Cardiovascular Risk Prediction

Samiksha Bandbuche
Data science trainees,
Almabetter, Bangalore

Abstract:

The Framingham Heart Study is a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham, Massachusetts. The study began in 1948 with 5,209 adult subjects from Framingham and is now on its third generation of participants. Prior to the study almost nothing was known about the epidemiology of the hypertensive or arteriosclerotic cardiovascular disease. Much of the now-common knowledge concerning heart diseases, such as the effects of diet, exercise, and common medications such as aspirin, is based on this longitudinal study. It is a project of the National Heart, Lung, and Blood Institute, in collaboration with (since 1971) Boston University. Various health professionals from the hospitals and universities of Greater Boston staffed the project.

1. 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 3,000 records and 17 attributes.

2. Dataset 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.
- 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 a 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") – DV

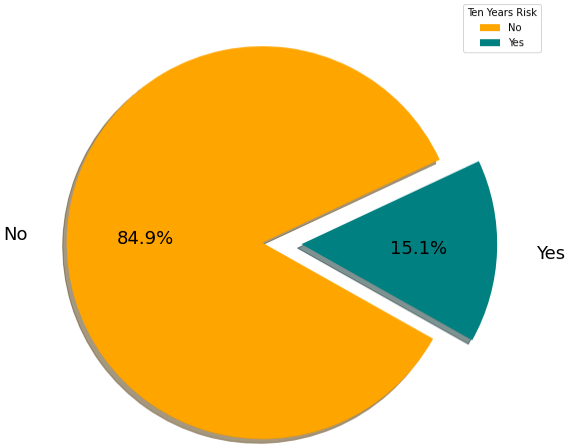
3. Missing Value Treatment:

Missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Continuing to that we found missing observations in seven columns which we further treated with the median value that corresponds to that column.

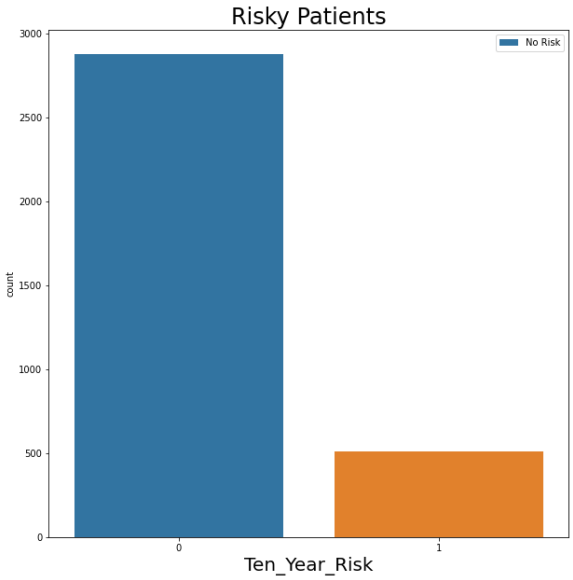
4.Exploratory Data Analysis (EDA):

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.
This section consists of details regarding the visual results:

4.1 Analysis Of Risky Patients:

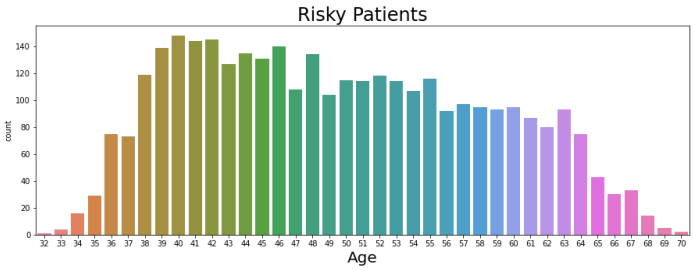


Here we can see there are 15.1% of people in our dataset are is risk for cardiovascular disease(ten-year risk)and 84.9% of people are safe.



