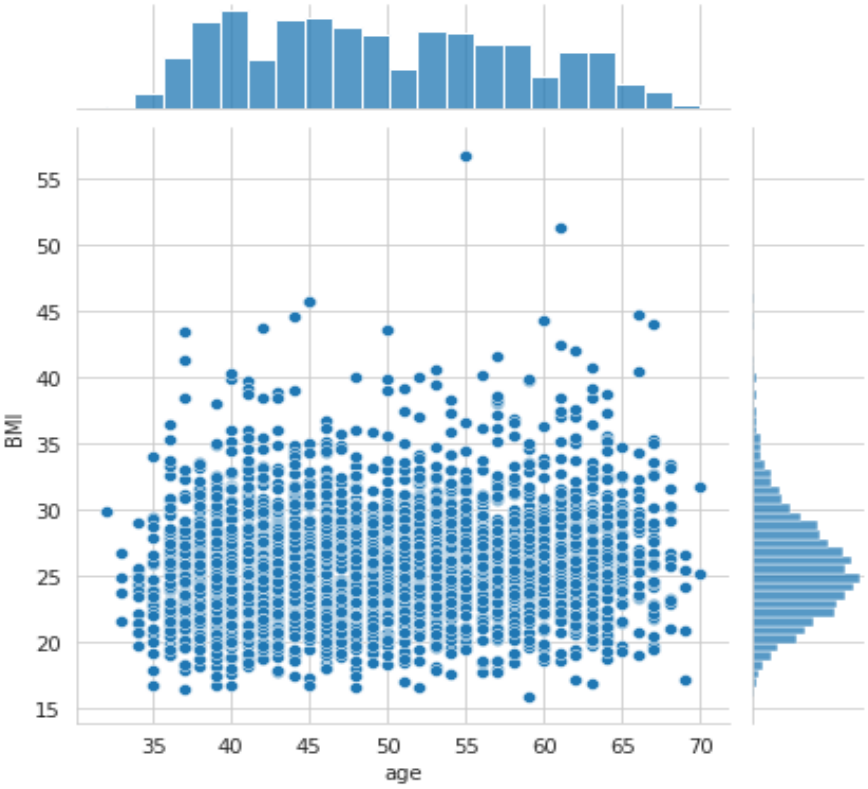
Most people smoke between 0 and 10 cigarettes a day.

