Heart Disease Prediction: Milestone Report

### Problem statement:

### Machine Learning is used across many spheres around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of locomotor disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

### Data Wrangling:

### Heart Disease UCI dataset is fetched from Kaggle.

### Reference URL: <https://www.kaggle.com/ronitf/heart-disease-uci>

### It's a clean, easy to understand set of data. However, the meaning of some of the column headers are not obvious. Below is the description of the columns

### Age: The person's age in years

### Sex: The person's sex (1 = male, 0 = female)

### cp: The chest pain experienced (Value 0: typical angina, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: asymptomatic)

### treetops: The person's resting blood pressure (mm Hg on admission to the hospital)

### chol: The person's cholesterol measurement in mg/dl

### fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)

### restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)

### thalach: The person's maximum heart rate achieved

### exang: Exercise induced angina (1 = yes; 0 = no)

### oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more [here](https://litfl.com/st-segment-ecg-library/))

### slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: down sloping)

### ca: The number of major vessels (0-3)

### thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)

### target: Heart disease (0 = no, 1 = yes)

### The Cleveland database is the only one that has been used by ML researchers to this date. The "target" field refers to the presence of heart disease in the patient. It have integers 0 or 1 which represents negative and positive scenarios of heart disease

### The dataset is model ready as they are no missing or null values found.

### We were able to determine the outliers for the cholesterol and resting blood pressure variable however we need to perform EDA on the column to understand the impact of the outliers on the data

### Exploratory Data Analysis:

### By applying the describe function on the heart disease dataset we are able to get some basic analysis on the features.

### 

### Minimum Age is 29 and Maximum age is 77 and the average being 54. It makes sense as we occurrence of heart disease is generally more prominent after age 25-30 .Also it shows that there is balanced Normal distribution with mean and 50% inter quartile range of age.

### The mean of the sex column is 0.68 stating they are more male records that female in the dataset with ratio 68:32 (male: female).

### The statistics of chest pain is hard to analyse from the displayed information as it is multi-categorical. But by mean value we can predict that most of them come under typical and atypical angina.

### Person’s resting blood pressure minimum value is 94 Maximum value is 200 where 75% of the data falls below 140 indicating there are some outliers.

### Person's cholesterol level minimum value is 125 Maximum value is 564 where 75% of the data falls below 274 indicating there could be some outliers.

### Visualizing the target count in the heart dataset

### C:\Users\611242028\Desktop\download.png

### Graphs between categorical variables and Target

### Sex vs Target

### C:\Users\611242028\Desktop\download.png

### Male seems more effected to heart disease than female the percentage of male records are significantly more than that of female

### Percentage of Male: 68.32%

### Percentage of female: 31.68%

### Chest Pain vs Target

### C:\Users\611242028\Desktop\download.png

### We can see that chest pain of type 2 (non-anginal pain) are more prone to Heart disease patients with chest pain of type 0 (typical angina) are not as significantly affected with heart disease as other cp types

### No of Major Vessels vs Target

### C:\Users\611242028\Desktop\download.png

### ca with type 1 are more to heart disease than other types

### ECG vs Target

### C:\Users\611242028\Desktop\download.png

### The below plot is column bar chart representing target vs ECG Measurements (Electro Cardio Gram). The below plot shows that the more number of patients not likely to suffer from heart diseases are having restscg value 0 whereas more number of people have restecg value 1 in case of more likelihood of suffering from a heart disease.

### Patient’s restecg of type 1 are most prone to heart disease.

### Thalassemia vs Target

### C:\Users\611242028\Desktop\download.png

### Patients with thal type 2 are more prone to heart disease

### Patients with thal type 3 are least effected by heart disease.

### C:\Users\611242028\Desktop\download.png

### Patients with slope of type 2 are most prone to heart disease.

### Patients with slope of type 1 are least effected to heart disease when compared to other slope types.

### Exercise induced angina vs Target

### C:\Users\611242028\Desktop\download.png

### Patients who have exang seems to be least affected by heart disease. Patients who doesn’t have exang are most affected by heart disease.

### C:\Users\611242028\Desktop\download.png

### Resting blood pressure and Cholesterol features seems to have outliers. But we need to understand how the outliers of these features are effecting the Target variable.

### C:\Users\611242028\Desktop\download.png

### The above Bivariate plot between cholesterol levels and target suggests that the Patients likely to suffer from heart diseases are having higher cholesterol levels in comparison to the patients with target 0(likely to not suffer from the heart diseases.

### Hence, we can infer from the above plot that the cholesterol levels plays an important role in determining heart diseases. We all must keep our cholesterol levels in control as possible.

### C:\Users\611242028\Desktop\download.png

### Resting Blood Pressure, The above Bivariate plot between tresbps(the resting blood pressure of a patient), and the target which says that whether the patient is suffering from the heart disease or not. The plot clearly suggests that the patients who are most likely to not suffer from the disease have a slightly greater blood pressure than the patients who have heart diseases.

### Creating the ML model:

### Converting the categorical variables to type category.

### Creating dummy variables of categorical attributes.

### Applying standardScalar on to remove the mean and also scales the data to unit variance.

### We will try to apply various Classification Machine learning Algorithms and will look ways to improve the model from their after. The Models are:

### Logistic regression

### Support vector machine (SVM)

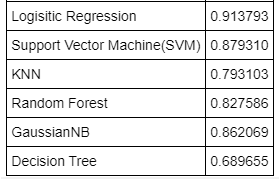
### K near neighbourhood (kNN)

### RandomForestClassifier

### GaussianNB

### DecisionTreeClassifier algorithms

### The Accuracy of the above models



### KNN have achieved maximum accuracy with 8 neighbours with accuracy 86%.

### Conclusion:

### Logistic regression have scored the highest accuracy followed by SVM and KNN

### We can achieve much better accuracy with deep learning models