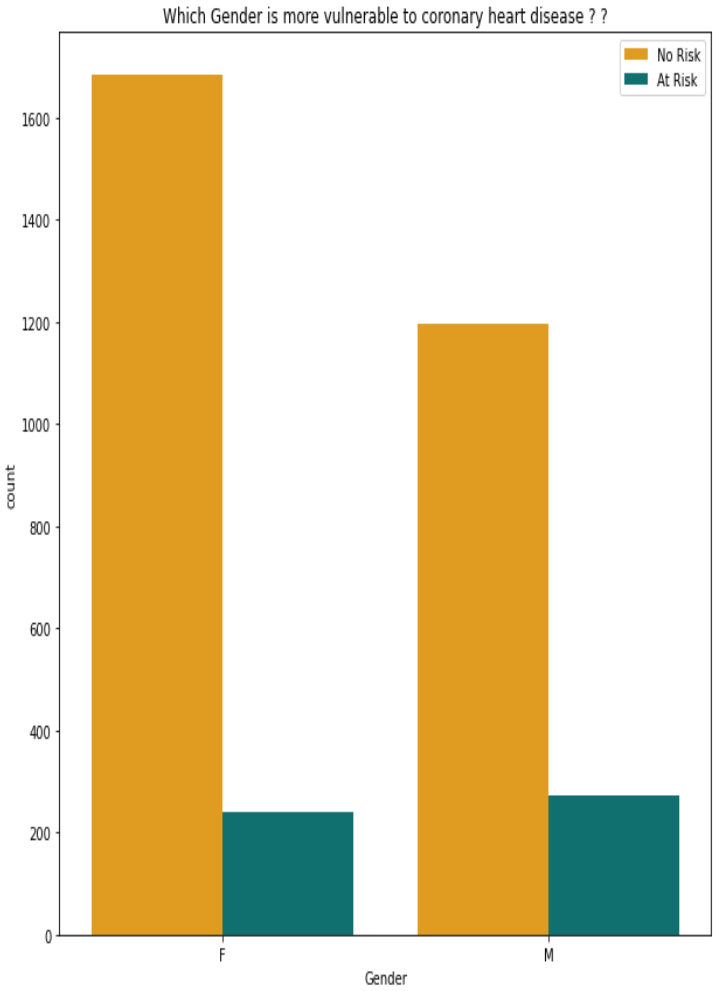
Above we can see the count of risky patient, around 500 patients are in risk and approx. 2800 patient are safe.15% data of one class and 85% of another class, this shows imbalance between the class.Goal of any classification algorithm wants to increase the accurcy by reducing the error. Thus, does not take into account of class proportion or Imbalaanced class.Imagine a scenario, here our model caps all observation as 0, though having accuracy of 85%. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored.This kind of misclassification will encountere in situation like, Crucial Disease detection, fraud detection in bank, Telecomm chrun analysis.Lets explain that situation in our task, if our model classifies a sample as risk, but actuall classification is Not risky. So in this sitution sample will get observation.But in case our model classified risky patient as not risky, will be the issue.We have balance the dataset before feeds to the model.

4.2 Analysis By Age Group:



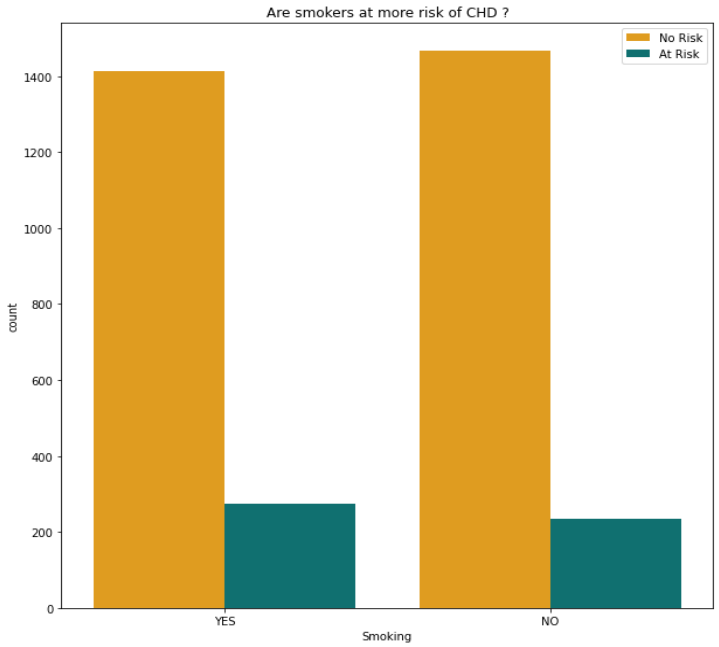
Here we can see there is more risk of cardiovascular disease in patients of age between 51 to 63.

