The visualizations in this section help you answer questions like these:

 What is the peak value of the distribution?

 How many peaks are there in the distribution (unimodality versus bimodality)?

 How normal (or lognormal) is the data? We’ll discuss normal and lognormal

distributions in appendix B.

 How much does the data vary? Is it concentrated in a certain interval or in a certain

category?

many summary statistics assume that the data is

approximately normal in distribution (at least for continuous variables), so you want

to verify whether this is the case.

**Data Analysis and Interpretation**

**Dataset**

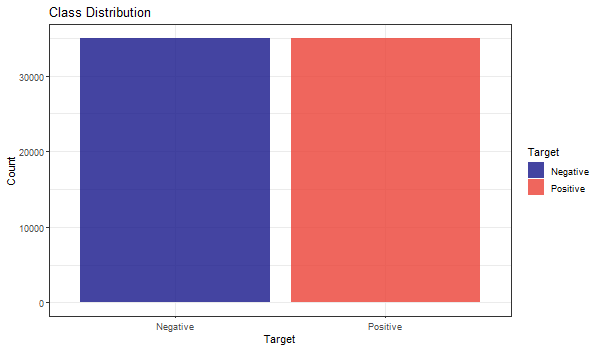
The dataset is obtained from Kaggle which consists a total of 70000 record of patient’s data. In this dataset, there is a total of 11 features which can be categorized into 3 types of input features, Objective, Examination and Subjective. Objective type is based on factual information, Examination type is from the medical examination results and Subjective type is information obtained from patient. There are two more attributes created in this project, they are BMI that was calculated using attributes height and weight and BloodPressure which uses systolic and diastolic blood pressure to categorize into lower than normal (-1), normal (0) and higher than normal (1).

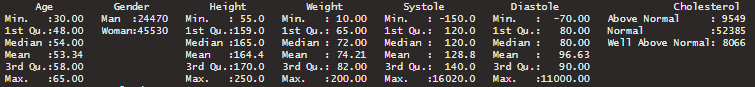
|  |  |  |
| --- | --- | --- |
| **Systolic blood pressure** | **Diastolic blood pressure** | **BloodPressure** |
| More or equal to 140 | More or equal to 90 | 1 |
| Less or equal to 90 | Less or equal to 60 | -1 |
| More than 90 and less than 140 | More than 60 and less than 90 | 0 |

The summarized data description is stated at the table below:

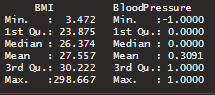
|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Attribute Name** | **Types of features** | **Data Type** |
| age | Age | Input, Objective feature | Integer (days) |
| gender | Gender | Input, Objective feature | Categorical code |
| height | Height | Input, Objective feature | Integer (cm) |
| weight | Weight | Input, Objective feature | Float (kg) |
| ap\_hi | Systolic blood pressure | Input, Examination feature | Integer |
| ap\_lo | Diastolic blood pressure | Input, Examination feature | Integer |
| cholesterol | Cholesterol | Input, Examination feature | 1: normal  2: above normal  3: well above normal |
| gluc | Glucose | Input, Examination feature | 1: normal  2: above normal  3: well above normal |
| smoke | Smoking | Input, Subjective feature | Binary |
| alco | Alcohol intake | Input, Subjective feature | Binary |
| active | Physical activity | Input, Subjective feature | Binary |
| cardio | Presence or absence of cardiovascular disease | Target Variable | Binary |
| BMI | BMI | Input, Objective feature | Float |
| BloodPressure | Blood Pressure | Input, Examination feature | Integer |

The class distribution of the target attribute, cardio is balanced as seen in the figure below.

****







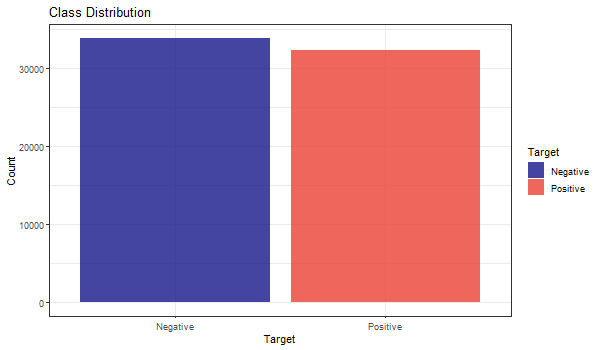
**Data Preparation**

**Missing or Null Data points**

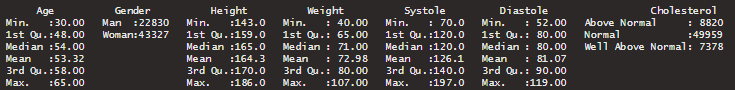
There are no missing values in the data, hence further data processing in not required.

**Unexpected Outliers**

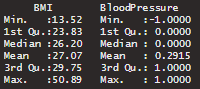
As seen in Figure xx (summary), there are outliers in some attribute such as Systole (ap\_hi) and Diastole (ap\_lo) that are negative in values which is impossible, weight attribute which has a minimum weight of 10kg and maximum value of 200kg and height attribute with a maximum of 250cm and minimum of 55cm which does not fit in the normal range. The outliers are handled by only retaining in the range of 25% to 75% quartile with a fixed multiplier of respective attribute. The difference between Negative count and Positive count in the Target class increased a little after the outlier removal process, but it is insignificant in impacting the modelling process.

****

With Figure xx as comparison, the attributes which have outliers like Height, Weight, Systole and Diastole looks way better after the outliers are handled which can be seen in the Figure below:







**Data Visualization**

|  |  |
| --- | --- |
|  | The peak density of the distribution is above 0.06 at the age range of 57 to 60. The distribution displayed multimodality characteristics with multiple peaks. It can be determined that the patients with cardiovascular diseases are more present in high age group. |
|  | The proportion of having cardiovascular diseases is not seen to be positively correlated to alcoholism as both groups have the proportion of negatives higher, but the contract is not pronounced. |
|  | Similar to Alcohol Distribution, it seems to be hard to put smoking as a powerful indicator due to same non-obvious distribution. |
|  | There is a trend that people who are not active in physical activity are more prone to have cardiovascular diseases. |
|  | There is no obvious trend from the distribution that target is correlated with gender. |
|  | The peak density of the distribution is at the height of 164cm. The distribution displayed multimodality characteristics with multiple peaks. There seems to be no trend in identifying whether patients have cardiovascular diseases in the height attribute. |
|  | Peak weight density of the distribution can be seen to be higher in the data where patients do not have cardiovascular disease. The distribution displayed multimodality characteristics with multiple peaks. As the weight increases starting from around 72kg, there is a higher chance the patient has cardiovascular disease. |
|  | Since BMI is calculated using attribute weight and height, we can see that the BMI distribution shows unimodality characteristic which is much more helpful. The peak density of the distribution is above 0.11 where patients do not have cardiovascular disease. It can be determined that patients with cardiovascular diseases are more present when BMI is at 28 onwards. |
| /home/shawn/Desktop/000009.png000009 | Cholesterol is seen an an obvious attribute and the level of it is positively correlated to cardiovascular diseases. |
|  | Glucose is seen as an obvious attribute as well. |
|  | The target is strongly correlated with Blood Pressure as those have lower-than-normal value have very low percentage being positive, the same for higher-than-normal instances. |

**Data Splitting**

Train test split will be used in this project with a fixed seed and ratio of 20% for test data and 80% for training data.