Shaan Dahar

Health Care Dataset ML Project

Preliminary Analysis

No rows had missing values, but there were some duplicates removed as well as a few outliers in the trestbps category. Also removed a few rows where ca was an outlier by age.

Average of dataset:

A screenshot of a computer

Description automatically generated

Mode of dataset:

A close up of a number

Description automatically generated

Categorical Variables:

Categorical variables are sex, cp, fbs, exang, and thal. 2/3rds of entries were men, CP trended towards 1, FBS towards 0, and thal towards 2.

Age and CVD:

This is what a countplot of age reveals.

A colorful graph with numbers

Description automatically generated with medium confidence

Here’s a barplot of age vs exang.A graph of different colored lines

Description automatically generated

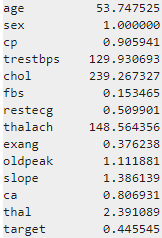
Here’s a barplot of age vs ca: A graph of different colored lines

Description automatically generated

we see a steady increase in CA and exang with age.

Composition of patients in relation to Sex category

Separating the dataset into two dataframes, one for males and one for females, we can then find the average/mean values. Here is the male average on left, and female average on right

 A screenshot of a computer

Description automatically generated

Women are older and have higher chest pain, resting bps, cholesterol, resting ecg, thalach, and slope on average. Men have higher fbs, exang, oldpeak, ca, and thal on average.

Can we detect heart attacks based on resting blood pressure anomalies?

What’s the relationship between cholesterol levels and target variable?

A graph of different colored bars

Description automatically generated

What’s the relationship between peak exercise and the occurrence of a heart attack?

Is Thalassemia a major cause of CVD?

How do the other variables effect the occurrence of CVD?

Pair Plot:

Baseline Model to predict heart attack risk

Logistic Regression

A screenshot of a graph

Description automatically generated

Random Forest

A screenshot of a computer screen

Description automatically generated



Correlation Analysis