

A
Project Report
on

Big Data Analytics

Heart Disease Prediction

Submitted by-

B.G.Riteesh Ram Chander

1800206C203

CSE

Venkata Sai Sreekar MP

1800190C203

CSE

Under the guidance of

Dr. Yogesh Gupta

Associate Professor



Department of Computer Science and Engineering
SCHOOL OF ENGINEERING AND TECHNOLOGY
BML MUNJAL UNIVERSITY GURGAON-122413, INDIA

05 May, 2021

Acknowledgement

I thank my fellow partner Venkata Sai Sreekar M.P for his assistance and interest in the project. I also thank Professor. Yogesh Gupta Sir for his continuous support in the fulfillment of this project. His guidance has helped us in gaining the knowledge and thorough understanding of this course work.

Thanking You

Bollavaram Golla Riteesh Ram Chander

Contents

S.No	Title	Page No
1.	Abstract	4
2.	Motivation	5
3.	Introduction	6
4.	Problem Statement	6
5.	Literature Review	6-8
6.	Methodology	8-19
7.	Results and Discussion	19
8.	Conclusion	19
9.	References	19

Abstract:

In US for every 40 seconds some one encounters heart stroke and also 1 out of every 19 deaths is from heart stroke/heart attack. By this information we can understand that how severe it is and we cannot neglect it. One can avoid heart attacks or take care of him self if he/she knows that whether they encounter with heart stroke in near future, so in this project we are going to predict whether the individual will suffer from heart stroke or not by considering various factors like chest pain, resting blood pressure, cholesterol, fasting blood sugar, ECG results, max heart rate achieved, exercise induced angina, ST depression, ST slope, no.of major blood vessels blocked and types of thalassemia including age and gender as attributes. We found a dataset in Kaggle with all the above mentioned attributes and for that dataset we visualized those attributes and came to an interpretation on their distributions, dependencies etc. Also we trained 6 different models like Logistic Regression, Naive Bayes, K-NN, SVM, Decision Tree and Random Forest. Among all those SVM performed with highest accuracy and also Random Forest and Decision Tree were at high accuracy. The further details of them will be discussed detailed in later sections.

Motivation:

Coronary heart is the primary organ in our body. We as a whole recognize that it siphons the blood at some point of our body to each one of the organs. Coronary illness is one of the major sicknesses which could set off diminish the existence expectancy of people nowadays. each year 17. Five million people are kicking the bucket due to coronary infection. lifestyles are problem to phase running of heart, on account that heart is crucial piece of our frame. Coronary contamination is an infection that have an effect on the potential of heart. A gauge of an individual's threat for coronary contamination is substantial for a few parts of wellness advancement and clinical medication. This persuaded us to participate inside the development of this forecast version which helps us in waiting for a character's hazard to those coronary sicknesses. those expectations help us in maintaining up proper precautionary measures to live away from the later odds of getting prompted by using these illnesses.

Introduction:

We have tried use some sklearn features while exploring and visualizing the heart disease data we have given. Basically, we tried to show distribution of data, relations between variables and target as well as correlations between each other then we have done a basic model building.

The data includes 303 patient level features including if they have heart disease at the end or not. Features are like:

1. Age is the age of candidate.
2. Sex has numeric values. 1 denotes male and 0 denotes female.
3. Chest Pain has values between 0-3. The types of angina that are described in the research paper. The higher the number, the lesser are the odds of heart attack.
4. Resting blood pressure is normal pressure with no exercise.
5. Cholesterol means the blockage for blood supply in the blood vessels.
6. Fasting Blood Pressure is blood sugar taken after a long gap between a meal and the test. Typically, it's taken before any meal in the morning.
7. Rest ECG results means ECG values taken while person is on rest which means no exercise and normal functioning of heart is happening.
8. The Maximum Heart Rate achieved.
9. Exercise induced angina is chest pain while exercising or doing any physical activity.
10. ST Depression is the difference between value of ECG at rest and after exercise.
11. ST Slope is the tangent to the depression value.
12. The number of major blood vessels supplying blood to heart blocked.
13. The Types of thalassemia.
14. Heart attack prediction where 1 denotes Heart attack occurred and 0 where it didn't occur.

Problem Statement:

There are three major problems we are solving in our project:

1. For a given persons data, the data consists of above mentioned 13 attributes for that data we need to predict whether he/she encounters with a heart stroke or not?
2. To perform various analysis on those attributes and analyse their distributions and relations on each other.
3. To build model for different classification algorithms and select the best classifying algorithm for our dataset.

These are the problems we are solving in our project.

Literature Review:

S.No	Research Paper	Author	Existing State of Art
1.	Prediction of Heart Disease using Multiple Regression Model.	K. Polaraju	Exceptional Linear Regression is proper for foreseeing coronary contamination possibility. The paintings are done utilizing preparing informational index accommodates of 3000 activities with thirteen distinct characteristics which has referenced earlier than. The informational

			index is separated into two sections this is 70% of the records are utilized for preparing and 30% utilized for testing. In mild of the outcomes, glaringly the grouping precision of Regression calculation is higher contrasted with exclusive calculation of algorithms.
are	Heart Disease prediction using KStar, j48, SMO, and Bayes Net and Multilayer perception using WEKA software.	Marjia	In view of execution from diverse issue SMO and Bayes net accomplish ideal execution than KStar, Multilayer discernment and J48 tactics using kfold cross approval. The exactness exhibitions carried out through one of the calculations are as yet not true. In the end, the precision's presentation is improved more to provide higher preference to end contamination.
3.	Heart Disease Prediction System using Data Mining Techniques.	Megha Shahi	WEKA programming applied for programmed conclusion of infection and to give traits of administrations in hospital treatment communities. The paper applied unique calculations like SVM, Naïve Bayes, affiliation rule, KNN, ANN, and selection Tree. The paper suggested SVM is successful and furnishes extra precision as contrasted and other records mining calculations.
4.	Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques.	Chala Beyene	The fundamental intention is to assume the event of coronary illness for early programmed locating of the sickness internal bring about brief time frame. The proposed manner is likewise basic in clinical services affiliation with professionals that don't have any more information and information. It makes use of one of a kind medical trends, as an instance, glucose and pulse, age and intercourse are a part of the traits are incorporated to understand if the character has coronary contamination or now not. Examinations of dataset are processed utilizing WEKA programming.
5.	Non-linear classification algorithm for heart disease prediction.	R. Sharmila	It is proposed to make use of bigdata devices, for example, Hadoop distributed report machine (HDFS), Map lessen alongside SVM for forecast of coronary infection with streamlined feature set. These paintings made an examination at the utilization of various statistics digging strategies for foreseeing heart illnesses. It proposes to utilize HDFS for placing away big records in diverse hubs and executing

