Prediction of Heart Disease

Analysis of Heart Disease Dataset

Sonal Jain 201961083 CS989:Big Data Fundamentals

Table of Contents

Table of Tables

Table of Figures

Introduction:

HEART DISEASES:

Heart Diseases rank first for maximum death counts in Australia, Canada, United Kingdom and United States. Heart diseases refer to illnesses caused my deformities in one's heart. While some hearts have issues since birth like Septal defects and Obstruction defects, some occur over time like Coronary artery disease where Coronary arteries are affected because of huge amounts of cholesterol while others could be genetic. There are symptoms which can help determine if a person's heart is weak like the amounts of chest pain, but one cant know for sure unless they get themselves checked up.

While medical science has evolved over the years, and has provided us with deeper knowledge and treatments it is still very crucial to detect the presence of heart disease in one's body as early as possible. The earlier its detected the better is the treatment. The motivation behind this project is to create a model which would facilitate this process of heart disease detection in humans based on simple factors so that people can be aware of the chances of their heart having a disease.

Aim:

The aim of this project is to predict if one has or can face a heart disease in a very simple and sophisticated way. User needs to input some values about their body and health, and the model would provide them with an answer of yes or no. The model strives to achieve maximum accuracy.

Who would be interested in this project?

Following are some of the domains where this project could be implemented -

1. Model can be used by medical institutions and hospitals to provide a quick response to patient's health report and predict whether the person's condition is favorable for any heart disease.
2. Model can be fit into a mobile application where the user can just input their results from their health report.
3. Could also be used by pharmaceutical companies for prescribing medicines.
4. Could have a collaboration with a fitness app.
5. General Public- People who just want to be sure about the status of their heart.

HEART DISEASE DATASET: PUBLIC HEALTH DATASET

Source-

This dataset was taken from Kaggle.com as comma separated file(csv). It dates back to 1988.The datasets of four countries namely: Cleveland, Hungary, Switzerland, and Long Beach V were combined to form this dataset. It originally had 74 attributes but were then narrowed to 14 for better analysis. The dataset was loaded into a pandas data-frame. It contains 1025 rows and 14 columns with indexes. It doesn't have any null or missing values (isnull().sum() function was used to check this). If it contained any null values they would have been replaced by the column mean as a part of data cleansing. The dataset mainly consists of integer values which are either continuous, categorical or discrete.

It contains the following 14 attributes-

Table 1

Attribute	Description	Type
age	Age Of The Patient.	Continuous.
sex	Gender Of The Patient.	Categorical- contains 2 values- {0: female, 1: male}
cp	Chest Pain Type	Categorical-contains 4 values- {0: typical angina, 1: atypical angina, 2: non-angina, 3: asymptomatic angina}
trestbps	Resting Blood Pressure	Continuous (mmhg on admission to the hospital)
chol	Cholesterol Level	Continuous (mg/dl)
fbs	Fasting Blood Sugar	Categorical- contains 2 values- {0: <= 120 mg/dl, 1: > 120 mg/dl}
restecg	Resting Electrocardiography	Categorical-contains 3 values-{0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy}
thalach	Maximum Heart Rate Achieved	Continuous.
exang	Exercise Induced Angina	Categorical- contains 2 values- {0: no, 1: yes}
oldpeak	ST Depression Induced By Exercise Relative To Rest	Continuous.
slope	Slope Of Peak Exercise ST Segment	Categorical-contains 3 values- {1: upsloping, 2: flat, 3: downsloping}
ca	Number Of Major Vessels Colored By Fluoroscopy	Discrete (0,1,2,3,4)
thal	Type Of Defect	Categorical-contains 3 values- {0: normal,1: fixed defect, 2: reversible defect}
target	Heart Disease	Categorical-contains 2 values {0: no, 1: yes}

Summary Statistics

First step into creating a model is to understand the dataset. This section contains various summary statistics of the dataset.

Table 2

No.	CHARTS
1.	Correlation heatmap.
2.	Count of heart diseases in males and females.
3.	Age distribution of patients.
4.	Relationship between type of Chest pain, Major Vessels Colored By Fluoroscopy and slope of the peak exercise with heart diseases.
5.	Levels of resting blood pressure and cholesterol with heart diseases.
6.	Distribution of Maximum Heart Rate