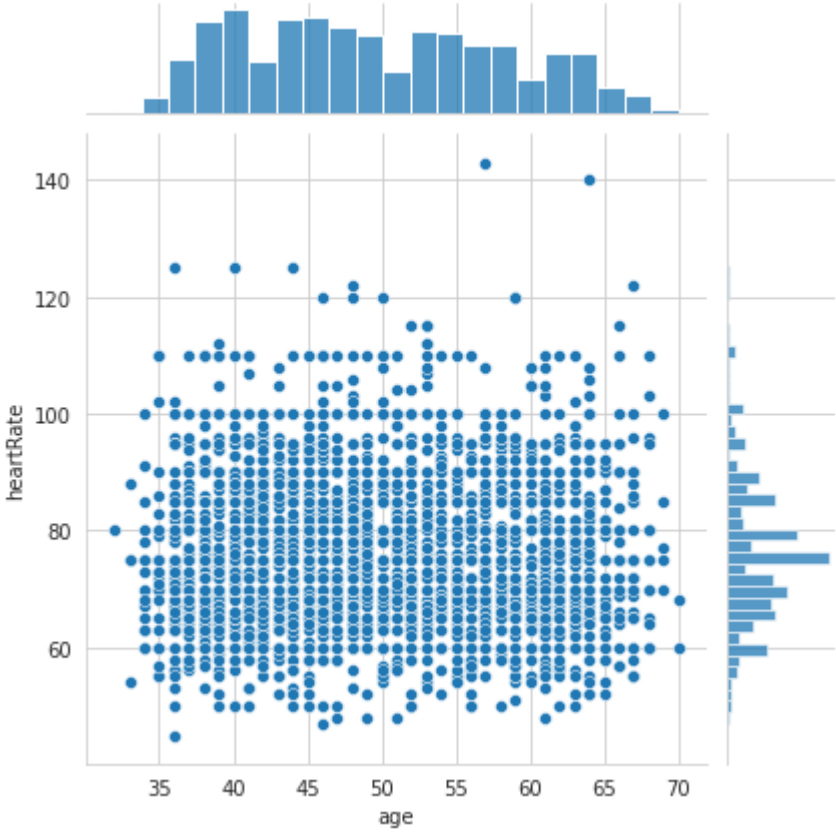


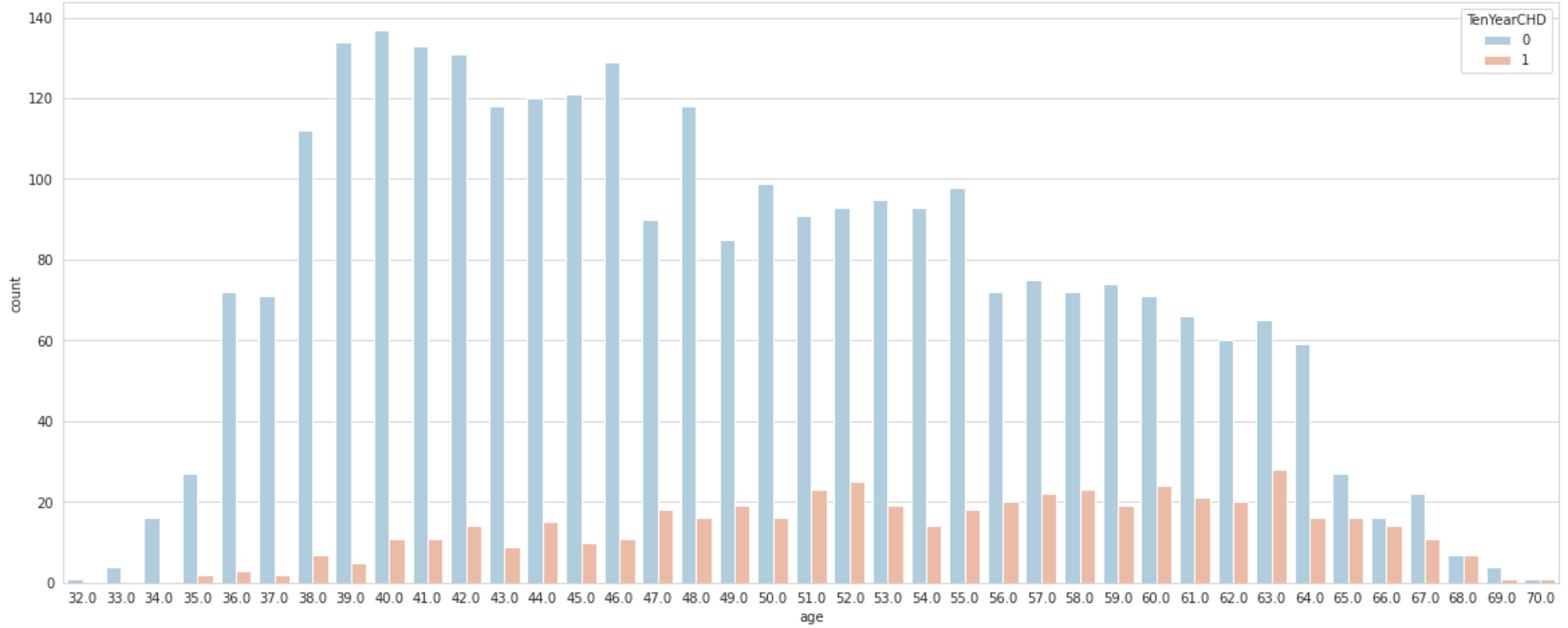
Histogram plot for the age of people in this dataset.


totChol
700
600
500
400
300
200
100
35
40
45
50
55
60
65
70
age

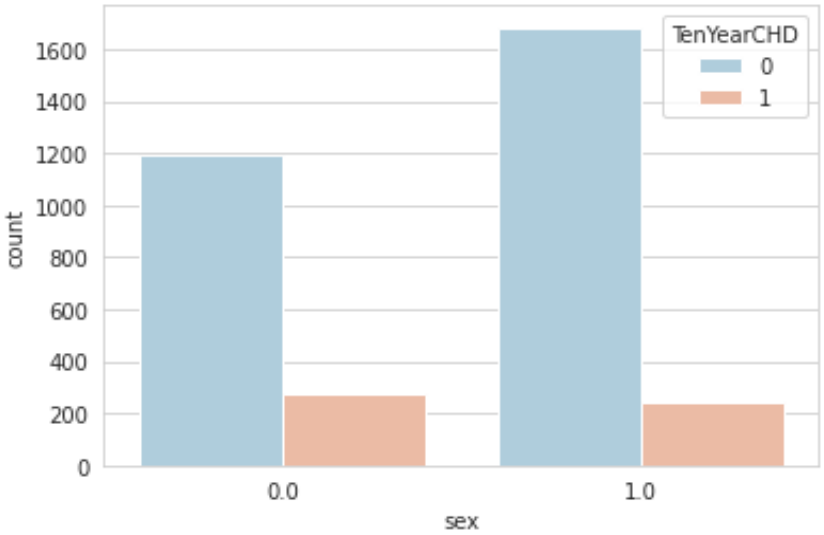

sysBP
300
250
200
150
100
35
40
45
50
55
60
65
70
age


BMI
age
15
20
25
30
35
40
45
50
55
35
40
45
50
55
60
65
70

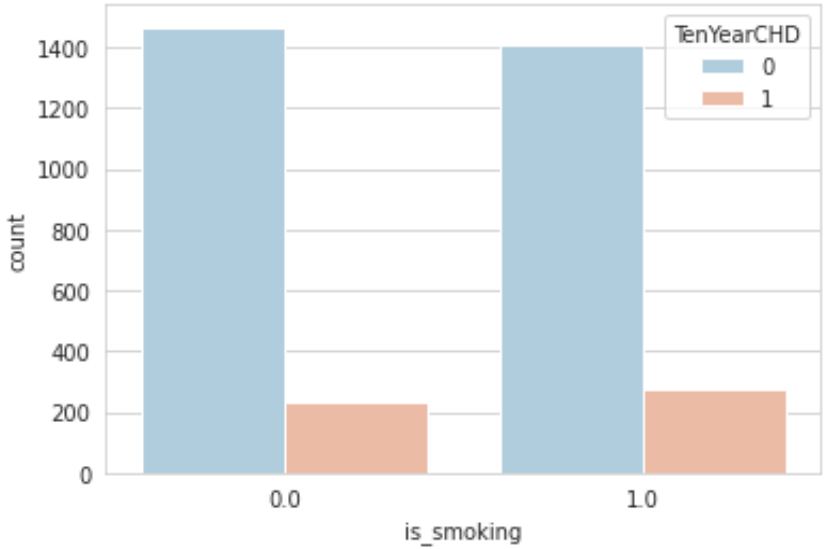

heartRate
age
60
80
100
120
140
35
40
45
50
55
60
65
70



The risk of coronary heart disease rises with age up to age 63 and then declines beyond that.

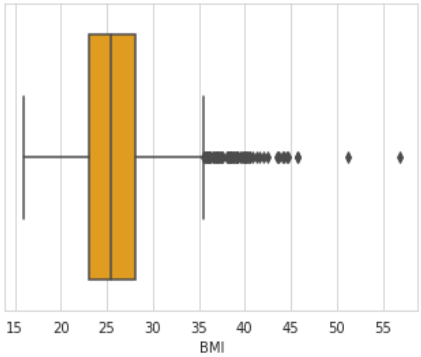


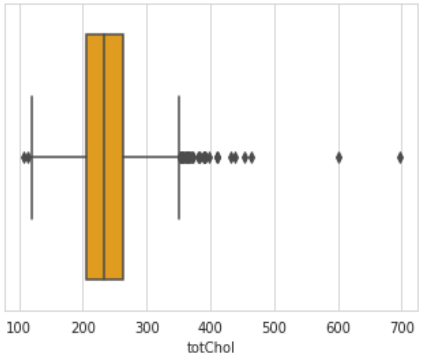
Men have a higher propensity for developing coronary heart disease (CHD) than women. Women, in contrast, are at a higher risk of stroke, which often occurs at an older age.

