MAKERERE UNIVERSITY

Machine Learning

Lebuga Bosco | 2024/HD05/25333U | 2400725333

Exploratory Data Analysis Process for Heart Disease Prediction

Problem Statement

According to WHO, Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age.

Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. This calls for improved early heart disease detection algorithms or machine learning models.

How best can we improve heart disease prediction models?

Questions Before

1. How do we predict if a patient is at high risk of heart disease?
2. Which factors are predominant in detecting heart disease?

Questions After

1. What is the average age for heart attack for males and females?
2. Which gender is more at a risk of heart disease?
3. What are the trends and relationships between different features of my data set?
4. Are there outliers in my dataset?
5. What are the missing values and duplicates in my dataset?

Step1: Gathering data

My dataset for heart disease was obtained from Github from the below address. It was then imported and read using the panda’s data frame library.

https://github.com/kb22/Heart-Disease-Prediction/blob/master/dataset.csv

Step1: Understanding the data

My dataset for heart disease prediction has 14 features and 303 records. The predicted attribute ('target') shows the presence of heart disease in the heart (where by assume its integer value 1 = heart disease and integer value 0 = no heart disease). The output is predicted from the other 13 features of the dataset. The different major columns that can predict heart disease are

- **"age"** The age of the person at the time
- **"sex"** The gender of the person as at birth (has 2 values, 1= male; 0 = female)
- **"cp"** The chest pain type (4 values); 0= Typical angina, 1 = Atypical Angina, 2= Non-Anginal, 3 = Asymptomatic
- **"trestbps"** The resting blood pressure of the person
- **"chol"** Serum cholesterol in mg/dL
- **"fbs"** Fasting blood sugar > 120 mg/dL, 0 = normal, 1= diabetic
- **"restecg"** Resting electrocardiographic results (3 values, 0 = normal, 1 = wave abnormality, 2 = probable left ventricular hypertrophy)
- **"thalach"** Maximum heart rate achieved
- **"exang"** Exercise induced angina(0 = no, 1 = yes)
- **"oldpeak"** Old peak = ST depression induced by exercise relative to rest(0 = upsloping, 1 = flat, 2 = downsloping)
- **"ca"** Number of major vessels (0-3) colored by fluoroscopy 0 = normal; 1 = fixed defect; 2 = reversible defect.
- **"thal"** The inherited blood disorder caused when the body does not make enough hemoglobin (4 values)
- **"target" Is out target variable or diagnosis of heart disease (0 = no heart disease, 1 = heart disease)**

Observations from understanding the dataset

- In the dataset, 13 columns are of the data type Int64 except only 'oldpeak' column which is of float64 data type. There are 302 entries from 0 to 302, and memory usage = 33.3KB
- All the features of the dataset are numerical in nature and there are no categorical values.
- Some standard deviations are slightly higher than the mean values for the features of 'cp', 'exang', 'oldpeak', 'ca', and this indicates high variations between values and abnormal distribution of data.
- The display shows that the feature "chol" has min =126 and max =564, and very high standard deviation of 51.83 and this implies there may be potential outliers in that column
- Checking the count of records, its observed that 1 record appeared twice(male patient of age 35, resting bps 138, chol 175, maximum bps 173 and has heart disease presence(1))

Step3: Data Cleaning and Data Wrangling

This process started with checking the dataset features which are not useful for the prediction of heart disease, checking the datatypes, renaming of features, checking for null values, duplicates and potential outliers in our dataset.

- To add more meaning to the dataset, I have renamed some of the features like 'chol' : 'cholesterol', and more others as can be seen in the notebook
- All the 13 features were very relevant for predicting presence or absence of heart disease and so none was dropped off. The datatypes are all numerical so no need for conversion.
- Checked **null values** and the results shows there are no empty records or null values in our dataset.
- Checked and detected **1 duplicated record** in our dataset (record 164 exactly matches record 165 of the dataset). Both records have the same values throughout for all the features. This might have been due errors in data capturing or data entry.
- After carefully removing the duplicates, the shape of our new dataset becomes **302** rows and 14 columns implying 1 record has been dropped from the dataset.
- Checked **potential outliers** using the boxplot for the columns of age, rest_bps, cholesterol, max_bps, and oldpeak and the findings are that 'age' column has no outliers, the 'max_bps' values below 80 are outliers, for 'rest_bps' most of the values above 170 are outliers, for 'cholesterol' column, values above 380 are outliers, for 'oldpeak' all values above 4 are outliers
- Those outliers were removed by **Trimming method** where by all rows which have outliers values are completely dropped and giving a new dataset.
- After removing all the potential outliers, our finally dataset has a shape of 293 less of 302 which means about 302-293 =9 outliers were detected and removed from the dataset
- While number of outliers for the affected columns include (old peak =5, max_bps =3, rest_bps = 2, cholesterol = 4), totaling to 14 outliers and yet only 9 removed. This implies the remaining 4 outliers have minimal effect on the output and hence not considered for removal by that trimming method
- It is also observed that our dataset has both categorical and continuous variables separated as continous_var(age, rest_bps, cholesterol, max_bps, oldpeak) and categorical_var(sex, chest_pain, fasting_bps, restecg, exercise_angina, slope, major_vessels, blood_disoder), target_varibale(target)

