# Heart attack analysis report

## Introduction

A heart attack occurs when an artery supplying your heart with blood and oxygen becomes blocked. A blood clot can form and block your arteries, causing a heart attack. This Heart Attack Analysis helps to understand the chance of attack occurrence in persons based on varied health conditions.

## Dataset

The dataset is Heart\_Attack\_Analysis\_Data.csv.   
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

## Group 133

|  |  |  |
| --- | --- | --- |
| Name | Bits ID | Contrubition |
| Chavhan Kalpesh Shantilal | [2020sc04653@wilp.bits-pilani.ac.in](mailto:2020sc04653@wilp.bits-pilani.ac.in) | Equal (100%) |
| Nishant Rastogi | [2020sc04875@wilp.bits-pilani.ac.in](mailto:2020sc04875@wilp.bits-pilani.ac.in) | Equal (100%) |
| Manikandan P | [2020sc04643@wilp.bits-pilani.ac.in](mailto:2020sc04643@wilp.bits-pilani.ac.in) | Equal (100%) |

## Scope

Scope of document is limited to heart attack analysis and implementation strategies.

## Deliverables

|  |  |
| --- | --- |
| Filename | Description |
| G133\_Heart\_attack\_analysis\_report.docx | Support document for design and strategy used. |
| G133\_Heart\_Attack\_Analysis.ipynb | Jupiter python notebook file for model evaluation |
| Readme.txt | Documentation to run python / Jupiter notebook and how to provide dataset. |
| Heart\_Attack\_Analysis\_Data.csv | Dataset |

## Problem Statement

We have to build the predication model using given data set, It’s 0,1 Classification problem where Target values are 0 = less chance and 1 = more change, We have to label tain/test/unseen data as 0,1 using predictive model

## List of Classifier used.

* Nearest Neighbors
* Linear SVM
* RBF SVM
* Gaussian Process
* Decision Tree
* Random Forest
* Neural Net
* AdaBoost
* Naive Bayes
* QDA
* LogisticRegression

## Solution

We have 303 records available in data set. we have performed following steps and strategies to select best model.

1. **Data analysis using statistics and visualization**

As per give dataset and problem statement we have identified feature datatype. There exist categorical and numeric features only.   
Data analysis steps

1. Check missing / invalid data   
   No missing/invalid entries in data set
2. **Outlier** visualization using box plot  
   boxplot help to identify existence of outliers in blood pressure, cholesterol, and max heart rate feature.
3. **Correlation** will work well with continuous data and we have used pearson method for correlation.

Chart, box and whisker chart

Description automatically generated A screenshot of a computer

Description automatically generated with medium confidence

1. **Data pre-processing**
   1. Since no missing/invalid data, we won’t need any missing value replacement strategy.
   2. Outlier removal

We use zscore method with threshold 3 to detetct outlier, we replace outlier with mean.

* 1. Normalization Method StandardScaler method ( normalize by zscore). All continuous entries are normalized zscore.

We prefer this technique since it help us to understand data distribution using graph.

* 1. **Dimensionality reduction**.  
     We prefer to use **PCA** for following reason (no of components =2)
     1. information distributed across a large number of columns is transformed into principal components (PC) such that the first few PCs can explain a sizeable chunk of the total information (variance). These PCs can be used as explanatory variables in Machine Learning models.
     2. Visualize Classes: Visualizing the separation of classes (or clusters) is hard for data with more than 3 dimensions (features). With the first two PCs itself, it’s usually possible to see a clear separation.

Data visualization after preprocessing and PCA

Chart, scatter chart

Description automatically generated

1. **Build and evaluate Model**

We have built 12 models using different classifiers. We have used Random train test split and strategfied KFlod cross validation for model selection.

* 1. **Random train test split steps**
     1. Split data into 80:20 ratio randomly
     2. Build model using train data
     3. Use test data and predict target
     4. Get Fscore, accuracy, recall and precision metrics
     5. Compare metrices for evaluation.
     6. Plot graph for decision boundary – it will help to understand model is overfit or underfit.

The Fscore, accuracy, recall and precision metrics will help us to select best model but it won’t provide any metrics to decide model is overfit or underfit.

Metrics for all classifiers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifiers | Accurracy | Prediction | Recall | F-score |
| Nearest Neighbors | 75.409836 | 71.428571 | 83.333333 | 76.923077 |
| Linear SVM | 78.688525 | 77.419355 | 80.000000 | 78.688525 |
| RBF SVM | 77.049180 | 72.222222 | 86.666667 | 78.787879 |
| Gaussian Process | 80.327869 | 78.125000 | 83.333333 | 80.645161 |
| Decision Tree | 62.295082 | 60.606061 | 66.666667 | 63.492063 |
| Random Forest | 77.049180 | 72.222222 | 86.666667 | 78.787879 |
| Neural Net | 78.688525 | 77.419355 | 80.000000 | 78.688525 |
| AdaBoost | 68.852459 | 64.864865 | 80.000000 | 71.641791 |
| Naive Bayes | 78.688525 | 75.757576 | 83.333333 | 79.365079 |
| QDA | 78.688525 | 77.419355 | 80.000000 | 78.688525 |
| SGDClassifier | 73.770492 | 69.444444 | 83.333333 | 75.757576 |
| LogisticRegression | 78.688525 | 77.419355 | 80.000000 | 78.688525 |

Background pattern

Description automatically generated

Decision Boundaries visualization

A picture containing building

Description automatically generated

* 1. **Stratified K Fold cross validation**

We have used cross verification using StratifiedKFold method for model selection. A model that fits the training set well but testing set poorly is said to be overfit to the training set and a model that fits both sets poorly is said to be underfit.

The strategy involves the following steps:

* + 1. split the dataset into training and test sets
    2. train the model with the training set
    3. test the model on the training and test sets
    4. calculate the **Mean Absolute Error (MAE)** for training and test sets
    5. plot and interpret results

**Mean Absolute Error (MAE) analysis**

We have observed MAE point (>30) for **Nearest Neighbors, RBF SVM, Decision Tree, Random Forest, Neural Net, AdaBoost, SGDClassifier**. So the model **underfits** the some portion of the test data.

We have observed MAE point (<20) for test and training set for **Linear SVM, RBF SVM, Decision Tree, Random Forest, AdaBoost, Naive Bayes, QDA.** SO the model is **overfits** the some portion of data.

We have observed all MAE point are between .30 and .20 for **Leaner SVM, Gaussian Process, QDA and Logistic regression.** So model is **best fit** for all portion of test and train data.

**Visualization of Mean absolute error.**

A picture containing shoji, building, window

Description automatically generated

## Conclusion

**Leaner SVM, Gaussian Process, QDA and Logistic regression** model is best fit and provide similar accuracy.

We recommend **Logistic regression** because its average MAE is lowest compare to best fit model