Data Science Project Report

Analysis of Heart Disease Patients

By: Aaron Soares (as04345) and Faizan Kazi (mk04339)

Introduction

According to the World Health Organization, an estimated 12 million deaths occur worldwide every year due to several heart diseases. Half of the deaths in the United States alone, as well as other developed countries, are due to cardiovascular diseases. This project aims to identify the most relevant factors contributing to heart disease and predict the overall risk of it. By doing so, we will be able to give an early prognosis of cardiovascular diseases and help aid individuals in making healthier life decisions in order to reduce and possibly prevent them from heart related complications later in their lives.

The dataset used for this project is publicly available on Kaggle. The data has been collected from individuals living in Massachusetts, USA. The dataset includes over 4000 records and 16 different attributes.

Questions to be asked

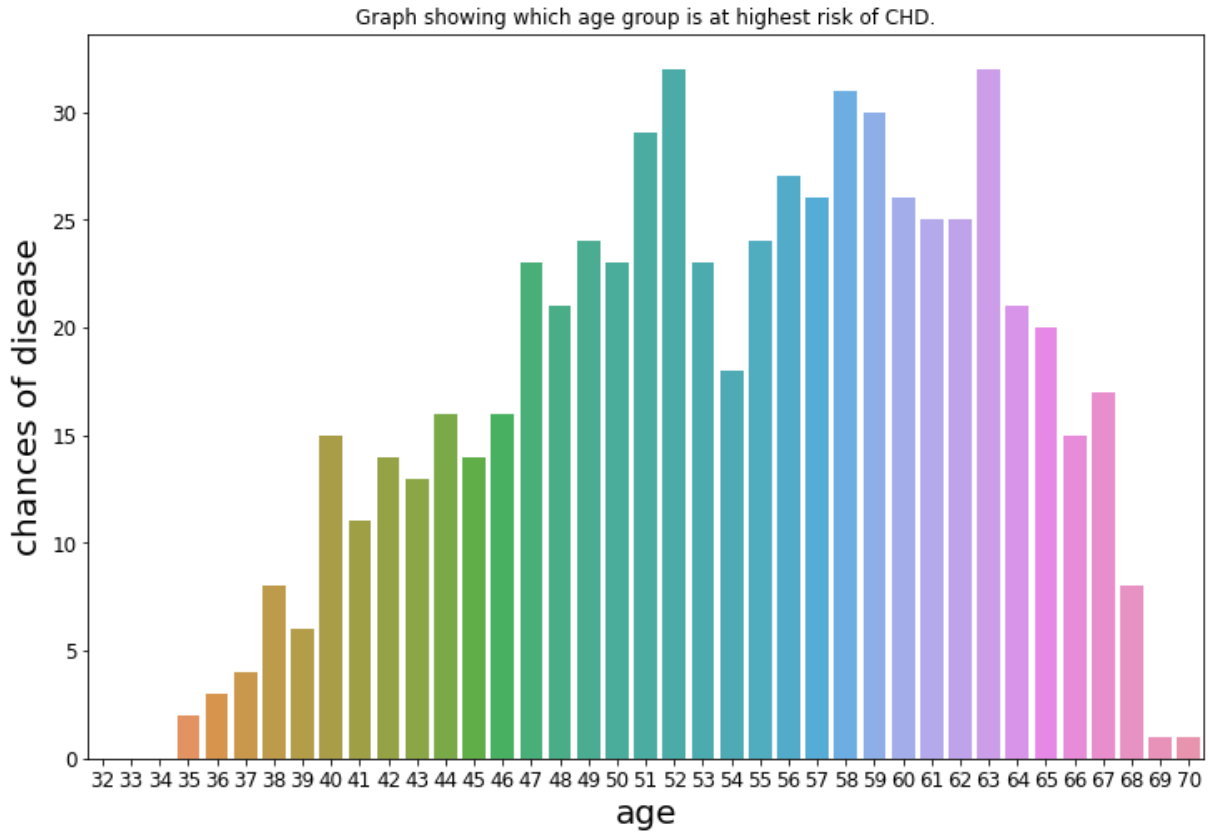
- Which factors contribute most towards cardiovascular diseases?
- Which age group is most susceptible to heart diseases?
- Are males or females at higher risk?
- How big a role does smoking play towards heart diseases?
- Does BMI play any role in heart diseases?
- Does the education of an individual affect how often they smoke?
- Cardiovascular diseases occur when the flow of blood to the heart becomes blocked. This may be due to high cholesterol levels which in turn increase blood pressure. Do high cholesterol levels in the blood lead to CHD?
- Can diabetes lead to a higher risk of CHD?