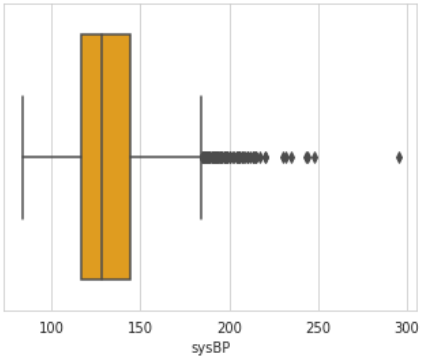


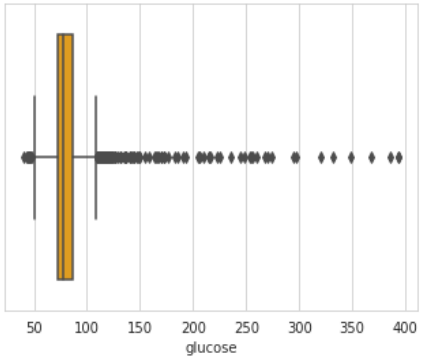
Compared to nonsmokers, smokers have a higher chance of developing coronary heart disease.

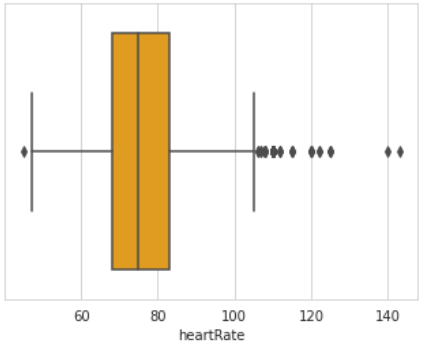
Treatment of outliers

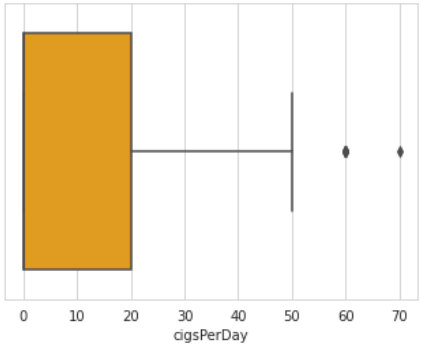


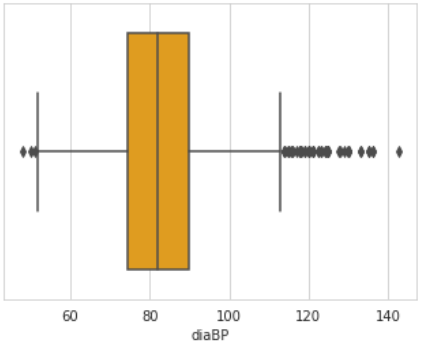


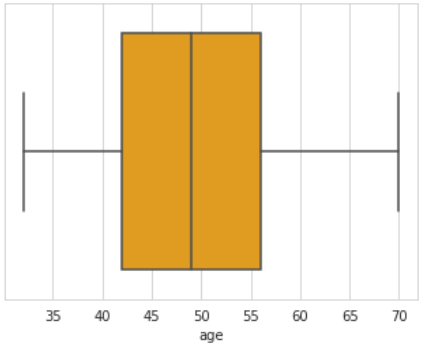






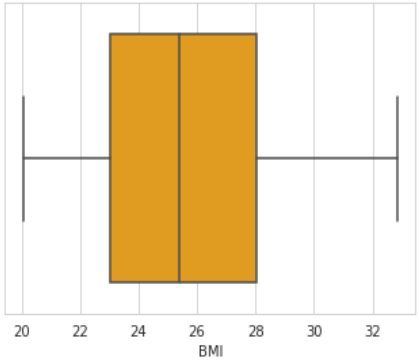


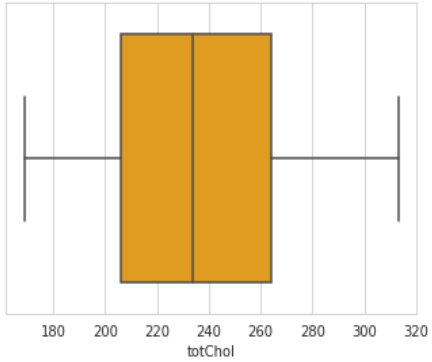


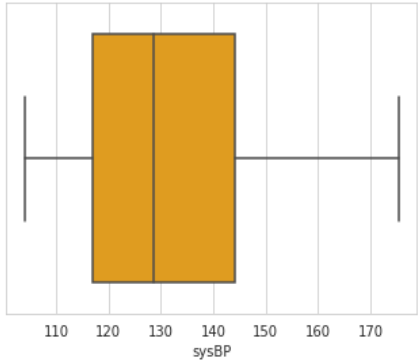


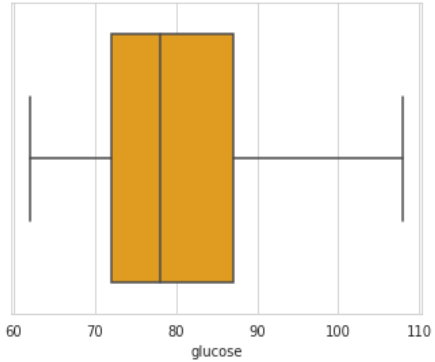
number of Outliers for feature BMI: 46
number of Outliers for feature totChol: 65
number of Outliers for feature sysBP: 104
number of Outliers for feature glucose: 148

