**Objective of the Project**

* **Objective:** The objective of this project is to perform the heart attack analysis and find the specific record (or person) potentially on high or low risk.
* **Methodology**: Apply Random Forest Classifier and Gradient Boosting. ***To perform smooth execution of project CRISP-DM methodology is used***

**Analysis of Dataset**

* **How many features**

As per the given data set excluding target feature, we have 10 features in data set.

* **Size of the dataset**

(303, 11)

* **Multiple files**

As per given data, entire data set is organized in single file

* **What kind of data – numerical or character**

After analysis of data, it is observed that the entire dataset is numerical/categorical

* **Balanced or imbalanced – what is the distribution**

After performing the data balancing on target label. It is observed that the split of data is in the ratio of 45:55 with respect to target label. Therefore, explicit data balancing is not performed. ({0:45%}, {1:55%})

* **Distribution of Training set, validation set, testing set**

Distribution of data of Training is 70% & test is 30%

* **Missing data and Preprocessing challenges**

After performing data exploration, the observations are as follows:

1. No Missing Data
2. No Null Data
3. All data types are same

Preprocessing Challenges:

1. Feature segregation & selection
2. Feature Scaling

**Feature Engineering Techniques**

* **Features removed**

After data exploration it was observed that the given sample size and feature counts are not bulky. Therefore, nothing was removed from data set. Although, during get\_dummies feature was broken down but its lossless.

* **Feature creation**

**Feature Engineering 1 (FE-1)**

1. No feature explicitly removed.
2. Performed get\_dummies for data manipulation
3. Converted categorical feature into indicators.
4. Standard Scaling performed to scale down numerical data

**Feature Engineering 2** **(FE-2)**

1. Feature Engineering performed by grouping the categorical features w.r.t. multiple continuous features and generated new features
2. Integrated new features with actual data set by performing join.
3. Actual data set now having more features than given data set

* **Feature ranking**

To perform feature ranking correlation analysis is done using method “**getCorrelatedFeature**(..)” and heatmap

* **Feature Selection**

1. Based on correlation results all features were selected for **feature engineering 1** and modelling was performed
2. Based on feature creations all features were selected for **feature engineering 2** and modelling was performed

* **Class imbalance treatment**

After performing the data balancing on target label. It is observed that the split of data is in the ratio of 45:55 with respect to target label. Therefore, explicit data balancing is not performed. ({0:45%}, {1:55%})

**Methodology**

The methodology followed is CRISP-DM, under that hood below tasks were performed for problem statement

* **Business Understanding:** Understand project objectives and requirements, objective is to identify potentially high/low risk heart attack patient.
* **Data Understanding:** Explored the data to find quality, obviousness, types & visualization etc.
* **Data Preparation:** Performed data cleaning, data segregation, feature selection/engineering & Scaling etc.
* **Modelling:** Performed data modelling based on the given scope and analyze the modelling params
* **Evaluation:** Identified issues which could be identified earlier, determine the results
* **Deployment:** Put the model into practice
* **The 2 classifiers used**

Random Forest and Gradient Boosting classifier are used for classification because it is mentioned in problem statement

* **Hyper-parameter tuning**

Parameter tuning is done on training data to get the best params for the classifiers rather than applying random params to get better results. Hyper tuning information can be checked in code, it is available with proper comment. It is observed that hyper tuned params gives better results when executed on training data but those same params are not giving better results on test data. Therefore, few param might look altered.

**Results**

* **Table for the evaluation metric for each ML technique used**



* **Plot of the curves**

Plots generate for data visualizations are shown in detail in code which contains the plots related with exploratory analysis and scaling.

* **Conclusion**

After the analysis of data, it is identified that the data is having lots of feature that ideally must contribute to the high risk of heart attack. But this is not the case for example:

1. We observed that the increasing average value of age and cholesterol is decreasing the risk of heart attack.
2. Also, we observed that the median value of both age and cholesterol is having very less difference.
3. Also, we observed that there is a positive correlation in case of feature “FamilyHistory”, if this feature is having more cases of heart attack chances are high and increasing.

Additionally, data is insufficient to put the model in practice. Apart from it, the truthness’ is a concern in the data because the health factors that contributes to heart attack are not contributing up to that level or it might be the case that the data is insufficient.

Also, it is observed that feature engineering performed implicitly & explicitly is not showing any impact in case of gradient-boosting, but random forest is showing decline in accuracy with FE-2

**Note**:

1. Results might change because random forest doesn’t create the same structure every time
2. Data is also available with name “Heart\_Attack\_Analysis\_Data”.