Analysis

We started our analysis by creating a heatmap with all 16 attributes. From it, we analyzed the following relationships:

1. A high systolic and diastolic blood pressure increases the risk of hypertension which lead to CHD.
2. Diabetes is directly related to the level of glucose in the blood
3. Age and CHD have the strongest positive correlation with each other.
4. Number of cigarettes per day and is directly proportional to the number of current smokers
5. High BMI and Cholesterol levels lead to high Systolic BP.

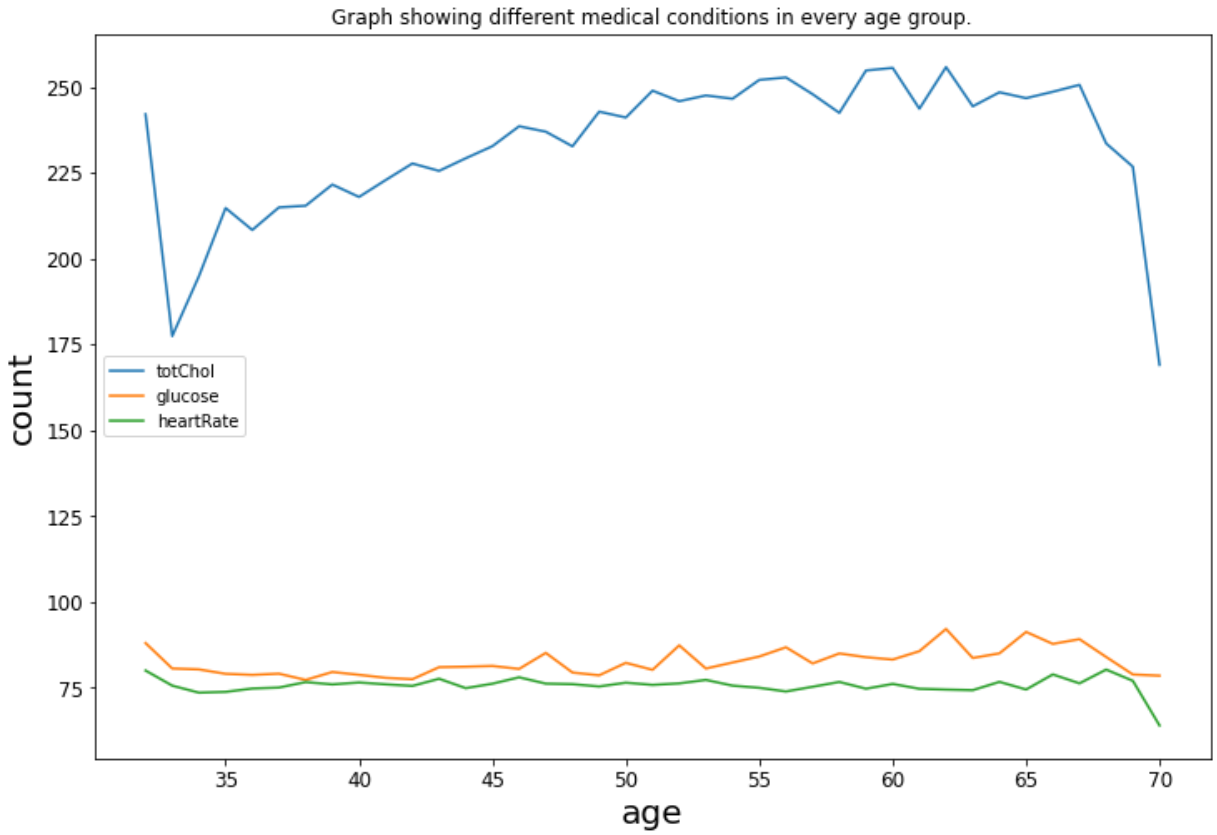
Which Age group is at greatest risk of Coronary Heart Disease (CHD)?