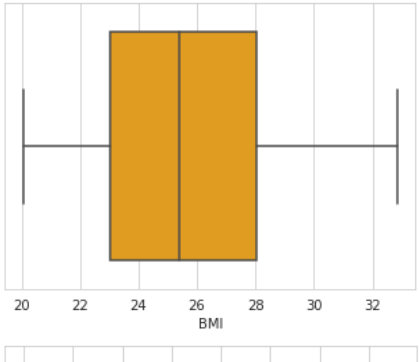
number of Outliers for feature heartRate: 167
number of Outliers for feature cigsPerDay: 184
number of Outliers for feature diaBP: 220
number of Outliers for feature age: 220

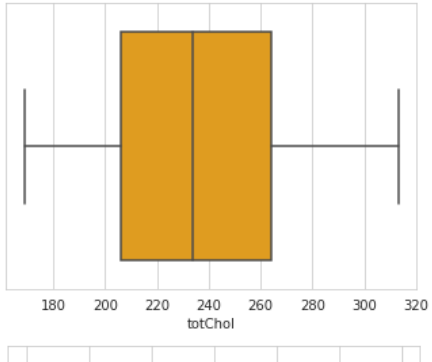

20
22
24
26
28
30
32
BMI

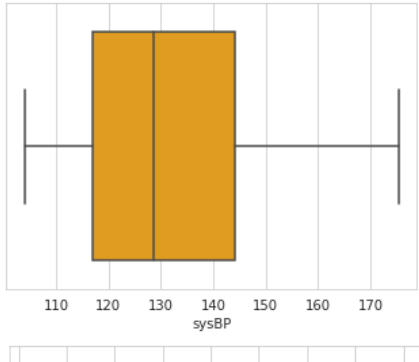

180
200
220
240
260
280
300
320
totChol

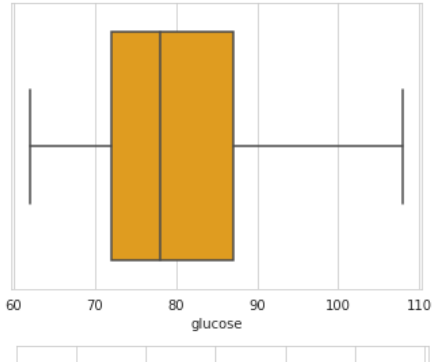

110
120
130
140
150
160
170
sysBP

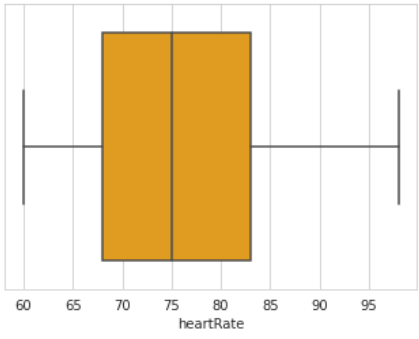

60
70
80
90
100
110
glucose

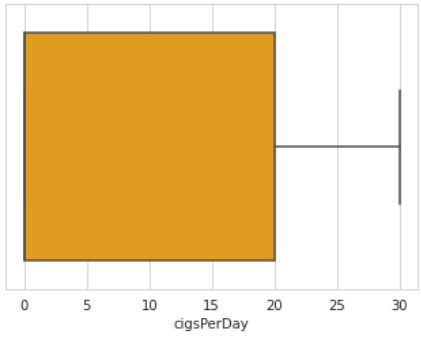

20
22
24
26
28
30
32
BMI

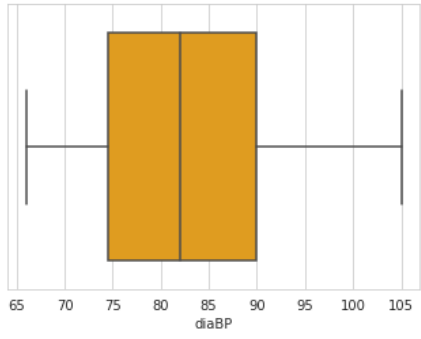

180
200
220
240
260
280
300
320
totChol

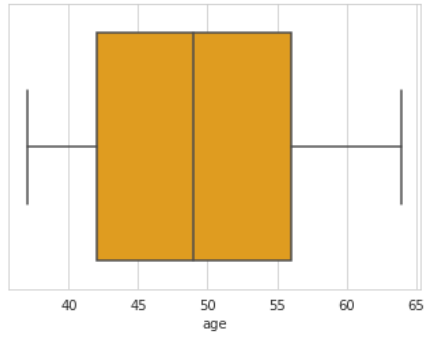

