**Project- HealthCare**

**Objective :** The aim of the project is to identify the features(causes) for the cardiovascular diseases, which are the leading cause of deaths in the world from the given set of many variables.

**Tasks performed:**

First the data file was Downloaded **CEP 1\_ Dataset.xlsx** using the link given in the **Healthcare** project problem statement.

1. Preliminary analysis:
   1. Preliminary data inspection & findings on the structure of the data, missing values, duplicates, etc.
   2. Removal of duplicates (if any) and treating missing values using an appropriate strategy.
2. There were no missing values found in the entire dataset. –

Function used- **df.isnull().sum()**

1. The shape , head & tail functions were used to explore the structure of the data- There are 303 records and 14 variables.

Function used- **df.shape**

1. There was 1 duplicated record found and was dropped from the dataset keeping the first record.

Function Used- **df[df.duplicated()]**

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

Functions used- **df.describe()**

* 1. Identify the data variables which are categorical and numerical:

Function used- **df.dtypes,**

All the variables were ‘integer’ or ‘float’ none of the variables were categorical type.

* 1. The occurrence of CVD across the Age category:

Distribution plot was plotted and explored.

The density of patients was more between 50 to 60 yrs of age.

.

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

Function used**- df["sex"].value\_counts()**

This gave number of Males and Females in the patients.

Female- 96 , Male- 206.

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

Function used- **fig, ax = plt.subplots(figsize = (10,6))**

**ax.scatter(df['trestbps'], df['target'])**

Scatter plot explored to see how many occurrence of CVD cases have happened at extreme ‘Trestbps’ points.

* 1. The relationship between cholesterol levels and a target variable-

Function used-

**df['target'].corr(df['chol'])**

To study how is cholesterol levels and target variable related statistically.

The Target variable does not have strong correlation with cholesterol levels.

* 1. The relationship between peak exercising and the occurrence of a heart attack explored-

Function used- **df['target'].corr(df['slope'])**

The heart attack and Peak exercising are correlated. It is a positive correlation so if the slope goes up, the heart attack also increases.

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

Functions used-

**df['target'].corr(df['thal']) -** It is Negatively correlated to Target variable.

**df.groupby('thal')['target'].mean().plot.bar()**

**plt.show()**

**df.thal.value\_counts(normalize=True).plot.barh()**

**plt.show()**

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

Functions Used-

**df.corr()**

**corr= round(df.corr()\*100,2)**

**sns.heatmap(corr)**

Correlation between all variable is checked along with the heatmap.

**Positively correlated** variables to Target variables- cp, restecg, thalach, slope

**Negatively correlated** variables to Target variable- age,sex,trestbps,chol, etc.

**Strongly correlated** variables to target variable- cp, oldpeak,ca,exang

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

Pairplot explored keeping ‘hue’ as target variable.

Data is non- linear in nature.

1. A baseline model to predict the risk of a heart attack using a logistic regression and random forest-
2. Skeweness of the data is checked.
3. Log Transformations are performed on Skewed variables.
4. Distribution is again checked.
5. Scaling is performed on the model.
6. Data is split into train and test
7. Logistic regression model is imported and the pre-processed data is passed through it.
8. **Prediction is done and the Accuracy is checked, which is 0.803.**
9. Confusion matrix is evaluated and False Negative is high.
10. ROC curve is applied to get AUC,
11. Optimal threshold is found and the model is run again on it.
12. The False Negative is reduced.
13. Random Forest classifier is imported and implemented to improve the results.

Exploring the results-

1. The results of Logistic regression and Random Forest show a little difference.
2. **As Accuracy of Random forest is 0.819.**
3. Still the accuracy is not up to the mark.
4. SMOTE is applied to correct class imbalance.
5. Cross Validation is applied to reduce Cross fitting in the model.
6. Over fitting is reduced.

Using correlation analysis and logistic regression (leveraging standard error and p-values from statsmodels) for feature selection-

1. Feature selection is performed to improve the performance of the model.
2. Correlation between variables is again checked.
3. Feature importance is checked.
4. Then ‘statsmodel’ is imported and applied on the data to check the Standard error and p-values of the features.
5. Any feature with p-value more than 0.05 is selected for dropping from the model.
6. The Affect of dropping these features is one by one checked on the ‘R-Squared’ and ‘Adj. R-Squared.
7. Finally Variables are identified with least impact on the model and are dropped.
8. Model is rebuilt using the selected features only.
9. **The accuracy of the model thus increases to 0.852.**
10. Finally the model and transformations are stored in Pickel format.
11. Predictions are made on the data which match the target variable.

**Conclusion:**

By performing the above mentioned tasks, the variables /features or the Causes which have a high impact on CVD are identified. The model built with these features can predict the target variable as 1 or 0 (1=yes, 0= no), chances of CVD in a patient.