The graph above shows us which age group is at highest risk of CHD. People within the age range of 50's to mid 60's is at the highest risk. Therefore, it is safe to assume that older individuals are more likely to get a heart disease.

Again, one may expect that the risk would be greatest within the oldest individuals in the dataset. However, the graph shows the lowest numbers for those aged 69-70. This may be due to a low number of individuals within this age range in our dataset.

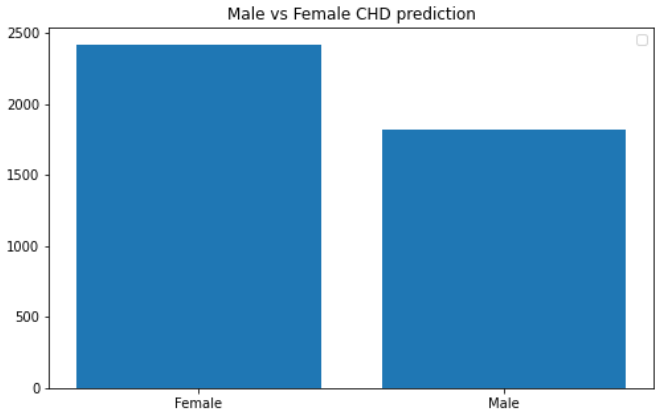
Relationship of Different Medical Conditions with respect to Age

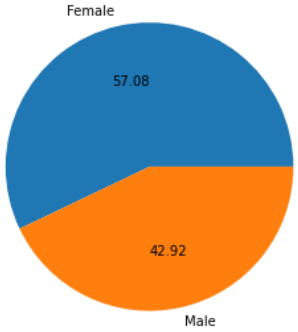


The graph above shows the relationship between the age of an individual and their cholesterol levels, glucose levels and heart rate. Our analysis does not show any significant changes in the glucose levels or the heart rate of an individual as they age. The heart rate stays relatively constant, but it does drop down below 75 bpm (beats per minute). This evidence isn't conclusive, and we would have to gather information about individuals older than 70 to reach a conclusion on this. The glucose levels also remain rather constant, within the range of 75-100 mmol/L (millimoles per liter).

On the other hand, cholesterol levels show a steady increase with age but drops drastically at the age of 70 (we can consider this to be an anomaly). Research suggests that over time, cholesterol builds up around the walls of the blood vessels. This decreases the space for blood flow. The cholesterol builds up can also break away from the walls of the blood vessels and be carried to smaller blood vessels/capillaries, where it may cause a blockage. This significantly increases the chance of CHD.

