Cardiovascular Risk Prediction

Sahil pardeshi, Pravin Bejjo, Kirtesh verma
Data science Trainee, almabetter Nashik

Introduction : Cardiovascular disease (CVD) is defined as any serious, abnormal condition of the heart or blood vessels(arteries, veins). Cardiovascular disease includes coronary heart disease (CHD), stroke, peripheral vascular disease, congenital heart disease, endocarditis, and many other conditions.

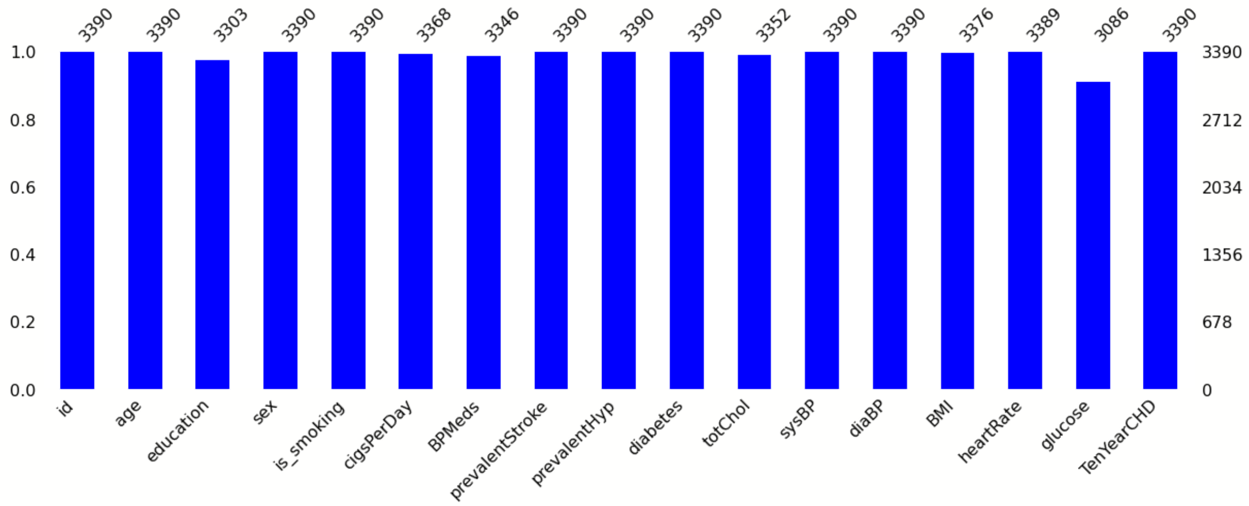
A type of disease that affects the heart or blood vessels. The risk of certain cardiovascular diseases may be increased by smoking, high blood pressure, high cholesterol, unhealthy diet, lack of exercise, and obesity.

Machine learning technique to develop screening tools. In this project we shall giving a path to the work through the development of a screening tool for predicting whether a patient has 10 years of developing coronary heart diseases(CHD) based on their present health condition using machine learning techniques.

Problem Statement : Heart disease is the leading cause of morbidity and mortality worldwide, killing more people each year than any other cause. In this project, we shall be giving you a walk through on the development of a screening tool for predicting whether a patient has a 10-year risk of developing coronary heart disease (CHD) using different Machine Learning techniques. The given dataset provides the patients' information. It includes over 3,390 records and 17 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors given for the analysis.

Visualizing Missing Data :

0



□ **Step Involved :**

1) Performing EDA and Data Preprocessing:

i) Exploring Head and tail of the data to get insights on the given data.

ii) Checking the Null values or missing values are present in the dataset or not.

iii) Check the duplicate values.

iv) Check the minority and majority class.