4.3 Analysis By Gender:



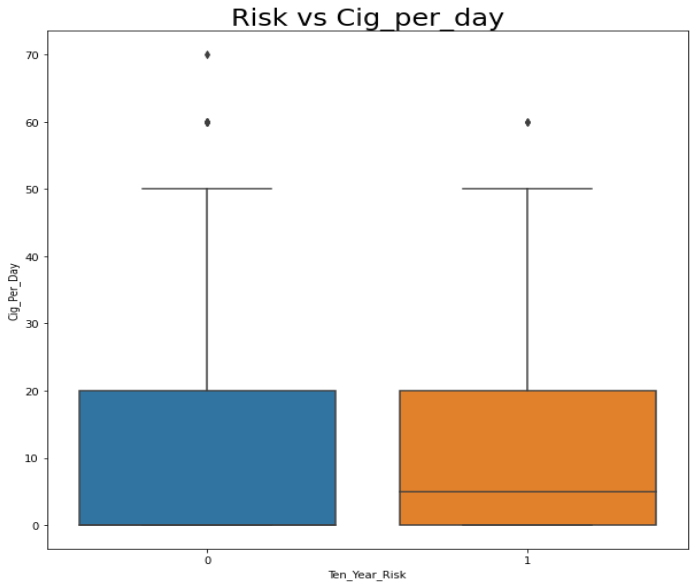
Here we can see the count of males and females are the same in risk, which is around 200 though females are more than males in our dataset.

4.4 Analysis Of Smokers:



Here we can see around 250 smokers are at risk and around 210 non-smokers are the risk for cardiovascular diseases. After seeing this we cannot directly say that only smokers are at risk.

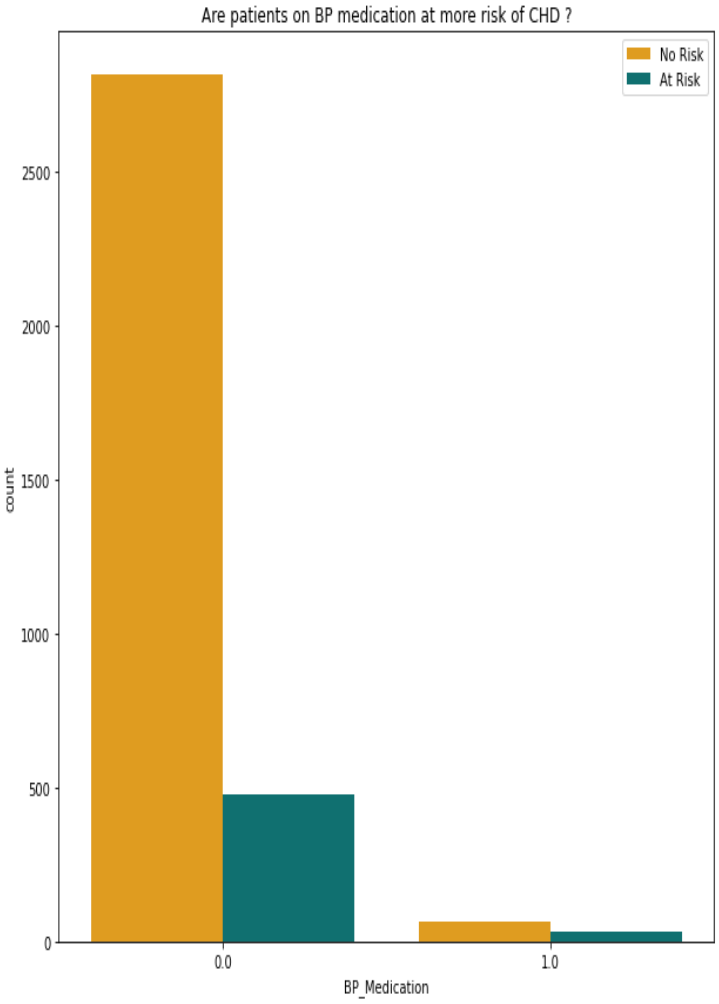
4.5 Analysis on Cigarettes:



From above information we cant evedentially state smoking will leads to haert disease, as we seen from count plot there is no huge difference between these to comune and also our extreme

