CAPSTONE PROJECT-3

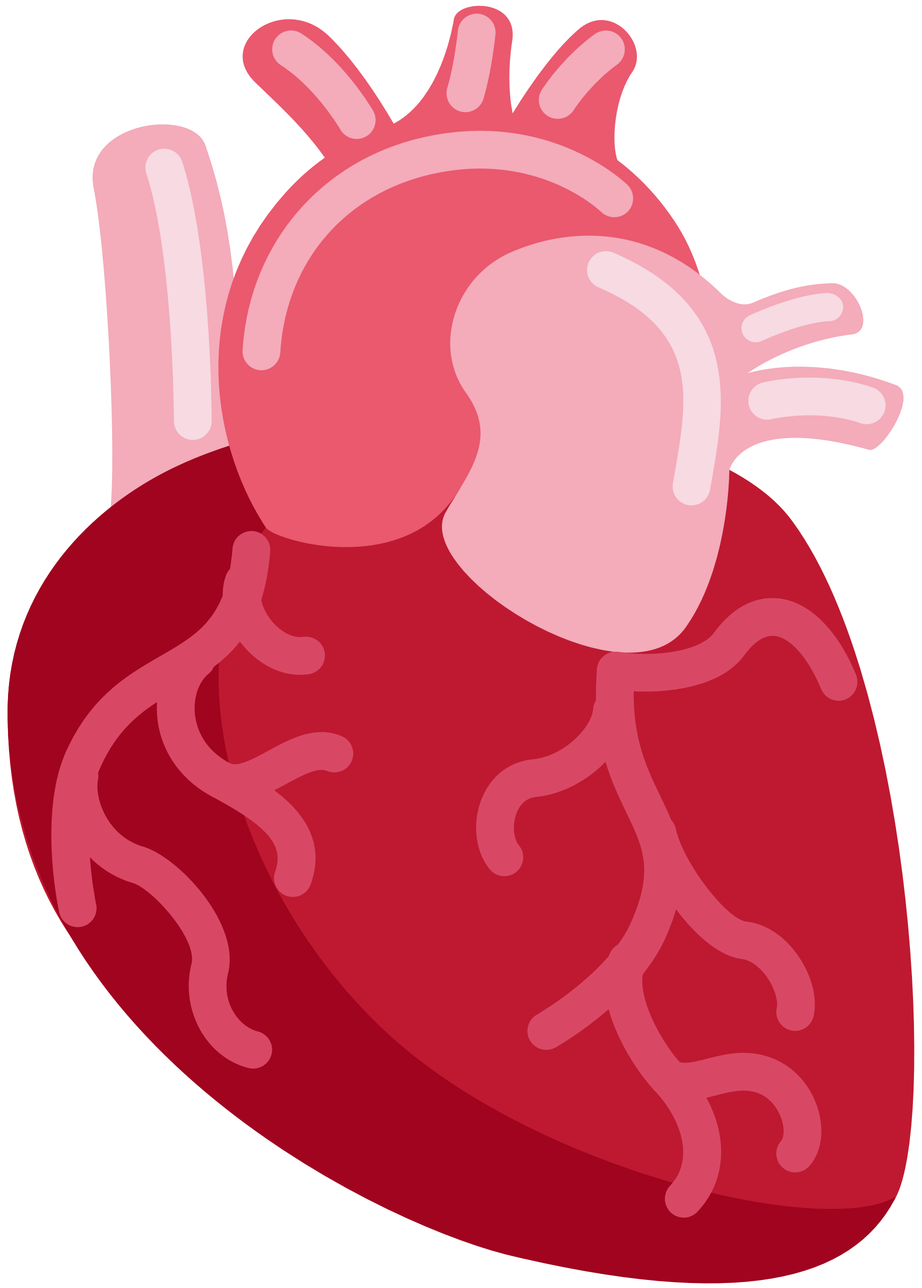
CARDIOVASCULAR DISEASE RISK PREDICTION

(SUPERVISED LEARNING-CLASSIFICATION)

Presented by:-

RAHUL JHA (COHORT VINDHYA)

CONTENT:-



1.Problem Statement.

2.Data Understanding.

3.Exploratory Data Analysis

- Data Visualization on univariate variable.
- Data Visualization on multivariate variable.

4.Data Cleaning

5.Model Evaluation.

6.Hyperparameter Tuning

PROBLEM STATMENT:-

The term heart disease is often used interchangeably with the term cardiovascular disease. Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke.

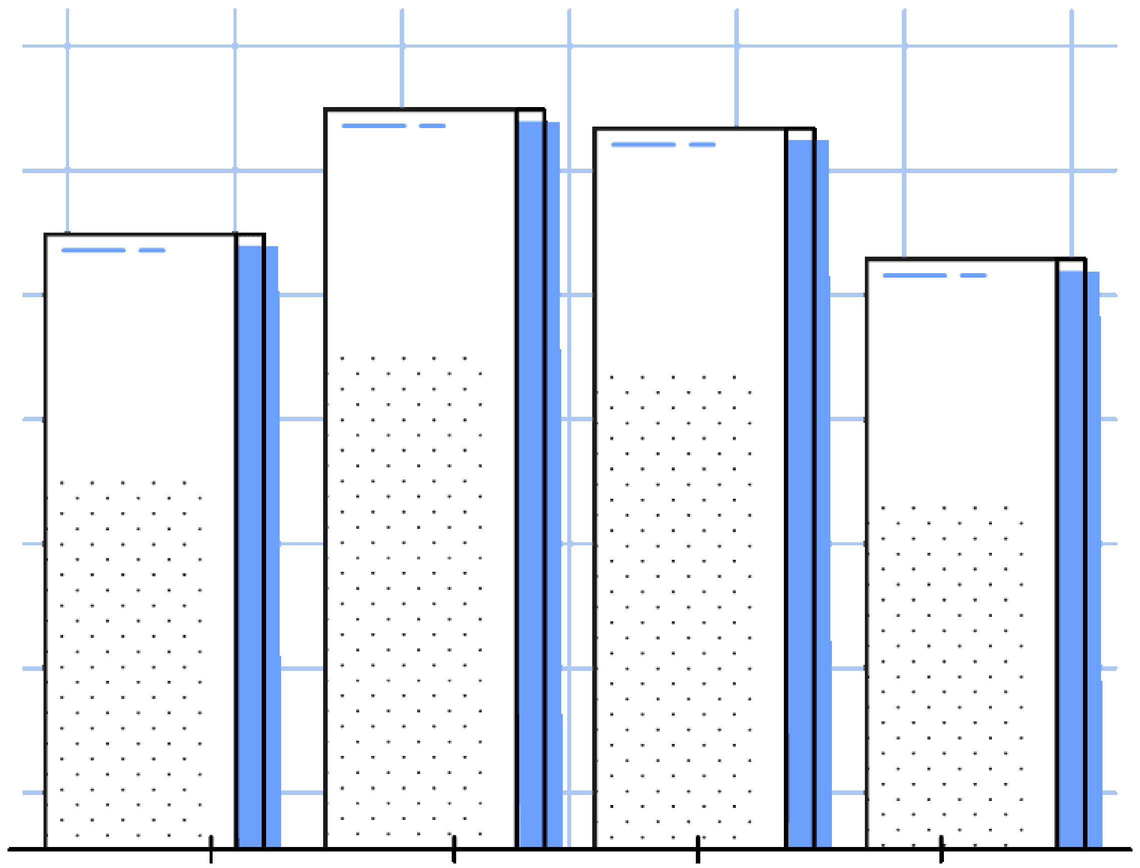
Risk factors are attributes, characteristics or exposures of a person that play a role in the development of cardiovascular disease, for example your smoking status or your blood pressure.

Our major objective is to analyze the dataset and determine whether someone is suffering from Cardiovascular disease using machine learning concepts.

ABOUT THE DATASET:-

- 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 3390 records and 15 attributes. Variables Each attribute is a potential risk factor.
- There are both demographic, behavioral, and medical risk factors.

Columns description:-



- **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.
- **BP Meds:** whether or not the patient was on blood pressure medication.
- **Prevalent Stroke: W**hether 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 Preprocessing

1.Missing value:-

- We use the median value to replace missing values in significant columns like heart rate, BMI, and others.
- And use the diabetic value to fill in the missing glucose value.