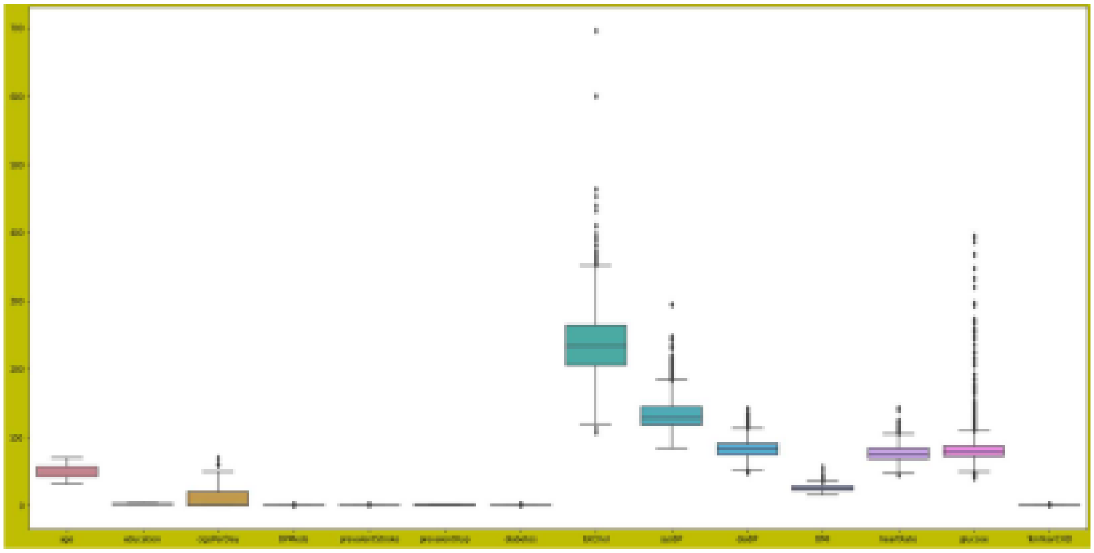
v) KNNImputer shall be used to impute the NaN values for continuous data.

vi) SimpleImputer shall be used to impute the NaN values for categorical data.

vii) missing_value_continuous function to handle missing values of continuous data.

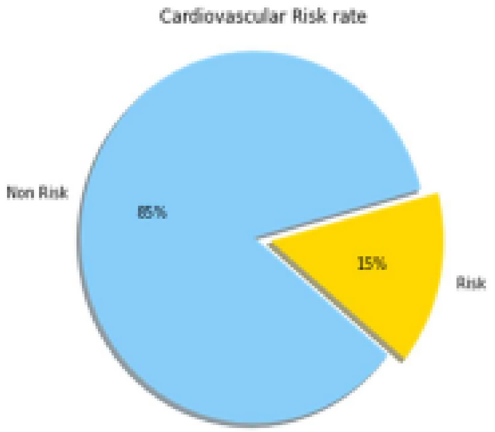
viii) variables missing_value_categorical function to handle missing values of categorical data.

Outlier Treatment:

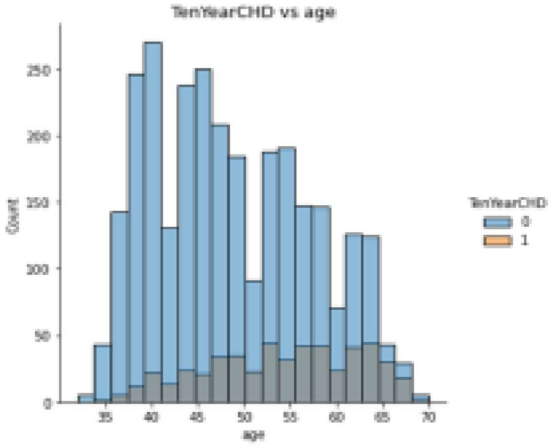


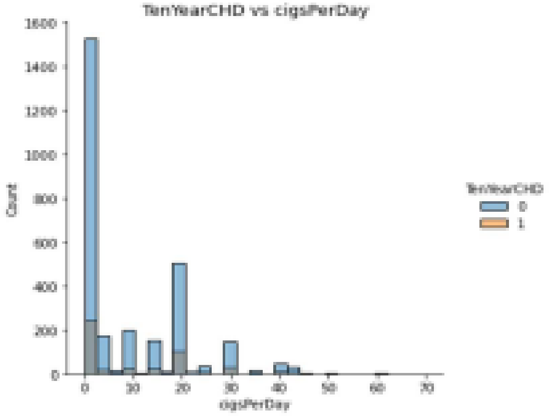
Drawing conclusion from the data :