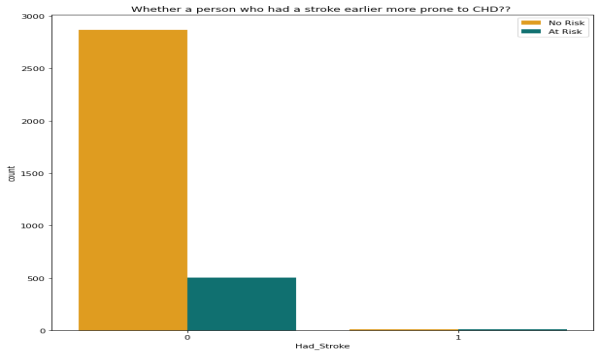
smkoker who smokes 70 cigarettes per day is not having ten year risk.

4.6 Analysis On BP Patients:



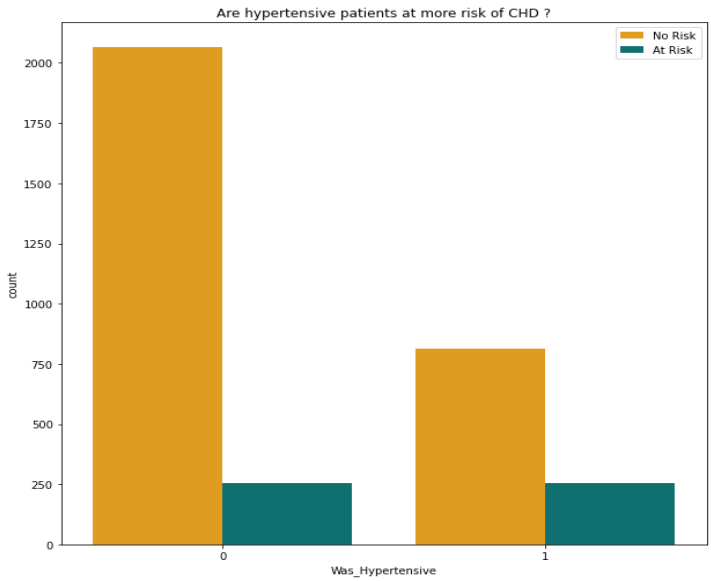
Here we can see there are very few people who are done with BP medication which is around 200 but many people have not taken any BP medication and they are around 3200. After seeing this trend we cannot say that after taking medication there is no risk.

4.7 Analysis Of Stroke Patients:



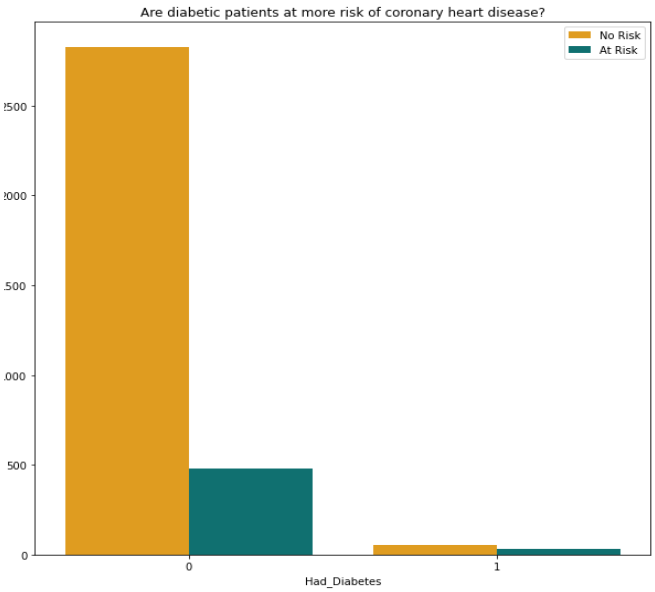
There are very few people had stroke.Here around 500 patients who did not had stroke yet and are at risk and around 2800 patients are safe.