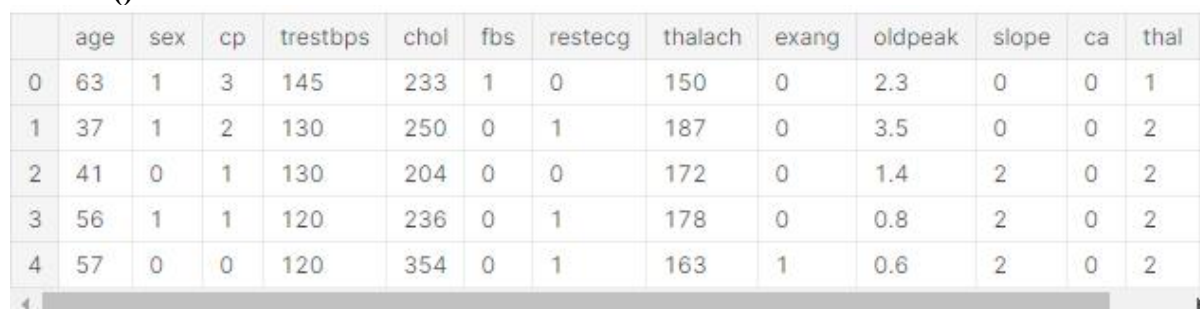
			the forecast calculation utilizing SVM in multiple hub on the same time making use of SVM. SVM is utilized in identical style which yielded desired calculation time over successive SVM.
6.	Heart disease prediction using data mining and machine learning algorithm.	Jayami Patel	The objective of this examination is to extricate protected up designs by applying statistics mining strategies. The first-rate calculation J48 depending on UCI information has the most noteworthy exactness charge contrasted with LMT.
7.	Heart Disease prediction system using data mining.	Purushottam	This framework assists medical specialist with selecting a hit dynamic dependent on the unique boundary. through trying out and preparing stage a specific boundary, it offers 86.3% precision in checking out stage and 87.3% in making ready level.

Methodology:

In this section firstly we are going to do the visualization on our data set and after that we perform data analysis by using groupby and cmap, univariate data analysis, bivariate data analysis, relative plots, box plot and correlations. After that we will discuss about those 6 Machine Learning models in detail, how we built them and how we evaluated those models.

Data Visualization:

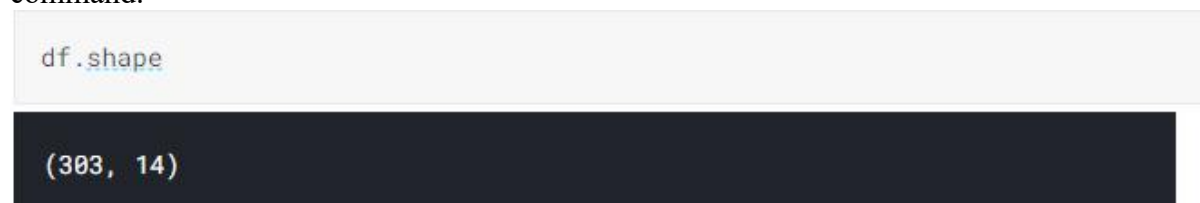
The first step in Data Visualization is to import data from the data set. In this case, we will be considering the dataset in a .csv file format. In order to view the contents in the dataset we will be using the command: **df.head()** as the data in the dataset is moved to the variable named **df**.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2

Fig.1

Here the next step is to find the shape of the data. To complete this step we will be using **df.shape** command.



df.shape
(303, 14)

Fig.2

As the factors or parameters in the dataset are not understandable we will be replacing the factors or parameters to general terms. This generalization is done as shown in **Fig.3**.


```
df.rename(columns ={'age':'Age','sex':'Sex','cp':'Chest_pain','trestbps':'R
esting_blood_pressure','chol':'Cholesterol','fbs':'Fasting_blood_sugar',
                'restecg':'ECG_results','thalach':'Maximum_heart_rat
e','exang':'Exercise_induced_angina','oldpeak':'ST_depression','ca':'Major_
vessels',
                'thal':'Thalassemia_types','target':'Heart_attack','slop
e':'ST_slope'}, inplace = True)
```

Fig.3

Now, we will be finding the null values present in the dataset which are represented beside each parameter in the dataset as shown in **Fig.4**.

```
df.isnull().sum()
```

```
Age                0
Sex                0
Chest_pain         0
Resting_blood_pressure  0
Cholesterol         0
Fasting_blood_sugar  0
ECG_results         0
Maximum_heart_rate  0
Exercise_induced_angina  0
ST_depression       0
ST_slope            0
Major_vessels       0
Thalassemia_types   0
Heart_attack        0
dtype: int64
```

Fig.4

Data Analysis:

Analysis of data by GroupBy and CMap:

Pandas GroupBy is a powerful and versatile function in Python. It allows the data into separate groups to perform computations for better analysis.

This can be achieved by using the **groupby** command in pandas. Also, **cmap** is used to designate a colour to the data which is represented by the level of darkness in the colour.

This helps us to estimate the maximum and minimum values in the analysed data values.

In the figures below, we will be analysing the data between different parameters along with the number of heart-attacks. Also, the data values are grouped in descending order of heart-attacks using GroupBy in pandas.