Correlation Matrix

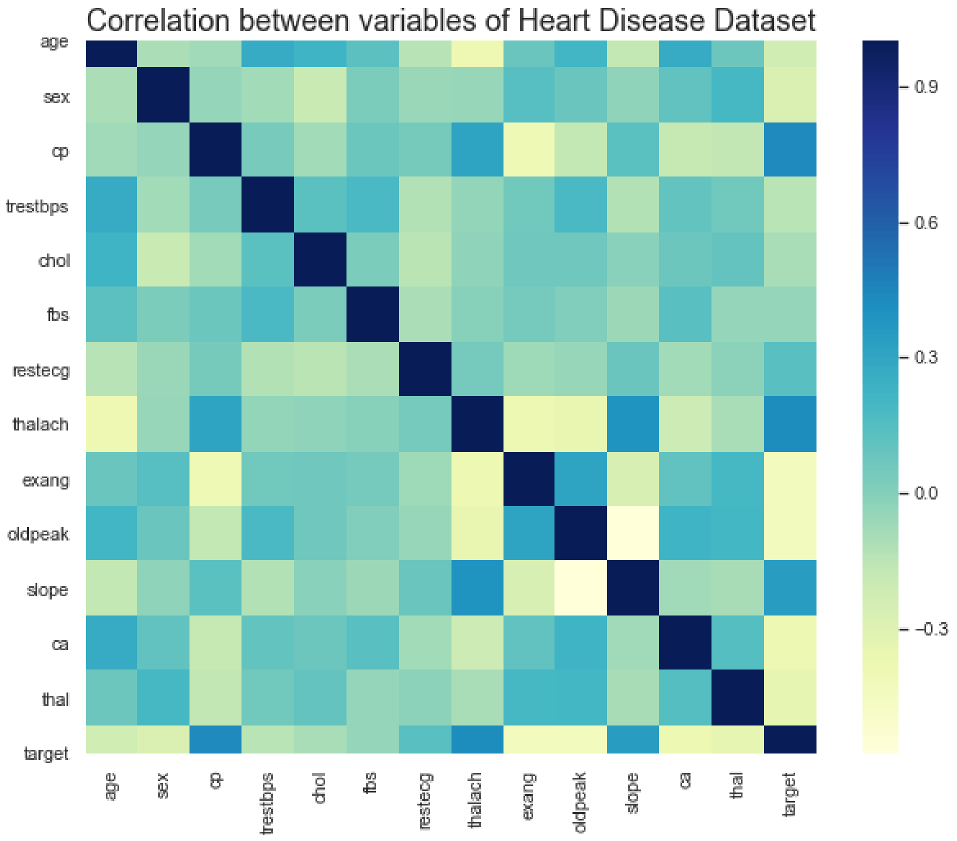
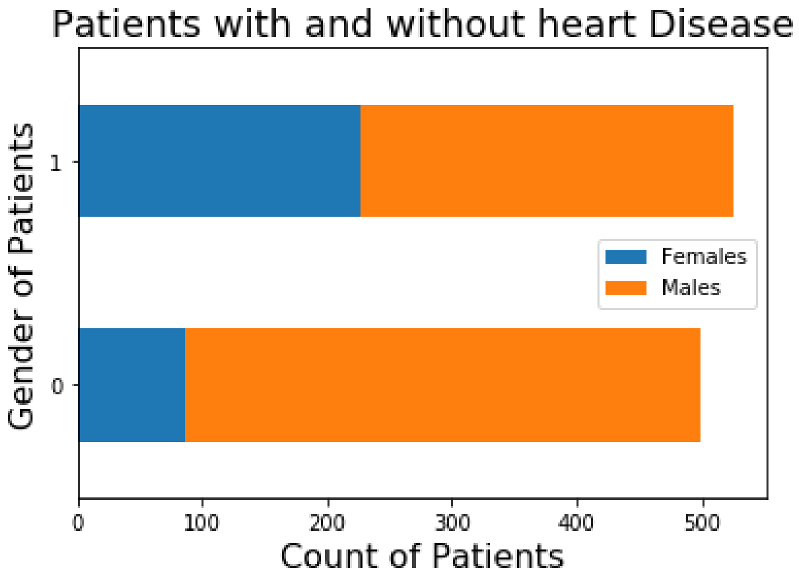


Figure 1

Above heatmap is created with seaborn library. It shows the correlation between different variables of the dataset. Darker the colour higher the correlation. We can see that ‘target’ – the variable that represents the presence of heart disease has a positive linear correlation with variable ‘cp, ‘thalach’, ‘slope’ which are Types Of Chest Pain, Maximum Heart Rate Achieved and Slope Of Peak Exercise ST Segment respectively. It has negative correlation with ‘exang’ and ‘ca’ and which are Exercise Induced Angina and Number Of Major Vessels Coloured By Fluoroscopy respectively.

Following visualizations aim to study these positive and negative correlations better.

COUNT OF HEART DISEASES IN MALES AND FEMALES.



1	With heart disease
0	Without heart disease

Figure 2

Above chart was made with pyplot package from matplotlib library. It’s a stacked horizontal bar chart showing the gender distribution of patients. On the x axis we have the number of patients and on the y axis we have 2 values where 0 represents patients without heart disease and 1 represents patients with heart diseases.