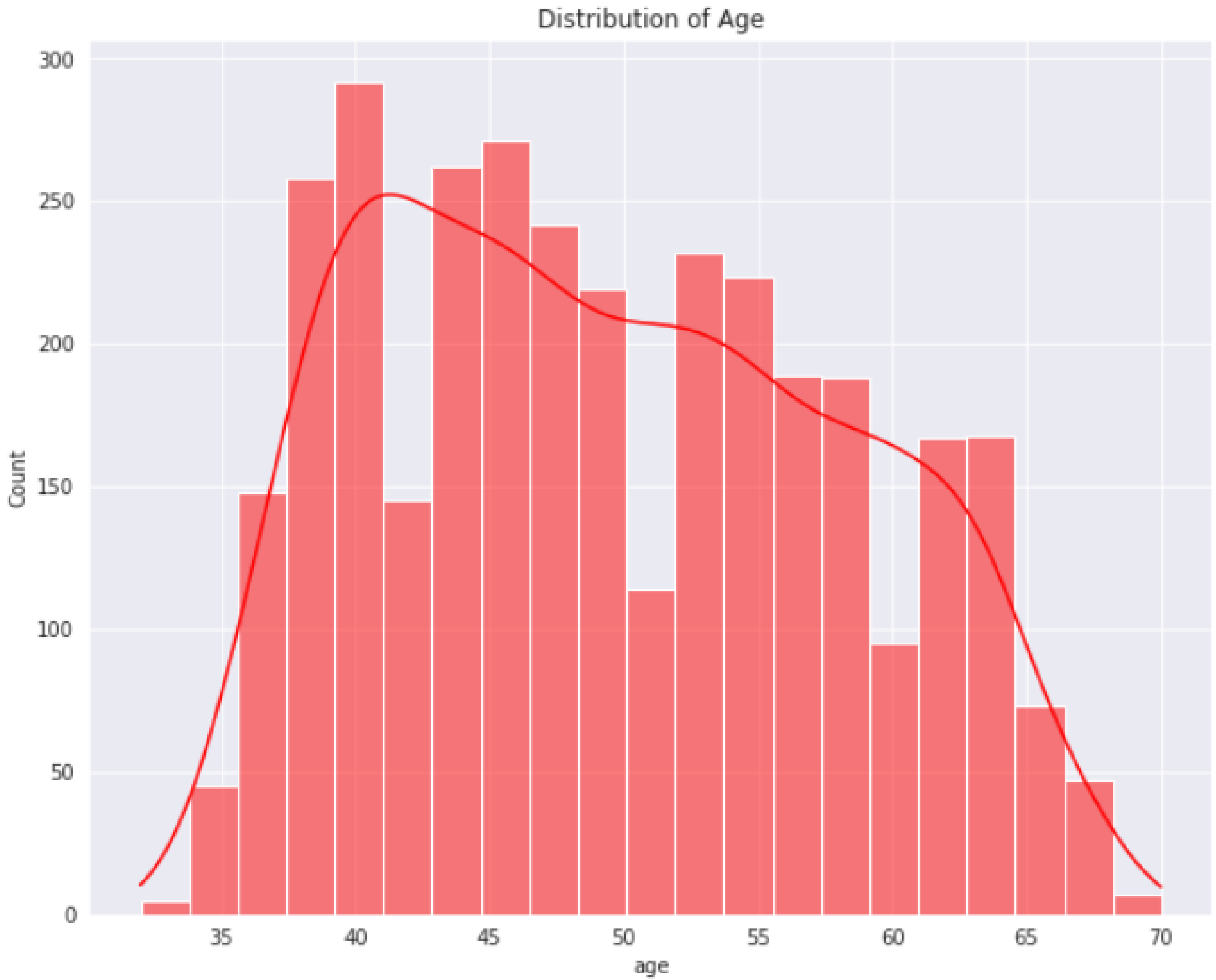
2.Label Encoder

- **We replace the categorical value of sex,and smoking with numerical value.**
- **Continuous value like cholestral and cigarette be converted into categorical group to increase the accuracy.**
- **One hot encoding 'education','cholestral_level' and 'cig_group'.**

3.There were no duplicate value in the dataset.

Exploratory Data Analysis

1.Data Visualization on univariate variable.

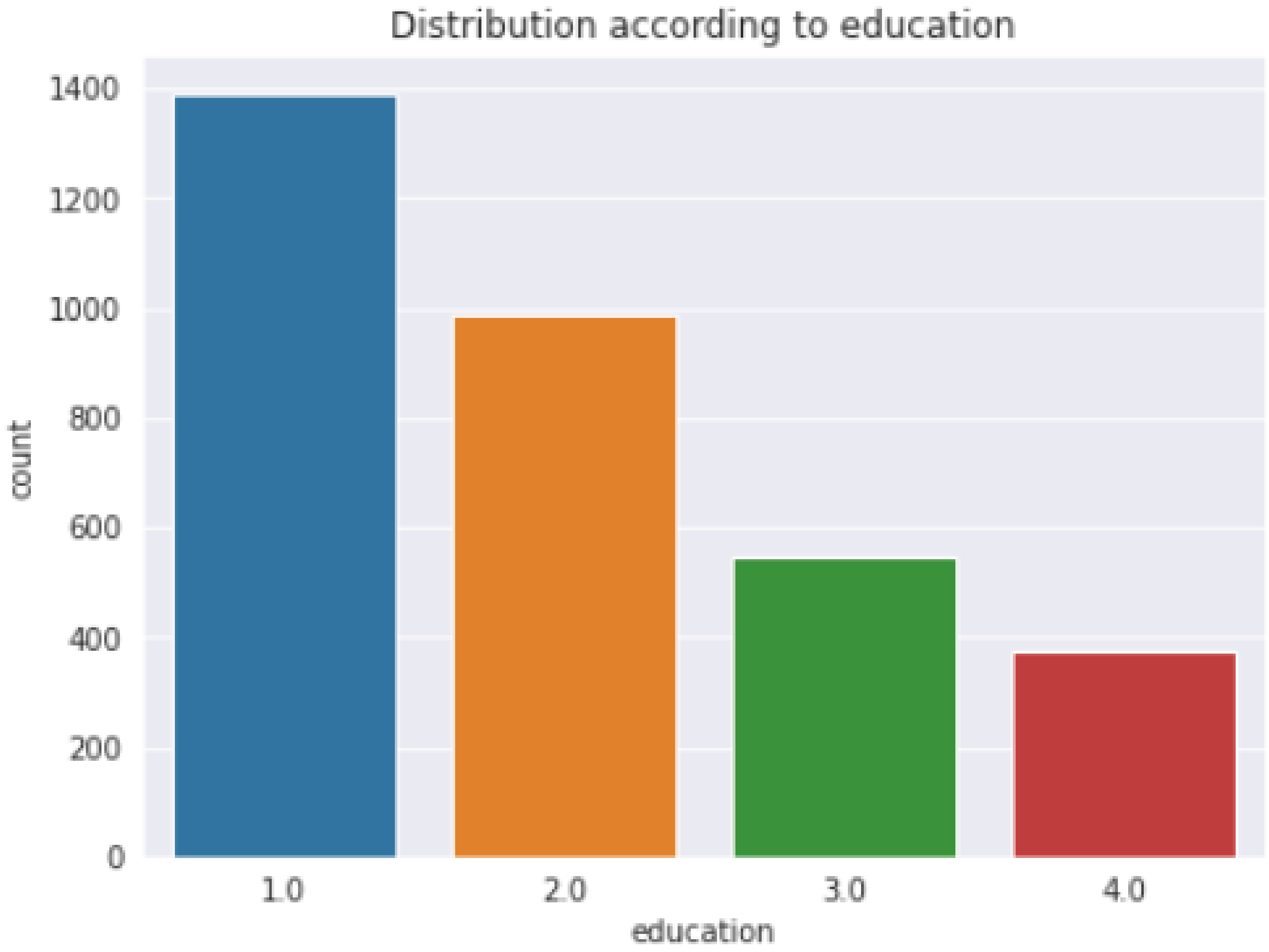


Distribution of age

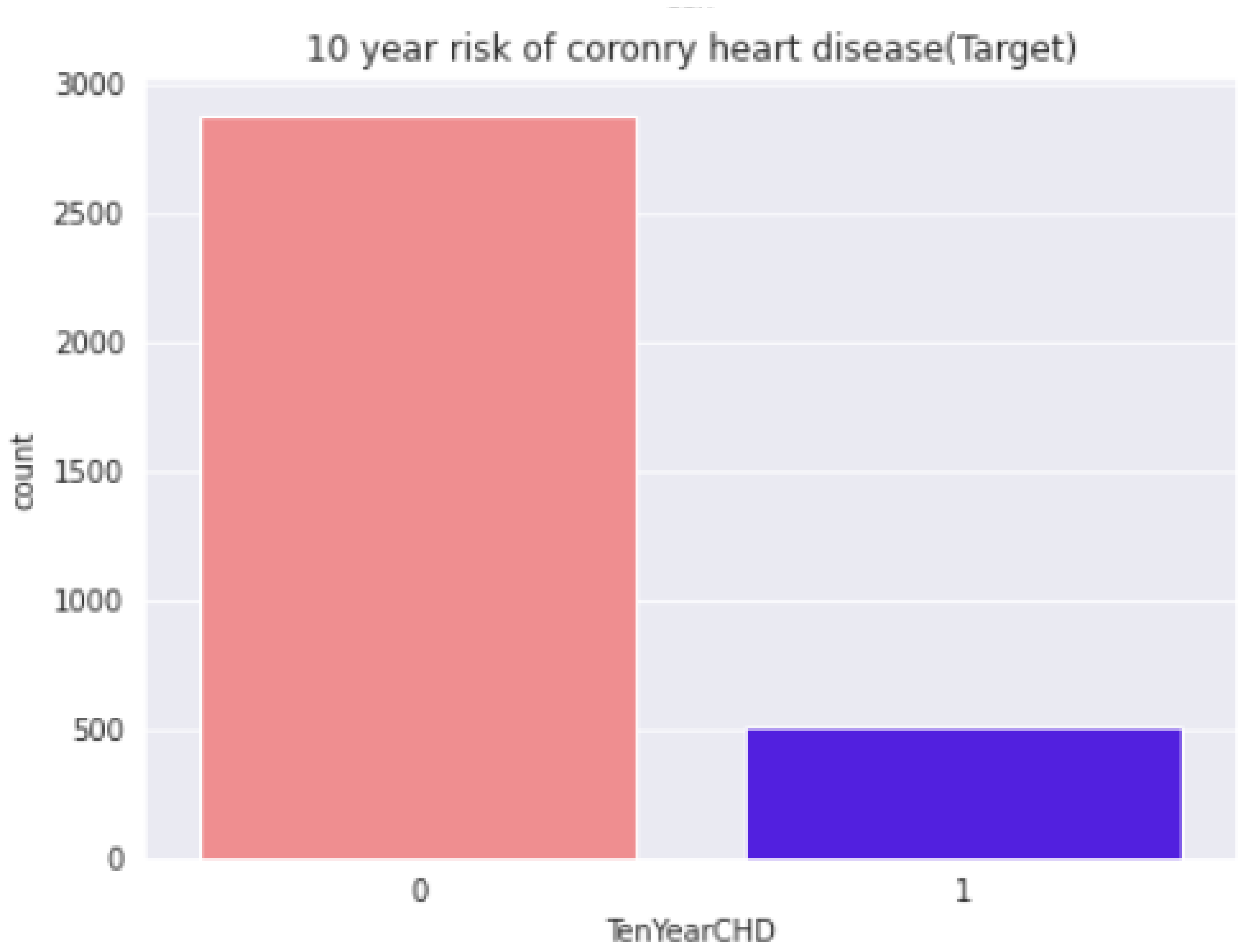
Conclusion:-

- **Most of the distribution fall between 38-50.**
- **And between 52-60 we see a small peak in the distribution.**