	Sex	Age	Heart_attack
59	Male	58	13
58	Male	57	13
60	Male	59	13
53	Male	52	12
55	Male	54	11
45	Male	44	9
52	Male	51	8
57	Male	56	8
61	Male	60	7
65	Male	64	7
24	Female	62	7
43	Male	42	6
49	Male	48	6
62	Male	61	6
44	Male	43	6
68	Male	67	6
20	Female	58	6
42	Male	41	6
25	Female	63	5
54	Male	53	5

	Sex	Chest_pain	Heart_attack
4	Male	0	104
6	Male	2	52
0	Female	0	39
2	Female	2	35
5	Male	1	32
7	Male	3	19
1	Female	1	18
3	Female	3	4

Fig.4

Fig.5

This suggests the entire range of coronary disasters that have befall for a selected age if there should rise up a prevalence of male and female. Inside the quality 20 tally of coronary failure, guys have seen extra coronary episodes for their ages.

The men having chest torment kind 0 persevered the most coronary episodes and sort 2 the second one maximum noteworthy. This sample is equal for women.

Sex	Resting_blood_pressure	Heart_attack	Sex	Cholesterol	Heart_attack
Male	120	27	Male	212	5
Male	130	24	Male	204	4
Male	140	22	Male	234	4
Male	110	16	Male	233	4
Female	130	12	Male	282	4
Male	125	11	Male	254	4
Female	140	10	Female	269	4
Male	150	10	Male	246	3
Female	120	10	Male	245	3
Male	128	9	Male	243	3
Female	150	7	Male	175	3
Male	138	7	Male	240	3
Male	160	7	Male	239	3
Male	112	6	Male	274	3
Female	138	6	Male	231	3
Male	118	5	Male	229	3
Male	132	5	Male	226	3
Male	145	4	Male	230	3
			Male	197	3
			Male	309	3

Fig.6

Fig.7

The resting pulse and the instances of coronary episode for such pressing factor if there should be an occurrence of males and females.

Large men having elevated cholesterol had higher instances of cardiovascular failures than females at a similar Cholesterol.

Heat Map:

A HeatMap is an information notion manner that shows size of the shading in measurements. The range in shading is probably through tint or energy, giving clean viewable alerts to the in keeping with consumer approximately how the surprise is bunched or fluctuates over area.

We eliminate the discrete variables values after copying the same dataset into a temporary variable. We apply HeatMap to the remaining variables containing continuous data. This lays an estimation of the relation between the continuous variables.

This relation is visualized in the Fig.8 mentioned below:

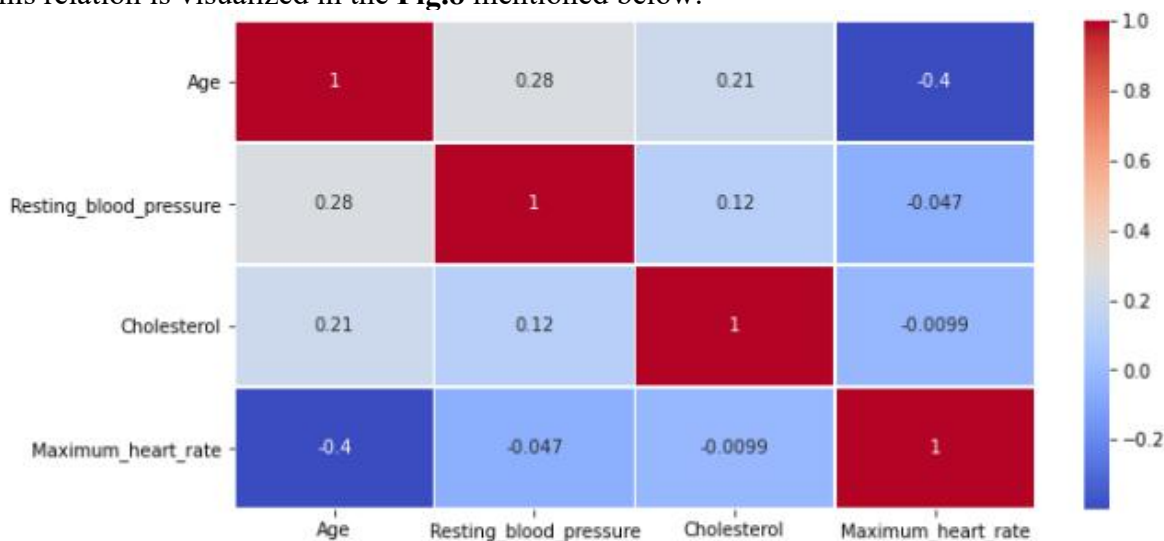


Fig.8

Univariate Data Analysis:

Univariate research investigates each component in an informational index, independently. It takes a gander at the scope of traits, just because the focal propensity of the features. It depicts the example of response to the variable. It portrays everything all on my own. Illustrative measurements depict and sum up statistics. The following figures display various values in the parameters based on their frequency.

```
sns.set(rc={'figure.figsize':(20,5)})
df['Age'].plot.hist(bins = 15, color = 'skyblue')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f987d71b410>
```

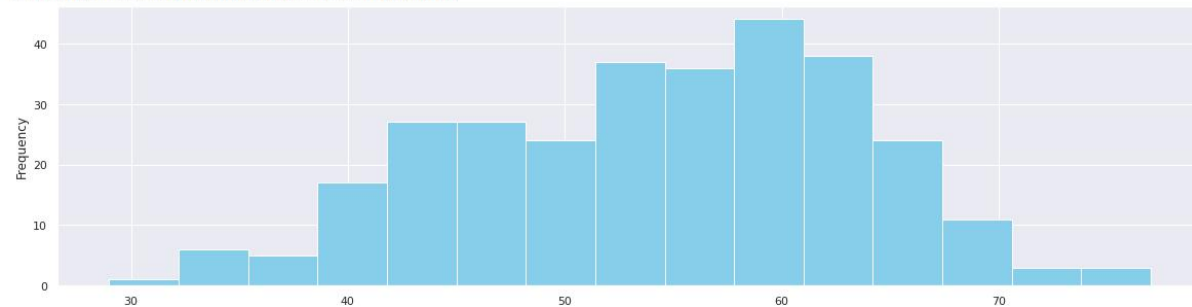


Fig.9

```
df['Resting_blood_pressure'].plot.hist(bins = 15, color = 'green')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f987360ce10>
```