Step4: Exploratory Analysis and Visualization

Here I look at univariate, Bivariate and Multivariate analysis to visualize the relationship and establish trends of variations of features

a. **Univariate analysis**

CountPlots for each of the categorical variables in the data set depicted the following

- There are more Male patients than females in the dataset
- Majority of the patients in the dataset are suffering from Typical Angina chest pain followed by those with Non-Anginal chest pains. Fewer pains are affected by asymptomatic chest pain type.
- Most of the patients have normal blood sugar levels of <120mg/dL. Less thab 49 patients have fasting blood sugars >120mg/dL and are diabetic.
- Many of the patients do not have exercise induced angina
- Majority of the patients have normal blood vessels(0 major blood vessel)
- Many of the patients have heart disease

A univariate analysis of the continuous variables reveals the following

- Majority of the patients are within the age bracket of around 41-61.(plot 2.1)
- Most of the patients have rest blood pressure of the range 120-140(plot 2.2)
- Many patients have cholesterol levels around 220-260
- Most patients have heart rates of 136 – 165
- Most of the patients are within the old pick values of 0 and 1 (upsloping and flat)

b. **Bivariate Analysis**

We then perform an analysis of features in relations to the target variables and the following observations were made

- There is no much significant relationship between age and heart attack as depicted in plot 3.1. Advancements in age does not directly correlate with heart attack
- Increase in blood pressure doesn’t directly contribute to heart attack as depicted in plot 3.2
- In polt3.3 cholesterol has no direct contribution to heart attack although there are some cases of cholesterol around 200 have heart attacks
- Patients with very high heart rates are more prone to heart attack than those with lower heart rates
- Most of the patients with Typical Angina have no heart attacks while those patients with Non-Anginal are prone to heart disease
- Patients with wave abnormality of electrocardiograph are more prone to heart attacks

Correlation between the Variables

As depicted in the plot4.1, there is no strong correlation between all the variables of the dataset. Only very few features are able to show a positive and negative correlations of around 0.4. These variables are chest_pain(0.4), max_heartRate(0.4), exercise_angina(-0.4), oldpeak(-0.4), major_vessels(-0.4)

c. Multivariate Analysis

According to plot 5.1, following observations are made;

- Males are more prone to heart attacks than females
- When fasting blood sugar <120, patients become less prone to heart attacks
- Oldpeak at 0 is more prone
- Slope at is more prone
- Major blood vessel a 1 is more prone
- Blood disorder at 2 is more prone
- More prone when maximum heart rate >200

General Observations and Conclusions

- There are no Null values in our dataset
- There was 1 duplicated record found and removed from the dataset.
- As depicted in plots 2.1 to 2.5, there appears to be outliers in every continuous variable
- The notion that an increase in age leads to increase in heart attack seems false here since there is no strong relationship between age and heart attack as depicted in plot 3.1
- In plots 3.2, and 3.3, its noticeable that blood pressure and cholesterol do not contribute much to heart attack
- In plot3.4, people with high heart rates are more prone to heart disease
- In plot 3.5, patients with Non-Anginal chest pains are more likely to get heart attacks.
- Patients having abnormal wave of electrocardiograph(restecg =1) are more prone to heart attacks
- Patients with no exercise induced angina are more prone to heart disease plot 3.8
- According to the heatmap plot 4.1, there is currently no strong linear correlation between all the continuous variables

Inferences

From the analysis of the dataset for heart disease, it can be deduced that the following factors can directly contribute to heart attacks

- Cheskpain(non-anginal),
- fasting blood pressure(?120mg/dL0
- electrocardiograph(with abnormal wave option)
- high heart rates (>200)
- exercise induced angina
- old peak
- major vessels(value 1)
- oldpeak
- slope

Meanwhile the following factors do not much affect the heart attack;

- age,
- sex,
- blood pressure,
- cholesterol