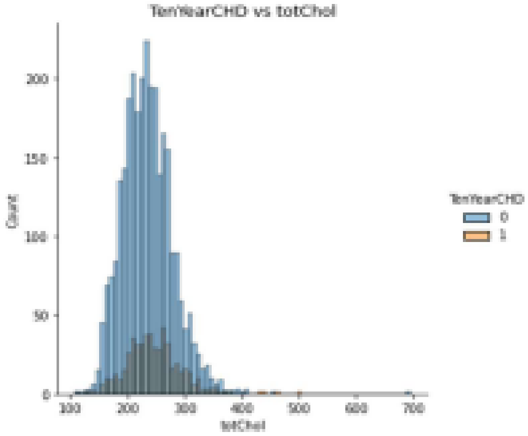
1) **Plotting the majority and minority sets of targeted variables.**

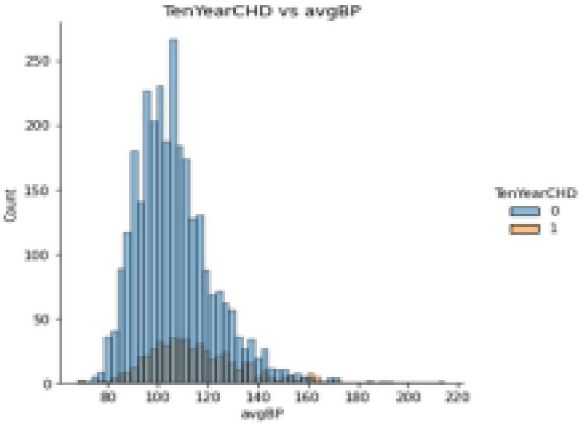


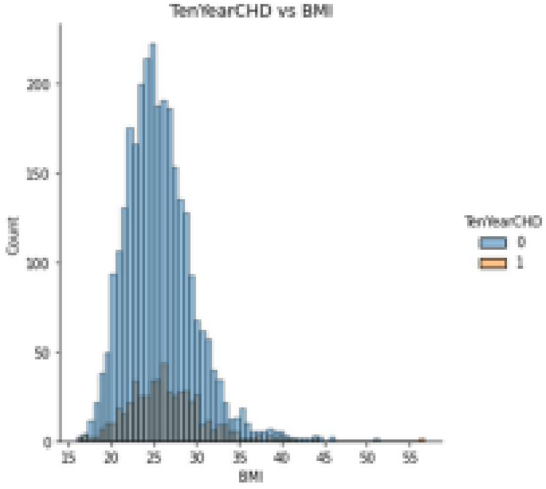
2) EDA on Continuous features Analysis :

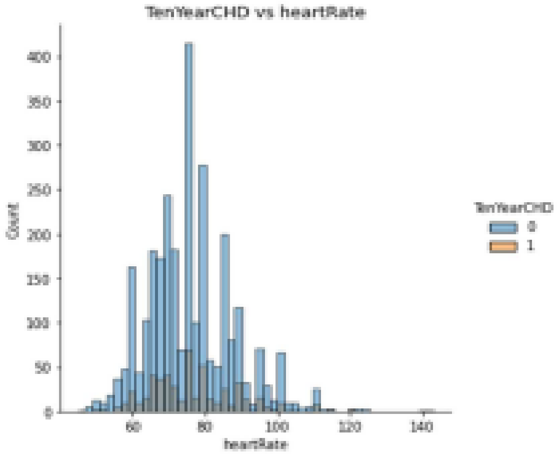


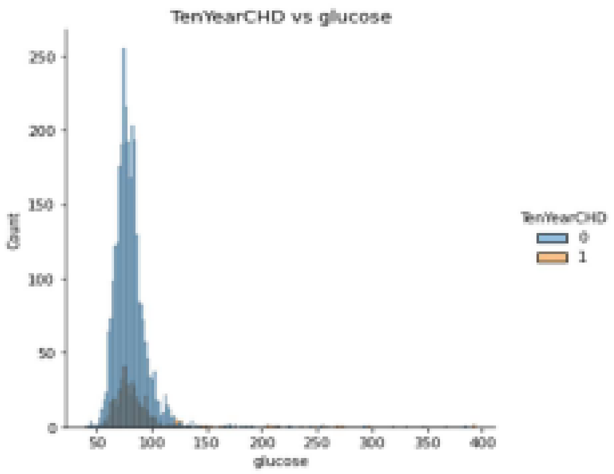



TenYearCHD vs totChol
Count
0
50
100
150
200
100
200
300
400
500
600
700
totChol
TenYearCHD
0
1