110
120
130
140
150
160
170
sysBP


60
70
80
90
100
110
glucose

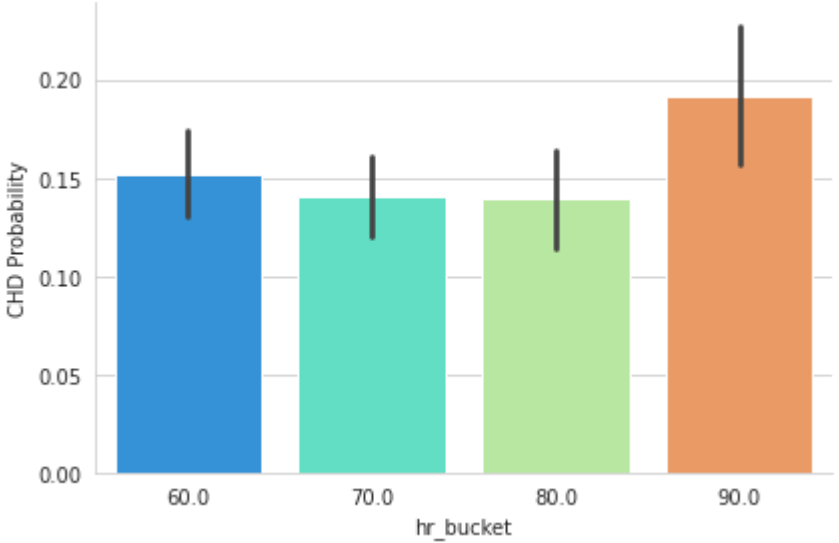

60
65
70
75
80
85
90
95
heartRate

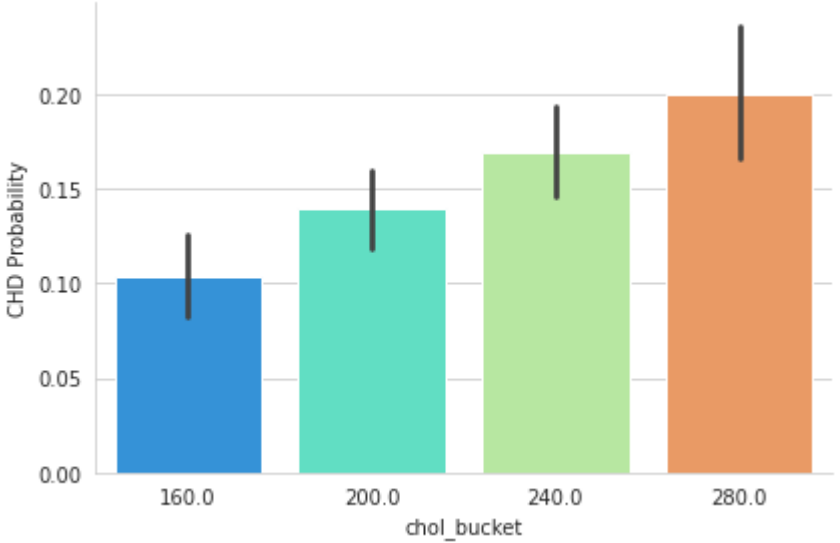

0
5
10
15
20
25
30
cigsPerDay


65
70
75
80
85
90
95
100
105
diaBP


40
45
50
55
60
65
age

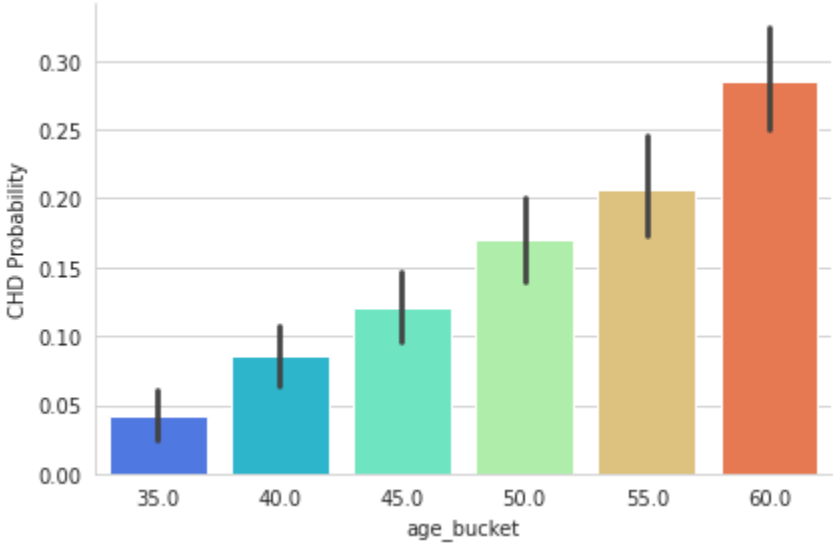
Insights Using Feature Engineering

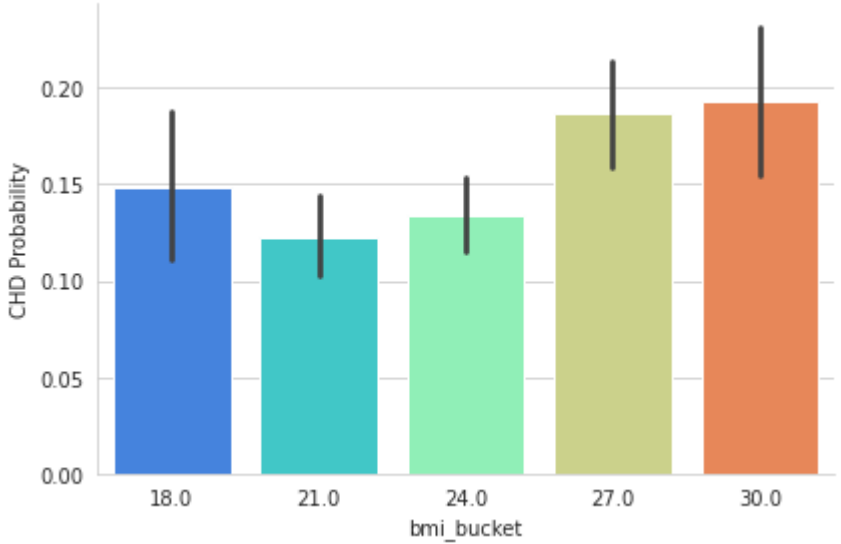




Impact of heart rate on the target variable: People with high heart rates have a high risk of having CHD in the next 10 years.

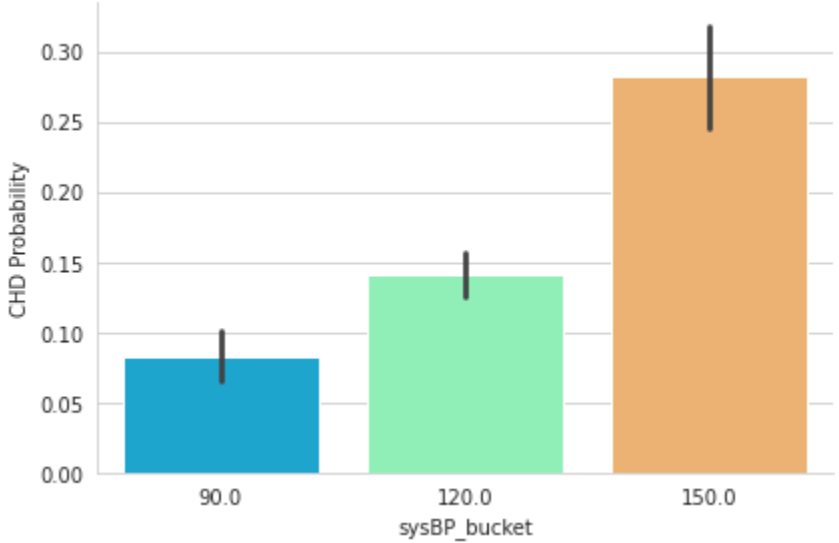
Impact of cholesterol level on the target variable: People with high cholesterol levels have a high risk of having CHD in the next 10 years.