Countplot of categorical value:-

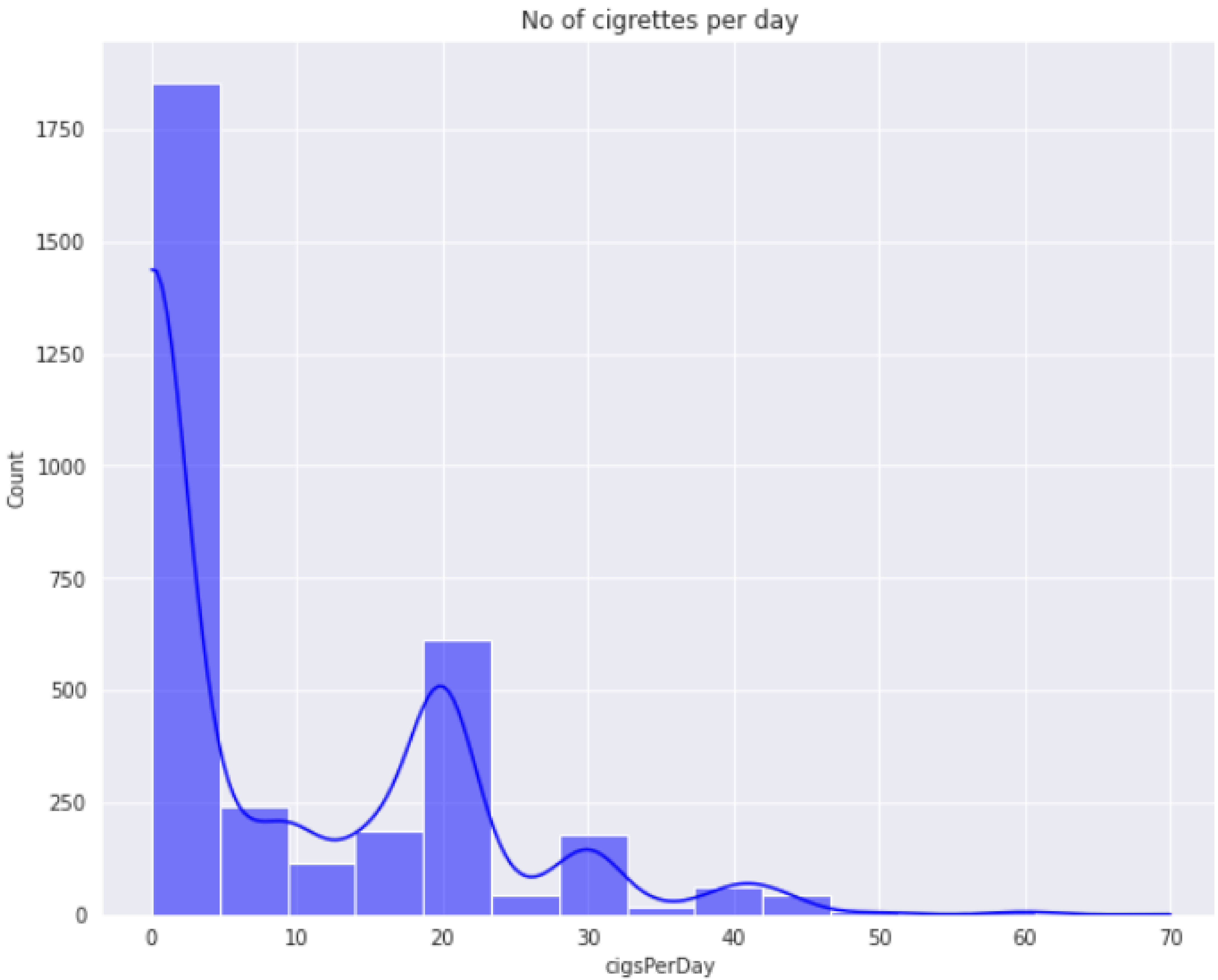


Education



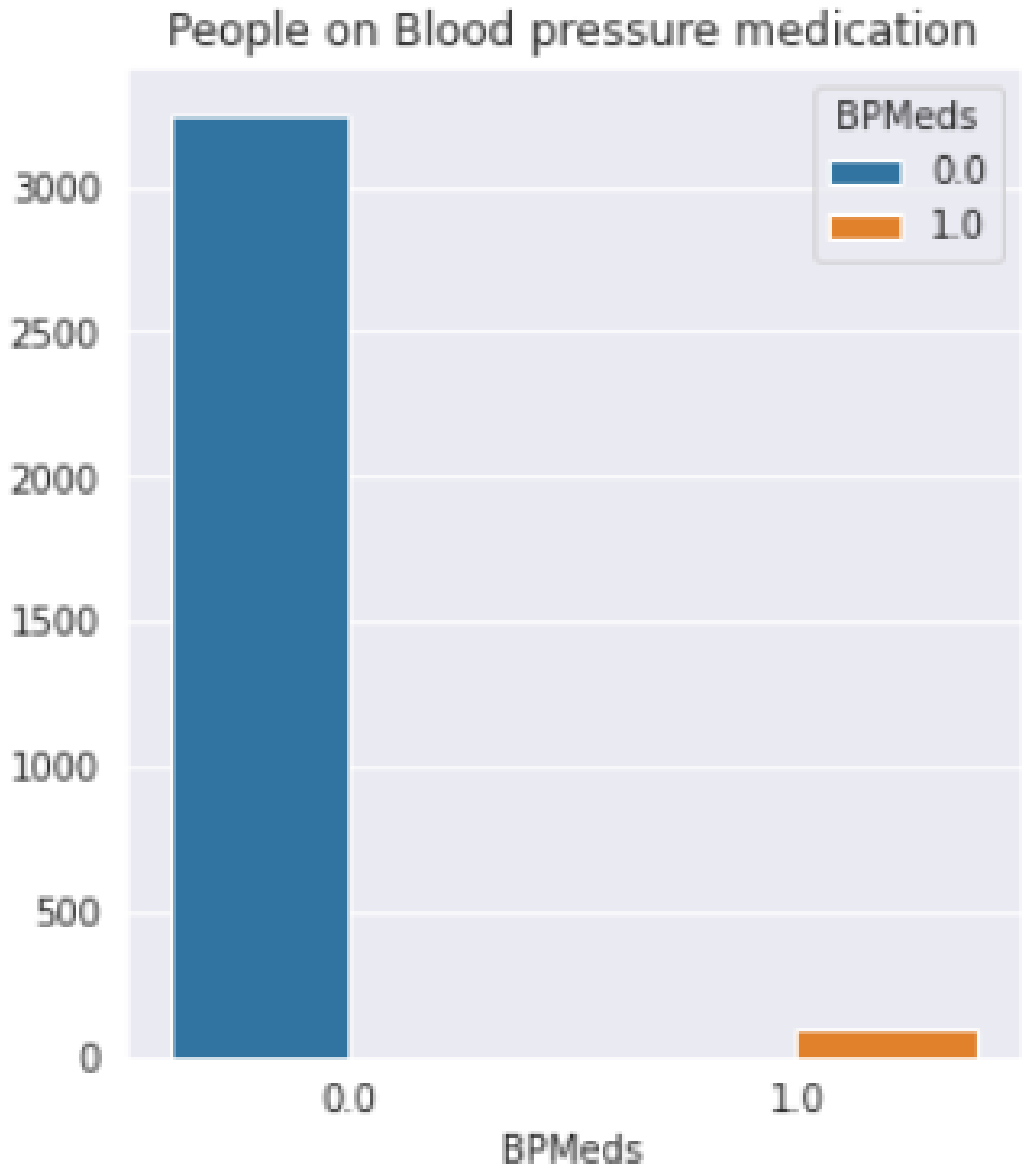
Distribution of people having heart disease.

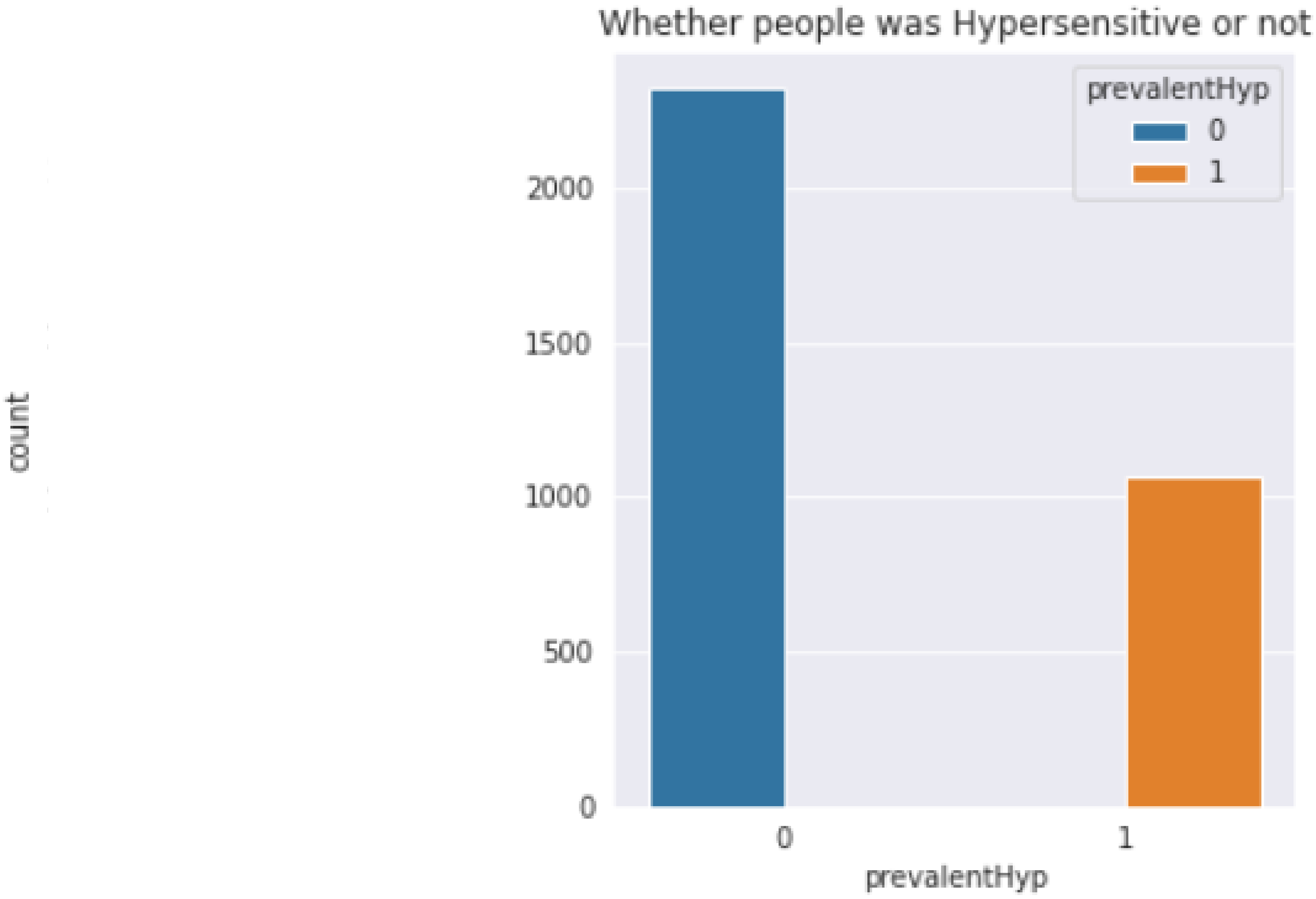
Distribution of number of cigarette people drink each day



