* Classification—Deciding if something belongs to one category or another
* Finding relations—Finding correlations or potential causes of effects seen in the data
* Characterization—Very general plotting and report generation from data
* Distribution – Too narrow Vs. Too wide, good predictor if wide, varies. You should rely on information about the problem domain to judge if the data range is narrow, but a rough rule of thumb is the ratio of the standard deviation to the mean. If that ratio is very small, then the data isn’t varying much.

**Exploratory Analysis**

**Model Selection**

Our aim is to predict whether a person is at risk of developing cardiovascular disease, which is a classification problem. Because we are classifying patients into two groups, that is positive and negative, this problem is a binary classification problem. There are many algorithms available for binary classification problems. For example, Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Machine, etc. For this project, we have chosen to use all of the aforementioned algorithms to select the best one.

A. Support Vector Machine

SVM is a non-parametric model and makes less assumptions about the data. For this reason, even if the real-world data do not follow the training data distributions in future, it will still give a fair result.

B. Naïve Bayes

In contrast, Naïve Bayes is a parametric model and has several assumptions about the data, for example, it assumes that the features are independent of each other.

**Model Evaluation**

For model evaluation, we, first, established the null model, which is the lower bound of the model. As it is a classification problem, we selected null model to be the most common of all target classes. Then we calculated the Bayes rate which is the upper bound of the model. We also constructed the best single variable model possible and compared it against our final models. For performance measurement, we constructed confusion matrices and calculated accuracy, precision, recall, f1 score, specificity, and sensitivity for all the models. However, in this case, misclassification of someone who is not at risk of developing disease into at risk or positive would not be much of a problem because taking preventive measures are not discouragable. In contrast, if we classify somone who is indeed at risk into negative, it would be a problem. So, we wanted the precision or sensitivity to be as high as possible.

**Documentation**

Steps were taken to ensure proper documentation and flawless reproducibility of the project from the start. For documentation, the codes were written in R Markdown file using the knitr package, which is basically a stream of texts and codes. This ensured that the documentation was synchronized with codes, results were in sync with data, and it provided efficient means of collaboration between the project contributors. The project history was tracked from the beginning using Git, the most popular version control system, and GitHub, the most popular code sharing platform based on Git. All the data were saved in non-proprietary format to ensure reproducibility.