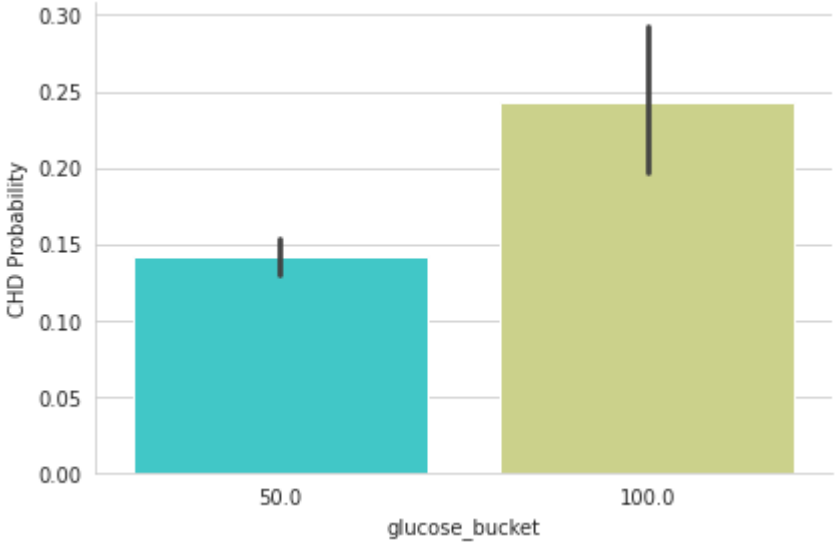


Impact of age on the target variable:
Older people have a high risk of having CHD in the coming 10 years.

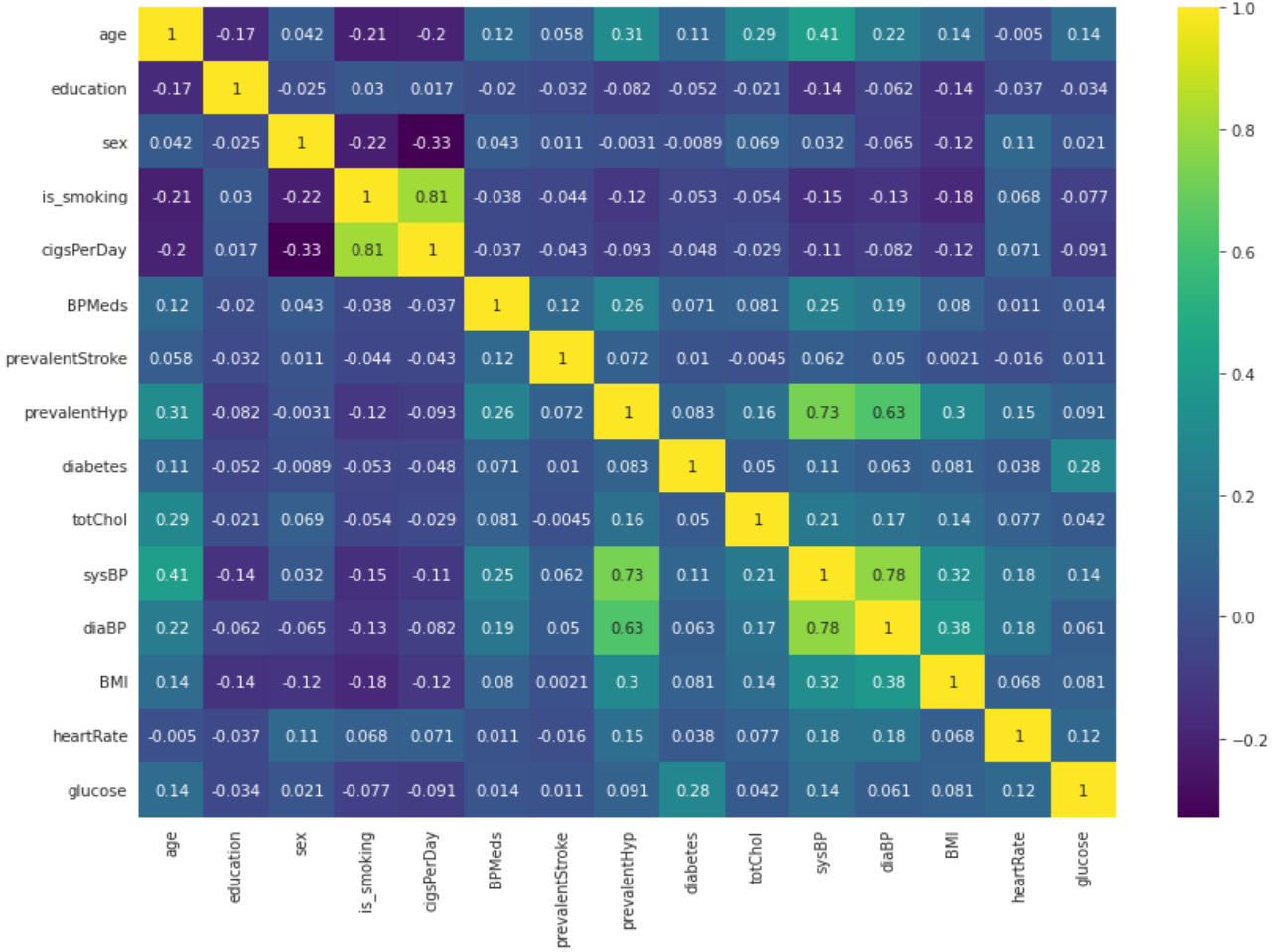
Impact of Body Mass Index on the target variable:
People with high body mass index have a high risk of having CHD in the next 10 years.



Impact of systolic blood pressure on the target variable:
People with high systolic blood pressure have a high risk of having CHD in the next 10 years.



