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.

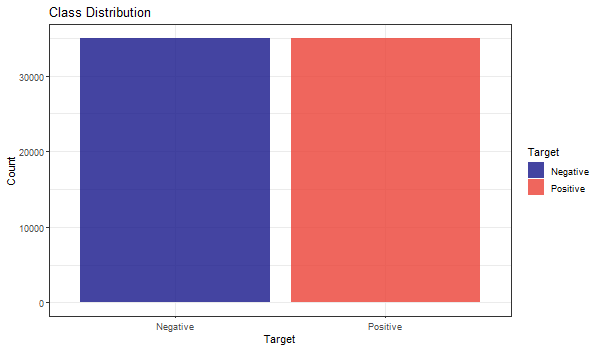
**Data Analysis and Interpretation**

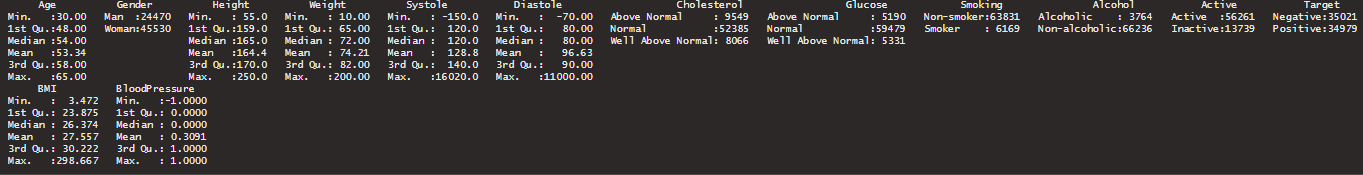
**Dataset**

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.

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 |

****

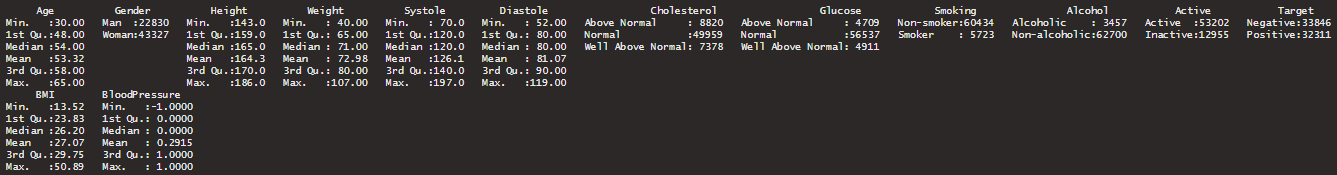


**Data Preparation**

**Missing or Null Data points**

**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, and the weight attribute has a minimum weight of 10kg and maximum value of 200kg which does not fit in the normal weight range. The outliers are handled by only retaining in the range of 25% to 75% quantile of respective attribute.



**Data Visualization**

**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.