4.8 Analysis Of Hypertensive Paitents:



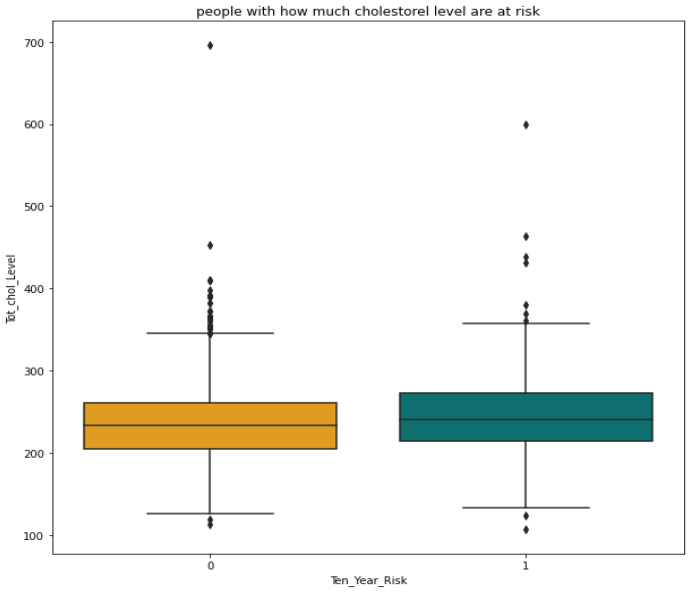
Here around 250 people with hypertension are at risk and around 255 people with no hypertension are at risk.

4.9 Analysis Of Diabetic Patients:



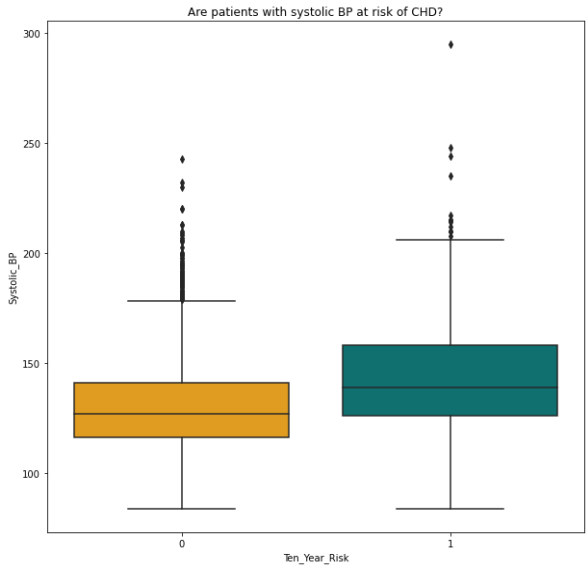
Here, we can see around 500 people who did not have diabetes are at risk.And there are very few people who had diabetes are at risk.So diabetes feature is not helping that much in ten years risk.

4.10 Analysis On Cholesterol Level:



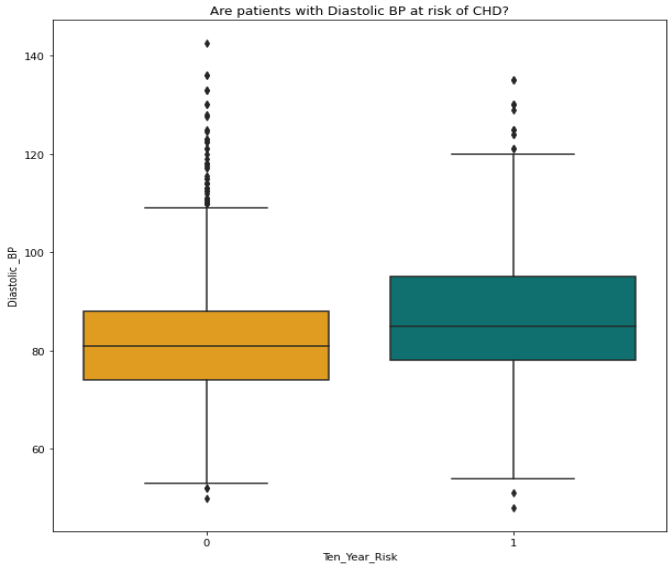
Here we can see most of the people who are not at risk their Cholesterol level lies between 210 to 280 and for people who are at risk their cholesterol level lies between 215 to 285 there in no huge difference it is quite normal.