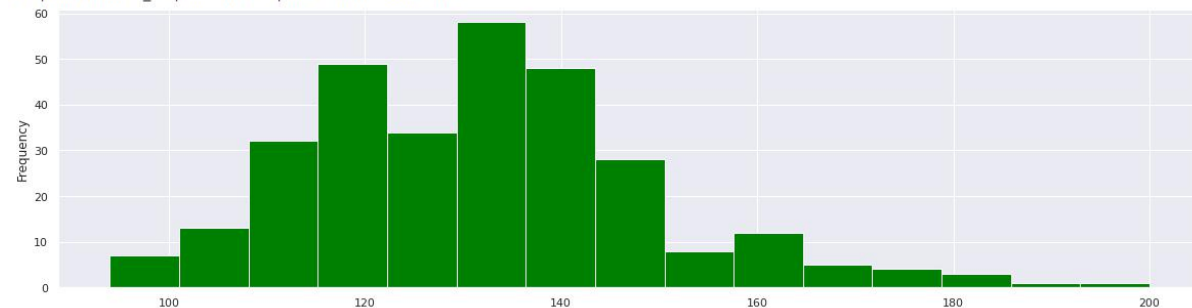


Fig.10

```
df['Cholesterol'].plot.hist(bins = 10, color = 'lightgrey')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9873546310>
```

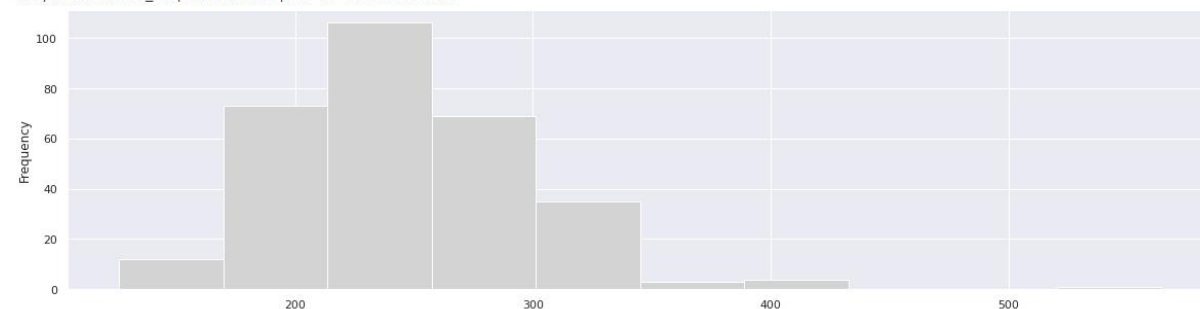


Fig.11

```
df['Maximum_heart_rate'].plot.hist(bins = 20, color = 'lightcoral')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f98734cde90>
```

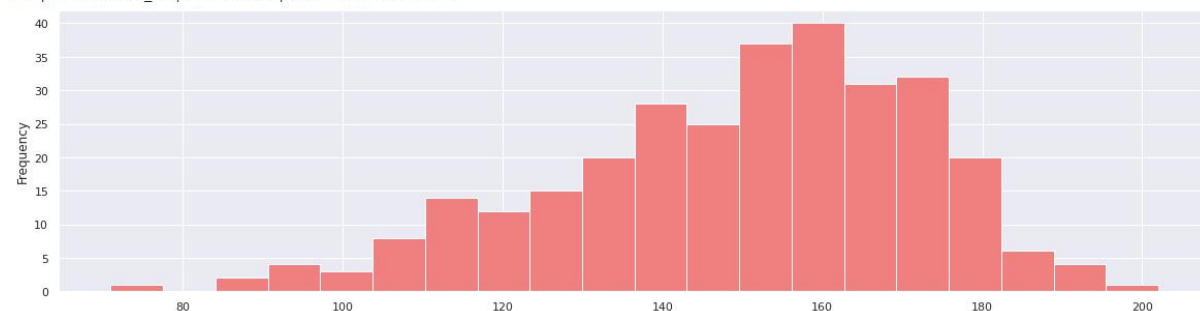


Fig.12

1. This shows age has some reliance on Resting Blood Pressure and Cholesterol.

2. There is basically no reliance of Maximum pulse on Age.

Resting pulse and Cholesterol likewise has a low reliance. Most extreme pulse and Resting circulatory strain practically no reliance.

Bivariate Data Analysis:

Bivariate research might be the easiest form of quantitative (actual) exam. It consists of the investigation of factors (often indicated as X, Y), to determine the experimental connection among them. Like univariate examination, bivariate research may be enlightening or inferential.

The following figures are obtained by establishing a relation between two different variables.

```
sns.countplot(x = 'Age', hue = 'Chest_pain', data = df, color = 'green')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f98733f9610>
```

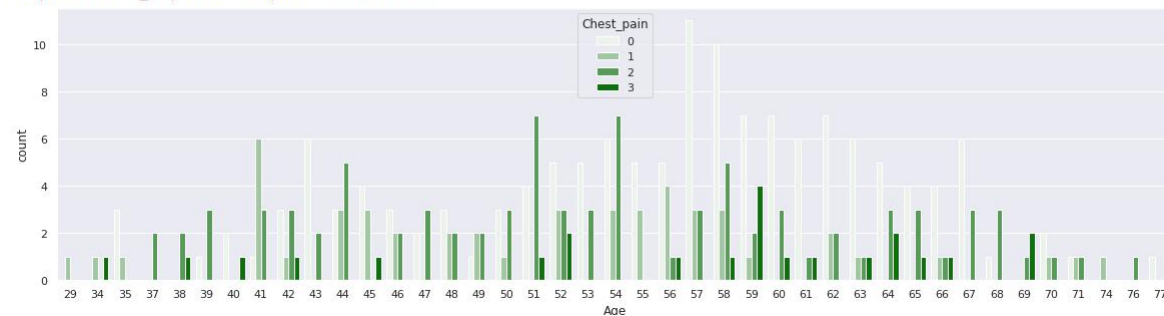


Fig.13

```
sns.countplot(x = 'Age', hue = 'Fasting_blood_sugar', data = df, palette="Set3")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9871e13cd0>
```