From this we can gather that the number of females having heart problems is slightly more than males having heart problems. This graph gives a very basic idea but definitive conclusions cannot be drawn from it.

AGE DISTRIBUTION OF PATIENTS.

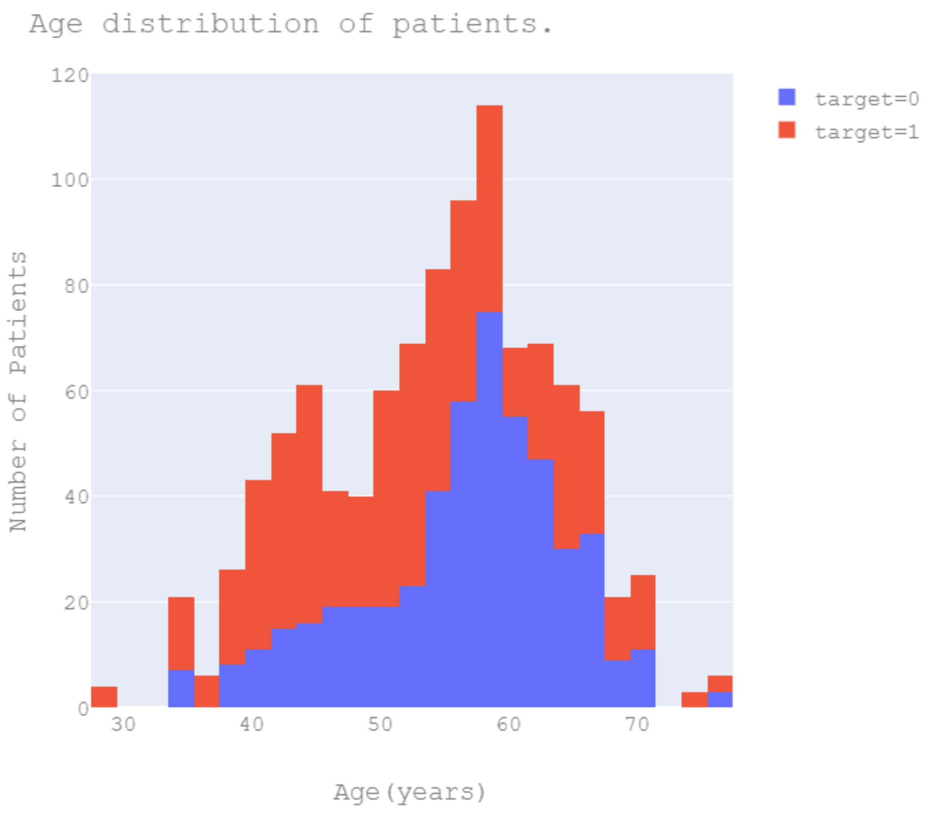


Figure 3

This histogram was made with plotly library. On the x axis it shows the age while on y axis it shows the number of patients. The area in blue represents patients without health diseases while area in red represents patients with health diseases. From the above chart we can infer that age is an important factor while considering the conditions of heart because maximum patients with heart diseases tend to fall under the age bar of 40 to 65 years of age.

RELATIONSHIP BETWEEN TYPES OF CHEST PAIN AND HEART DISEASE.

[0: typical angina, 1: atypical angina, 2: non-angina, 3: asymptomatic angina]

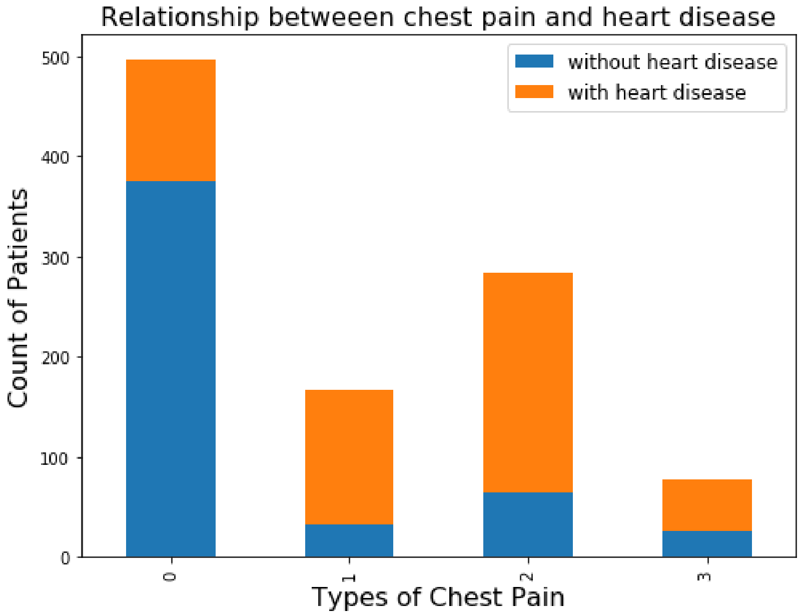


Figure 4

The above stacked bar chart is made with pyplot package from matplotlib library. The x axis shows the types of chest pain and the y axis shows the number of patients. From the chart we can see that patients with chest pain type 1 (atypical angina), 2 (non-angina) and 3(asymptomatic angina) are more prone to heart disease as compared to chest pain type 0 (typical angina). We can say that the type of chest pain is an important factor while considering the conditions of heart as the values are discrete and they give discrete results.