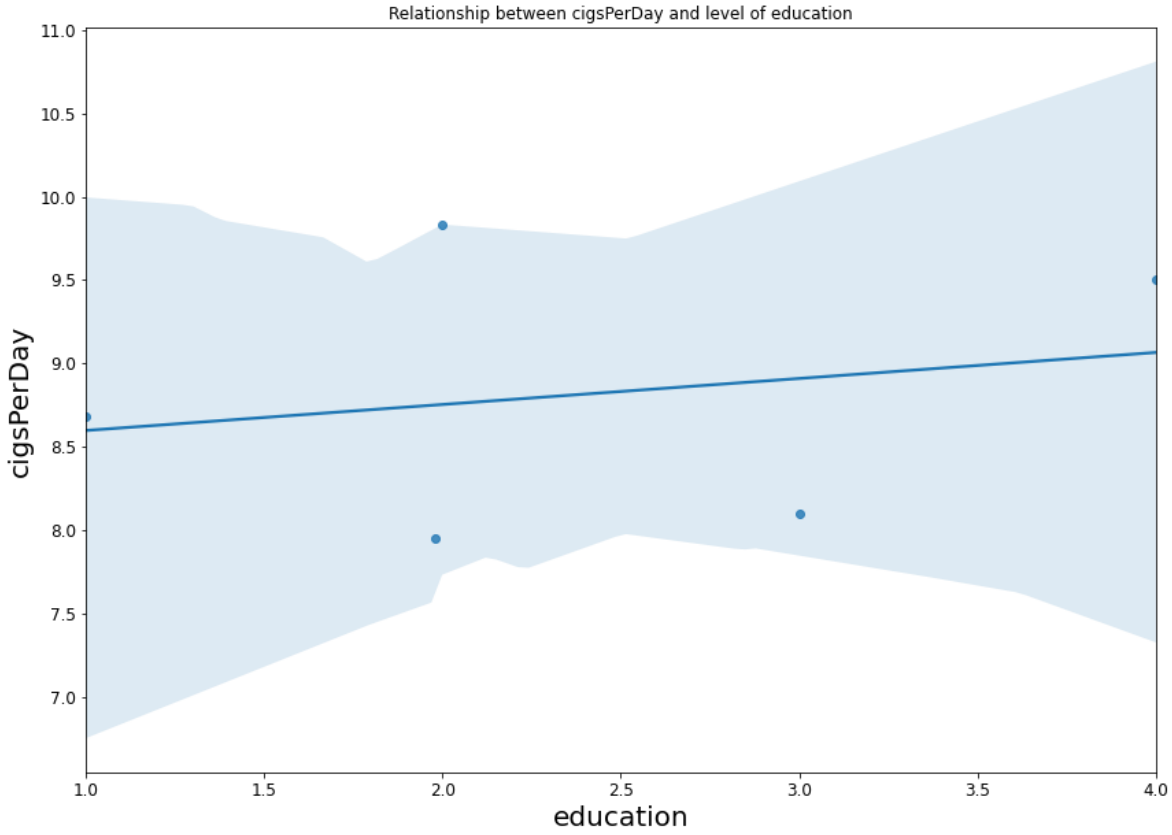
Are Males or Females at higher risk of CHD?