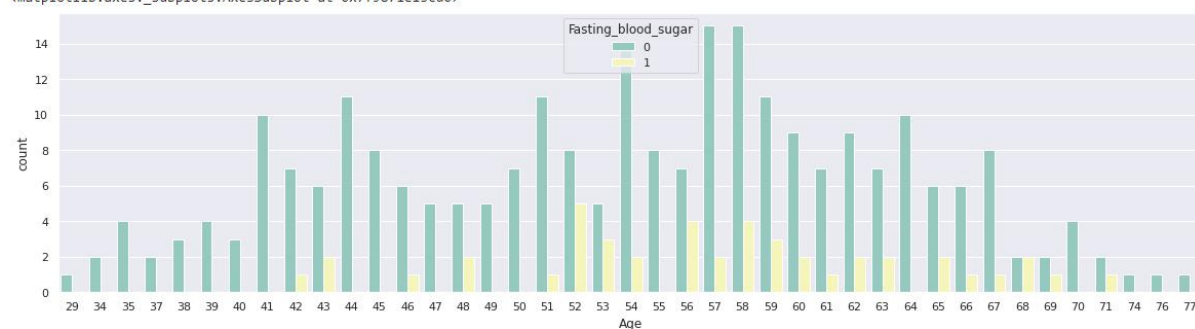


Fig.14

```
sns.countplot(x = 'Age', hue = 'ECG_results', data = df, palette="Set2")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9871c56c10>
```

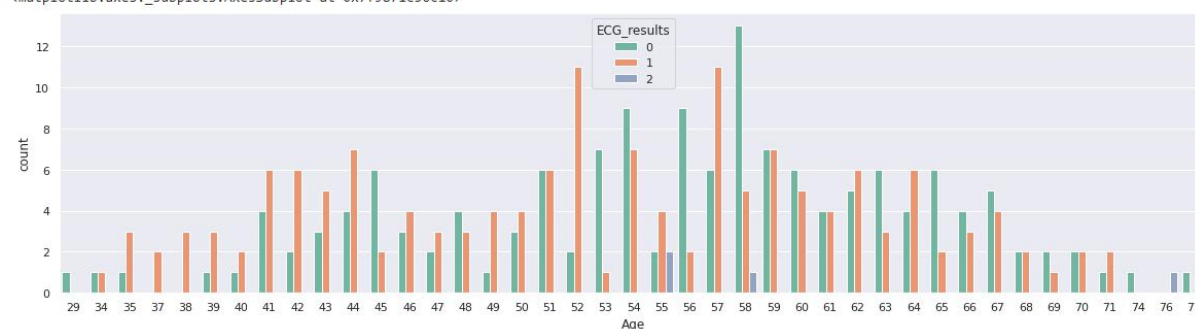


Fig.15


```
sns.countplot(x = 'Age',hue = 'Exercise_induced_angina', data = df,palette="Set1")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f98734ca190>
```

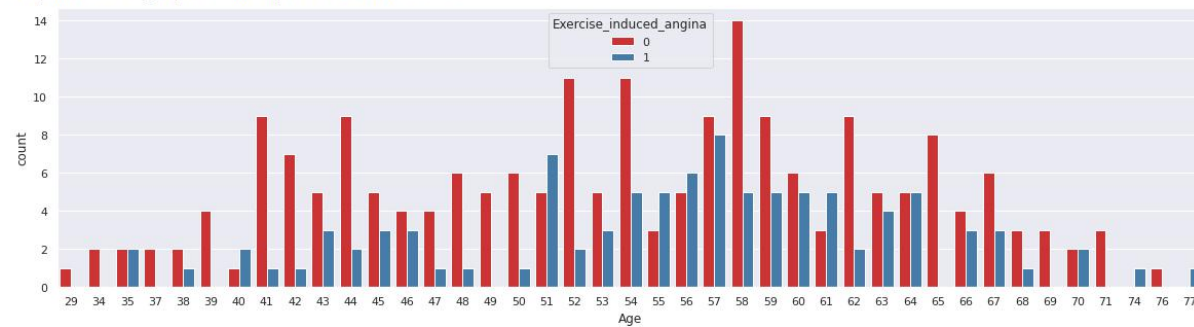


Fig.16

Relative Plot:

For more classification we plot two different graphs based on the gender. We will be plotting a relative plot in the following figures:

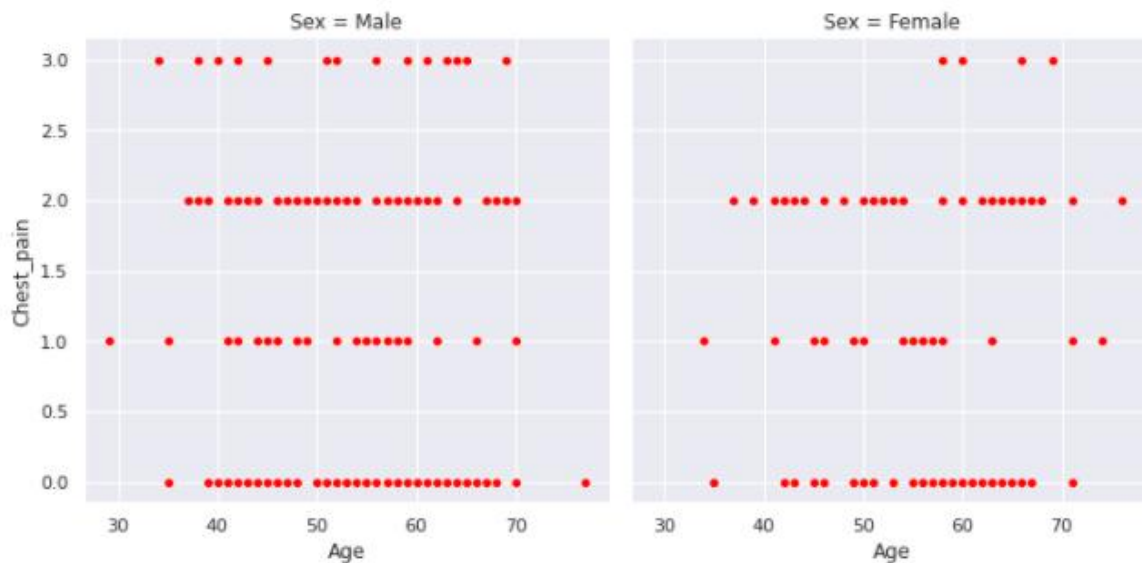
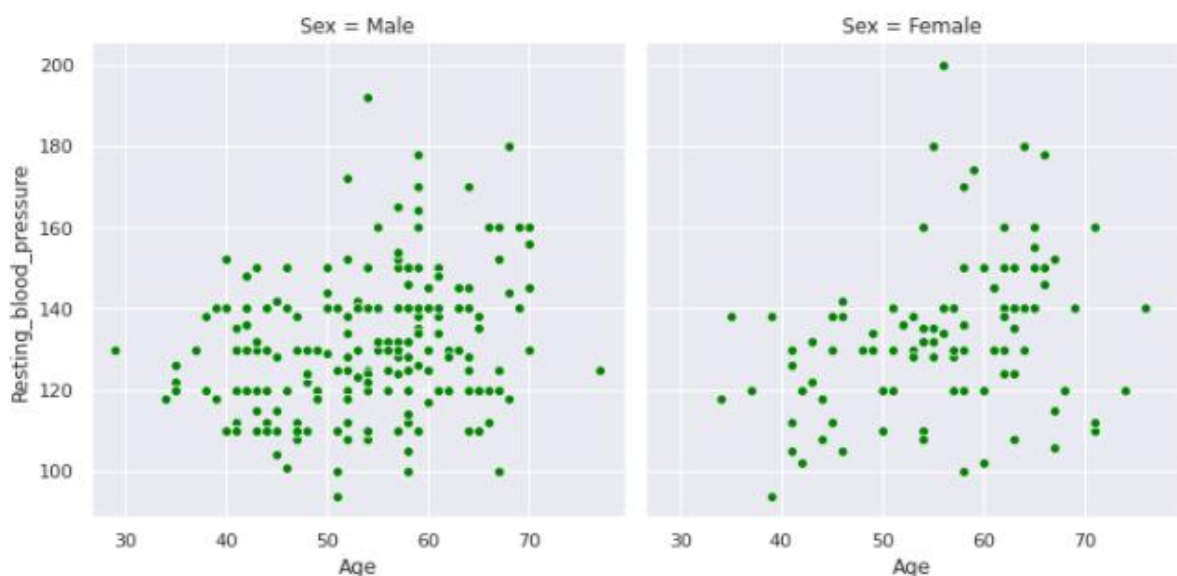


Fig.17

In Fig.17, we notice that the plot has fixed line of points marked on it as the levels of chest pain in the data set are discrete integers which results us in a fixed line of points at every integer at the chest pain parameter.



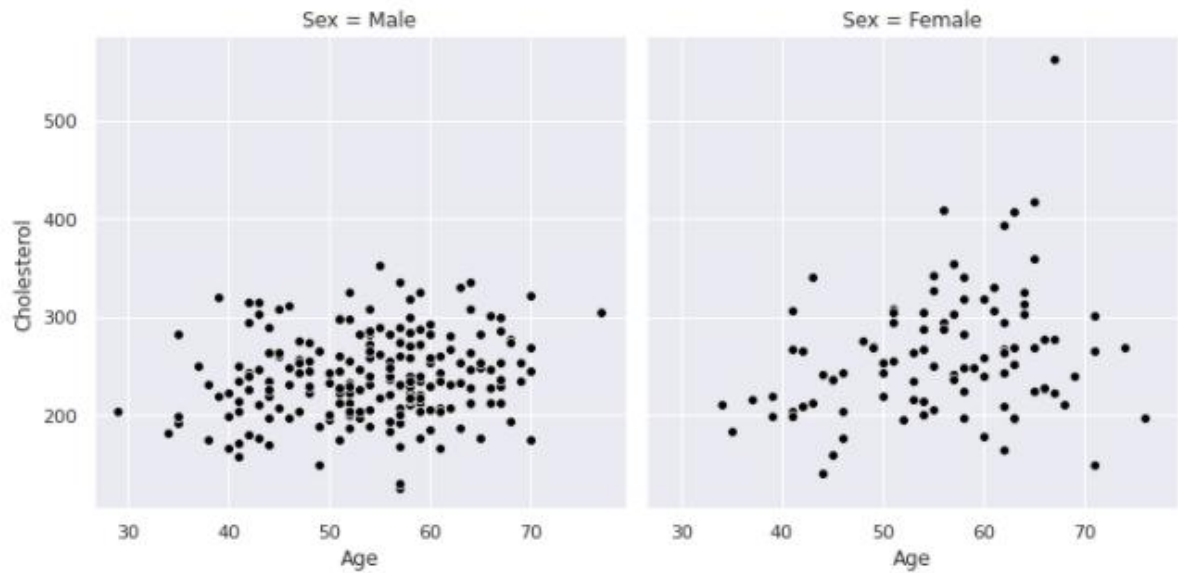


Fig.18

Fig.19

Box-Plot:

A boxplot is a normalized approach of showing the appropriation of information depending on a 5 numbers rundown ("least", first quartile (Q1), middle, 1/3 quartile (Q3), and "maximum extreme"). it could train you concerning your anomalies and what their characteristics are.