Conclusion:-

- Most number of people do not drink cigarette.
- There is a peak representing approx 500 people drink atleat 20 cigaretee
- We can see a small peak of distribution of people drinking cigarette about 30 and 40.

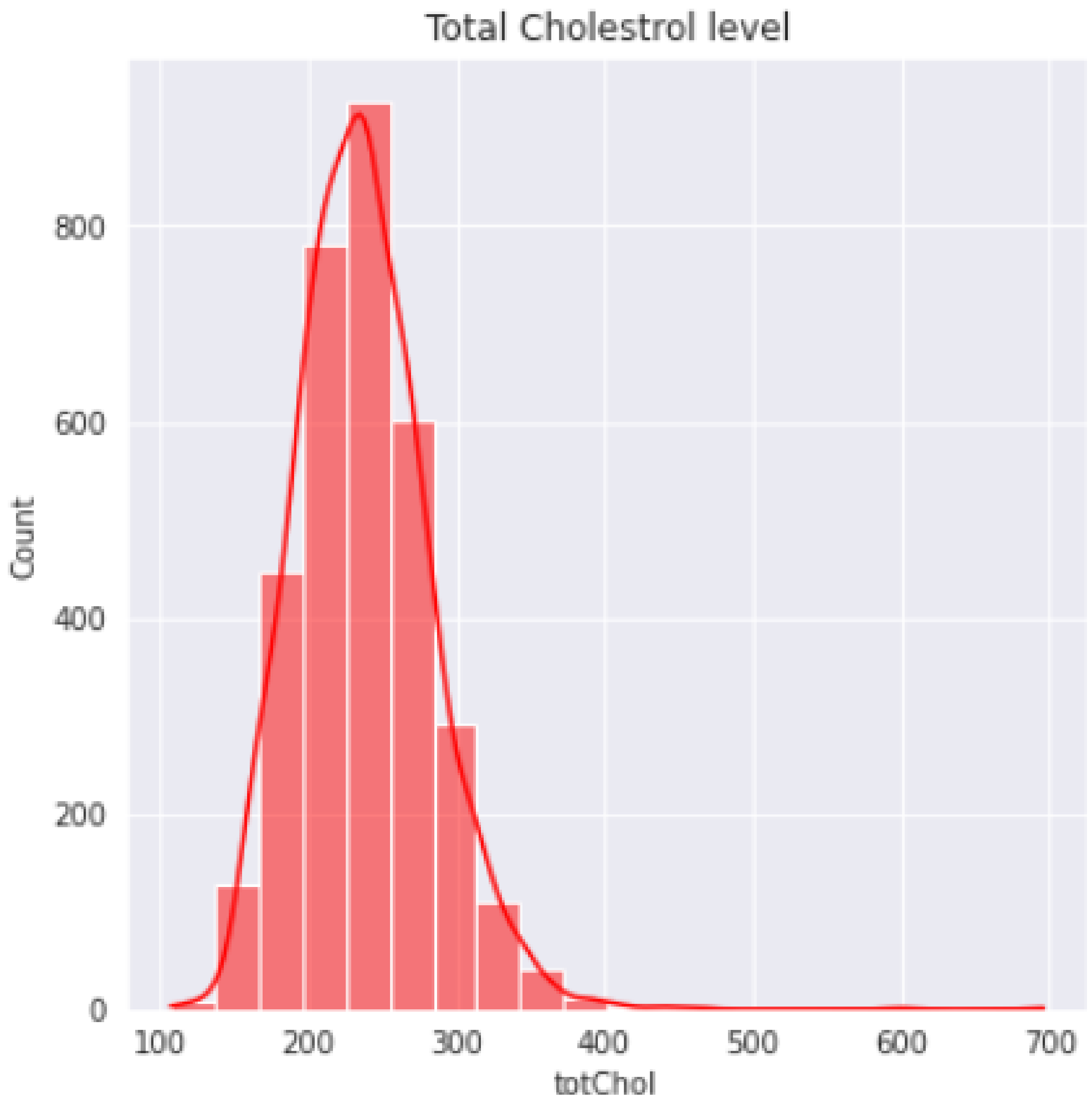




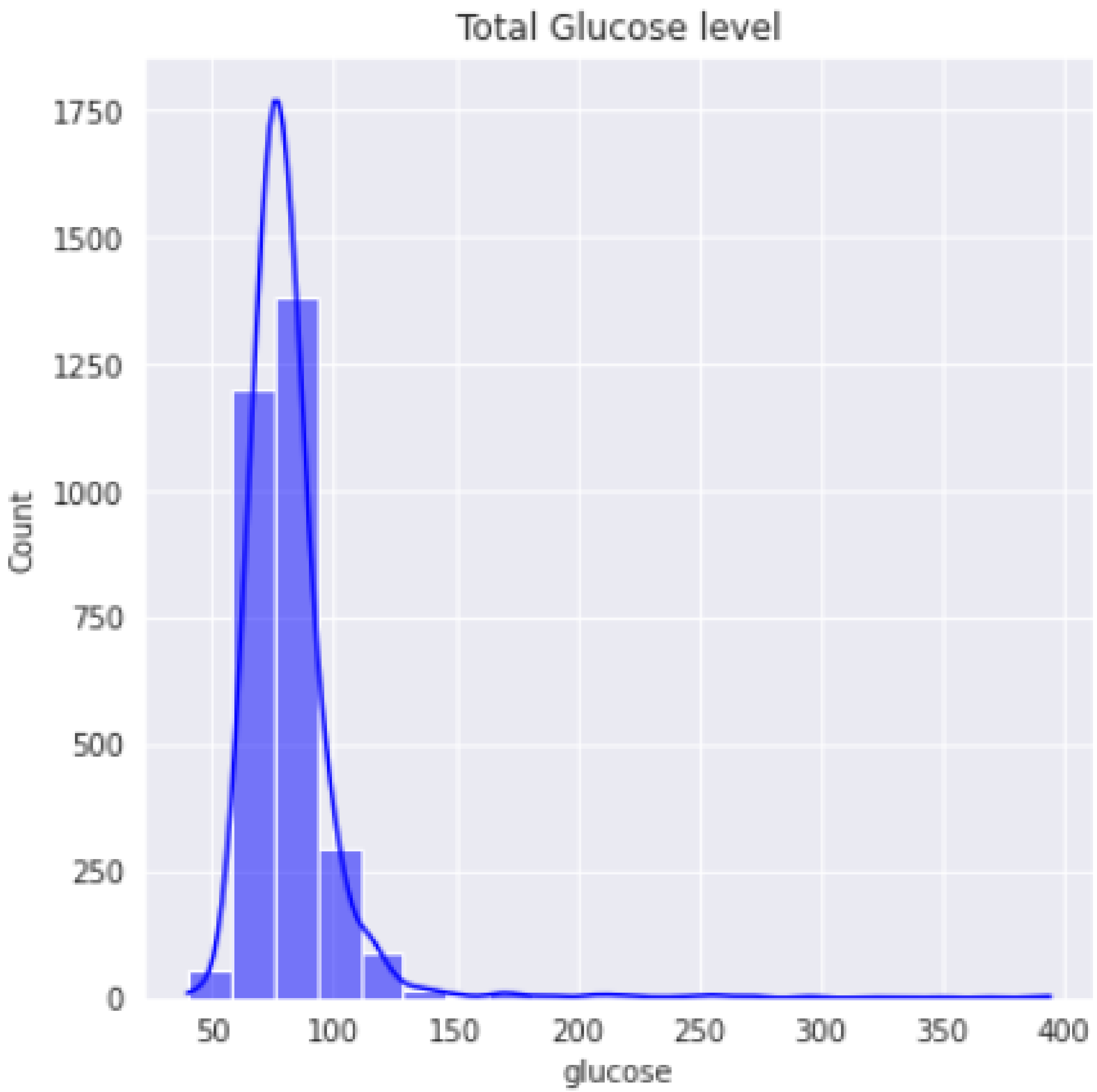
1.About less than 1% people has blood pressure medication.

2.About 68% of people are hypersensitive.