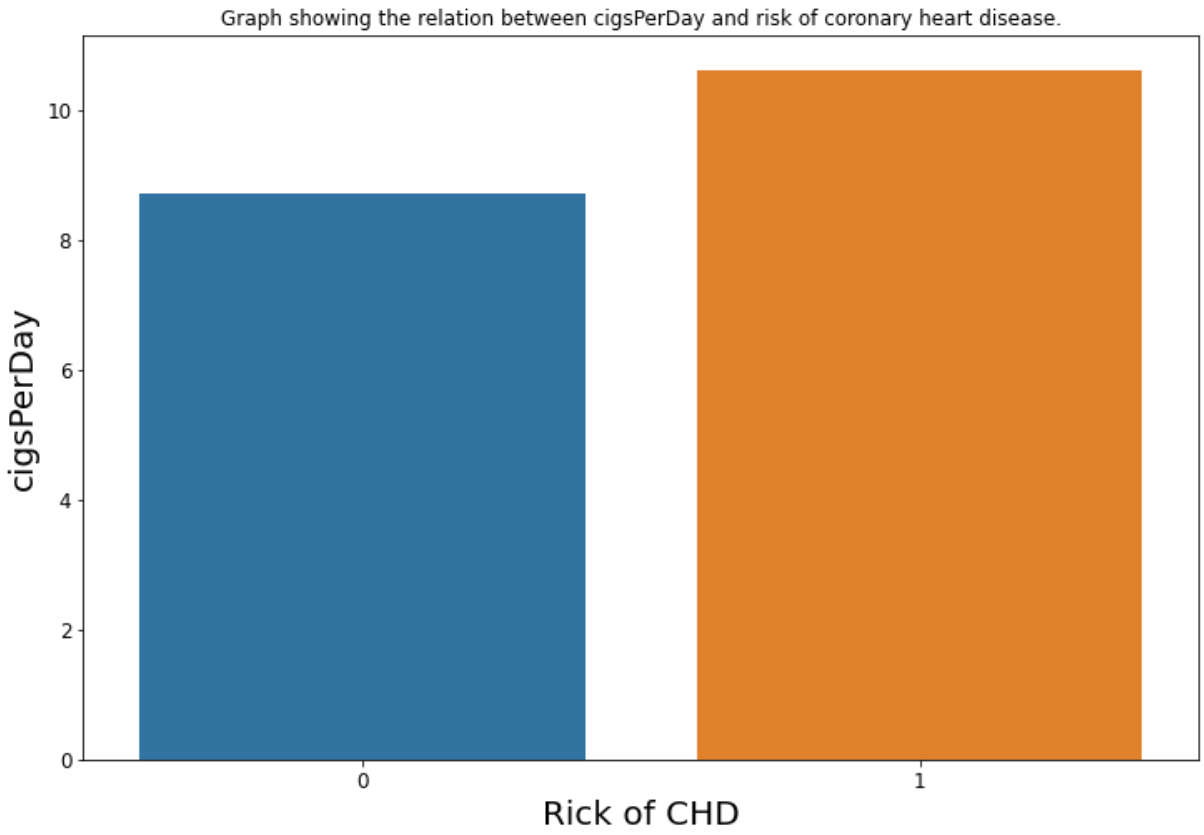
Our analysis shows that females are at higher risk of CHD. Research on this is mixed. Many people believe that women are less likely to develop heart diseases, but this is not true. Estrogen in women does offer some protection, but after menopause, women are more likely to develop heart diseases. Although our analysis suggests women are at higher risk, there is not enough conclusive evidence to support this.

Relationship between Cigarettes per day and Education



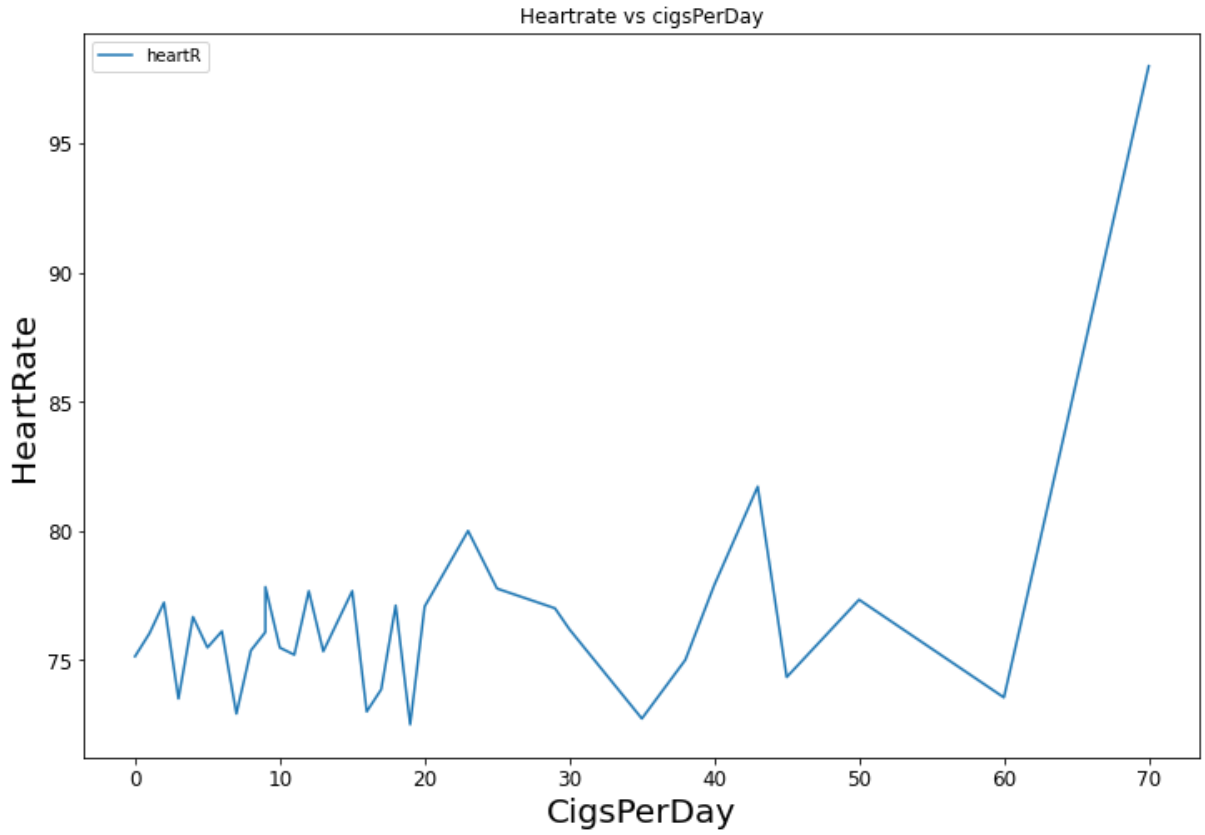
While one might assume that an individual with a lower level of education may be unaware of the negative effects that smoking had on one's health. However, the available dataset to us above does not show any such relation. The education of an individual has no effect on the number of cigarettes they smoke per day.