Relationship between Number of Major Vessels Colored By Fluoroscopy and Heart Disease

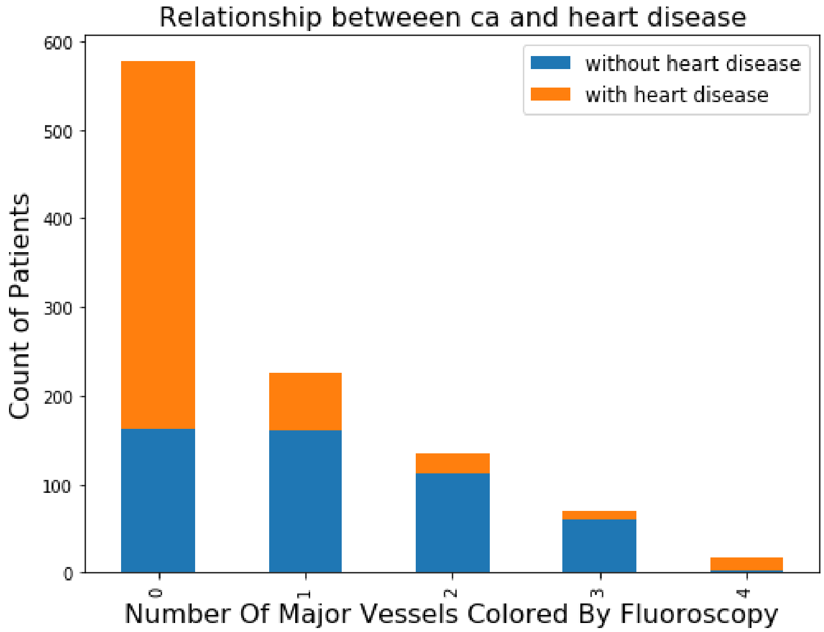


Figure 5

The above stacked bar chart is created with pyplot package of matplotlib library. On the x-axis it shows the discrete variables- 0,1,2,3,4 of Number of Major Vessels Colored By Fluoroscopy. From the chart we can infer that patients with 1, 2 and 3 seemed to have less chances of getting a heart disease as compared to 0 and 4. Also, among the sample size the maximum people who suffered from heart disease didn't have any major vessels colored by Fluoroscopy. But those who had all 4 vessels colored by Fluoroscopy seemed to have a definite chance of heart disease. Hence, this variable plays an important role is determining heart diseases.

RELATIONSHIP BETWEEN SLOPE AND HEART DISEASE.

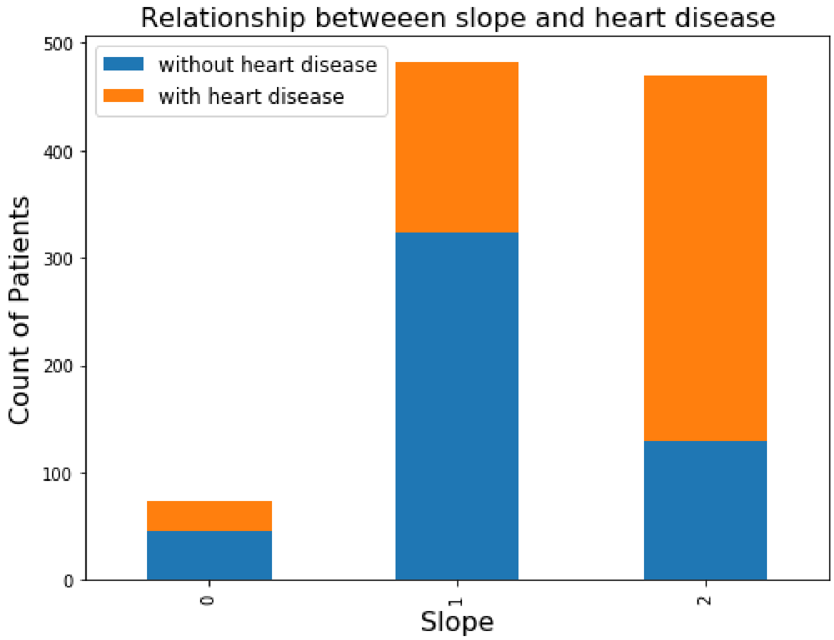


Figure 6

Above chart has been made with pyplot from matplotlib library. X-axis represents Slope and Y-axis represents the number of patients. The variable ‘slope’ represents the Slope Of Peak Exercise ST Segment. It has a positive linear correlation with heart disease. Higher the value of slope more are the number of patients with heart disease. This makes slope an important factor while considering the conditions of heart.