TenYearCHD vs avgBP
Count
0
50
100
150
200
250
80
100
120
140
160
180
200
220
avgBP
TenYearCHD
0
1

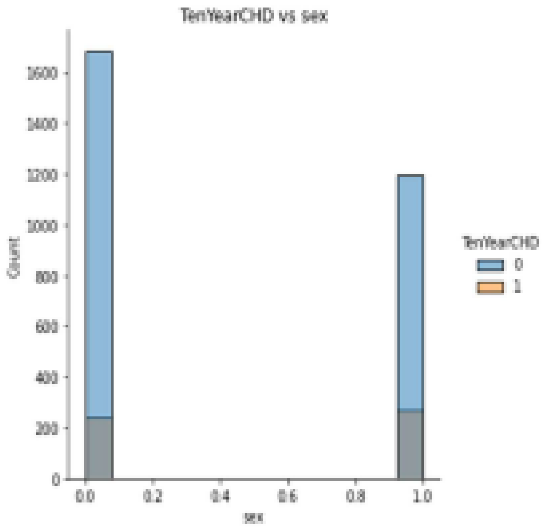

TenYearCHD vs BMI
Count
0
50
100
150
200
15
20
25
30
35
40
45
50
55
BMI
TenYearCHD
0
1

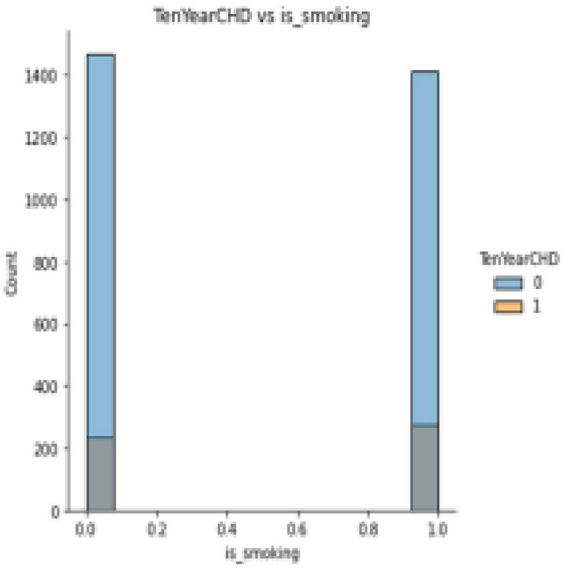

TenYearCHD vs heartRate
Count
0
50
100
150
200
250
300
350
400
60
80
100
120
140
heartRate
TenYearCHD
0
1

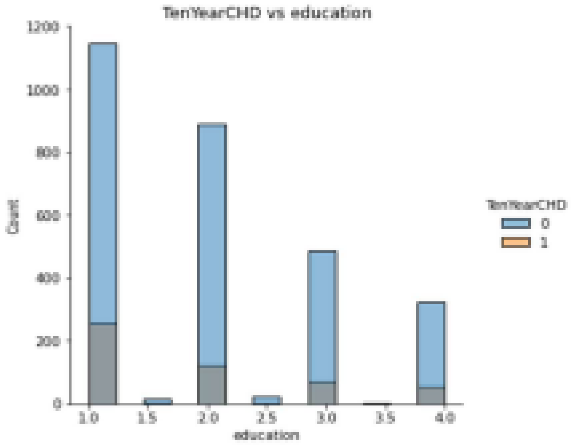


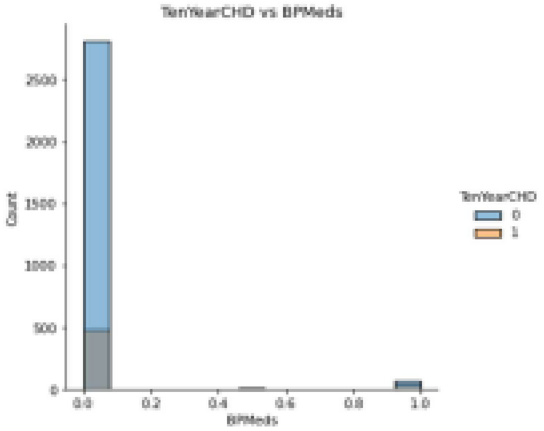
Cigsperday is not following gaussian/ Normal distribution and from the displot as well as kde plot.