Impact of Glucose on the target variable:
People with high Glucose have a high risk of having CHD in the next 10 years.



Heatmap helps to find the correlation between the features. While implementing Logistic Regression, features having high collinearity will be removed.

From the correlation heatmap, it is clear that these columns are highly correlated.

- Cigs Per Day and is_smoking are highly correlated.
- SysBP and Prevalent Hyp are highly correlated.
- DiaBP and SysBP are highly correlated.

Combined SysBp and DiaBP to denote a new feature pulse rate.

Data preparation

The high correlation between :

1. Cigs Per Day and is_smoking
2. SysBP and Prevalent Hyp
3. DiaBP and SysBP

Combined SysBp and DiaBP to denote a new feature pulse rate.

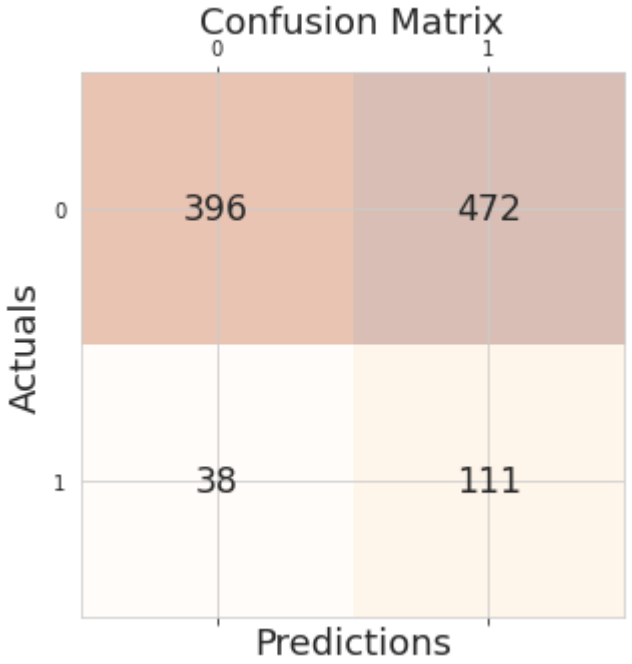
Dropping Cigs Per Day, is_smoking, SysBP, Prevalent Hyp, and DiaBP columns from the dataset.

Added the columns like pulse pressure, age bucket, and BMI bucket.

The dependent column will be predicted as that is the target variable named “Ten Year CHD.

Logistic Regression

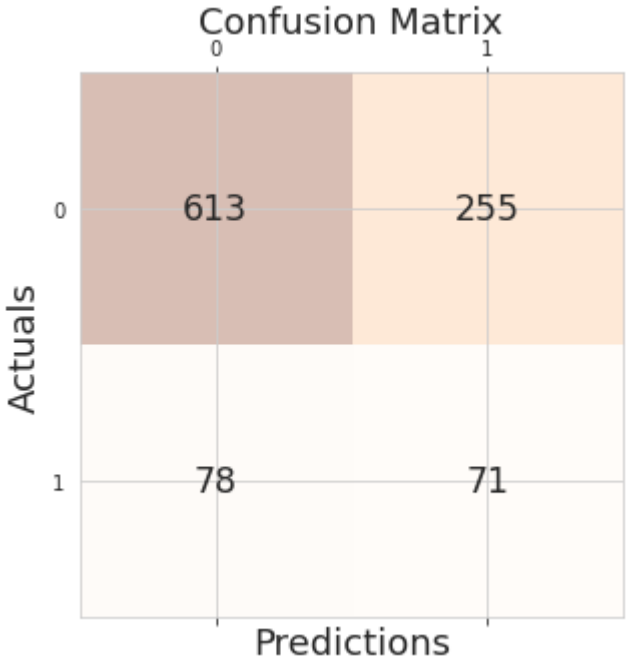
	precision	recall	f1-score	support
0	0.91	0.46	0.61	868
1	0.19	0.74	0.30	149
accuracy			0.50	1017
macro avg	0.55	0.60	0.46	1017
weighted avg	0.81	0.50	0.56	1017



1.Precision = 19%, which means 19 percent of your predictions were correct.
2.Recall = 74.5%, which means 74.5 percent of the positive cases model catch.
3.Accuracy = 49.9%, which means the model can predict sick people 49.9% of the time.
4.F1 score = 30 %, means 30 percent of positive predictions were correct.

Decision Tree

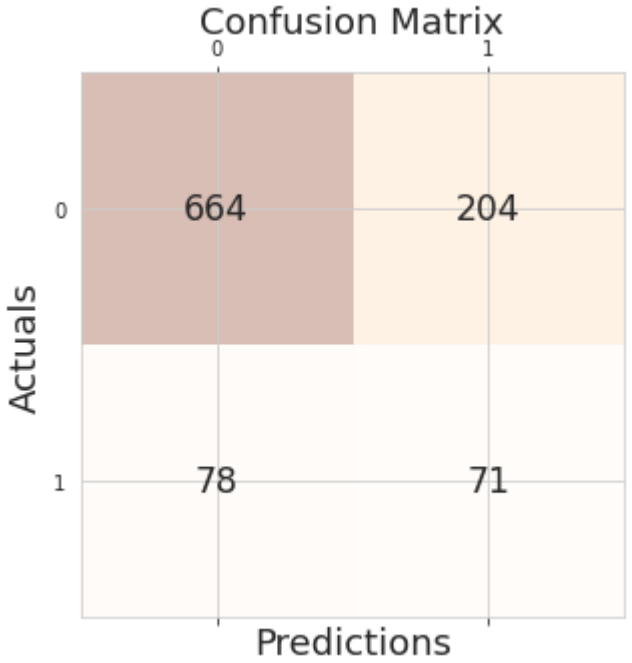
	precision	recall	f1-score	support
0	0.89	0.71	0.79	868
1	0.22	0.48	0.30	149
accuracy			0.67	1017
macro avg	0.55	0.59	0.54	1017
weighted avg	0.79	0.67	0.71	1017



1.Precision = 21.4%, which means 21.4 percent of your predictions were correct.
2.Recall = 49 %, which means 49 percent of the positive cases model catch.
3.Accuracy = 66.2%, which means the model can predict sick people 66.2% of the time.
4.F1 score = 30 %, which means 30 percent of positive predictions were correct.