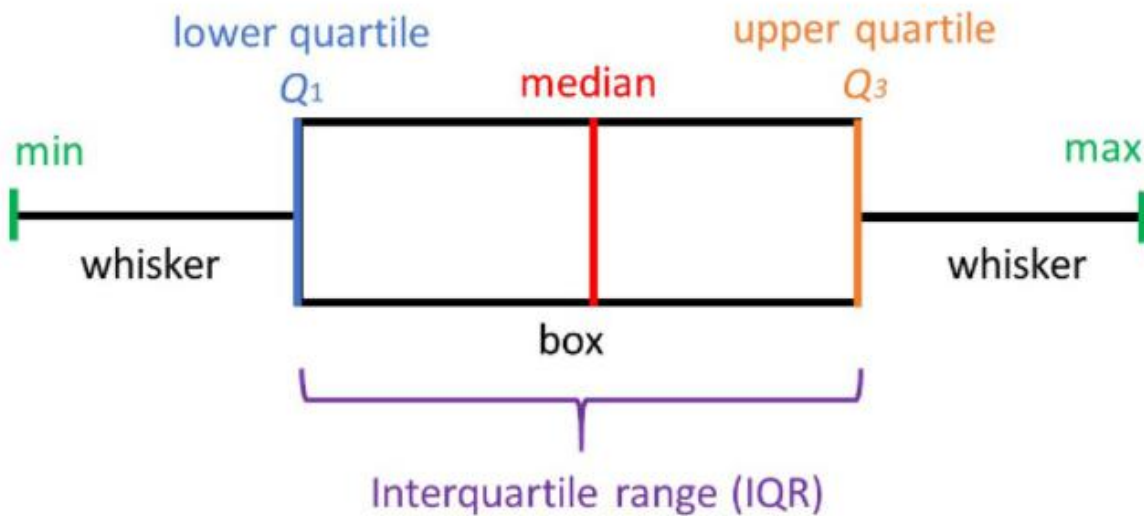


Fig.20

Based on Fig.20 box plot reference, we will be plotting a four relational box-plots between the Heart-attacks and age for both the genders (Male and Female) which is represented in **Fig.21**.

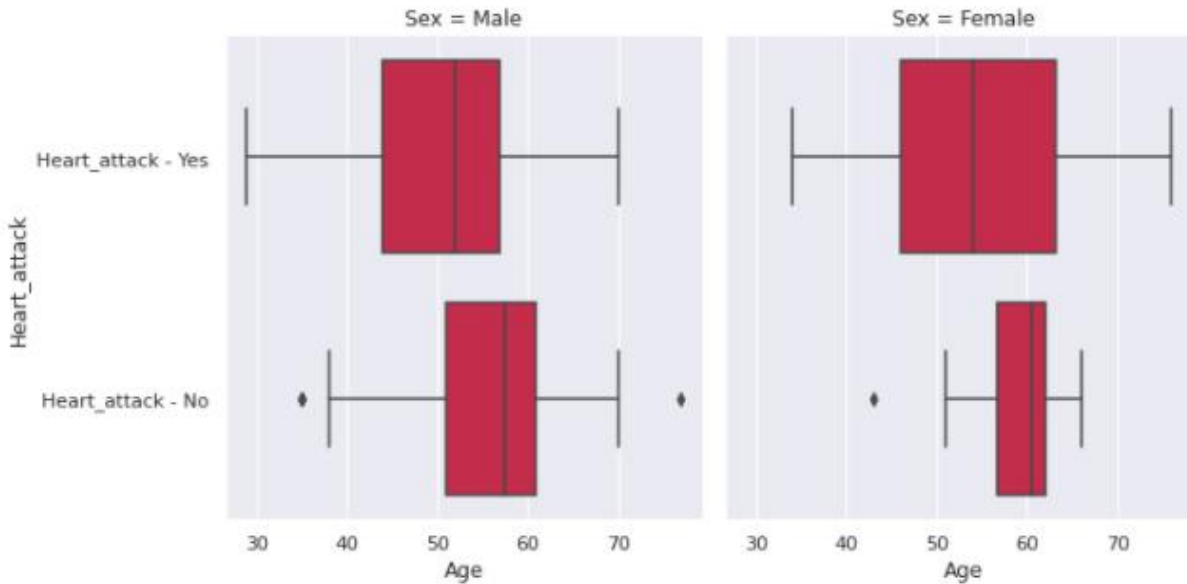


Fig.21

Model Building:

Now going to building and training of different classification models like Logistic Regression, Naive Bayes, K-Nearest Neighbours, Support Vector Machine, Decision Tree and Random Forest classifier for our dataset. Let's talk about the preprocessing that we have done to build any type of model. Firstly, we separated the target variable from the rest of other variables, after that we split the data into x train, x test, y train, y test. Here the test size is 0.3 (that means 70% data used for training purpose and 30% data used for testing purpose). x train, x test stores all those 13 variables/attributes data with 70% and 30% respectively, and y train and y test stores target variable data with 70% and 30% respectively. Finally, we applied Standard Scaler to our x train and x test data, where it removes mean and scales every variable to unit variance. Now we are ready to build our models.

Logistic Regression:

It is a model that predicts output for binary classes and here the target variable would be categorical in nature. Also, using MLE (Maximum Likelihood Estimate) the logistic regression will be estimated; it is a likelihood maximization method. Logistic function, which is also called as sigmoid function, gives an S-shaped curve where it maps any real-valued number into a number between 0 and 1. If the output of the logistic function is >0.5 , then it is classified as 1; otherwise, it is classified as 0. In implementation, we declared a logistic regression object with the help of Sklearn, trained with x train and y train data, and predicted output for x test data. Also, we built a confusion matrix to find TP, TN, FP, FN, and finally, we calculated the accuracy of the logistic regression model, which is at 0.8791.

The below shown image is the function of logistic regression

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

Fig.22

Naive Bayes:

It is a classification technique based on Bayes' theorem; it is accurate, fast, reliable, and simple supervised learning algorithm. Firstly, for given class labels, it calculates prior probability, then it finds likelihood probability for every variable with each class, by these values, it calculates posterior probability and finally measures which class has the higher probability for input data. In implementation, the steps followed are similar to that we have done in logistic regression, and we got a model with accuracy at 0.8461.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Fig.23

K-Nearest Neighbours:

It is a lazy learning algorithm where it doesn't need training data for model generation all the training process will be done at testing phase. In lazy approach the training would be faster but testing would be slower and costlier. Here before building the model we first we have to decide the value of K, in our case we assigned it as 3. For test data we will find three nearest neighbours to that test data based on the voting of classes, the highest voted class will be considered and that test data will be assigned to that class. Coming to implementation the steps followed would be similar to above both models but we use KNeighboursClassifier and parameter is 3(no.of neighbours). The accuracy of K-NN model is at 0.8571.

