Devanshu Tayal

HEART ATTACK ANALYSIS

# Overview

|  |  |
| --- | --- |
| Objective | Methodology |
| From the given dataset we will need to analyse and identify the most relevant parameters that leads to heart attack. Answer the questions like:  ● Increasing in age has any effect towards heart attack.  ● Does increase in cholesterol level in body have any effect towards the heart attack  ● Increase in blood pressure have any relation with heart attack  ● Does family history have any effect towards heart attack | We will adopt the CRISP-DM methodology to perform the required analysis with the following workflow. For the purpose of this assignment, we would be going through the below steps:   * Understanding of the parameters required to predict Heart Attacks * Exploratory Data Analysis of the given data set * Check for Missing values and perform any required pre-processing required * Perform Feature Engineering to select the right features required for modelling * Use different models for evaluation of the given data set – features obtained through feature engineering * Once good accuracy is achieved - deploy the model for prediction |

# Dataset

Based on the EDA performed on the given data set, below are the observations:

|  |  |
| --- | --- |
| Features | Distribution of data w.r.t Target variable: |
| This dataset contains data about some hundreds of patients mentioning Age, Sex, Exercise Include Angia (1=YES, 0=NO), Chest Pain Type (Value 1: typical angina, Value2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic), ECG Results, Blood Pressure, Cholesterol, Blood Sugar, Family History (Number of persons affected in the family), Maximum Heart Rate, Target -0=LESS CHANCE , 1= MORE CHANCE | We have a good split of data between patients who have less vs more chances of heart attack – about 54:46 split. |
| Feature Properties | Distribution of Training set, validation set, testing set |
| * **Shape**: (303, 11) * The **categorial Columns** are: ['Sex', 'CP\_Type', 'BloodSugar', 'ECG', 'ExerciseAngia', 'FamilyHistory'] * The **continuous Columns** are : ['Age', 'BloodPressure', 'Cholestrol', 'MaxHeartRate'] * The **target** variable is : ['Target'] * All data types are of type int– as the categorical columns are label encoded. | As part of this assignment, we would be going with a 75:25 split for Training and Testing the model accuracy. Data Wrangling and Pre-Processing Based on the exploratory analysis performed, there were no missing values found – hence no pre-processing was necessary. Data looked clean.  Also from modelling standpoint, all the data presented were already encoded into desired format, hence no necessary for Label Encoding was necessary. |

# Feature Engineering Techniques

We will be using the below two filter feature selection engineering techniques.

1. ANOVA: [f\_classif()](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html)
2. Chi-Squared: [chi2()](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html)

Post application of the above features, we will rank them to identify the Top 6 features that will be further used in the models to predict the chances of heart attack.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA | | |  | CHI2 | | |
| **CID** | **Feature** | **Score** |  | **CID** | **Feature** | **Score** |
| **8** | ExerciseAngia | 70.95244 |  | **7** | MaxHeartRate | 188.32047 |
| **2** | CP\_Type | 69.77227 |  | **2** | CP\_Type | 62.598098 |
| **7** | MaxHeartRate | 65.1201 |  | **8** | ExerciseAngia | 38.914377 |
| **1** | Sex | 25.79219 |  | **4** | Cholestrol | 23.936394 |
| **0** | Age | 16.1167 |  | **0** | Age | 23.286624 |
| **3** | BloodPressure | 6.458169 |  | **3** | BloodPressure | 14.823925 |

# Methodology

As part of the modelling, we will be using the below two models required to predict the Target class for any given patient:

|  |  |
| --- | --- |
| Gradient Boosting Classifier | Random Forest |
| Gradient Boosting Classifier is one of the most powerful algorithms in the field of machine learning. It is used for predicting continuous target variable (as a Regressor) and also in predicting the categorical target variable (as a Classifier) – which is the case in this proposal.  Gradient boosting classifiers are the AdaBoosting method combined with weighted minimization, after which the classifiers and weighted inputs are recalculated. The objective of Gradient Boosting classifiers is to minimize the loss, or the difference between the actual class value of the training example and the predicted class value. It isn't required to understand the process for reducing the classifier's loss, but it operates similarly to [gradient descent](https://en.wikipedia.org/wiki/Gradient_descent) in a neural network. | Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.  In simple terms, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.  One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. |

# Results

|  |  |
| --- | --- |
| Evaluation metric for each ML technique used | Conclusion |
| Training / Testing and Accuracy levels achieved through each of the models / features utilized:   |  |  |  |  | | --- | --- | --- | --- | | **Model** | **Train Score** | **Test Score** | **Accuracy** | | Random Forest using Chi2 based features | 0.77 | 0.81 | 82% | | Gradient Boosting using Chi2 based features | 0.87 | 0.75 | 75% | | Random Forest using ANOVA based features | 0.80 | 0.75 | 75% | | Gradient Boosting using ANOVA based features | 0.88 | 0.75 | 75% | | * Conclusions from the EDA: There is no apparent linear correlation between continuous variable according to the heatmap. The heatmap matrix suggests that there might be some correlation between Target and Cp\_Type, MaxHeartRate and ExerciseAngia. With increase in Age we would assume the chances of heart attack increases, however with the current data set it could be a possibility for almost all ages. Refer the ipynb file for more details on EDA. * Filtering based on Chi2 gives better accuracy in terms of predicting Heart Attacks with about 82% accuracy. * Both Gradient Boosting classifier and Random Forest have similar outcome in terms of prediction, with Random Forest giving better accuracy in comparison with Gradient Boosting Classifier. * If there are a greater number of instances provided for training - the accuracy of the model can be improved. |