LEVELS OF RESTING BLOOD PRESSURE AND CHOLESTEROL TO HEART DISEASE.

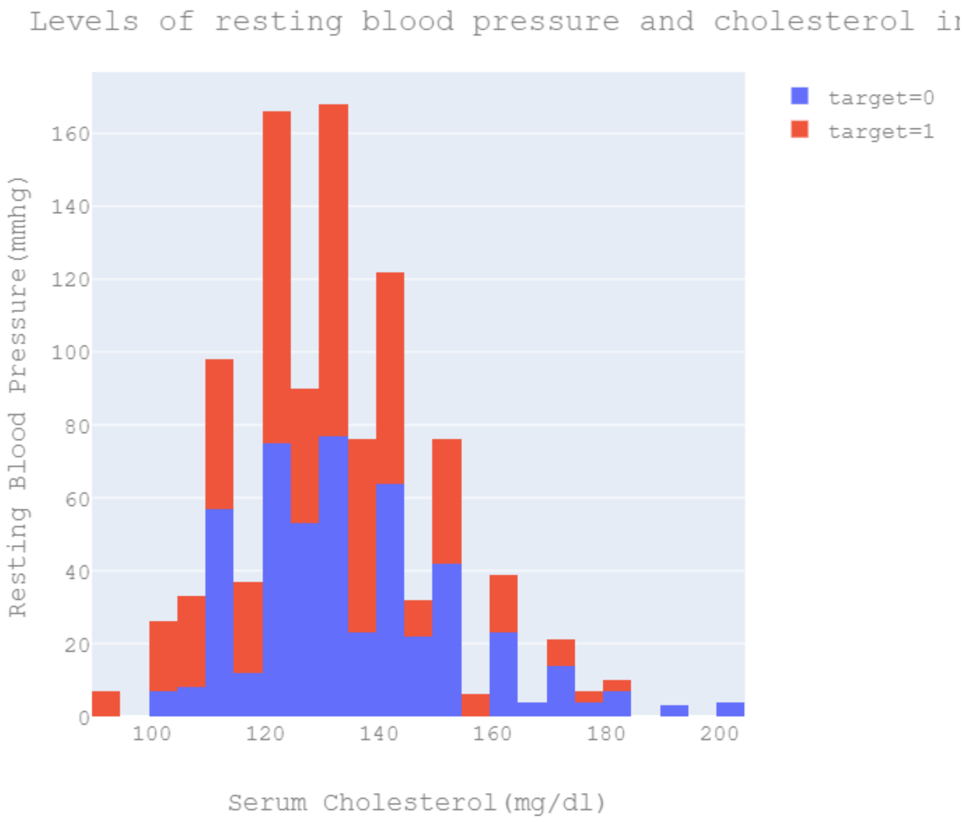


Figure 7

The histogram depicts the resting blood pressure(trestbps) on the x axis and serum cholesterol level on y axis. The area in blue shows the number of patients without heart disease and red shows the number of patients with heart disease. From this it can be inferred that maximum patients with heart disease tend to have high levels of cholesterol(mg/dl) and resting blood pressure higher than 120 mmhg as compared to people without. It can be said that Serum Cholesterol and Resulting Blood Pressure are crucial variables that need to be considered while examining the conditions of the heart.