Does Smoking increase the risk of CHD?



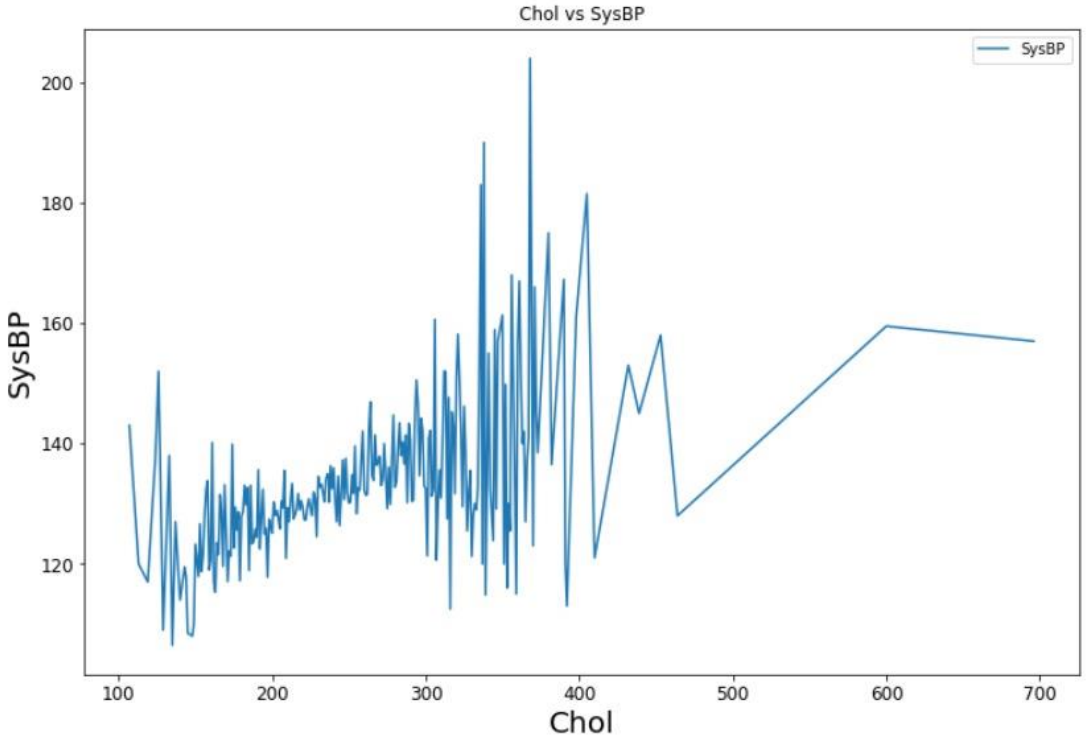
Smoking does increase the chances of heart disease. The average number of cigarettes smoked by individuals at risk of CHD is higher than the average number of cigarettes smoked by those who aren't at risk. Why is this so? Smoking causes the buildup of plaque and clots which narrow the blood vessels. Chemicals from the smoke cause the blood to thicken and clot. These blockages lessen or even stop the flow of blood to the heart which leads to heart attacks and other heart diseases.