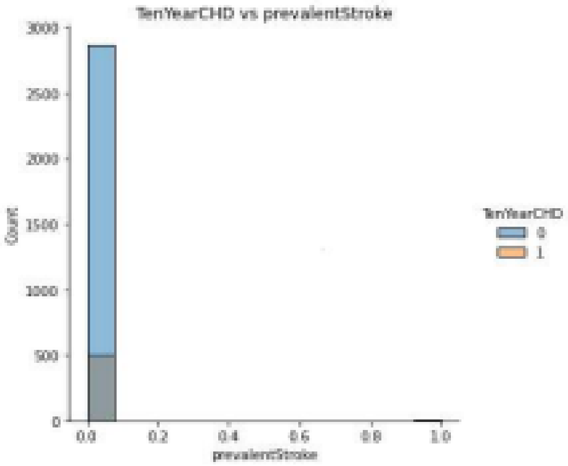
3) EDA on discrete features :

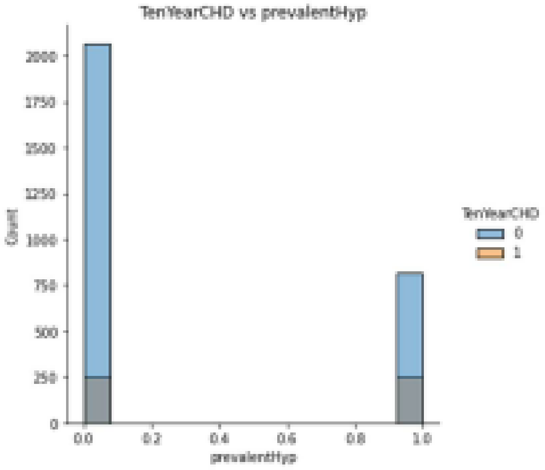


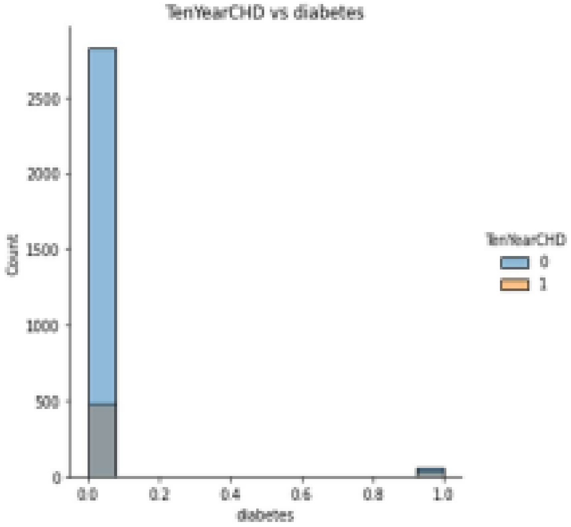

TenYearCHD vs is_smoking
Count
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0
is_smoking
TenYearCHD
0
1


TenYearCHD vs education
Count
0
200
400
600
800
1000
1200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
education
TenYearCHD
0
1