Distribution of Maximum Heart Rate

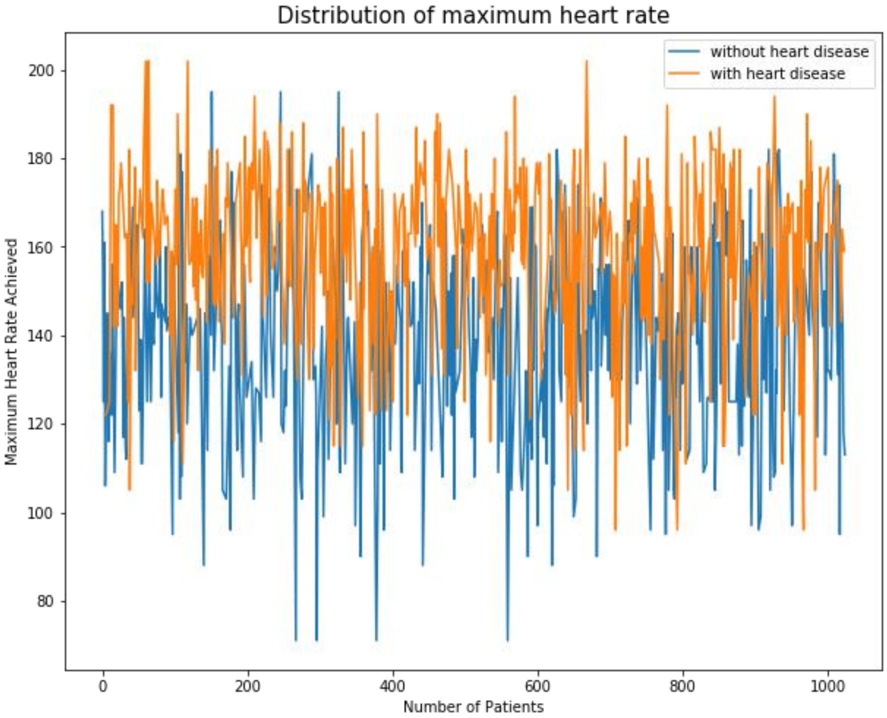


Figure 8

The above graph is made with pyplot from matplotlib library. This line plot shows the distribution of patients and their maximum heart rate. From above we can see a linear relationship i.e. as the maximum heart rate increases the number of patients with heart disease also increases or patients without heart diseases mostly have lower values of maximum heart rates. Hence making ‘thalach’ an important deciding factor determining the presence of heart disease.

Summary:

From the above visualizations following observations can be drawn-

I. Gender doesn’t seem to have a strong impact on the dataset as the values remain fairly similar for both males and females. This is because the sample size is not large enough or equal.
II. Age group of 40-65 years were more likely to have a heart disease, hence age is an important factor for determining the presence of heart disease.
III. Type of Chest Pain, Slope and Maximum Heart Rate have a strong positive correlation with ‘target’ (heart disease) and hence are important factors for determining the presence of heart disease.
IV. ‘exang’,’thal’,’ca’,’oldpeak’ are the variables that show negative correlation with ‘target’(heart disease) and hence are important factors for determining the presence of heart disease.
V. ‘trestbps’,‘chol’,’fbs’,’restecg’ also seem to have an impact in the decision making of the presence of heart disease as they have positive correlation with ‘target’.

Now that we know which variables are important we will use them to create models through supervised and un supervised learning to predict the presence of heart disease.

UNSUPERVISED LEARNING

Unsupervised is a machine learning technique used to analyze unlabeled data.

"Unsupervised learning is defined as the task performed by algorithms that learn from a training set of unlabeled or unannotated examples, using the features of the inputs to categorize them according to some geometric or statistical criteria". - *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. Igual L., Seguí S. Unsupervised Learning Chapter*

CLUSTERING

Clustering is a unsupervised learning technique of identifying similar objects and putting them in clusters or groups. Here, the labels are not defined beforehand hence the name unsupervised.

"In machine learning, unsupervised refers to the problem of finding hidden structure within unlabeled data."- *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*

K-MEANS

K-means is a clustering algorithm which creates 'k' specified amount of clusters in such a way that most similar data are put in the same cluster. Every element is assigned one cluster and the every cluster is defined by its mean. This mean is also known as centroid.

K-means clustering was chosen for analyzing this dataset as it is known for its application in the field of Medical. The clusters created are widely used for making decisions regarding clinical trials and also for classification of diseases.

METHODOLOGY-

First we determine which values we need to work on and what output we want to predict. Here we will predict the 'target' which is the presence of heart disease and our attributes for this prediction will be all the other 13 variables in the dataset.

Table 3

Attributes	Age, Sex, Type Of Chest Pain, Resting Blood Pressure, Serum Cholesterol Level, Fasting Blood Sugar, Resting Electrocardiography, Maximum Heart Rate Achieved, Exercise Induced Angina, ST Depression, Slope, Vessels Colored By Fluoroscopy And Thal.
Outcome	Presence of heart disease.

Scaling the Attributes- Its essential to preprocess the data by transforming and scaling it. This avoids having distorted data modelling results and helps in fitting the attributes well in the model. Scale package from sklearn preprocessing was used to scale the attributes for both unsupervised as well as supervised learning.

Different combinations were tried for the kmeans attributes such as-

a. Taking all the variables as attributes except sex.
b. Taking only the variables that showed strong correlation -positive and negative with target.
c. Taking only continuous variables of the dataset.
d. Taking all the variables as attributes(except target because that is the outcome to be predicted).

To check which set of attributes would best suit the model -completeness score, homogeneity score and silhouette score are used. And the best scores were obtained for combination d. The scores are shown in the table below-

Table 4

Score Type	Scores(rounded up)
Completeness	0.11
Homogeneity	0.40
Silhouette	0.16

Homogeneity Score- Score 1 shows that class of members is similar among the cluster.