The Effect of Smoking on Heart Rate



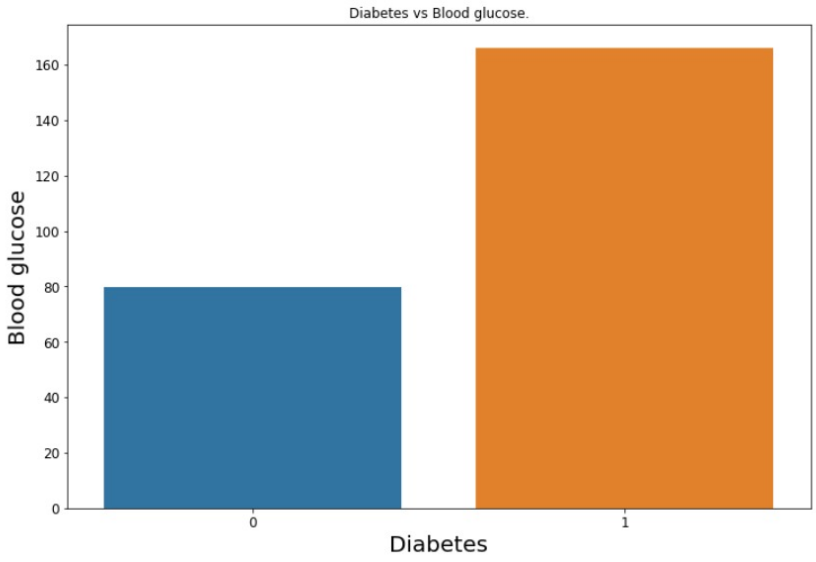
Our findings for the relationship between heart rate and the number of cigarettes per day is rather interesting. The graph above shows an increasing trend and between 60-70 cigarettes per day, an individual's heart rate increases above 95 bpm. This supports our earlier findings. With the narrowing of blood vessels, an individual's heart needs to beat faster to pump out enough blood to the rest of the body.