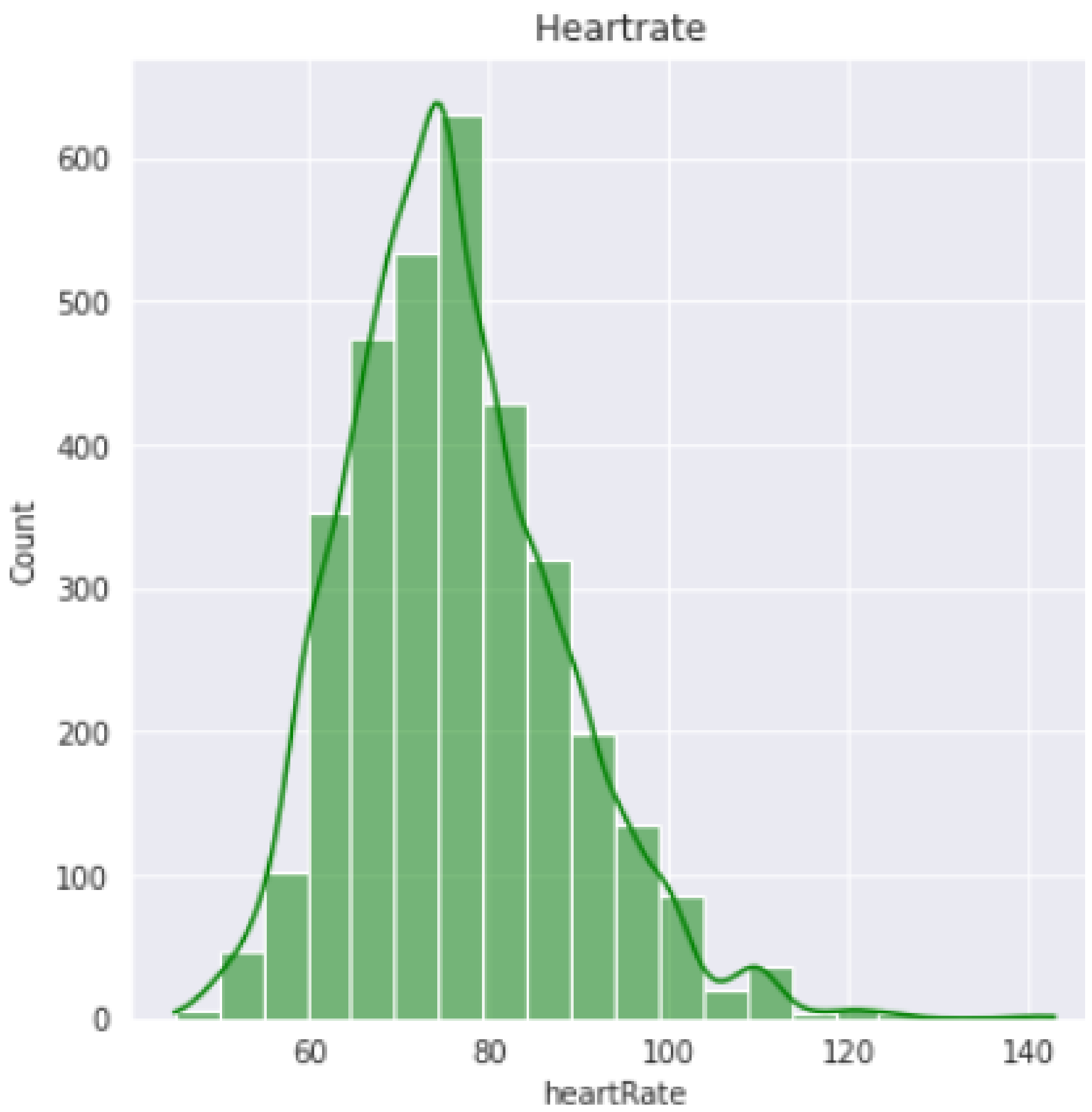
Distribution of different continuous column:-