Completeness Score- Score 1 shows that the members of the class belong to the same cluster.

Silhouette Score- Score 1 shows that the clusters are dense and well separated from other clusters.

Findings from Unsupervised Learning

As our outcome was predicting if the patient is suffering from a heart disease, the kmeans clustering model produced 2 clusters- one for the presence of heart disease and the other for no heart disease.

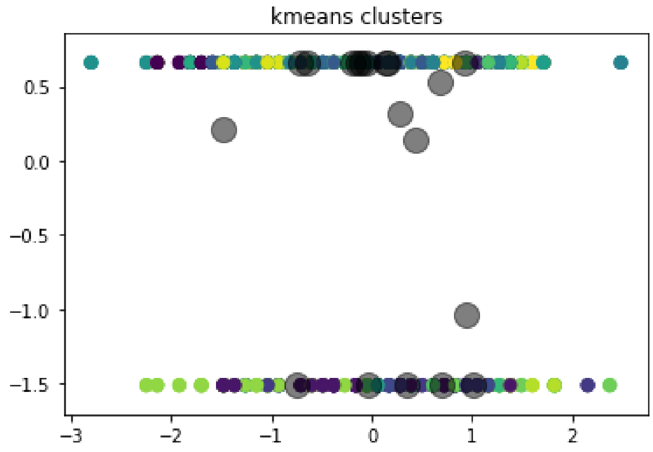


Figure 9

Above visualization was created using pyplot package from matplotlib library. It shows the clusters created by k-means clustering model and its centroids.

It is possible to get a idea about the conditions of one's heart by just inputting the Electrocardiography report results which consist of most of the variables used for this model. The accuracy of the model is not 100% but still helps in determining whether one needs to worry about their health or need to take any precautionary measures such as visiting a heart doctor or changing their lifestyle to a healthier one.

SUPERVISED LEARNING:

“Algorithms which learn from a training set of labeled examples (exemplars) to generalize to the set of all possible inputs.”- *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. Igual L., Seguí S. Supervised Learning Chapter*

Supervised Learning is a technique of Machine Learning which is used to predict outcomes , but unlike unsupervised this technique uses labeled data. Here we train the model and test the results while in unsupervised learning the model self-learns and predicts the outcomes.

METHODOLOGY-

First we need to determine our attributes and outcomes and then scale the attributes. This method is similar to that of unsupervised and we are using the same values for attributes and outcome.

Next we divide the data to train and test sets. As the names suggest train set is the part of data which the user trains or manipulates while test set is the data on which the model works to predict outcomes based on the training given.

Many alternatives were tried for splitting the data such as-

a. 80-20 split (80% testing set and 20% training set)
b. 70-30 split (70% testing set and 30% training set)
c. 65-35 split (65% testing set and 35% training set)

All the alternatives were worked upon and the results were evaluated. It was found that 70-30 split gave the best results.

Supervised Learning consist of many methods and classifiers. Some of them are

1. Logistic Regression
2. Decision Tree Classifier
3. K-Nearest Neighbors
4. Naïve Bayes
5. Random Forest Classifier

Since our outcome was to predict a yes or no result it would be a wise choice to use a classifier as they have to predict a discrete number of results. For our model Decision Tree Classifier and K-Nearest Neighbor seemed the best fit. Naïve Bayes was also considered and evaluated but it didn’t give the best results as compared to the ones mentioned above. Also, Random Forest Classifier gives good results but only for large data sets and our dataset is comparatively small. Decision Tree Classifier and K-Nearest Neighbor proved to be more accurate to predict the presence of heart disease as they had the least deviation from the exact values.

DECISION TREE CLASSIFIER

“A decision tree (also called prediction tree) uses a tree structure to specify sequences of decisions and consequences.”- *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data - Chapter 7 Advanced Analytical Theory and Methods: Classification.* They work best for continuous and categorical variables.

307 values were trained and 718 tested in the Decision Tree Classifier. The results were put into a data-frame called heart_test_resuts and then compared with the actual outcomes. Following bar graph made from matplotlib pyplot shows that 665 values were predicted correctly while 53 were incorrect.