Random Forest

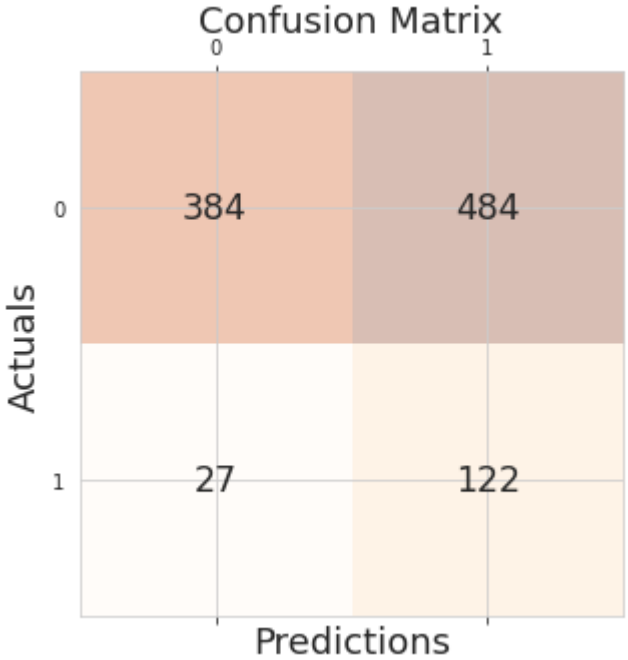
	precision	recall	f1-score	support
0	0.89	0.76	0.82	868
1	0.26	0.48	0.33	149
accuracy			0.72	1017
macro avg	0.58	0.62	0.58	1017
weighted avg	0.80	0.72	0.75	1017



1.Precision = 26.2%, which means 26.2 percent of your predictions were correct.
2.Recall = 49 %, which means 49 percent of the positive cases model catch.
3.Accuracy = 72.3%, which means the model can predict sick people 72.3% of the time.
4.F1 score = 34 %, which means 34 percent of positive predictions were correct.

Support Vector Machine

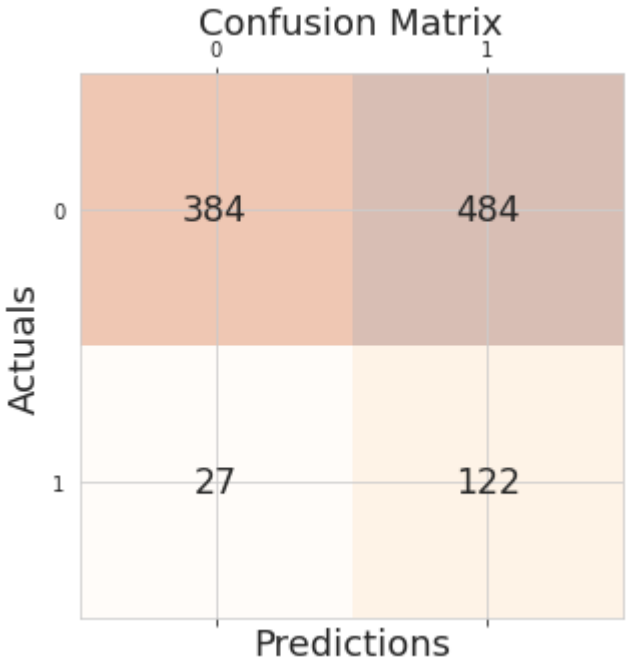
	precision	recall	f1-score	support
0	0.93	0.44	0.60	868
1	0.20	0.82	0.32	149
accuracy			0.50	1017
macro avg	0.57	0.63	0.46	1017
weighted avg	0.83	0.50	0.56	1017



1.Precision = 20.1%, which means 20.1 percent of your predictions were correct.
2.Recall = 81.9 %, which means 81.9 percent of the positive cases model catch.
3.Accuracy = 49.8%, which means the model can predict sick people 72.3% of the time.
4.F1 score = 32 %, which means 32 percent of positive predictions were correct.

Grid Search CV

	precision	recall	f1-score	support
0	0.87	0.78	0.83	868
1	0.20	0.32	0.25	149
accuracy			0.72	1017
macro avg	0.54	0.55	0.54	1017
weighted avg	0.77	0.72	0.74	1017



1.Precision = 20.1%, which means 20.1 percent of your predictions were correct.
2.Recall = 81.9 %, which means 81.9 percent of the positive cases model catch.
3.Accuracy = 49.8%, which means the model can predict sick people 72.3% of the time.
4.F1 score = 25 %, which means 25 percent of positive predictions were correct.

Conclusion

Since our aim was to lower the false-negative value so that patients do not get detected improperly and are demonstrated to be safe, I used the recall score as the evaluation matrix. The patient's health may suffer greatly as a result of this.

Data were resampled because they weren't balanced. High accuracy can be achieved with imbalanced data, however, in these situations, recall, precision, and F1 score must be considered.

Used a KNN-imputer to perform missing value imputation, and processed data to remove outliers. To solve the issue of class imbalance, SMOTE boosting was used to over-sample the minority class observations.

Newer elements like pulse pressure, age bucket, and BMI bucket that helped to explain the separation in the Risk were created using the information from EDA.

Due to the parametric relationship in the data, a logistic regression model was implemented, and it was successful in achieving a Recall of 74.5%. Even though the recall score for SVM was 81.9 %, SVM is not an interpretable model, thus I chose an interpretable model for this situation.

All measures, including Precision, Recall, Accuracy, and F1 score were evaluated for each model.

Based on this analysis,
Logistic regression can identify positive cases with a 74.5% Recall.
Using a decision tree, positive cases may be predicted with a recall of 49%.
With the help of Random Forest, positive cases may be predicted with a 49% Recall.
Using a Support Vector Machine, positive cases can be predicted with an 81.9% Recall.
Using Grid Search CV, positive cases can be predicted with 81.9% Recall.

THANK YOU