Cholestral level



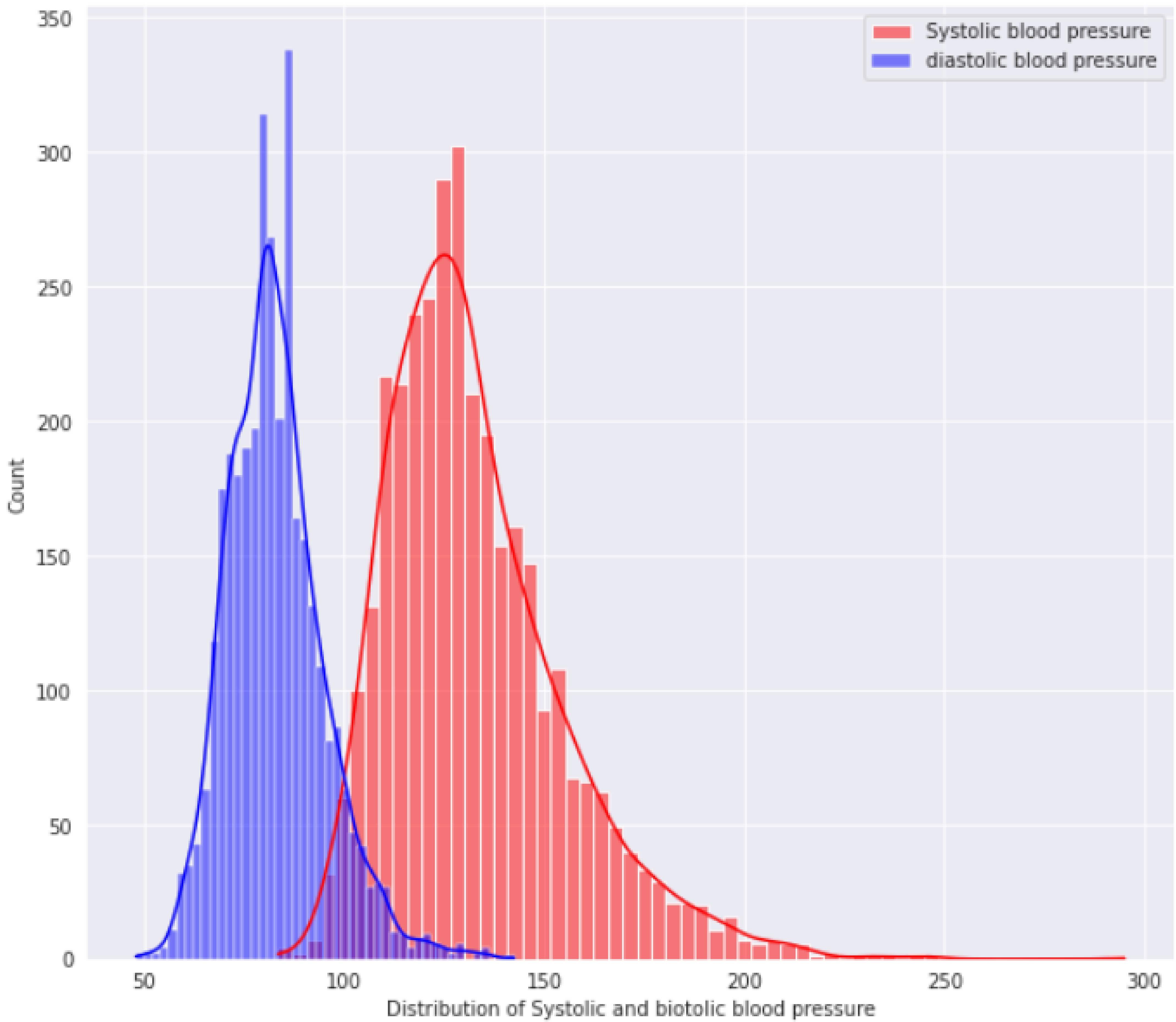
Glucose



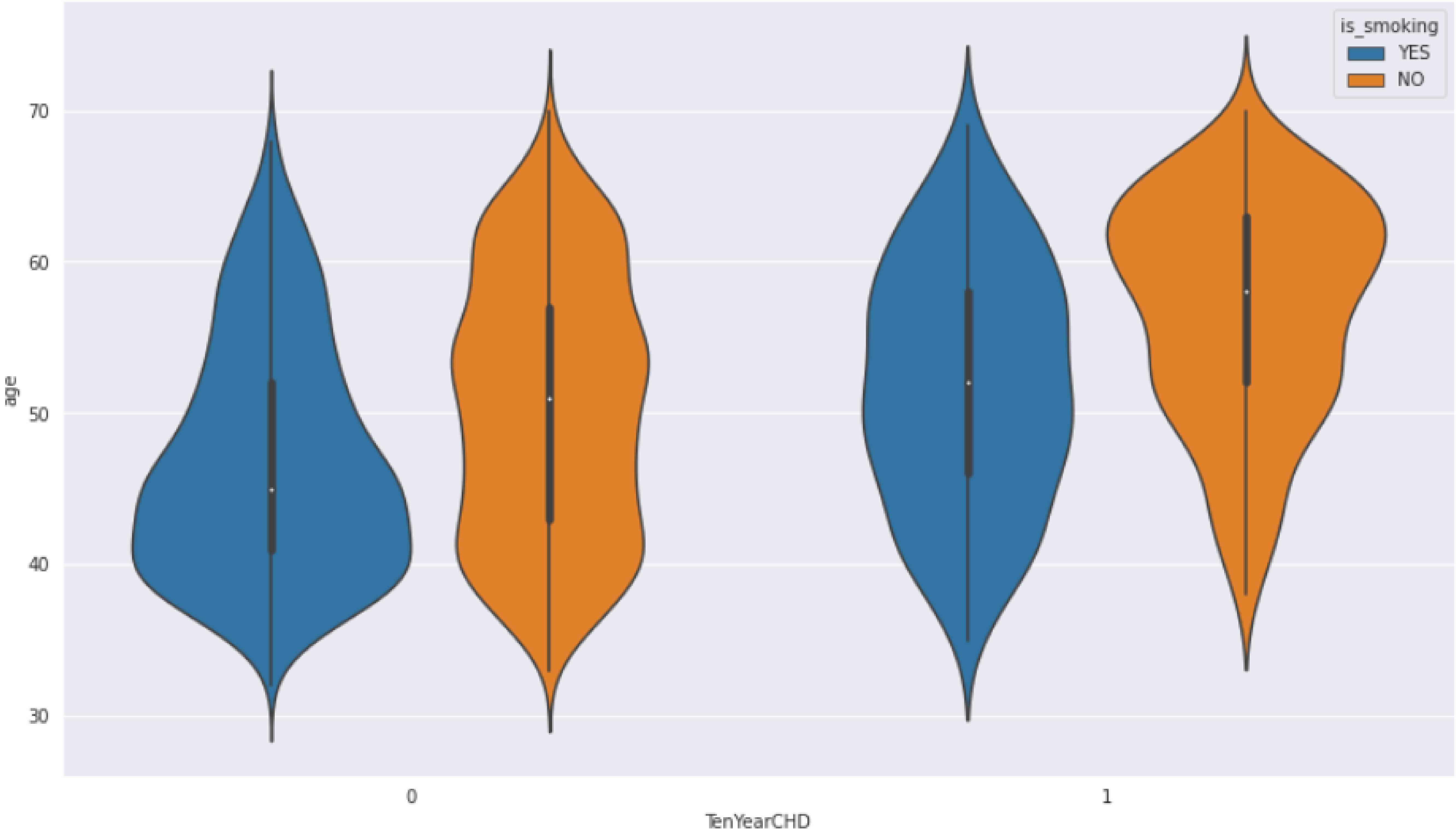
Heartrate

- More cholestral level are between 100-400.
- Glucose level are mostly left skewed between 50-150.
- Heartrate represent a better normal curve tilted leftwards a little.

Distribution of sysBP and diaBP.

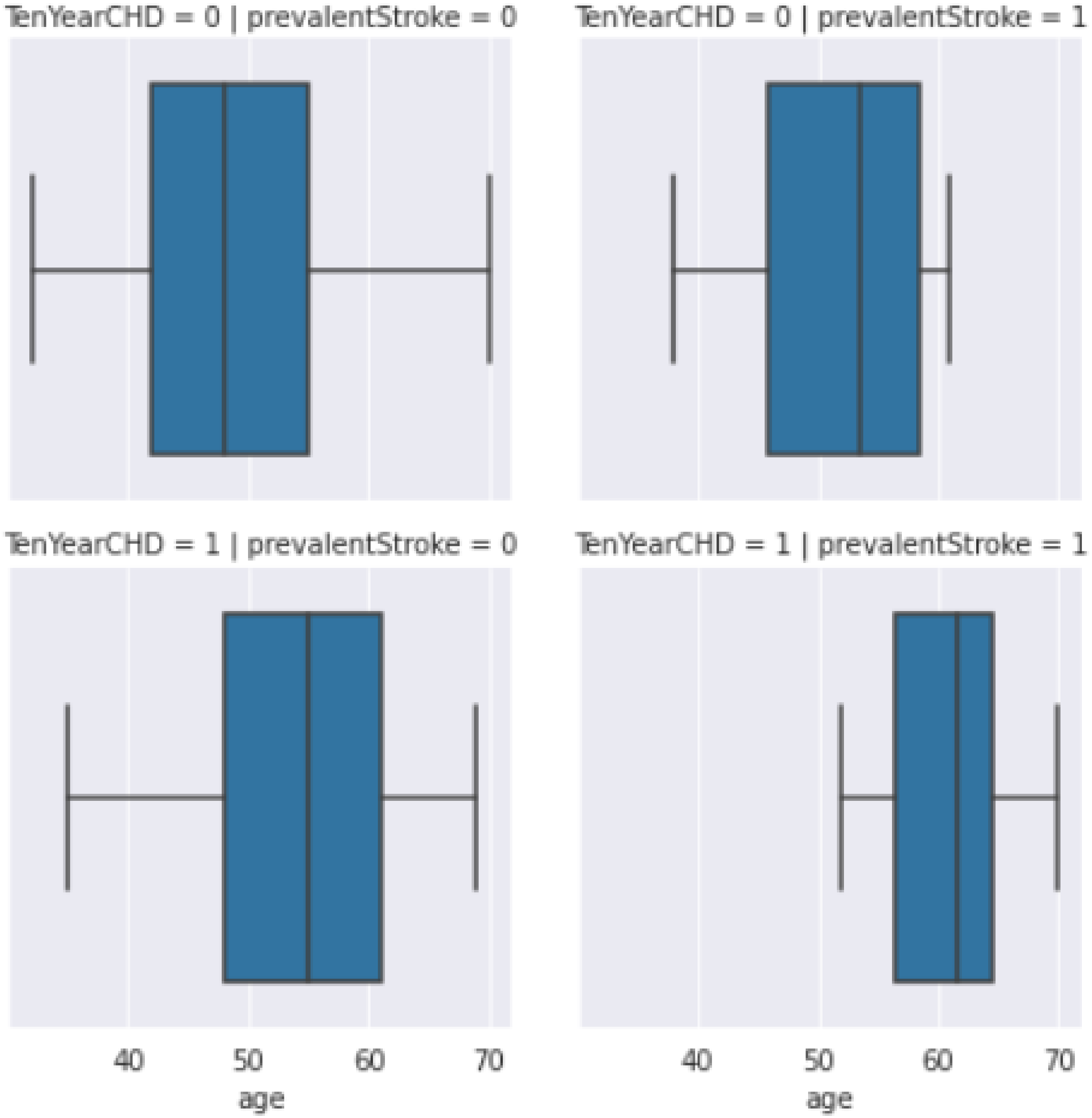


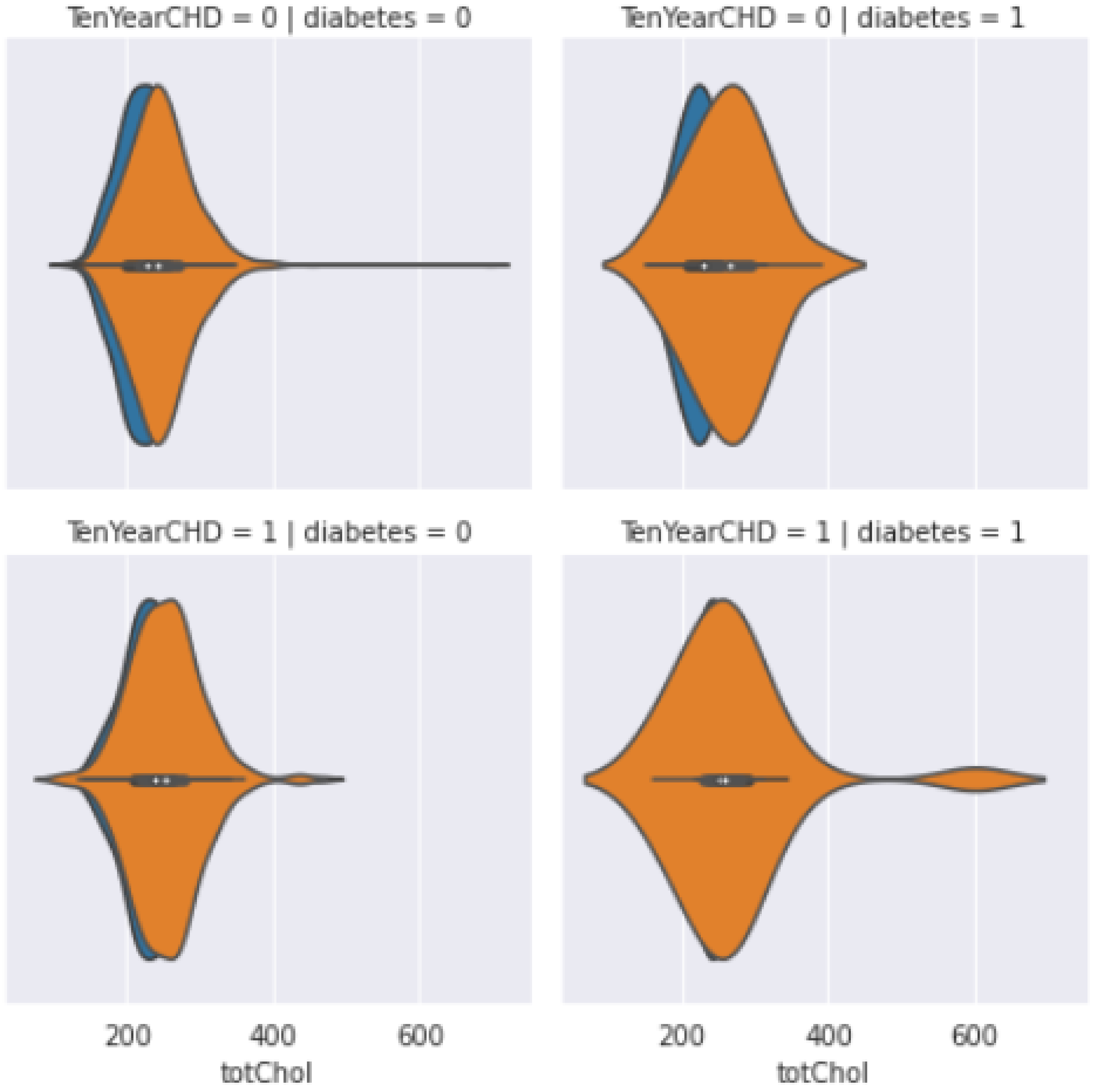
Violin plot of distribution of age on basis of smoking and cardiovascular disease.

