**Healthcare Project – Summary Report**

Problem Statement:

Cardiovascular diseases are the leading cause of death globally. It is therefore necessary to identify the causes and develop a system to predict heart attacks in an effective manner. The data below has the information about the factors that might have an impact on cardiovascular health.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Age | Age in years |
| Sex | 1 = male; 0 = female |
| cp| | Chest pain type |
| trestbps | Resting blood pressure (in mm Hg on admission to the hospital) |
| chol | Serum cholesterol in mg/dl |
| fbs | Fasting blood sugar > 120 mg/dl (1 = true; 0 = false) |
| restecg | Resting electrocardiographic results |
| thalach | Maximum heart rate achieved |
| exang | Exercise induced angina (1 = yes; 0 = no) |
| oldpeak | ST depression induced by exercise relative to rest |
| slope | Slope of the peak exercise ST segment |
| ca | Number of major vessels (0-3) colored by fluoroscopy |
| thal | 3 = normal; 6 = fixed defect; 7 = reversible defect |
| Target | 1 or 0 |

Solution:

1. **Preliminary analysis:**
   1. **Perform preliminary data inspection and report the findings on the structure of the data, missing values, duplicates, etc.**
   2. **Based on these findings, remove duplicates (if any) and treat missing values using an appropriate strategy**

The data comprised 303 records and 14 features, of which one record was duplicate. The same was removed.

The data comprised no missing values. Out of the 302 records, 165 patients had heart attacks, while the remaining 138 did not had any heart attack.

1. **Prepare a report about the data explaining the distribution of the disease and the related factors using the steps listed below:**
   1. **Get a preliminary statistical summary of the data and explore the measures of central tendencies and spread of the data**

Graphical user interface, Word

Description automatically generated

* 1. **Identify the data variables which are categorical and describe and explore these variables using the appropriate tools, such as count plot**

Chart

Description automatically generated

* 1. **Study the occurrence of CVD across the Age category.**

Graphical user interface

Description automatically generated with medium confidence

The age category 40-60 was more prone to CVD as compared to other age groups.

* 1. **Study the composition of all patients with respect to the Sex category**

A picture containing graphical user interface

Description automatically generated

Females were more prone to CVD as compared to Males.

* 1. **Study if one can detect heart attacks based on anomalies in the resting blood pressure (trestbps) of a patient**

Since amongst the outliers i.e. anomalies in resting blood pressure, at similar trestbps, there are both cases of people having CVD and people not having CVD.

Hence, it is inconclusive to detect heart attacks only based on anomalies in resting blood pressure.

* 1. **Describe the relationship between cholesterol levels and a target variable**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

People having cholestrol levels between 200 to 300 were more prone to having CVD.

Since amongst the outliers i.e. people having very high cholestrol levels, there are both cases of people having CVD and people not having CVD.

Hence, it is inconclusive to detect heart attacks only based on very high cholestrol levels.

* 1. **State what relationship exists between peak exercising and the occurrence of a heart attack**

Graphical user interface, application

Description automatically generated

People with slope = 2 i.e. peak exercise ST segment were more prone to having CVD.

* 1. **Check if thalassemia is a major cause of CVD**

Chart

Description automatically generated with medium confidence

W.r.t. Thalesima, people with thal = 2 were more prone to having CVD.

* 1. **List how the other factors determine the occurrence of CVD**

Graphical user interface, text, application

Description automatically generated

* 1. **Use a pair plot to understand the relationship between all the given variables**

1. **Build a baseline model to predict the risk of a heart attack using a logistic regression and random forest and explore the results while using correlation analysis and logistic regression (leveraging standard error and p-values from statsmodels) for feature selection**
   1. **Using Logistic Regression**

Graphical user interface, text, application, email

Description automatically generated

* 1. **Using Random Forest**

Graphical user interface, text, application, email

Description automatically generated

**Accuracy score of 82% using the Random Forest model was providing the best prediction as compared to Logistic Regression.**

**Under Random Forest, the parameters were fine-tuned using GridSearchCV.**

Graphical user interface, application

Description automatically generated

The model accuracy did not improve.

**Logistic regression was carried out using Stats Model**

Graphical user interface, text, application

Description automatically generated

The model accuracy was 70%.

**Using feature selection by dropping features having p-value <0.05, the model accuracy improved to 82%.**

Graphical user interface, text, application

Description automatically generated

**Prediction of CVD after using feature engineering in Logistic Regression, using the StatsModel, had an accuracy score of 82%.**