4.11 Analysis On Systolic BP Paitents:



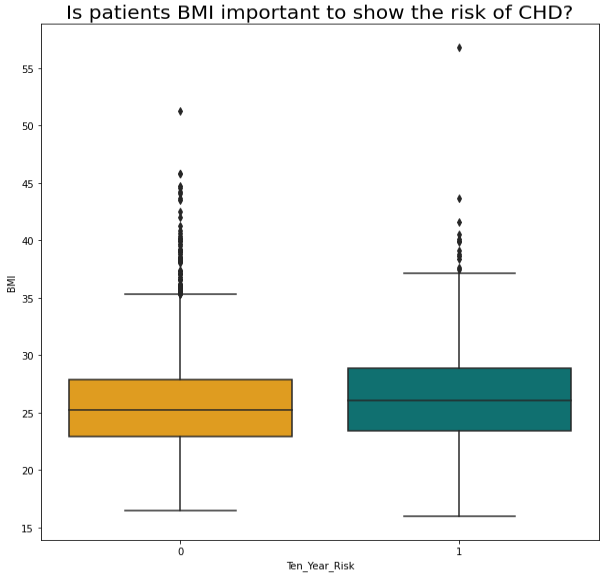
Here we can see most people who are not at risk their systolic BP lies between 110 to 140 and for people who are at risk their systolic BP lies between 125 to 160.So, we can say that people with high systolic BP are at risk.

4.12 Analysis On Dialostic BP Patients:



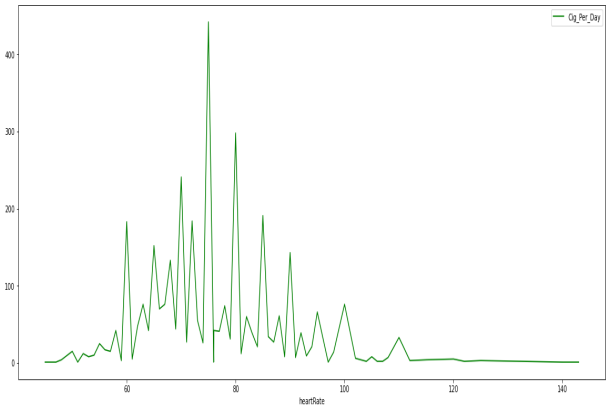
Here we can see most people who are not at risk their diastolic BP lies between 75 to 85 and people who are at risk their diastolic BP lies between 89 to 90. So, we can say there is a slight increase in diastolic BP of people who are at risk.

4.13 Analysis of Paitent's BMI:



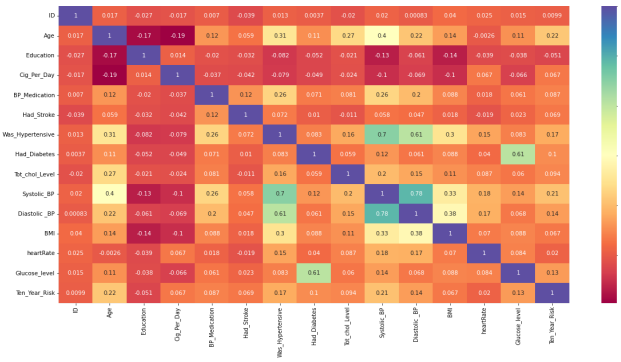
Here we can see most people who are not at risk their BMI lies between 22 to 28 and people who are at risk their BMI lies between 23 to 29 approx.We cannot see much difference in BMI is approx the same for risky and not risky people.

4.13 Analysis on Heart Rate:



Here we can see most people smoke cigarettes between 1 to 10 approx. and their heart rate lies between 60 to 100.

4.14 Correlation Analysis



Here we can see systolic BP and was hypertensive variable are correlated, diastolic BP and was hypertensive variable are correlated, Glucose level and diabetes variable are correlated with diastolic BP and systolic BP are highly Correlated.

5. Model Building:

In this project, we used machine learning algorithms to get better result and to know which machine algorithm will be the best fit for this project.Machine Learning we used in this project are as follows:

5.1 Logistic Regression:

Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or interval type.