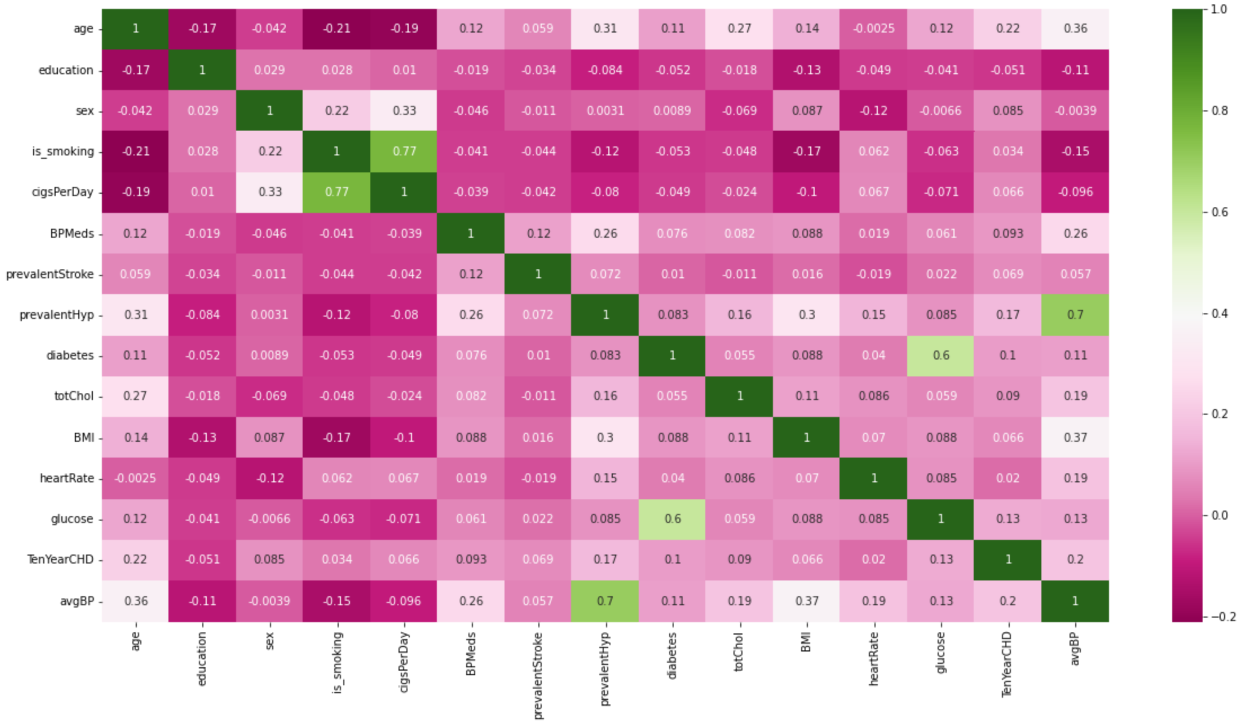

TenYearCHD vs BPMeds
Count
0
500
1000
1500
2000
2500
0.0
0.2
0.4
0.6
0.8
1.0
BPMeds
TenYearCHD
0
1


TenYearCHD vs prevalentStroke
Count
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0
prevalentStroke
TenYearCHD
0
1

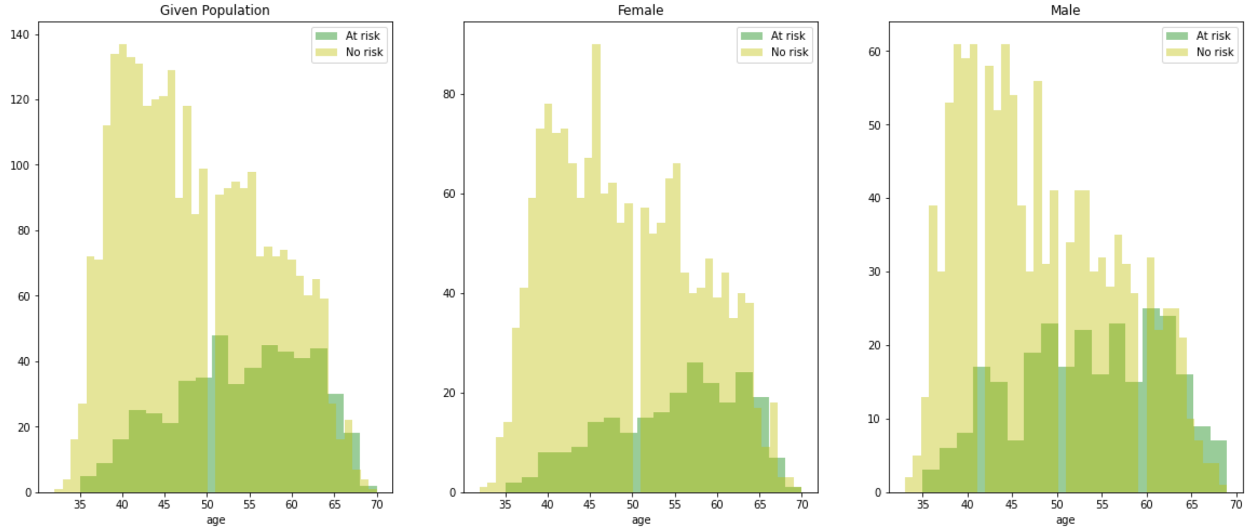

TenYearCHD vs prevalentHyp
Count
0
250
500
750
1000
1250
1500
1750
2000
0.0
0.2
0.4
0.6
0.8
1.0
prevalentHyp
TenYearCHD
0
1