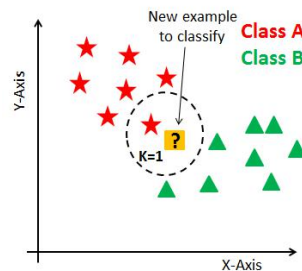


Fig.24

Support Vector Machine:

It builds hyperplane to separate classes in a multidimensional space. For a dataset SVM will find the maximum margin hyperplane which best divides the classes, There are Support vectors(data points) which are closest to the hyperplane and margin is the distance between two lines on closest class points. In case of implementation for SVM we considered parameters like kernel = 'linear', C = 30, gamma = 0.03, Kernel means it takes input lower dimensional space and transforms it into higher dimensional space so by adding dimensions the non linear separable problem into linear separable. Linear kernel will find the normal dot product for two observations. In case of C it is a regularization parameter small value of C represents small margin hyperplane and vice versa. In case of gamma the higher value of it will tightly fit the training dataset. The accuracy of SVM model is 0.8901

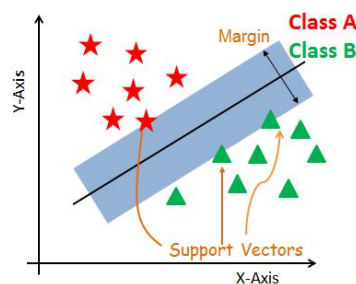


Fig.25

Decision Tree:

It will be in tree structure the topmost node is called as root node and here the internal nodes specifies attributes, leaf specifies outcome and branch specifies decision rules. Decision tree will partition data based on attributes value. Decision tree will first select the best attribute by using methods like Information Gain or Gain ratio or Gini Index, after that attribute behaves like a decision node and divides the data in subsets and the same process will be repeated until there is no more leftover attributes. In implementation for DecisionTreeClassifier the parameters taken were criterion = 'entropy', max_depth = 4, here entropy represents randomness in data and high Information Gain means low entropy. Information Gain can be calculated as difference between before split entropy and after split average entropy for the given attribute

values. Max_depth specifies the length or height of the decision tree and in our case it is 4. The accuracy of the decision tree model is 0.8852.

Formula for Information Gain given below

$$\text{Info}_A(D) = \sum_{j=1}^V \frac{|D_j|}{|D|} \times \text{Info}(D_j)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

Fig.27

Random Forest:

Random Forest is the collection of decision trees, each and every tree will be trained independently by selecting best attribute to be next by using indicators like IG, gini index, gain ratio etc. Also for each tree the given random samples are independent. Here each tree vote and most voted class will be considered and the test data will be assigned to that class. In implementation for RandomForestClassifier the parameters were n_estimators = 26, criterion='entropy', max_depth = 4 here last two parameters were same as decision tree parameters that we have discussed above and in case of n_estimators it represents that how many no.of decision trees we should consider to get more accurate output, in our case it was 26 and we test for nearly 100 values and out of that 26 was optimal. Therefore we considered 26 as n_estimators. The accuracy of random forest is at 0.8852.

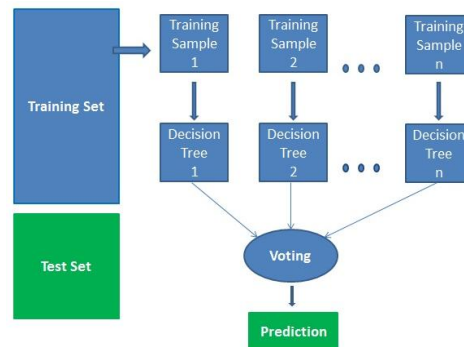


Fig.21

We implemented the entire project on Google Colaboratory which is called as Google Colab. This Colab files will be helpful in prototyping ML models on powerful hardware devices like TPU's and GPU's. We implemented the entire code in Python language and used the libraries like Pandas, matplotlib, seaborn, scikitplot, sklearn, StandardScalar and classification_report.

pandas: Used to create dataframe, do manipulations and changes accordingly to the dataframe.

matplotlib: Used to plot the graph for finding the optimal K value in doing K-NN.

seaborn: Used in data analysis part and this library plays a key role in doing data analysis for our project.

scikitplot: Used for building the confusion matrix which is used for evaluation purpose.

sklearn: It is a ML library which has various clustering, classification and regression algorithms in it and in our case we are using classification algorithms only. We import the above built 6 models for this library.

StandardScalar: For the dataset(where every variable is numeric) it removes the mean and scales every variable/attribute to unit variance.

classification_report: Used to get the accuracy, precision and recall scores for every training model we built so that we can analyse the model performance.

Coming to the further implementation yes, there are many ways to implement the project like in case of visualization and analysis part people can use various analysis tools and techniques in different ways and they can be lead into different or new interpretation that we haven't done in this project. And in case of model building we have used only 6 classifiers but in Machine Learning there are some more classifiers that we haven't used in our project like Artificial Neural Networks(ANN), Perceptron, Stochastic Gradient Descent etc. Also In case of Decision Tree I used Information Gain as parameter but we can use other indicators like gini index and gain ratio as parameter like this we can do many variations for the models that

I have built. Even we can use K-fold cross validation in splitting the training data before feeding it to the model. Hence from this discussion we came to know that we can implement this project in various ways but ultimately the target for this dataset is to predict whether a person gets heart stroke or not.

Results and Discussion:

Model Name	Accuracy
Logistic Regression	0.8791
Naïve Bayes	0.8461
K Nearest Neighbours (KNN)	0.8571
Support Vector Machine (SVM)	0.8901
Decision Tree	0.8852
Random Forest	0.8852

Here, we can observe that SVM has the maximum accuracy. Hence, we prefer to establish SVM as our prediction model. We can also observe that the accuracies doesn't vary much in their predictions due to which this method can be applicable in various other models mentioned above.

Conclusion:

In this project we performed data analysis and visualization on the Kaggle Heart Disease UCI dataset and came to many interpretations on the attributes of the dataset. Even we built 6 classification models like Logistic regression, Naive Bayes, SVM, K-NN, Decision Tree and Random Forest. Out of all those models SVM performed well so we considered SVM as our predicting model. Therefore with this classification model we achieved our target that is to predict whether a person suffers from heart attack or not.

References:

1. Salhi, Dhai Eddine, Abdelkamel Tari, and M-Tahar Kechadi. "Using Machine Learning for Heart Disease Prediction." Advances in Computing Systems and Applications: Proceedings of the 4th Conference on Computing Systems and Applications. Springer International Publishing, 2021.
2. Marimuthu, M., et al. "A review on heart disease prediction using machine learning and data analytics approach." International Journal of Computer Applications 181.18 (2018): 20-25.