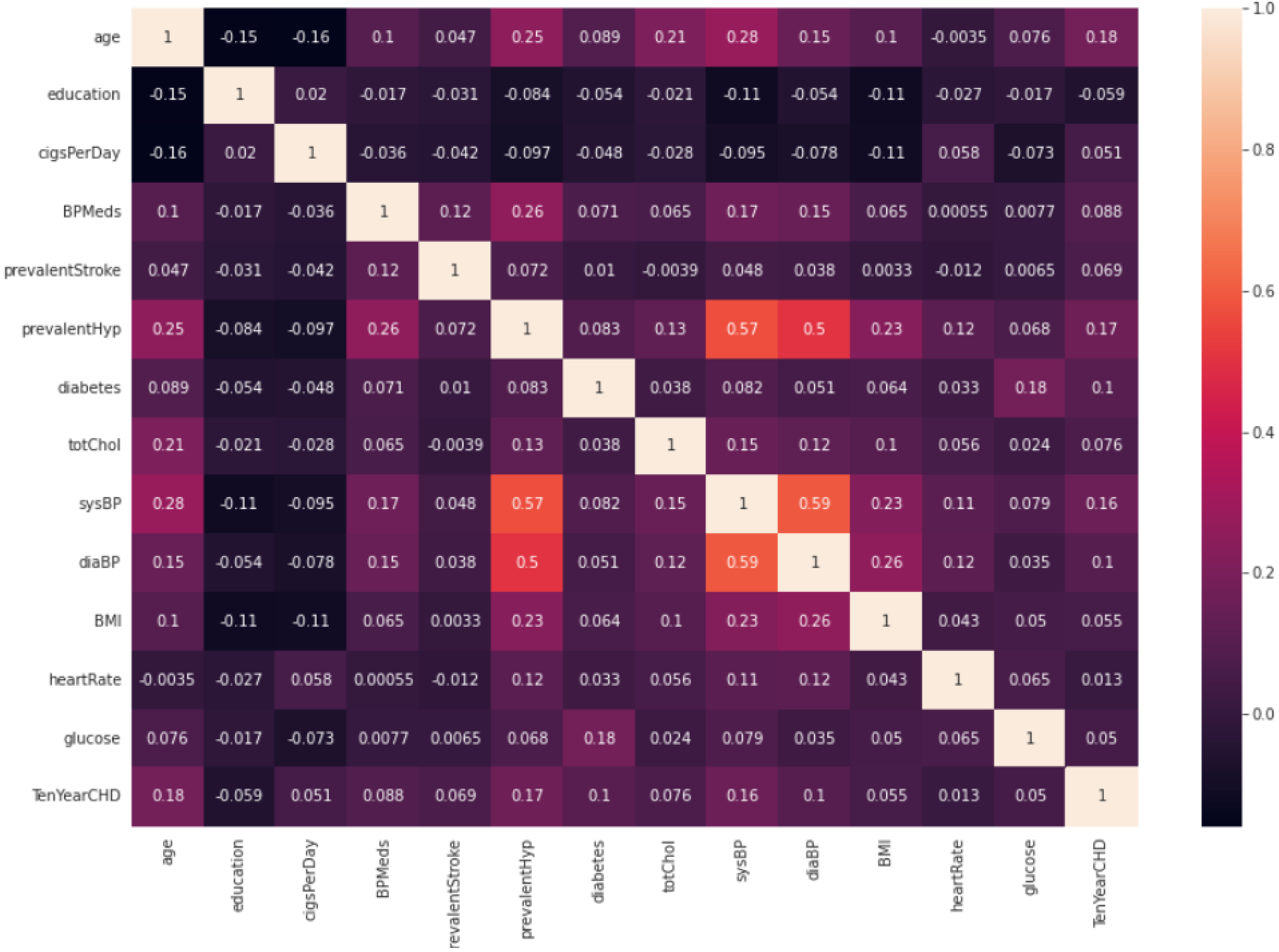


- More people above age 60 with diabetes has more chance of getting cardiovasular disease.
- A major distribution below age 40 has no type of disease.

Distribution of age and cholestrol on basis of stroke and diabetes alog with cardiovascular disease.







	age	education	cigsPerDay	BPMeds	revalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age	1	-0.15	-0.16	0.1	0.047	0.25	0.089	0.21	0.28	0.15	0.1	-0.0035	0.076	0.18
education	-0.15	1	0.02	-0.017	-0.031	-0.084	-0.054	-0.021	-0.11	-0.054	-0.11	-0.027	-0.017	-0.059
cigsPerDay	-0.16	0.02	1	-0.036	-0.042	-0.097	-0.048	-0.028	-0.095	-0.078	-0.11	0.058	-0.073	0.051
BPMeds	0.1	-0.017	-0.036	1	0.12	0.26	0.071	0.065	0.17	0.15	0.065	0.00055	0.0077	0.088
prevalentStroke	0.047	-0.031	-0.042	0.12	1	0.072	0.01	-0.0039	0.048	0.038	0.0033	-0.012	0.0065	0.069
prevalentHyp	0.25	-0.084	-0.097	0.26	0.072	1	0.083	0.13	0.57	0.5	0.23	0.12	0.068	0.17
diabetes	0.089	-0.054	-0.048	0.071	0.01	0.083	1	0.038	0.082	0.051	0.064	0.033	0.18	0.1
totChol	0.21	-0.021	-0.028	0.065	-0.0039	0.13	0.038	1	0.15	0.12	0.1	0.056	0.024	0.076
sysBP	0.28	-0.11	-0.095	0.17	0.048	0.57	0.082	0.15	1	0.59	0.23	0.11	0.079	0.16
diaBP	0.15	-0.054	-0.078	0.15	0.038	0.5	0.051	0.12	0.59	1	0.26	0.12	0.035	0.1
BMI	0.1	-0.11	-0.11	0.065	0.0033	0.23	0.064	0.1	0.23	0.26	1	0.043	0.05	0.055
heartRate	-0.0035	-0.027	0.058	0.00055	-0.012	0.12	0.033	0.056	0.11	0.12	0.043	1	0.065	0.013
glucose	0.076	-0.017	-0.073	0.0077	0.0065	0.068	0.18	0.024	0.079	0.035	0.05	0.065	1	0.05
TenYearCHD	0.18	-0.059	0.051	0.088	0.069	0.17	0.1	0.076	0.16	0.1	0.055	0.013	0.05	1

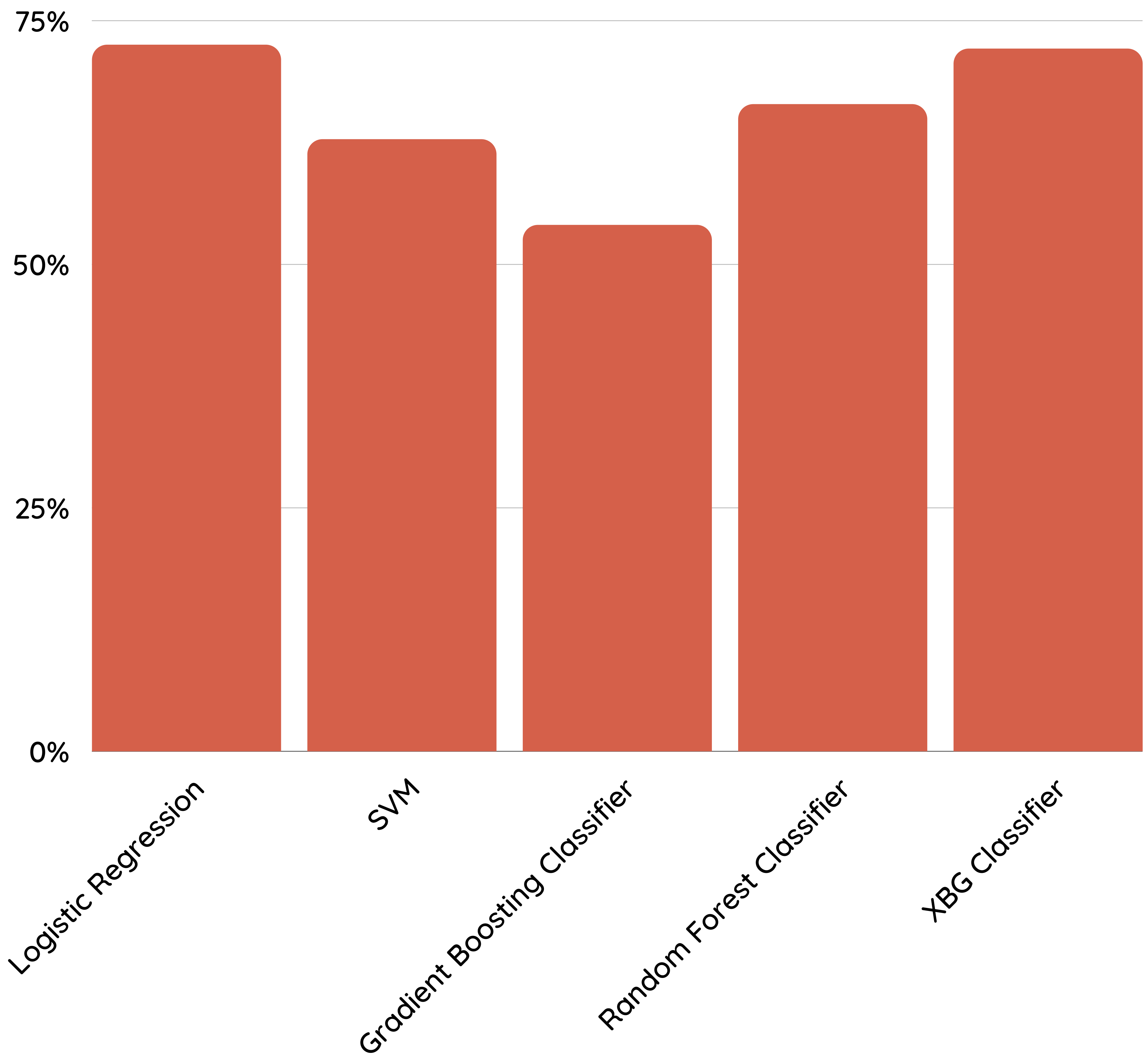
Correlation map between each columns

Model Evaluation

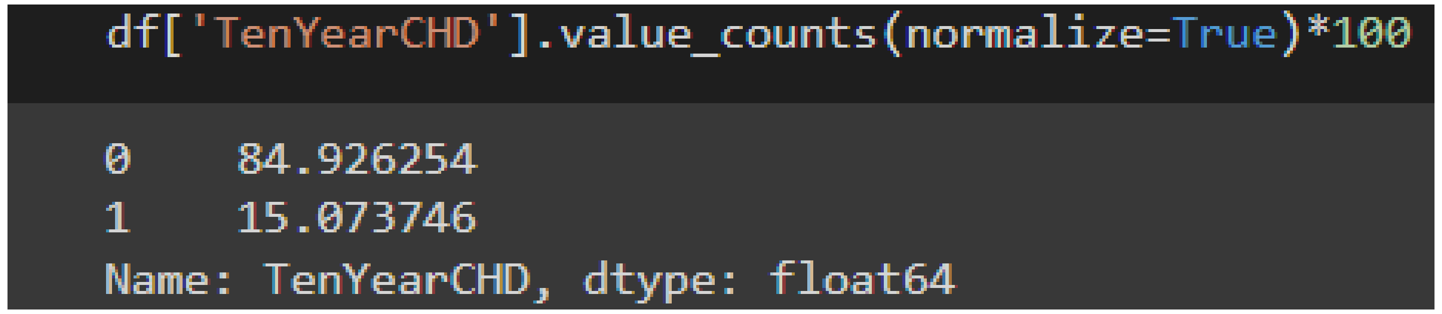
Method:We preprocess the data to train the dataset with different classification models:

- Regularized Logistic Regression;
- Support Vector Machine;
- Random Forest Classifier;
- Gradient Boosting Classifier.
- Linear Discriminant Analysis.

Comparison of different algorithm model on basis of roc-auc score:-



- **Logistic Regression has the better accuracy but 71% is too low .**
- **Since target column is imbalanced we have to make it balanced first to improve the accuracy of it.**



```
df['TenYearCHD'].value_counts(normalize=True)*100

0    84.926254
1    15.073746
Name: TenYearCHD, dtype: float64
```

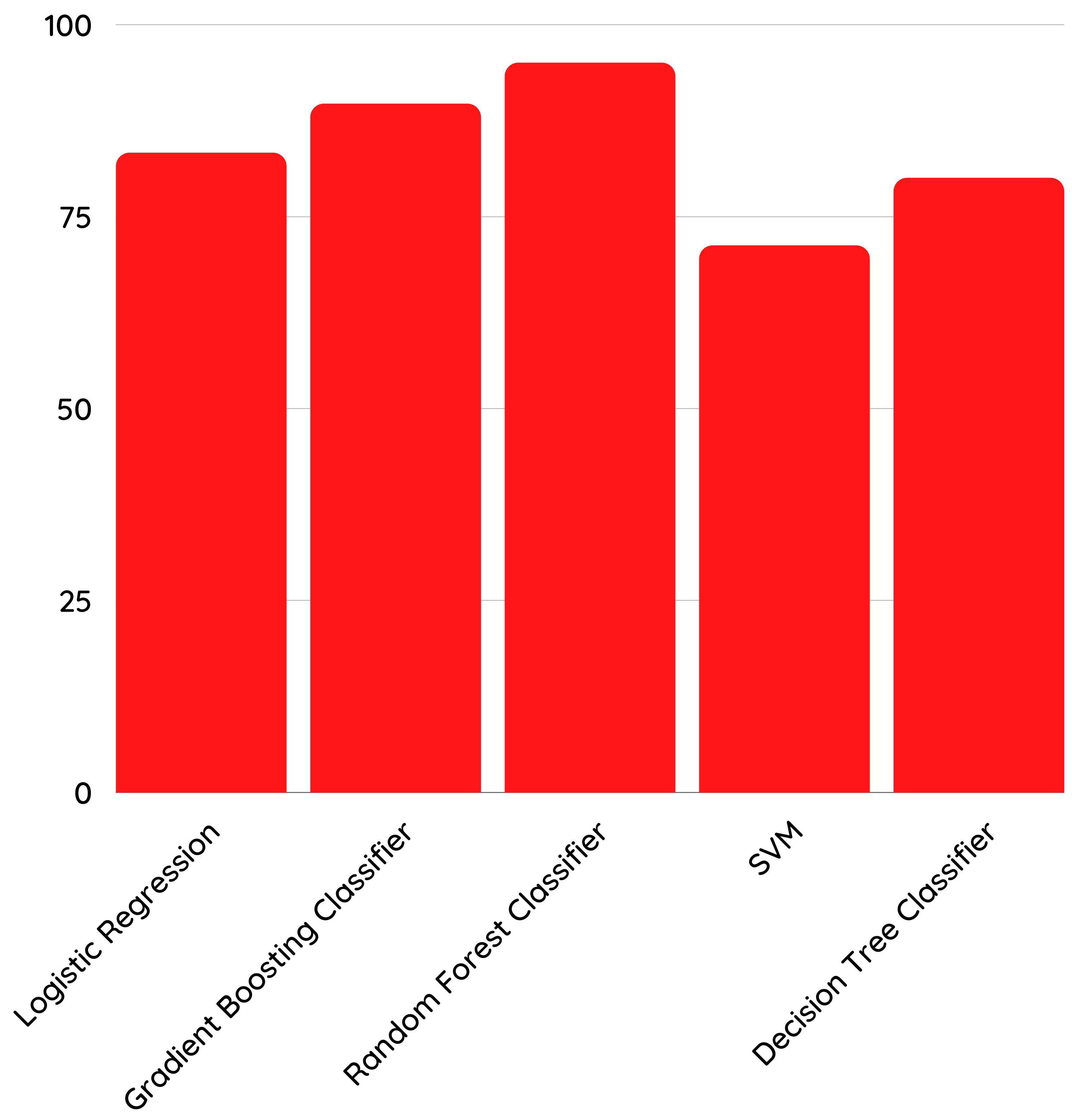
85:15

Handling imbalanced dataset:-

Standard machine learning techniques, such as Decision Tree and Logistic Regression, favour the majority class and neglect the minority. They tend to anticipate just the majority class, resulting in significant misclassification of the minority class in comparison to the majority.

- One of the most widely used oversampling approaches to overcome the imbalance problem is SMOTE (synthetic minority oversampling technique).
- Its goal is to achieve a more balanced distribution of classes by replicating minority class examples at random.
- The data is reconstructed after the oversampling procedure, and many classification models can be applied to the processed data.

Accuracy after Resampling of data:-



- Random Forest and Gradient Boosting Classifier are two of the models with most number of accuracy.

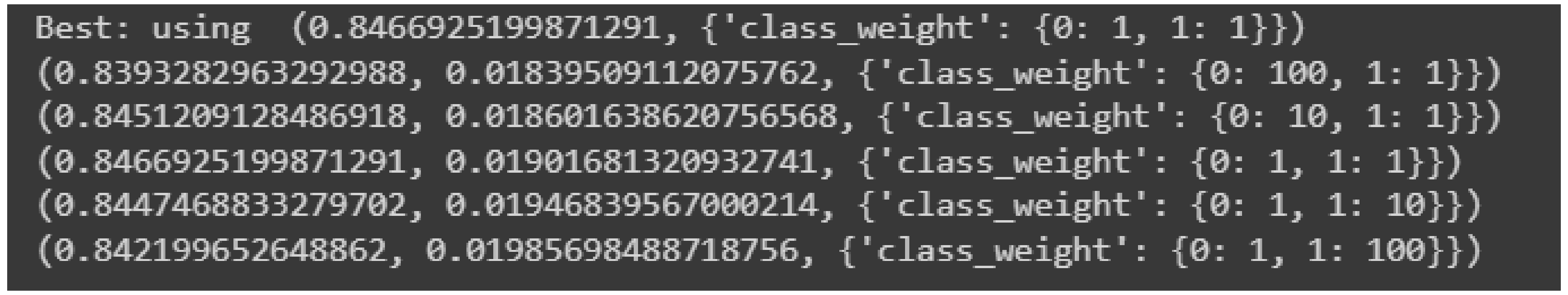
- But after comparing it against Logistic Regression it precision against minority is way less .

Hyperparameter Tuning of model

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins.

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. Two best strategies for Hyperparameter tuning are:

Hyperparameter tuning of Logistic Regression

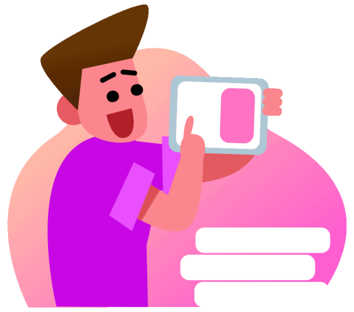


```
Best: using  (0.8466925199871291, {'class_weight': {0: 1, 1: 1}})
(0.8393282963292988, 0.01839509112075762, {'class_weight': {0: 100, 1: 1}})
(0.8451209128486918, 0.018601638620756568, {'class_weight': {0: 10, 1: 1}})
(0.8466925199871291, 0.01901681320932741, {'class_weight': {0: 1, 1: 1}})
(0.8447468833279702, 0.01946839567000214, {'class_weight': {0: 1, 1: 10}})
(0.842199652648862, 0.01985698488718756, {'class_weight': {0: 1, 1: 100}})
```

In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

Conclusion:-

- There were few missing values in the cardiovascular disease dataset, which were mostly replaced by the mean value.
- Most people only drink one or two cigarettes, but there was a time when the group drank up to 20 cigarettes. Max weighed around 70 cigarettes.
- Men without diabetes consume at least 14 cigarette every day.
- Men on blood pressure medication, on the other hand, drink at least 10 cigarettes a day.)
- Stroke patients frequently refuse to drink cigarettes.
- Finally, we may conclude that cigarettes do not have a significant role in cardiovascular disease, as those under the age of 60 who drink cigarettes have a lower risk of developing the condition than people over the age of 60 who do not.
- Because the target was unbalanced, we had to balance it using the SMOTE approach. In the end, Logistic Regression provided superior precision than both variables.



THANK YOU