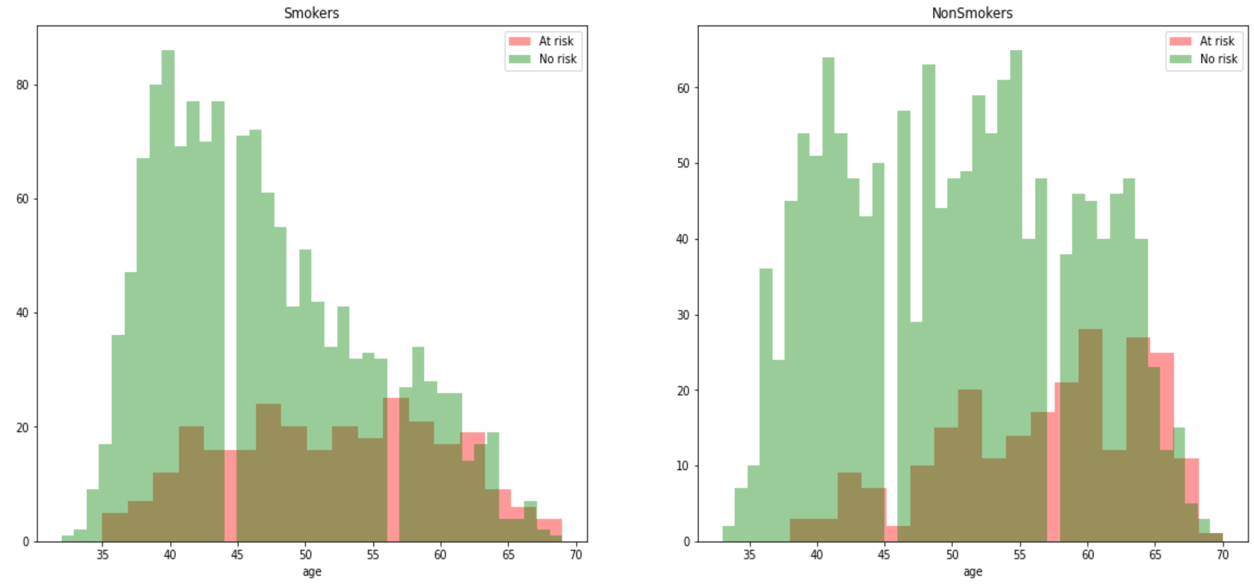

TenYearCHD vs diabetes
Count
0
500
1000
1500
2000
2500
0.0
0.2
0.4
0.6
0.8
1.0
diabetes
TenYearCHD
0
1

4)Correlation Heat Map :