The Relationship between High Cholesterol Levels and Blood Pressure



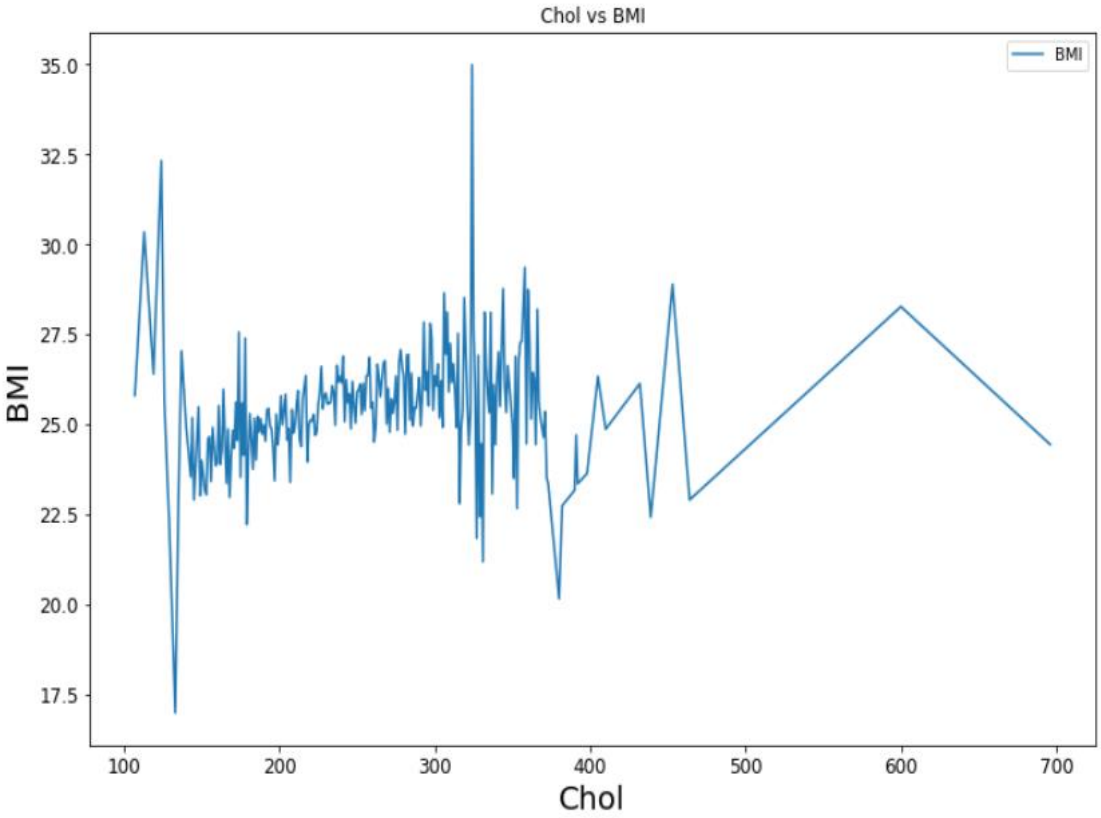
The graph above shows a steady increase in systolic blood pressure as cholesterol increases. After 400 mg/dL (milligrams per deciliter), there is no significant trend as there are too few records within this range. As discussed above, high cholesterol levels lead to the buildup of fatty deposits around the walls of the blood vessels which narrow the passage way for blood and increase blood pressure. This in turn increases the risk of CHD and heart attacks.

Can Diabetes lead to CHD?



The graph above shows that individuals having diabetes have a higher level of glucose flowing in their blood. This is because, diabetes is a condition in which the body is unable to convert the glucose in the blood into its stored form, glycogen. Therefore, the glucose levels will be higher. With higher levels of glucose in the blood, the blood begins to thicken, leading to a reduced flow of blood thereby increasing the risk of CHD.

Does BMI play any role in heart diseases?



Since the graph above shows a positive correlation between BMI and Cholesterol levels, it is safe to say that High BMI index indirectly results in high chances of CHD because earlier, we analyzed that high cholesterol level cause high systolic-BP, increasing the chances of CHD.

Using the information that we have gathered through our analysis; we created several models to predict the risk of CHD in individuals. Our learned model will be useful for individuals who are at higher risk as they can change their lifestyle and eating patterns to decrease or even possibly prevent the risk of early onset heart diseases.

Model Learning

Firstly, we split our dataset into train and test where the test data was 30%. We the scaled all our variables in the range (0,1) for better learning accuracies. Then, we fitted 4 learning models on out training data; Logistic Regression, Naïve Bayes, Decision tree and K-Nearest Neighbors. We applied Bagging and Boosting to improve our accuracy but there wasn't any noticeable improvement. Finally, we did K-Best features but couldn't achieve any massive improvement in accuracy of the models.

Conclusion

- Sys-BP and Age plays a big role in CHD.
- High BMI results in Systolic BP leading to CHD.
- Our analysis of the dataset suggests that women are at higher risk of CHD
- High blood pressure, high cholesterol levels and high blood glucose levels significantly increase the risk of CHD