5.2 Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

5.3 Random Forest Classifier:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

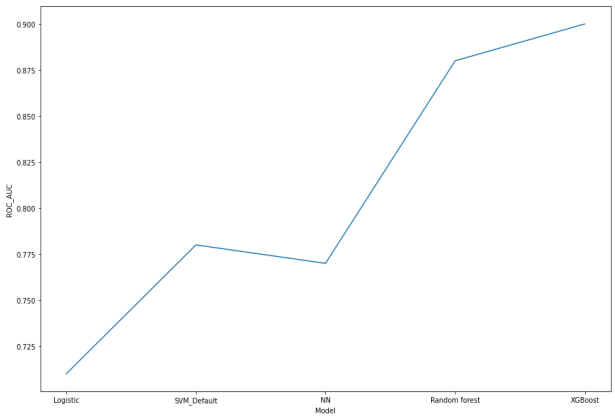
5.4 XGBoost Classifier:

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

5.5 Neural Network:

A neural network is a method in artificial intelligence that teaches computers to process

data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain**.**



XGBoost Classifier have performed really well and got the best scores with XGBoost Classifier as compared to other Models, so I conclude XGBoost is my optimal model for use and we can use this model for further in predicting Cardiovascular risk.

6. Technologies used:

Python: Python is a high-level interpreted language that supports different platforms like Windows, Linux, Mac, Raspberry Pi, etc. Python can be used for creating web applications, database systems, handling big data, and performing complex mathematical calculations. Python can be treated in an object-oriented, functional or procedural way.

Google Colab: Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

Python packages: Following are some of the python packages used in this project.

Matplotlib: Matplotlib is a visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualiz- ation library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram, etc.

Pandas: Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

NumPy: It provides structures for multiple dimensional array objects and tools for related operations. NumPy is usually used for high performance scientific computational tasks.

Seaborn: Seaborn is a visualization library for statistical graphics plotting in Python. It provides default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for the same variables for better understanding of the dataset**.**

6. Conclusion:

- There are 15.1 % people in our dataset are are at risk for cardiovascular disease and 84.9 % people are safe (ten year risk).
- There is more risk of cardiovascular disease in patients of age between 51 to 63.
- The count of male and female are same in risk which is around 200, though

females are more than males in our dataset.

- Around 250 smokers are in risk and around 210 non-smokers are is risk for cardiovascular disease.
- We can't evidentially state smoking will lead to heart disease, as we seen from count plot there is no huge difference between these to commune and also our extreme smoker who smokes 70 cigarettes per day is not having ten year risk.
- There are very few people who are done with BP medication which are around 200 but many people have not taken any BP medication and they are around 3200. We cannot say that after taking medication person are safe.
- Around 500 patients who did not had stroke yet and are at risk and around 2800 patients are safe.
- Around 250 people with hypertensive are in risk and around 255 people with no hypertensive are at risk.
- Here we can see people who did not had diabetes are more and around 500 people who did not had diabetes are at risk.And there are very few people who had diabetes are at risk.
- Most of the people who are not in risk their Cholestrol level lies between 210 to 280 and people who are in risk their cholestrol level lies between 215 to 285 there in not huge difference it is quite normal.
- Most people who are not in risk their systolic BP lies between 110 to 140 and people who are at risk their systolic BP lies between 125 to 160. We can say people with high systolic BP are at risk.
- Most people who are not in risk their diastolic BP lies between 75 to 85 and people who are at risk their diastolic BP lies between 89 to 90. We can say there is a slight increase in diastolic BP of people who are in risk.
- Most people who are not in risk their BMI lies between 22 to 28 and people who are at risk their BMI lies between 23 to 29 approx. WE cannot see any difference BMI is approx. same of risky and not risky people.
- Most people who are not in risk their heart rate lies between 68 to 83 and people who are at risk their heart rate lies between 68 to 84. which is same for risky and not risky people.
- There is not that difference between the glucose level of risky and non risky patients. glucose level lies between 70 to 80 for both risky and non risky patients.
- Most people smoke cigarettes between 1 to 10 approx. and there heart rate lies between 60 to 100.
- With logistic regression we got the accuracy score of 0.68 on train data and 0.66 on test data.
- With Support vector machine we got the train accuracy score of 0.79 and test accuracy score of 0.78.
- With Neural Networks we got the Train Accuracy score of 0.81 and test accuracy score of 0.77.
- With Random forest classifier we got the train accuracy of 0.97 and test accuracy of 0.89.
- With XGBoost classifier we got the train accuracy score of 0.99 and test accuracy of 0.90.

References:

1. GeeksforGeeks
2. Analytics Vidhya