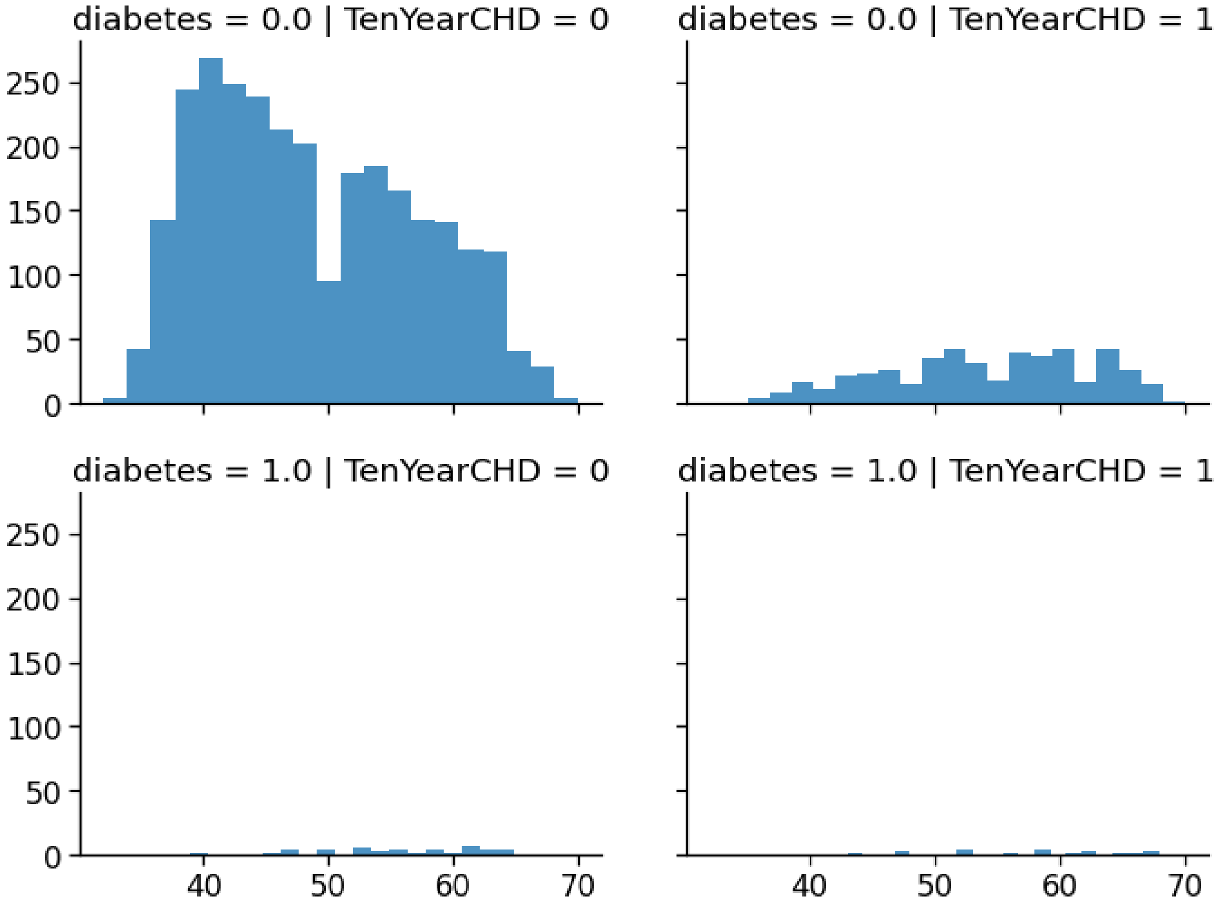


5)Age & sex vs Risk :





6)Diabetes & age vs risk :



- □ **Data modelling :**
  Data Modelling is the process of analyzing the data objects and their relationship to the other objects. It is used to analyze the data requirements that are required for the business processes. The data models are created for the data to be stored in a database.

- □ **Approach :** **we have checked the outliers and correlation matrix to overcome the noise in the datasets.The data was balanced using SMOTE data. We have used various techniques. As the CHD datasets defined the classification problems. We have decided to train the model such as Logistic regression, K nearest Neighbors, Decision Tree Classifier, Support Vector Machine, Random forest & Gradient boosting. Also, we used Hyperparameter Tuning for improvement in the model fitting to understand the better results of the model as well as the metrics.**

- □ **Scaling the dataset :** **we have used a standard scalar method to scale the dataset.**

- □ **Metric used :**
- □ **Classification Report: A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model.**
- □ **Accuracy: the proportion of total dataset instances that were correctly predicted out of the total instances**
- □ **Recall (sensitivity): the proportion of the predicted positive dataset instances out of the actual positive instances**
- □ **sensitivity=true positives/ (true positives+false negatives)**
- □ **F1 score: a composite harmonic mean (average of reciprocals) that combines both precision and recall. For this, we first measure the precision, the ability of the model to identify only the relevant dataset instances**
- □ **precision=true positives/ (true positives+false positives)**
- □ **The F1 score is estimated as**
- □ **F1=2×(precision×recall)/(precision+recall)**

Models :

1) **Logistic Regression:** **Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.**
2) **K- nearest neighbor :** **The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression**

problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

3) **Support vector machine :** Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well, it's best suited for classification. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.
4) **Decision tree classifier :** The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.
5) **Random forest :** Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.
6) **Gradient Boosting :** Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

□ **Challenges faced** :1) using appropriate metrics for comparison implemented machine learning algorithms.

2) Preprocessing datasets was one of the challenges.

□ **Conclusion :**

1) Male smokers are higher than females.
2) XGBoost performs the best amongst all other models with highest accuracy and f1 score.
3) Male and female both are equally prone to CHD.
4) People who suffer from previously heart attack had a chance to CHD