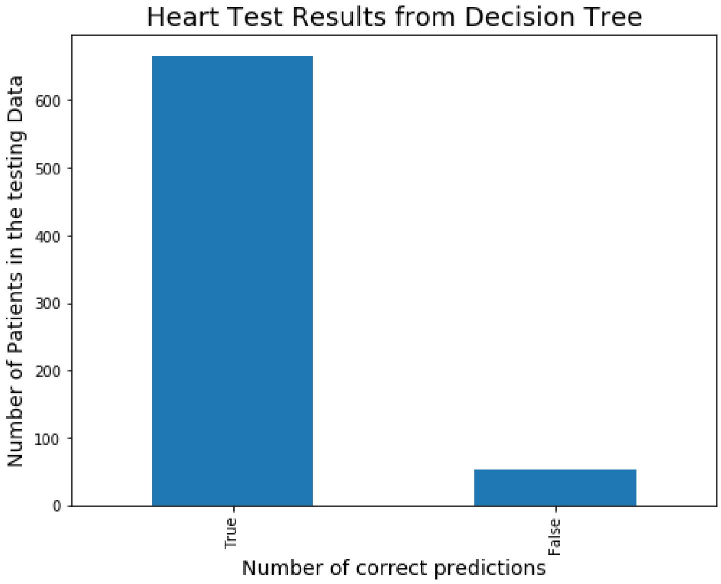


Figure 10

To check the accuracy and precision of the classifier we use ‘metrics.classification_report()’ and confusion matrix().

Findings-

1.Classification report gave us a precision accuracy value of 93%, which means our model prediction was 93% accurate.

2. Confusion Matrix results are shown in the table below

Table 5

	Predicted: Yes	Predicted: No
Actual: Yes	288	60
Actual: No	72	298

K-NEAREST NEIGHBOURS

KNN is a supervised learning algorithm where the learning of the classifier is based nearness of the data from the other. Nearness is based on Euclidean Distance.

307 values were trained and 718 tested in the Decision Tree Classifier. The results were put into a data-frame called heart_test_resuts and then compared with the actual outcomes. Following bar graph made from matplotlib pyplot shows that 586 values were predicted correctly while 132 were incorrect.

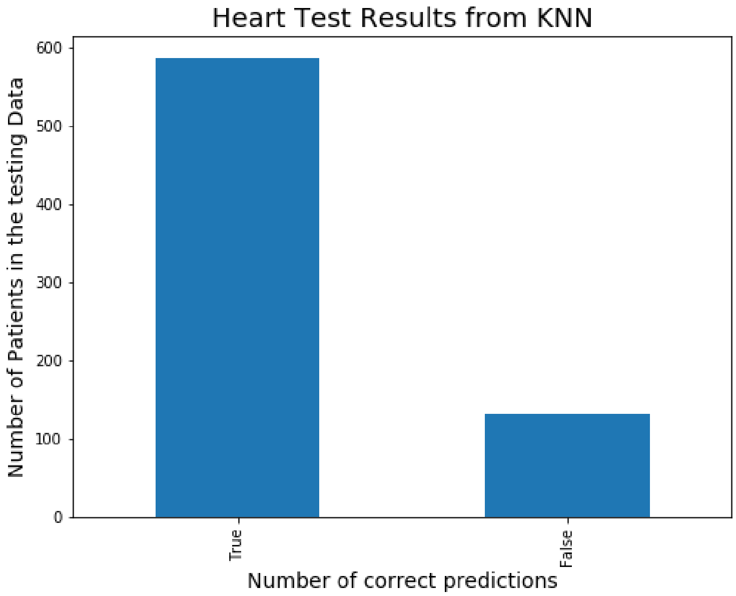


Figure 11

To check the accuracy and precision of the classifier we use 'metrics.classification_report()' and confusion matrix().

Findings-

1.Classification report gave us a precision accuracy value of 82%, which means our model prediction was 82% accurate.

2. Confusion Matrix results are shown in the table below Table 6

	Predicted: Yes	**Predicted: No**
Actual: Yes	288	60
Actual: No	72	298

Conclusion-

From supervised learning classifiers it can be said that the model makes good predictions with 93% and 82% precision. It can be practically implemented for predicting the presence of heart disease in a patient.

REFLECTIONS

1.The dataset was good enough for a very basic analysis, but to get more clear and accurate results it would be better to have a large sample size.

2.Also the dataset only dealt with a particular set of countries, the model would have wider applications if the dataset involved data from other countries as well to show country wise comparison.

3.The evaluation would be better if the dataset contained equal demographics based on sex and age.

4. The above mentioned points could help improve the k-means clustering scores, as in the current evaluation even though they were positive they weren’t that accurate.

5. Most of the variables were binary which made the classification and clustering easy, but if the dataset contained more continuous data techniques such as Regression and Random Forest could be applied.

ENVIRONMENT

Anaconda- Jupyter Notebook- Version-6.0.1

REFRENECES

https://www.kaggle.com/johnsmith88/heart-disease-dataset

https://www.medicalnewstoday.com/articles/237191.php#causes

https://matplotlib.org/

https://python-graph-gallery.com/

https://scikit-learn.org/

https://plot.ly/python

Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data

Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. Igual L, SeguíS.