RISK OF HEART DISEASE PREDICTION

Department of Statistics

University of Allahabad

PROJECT REPORT

Submitted as a Fulfillment of M.Sc. Final Year

Major Project 2022

Under Supervision

Dr. G. Madhusudan

Assistant Professor

Department of Statistics

University of Allahabad

Submitted By

Riya Vishwakarma

M.Sc. Final Year (2022)

Enrollment No.- U2075016

University of Allahabad

CERTIFICATE

This is to certify that the secondary data given in this report has been collected, tabulated, analysed and presented by Riya Vishwakarma, Student of M.Sc Statistics Session 2021-22 .

The Titled Project is:

“Risk of Heart Diseases Prediction “

Prof. Shekhar Srivastava
(HoD,Statistics & Dean of Science)
Department of Statistics
University of Allahabad

Dr. G. Madhusudan
(Assistant Professor)
Department of Statistics
University of Allahabad

Submitted By:

Riya Vishwakarma
M.Sc Statistics (2021-22)
Enrolment No – U2075016
Roll No – 22421906
University of Allahabad

ACKNOWLEDGEMENT

This project is the resultant of a successful effort guidance and it would have been impossible to complete this without the efforts of following people to whom I would like to extend my heartily gratitude to my Supervisor - Dr. G. Madhusudan- for his valuable guidance, suggestions, most importantly motivation and inspiration.

I am also thankful to our faculty members especially Dr. Pramendra Singh Pundir, Dr. Abhay Pratap Pandey, Dr. G Madhusudan, Dr. Priyanka Singh, previous teaching staff of Department for mentorship – Mr. Pulkit Srivastava, Mr. Prashant Kumar Sonker, & Prof Ranjita Pandey – Department of Statistics, University of Delhi. And my seniors, batchmates and juniors.

Thank You all for support and kind help.

- Riya Vishwakarma

ABSTRACT

Heart disease often called Cardiovascular disease mainly caused due multiple factors such as blood pressure, cholesterol level, more or less due to life styles are now days are changing day by day as world is becoming so fast and furious. Hence this project is an effort to develop and study the relationship of heart diseases and identification of factors affecting.

Analysis is Done by using Logistic regression and Exploratory Data Analysis (EDA) – for data visualization.

Comparative Analysis of accuracy of Logistic Regression Vs Random Forest Classifier.

Contents

OBJECTIVE

Of this project is to use this data set to:

1. To Classify Risk of Heart Disease according to gender and age.
2. Exploratory Data Analysis (EDA)
3. Compare Logistic Regression with random forest classifier.

Chapter – 1

INRODUCTION

There is a famous Quote “A Healthy Body Contains a Healthy Mind “but now days this Quote is little bit altered as per situation “A Healthy Body Not Only contains Healthy Mind but also a Healthy Hearts Too “Heart being an important part of our body and if our motive is to keep ourselves healthy we must take care of our heart. So it is necessary to identify the cause. We know that prevention is always better than cure.

1.1 Heart Disease

One of the most common complications of heart disease, heart failure occurs when our heart can’t pump enough blood to meet our body’s needs, Heart failure can result from many forms of heart disease including, heart defects, cardiovascular disease, valvular heart disease, heart infections or cardiomyopathy.

Doctor will perform a physical exam and ask about your personal and family medical history. The tests you'll need to diagnose your heart disease depend on what condition your doctor thinks you might have. Besides blood tests and a chest X-ray, tests to diagnose heart disease can include:

1.2 Diagnosis

- **Electrocardiogram (ECG or EKG).** An ECG is a quick and painless test that records the electrical signals in your heart. It can spot abnormal heart rhythms. You may have an ECG while you're at rest or while exercising (stress electrocardiogram).
- **Holter monitoring.** A Holter monitor is a portable ECG device you wear to continuously record your heart rhythm, usually for 24 to 72 hours. Holter monitoring is used to detect heart rhythm problems that aren't found during a regular ECG exam.
- **Echocardiogram.** This non-invasive exam uses sound waves to produce detailed images of your heart's structure. It shows how your heart beats and pumps blood.
- **Stress test.** This type of test involves raising your heart rate with exercise or medicine while performing heart tests and imaging to check how your heart responds.
- **Cardiac catheterization.** In this test, a short tube (sheath) is inserted into a vein or artery in your leg (groin) or arm. A hollow, flexible and longer tube (guide catheter) is then inserted into the sheath. Using X-ray images on a monitor as a guide, your doctor carefully threads the catheter through the artery until it reaches your heart.

  During cardiac catheterization, the pressures in your heart chambers can be measured, and dye can be injected. The dye can be seen on an X-ray, which

helps your doctor see the blood flow through your heart, blood vessels and valves to check for problems.

- **Cardiac computerized tomography (CT) scan.** In a cardiac CT scan, you lie on a table inside a doughnut-shaped machine. An X-ray tube inside the machine rotates around your body and collects images of your heart and chest.
- **Cardiac magnetic resonance imaging (MRI).** A cardiac MRI uses a magnetic field and computergenerated radio waves to create detailed images of your heart.

MATERIAL & METHODS

This data set is a secondary data imported from UCI machine learning . This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes.

- Cleveland: 303 observations

2.1 DATA PREVIEW AND DATA DESCRIPTION

- Age: Age of the patient
- Sex: Sex of the patient
- exang: exercise induced angina (1 = yes; 0 = no)
- ca: number of major vessels (0-3)
- cp: Chest Pain type chest pain type

Value 1: typical angina
Value 2: atypical angina
Value 3: non-anginal pain
Value 4: asymptomatic

- trtbps: resting blood pressure (in mm Hg)

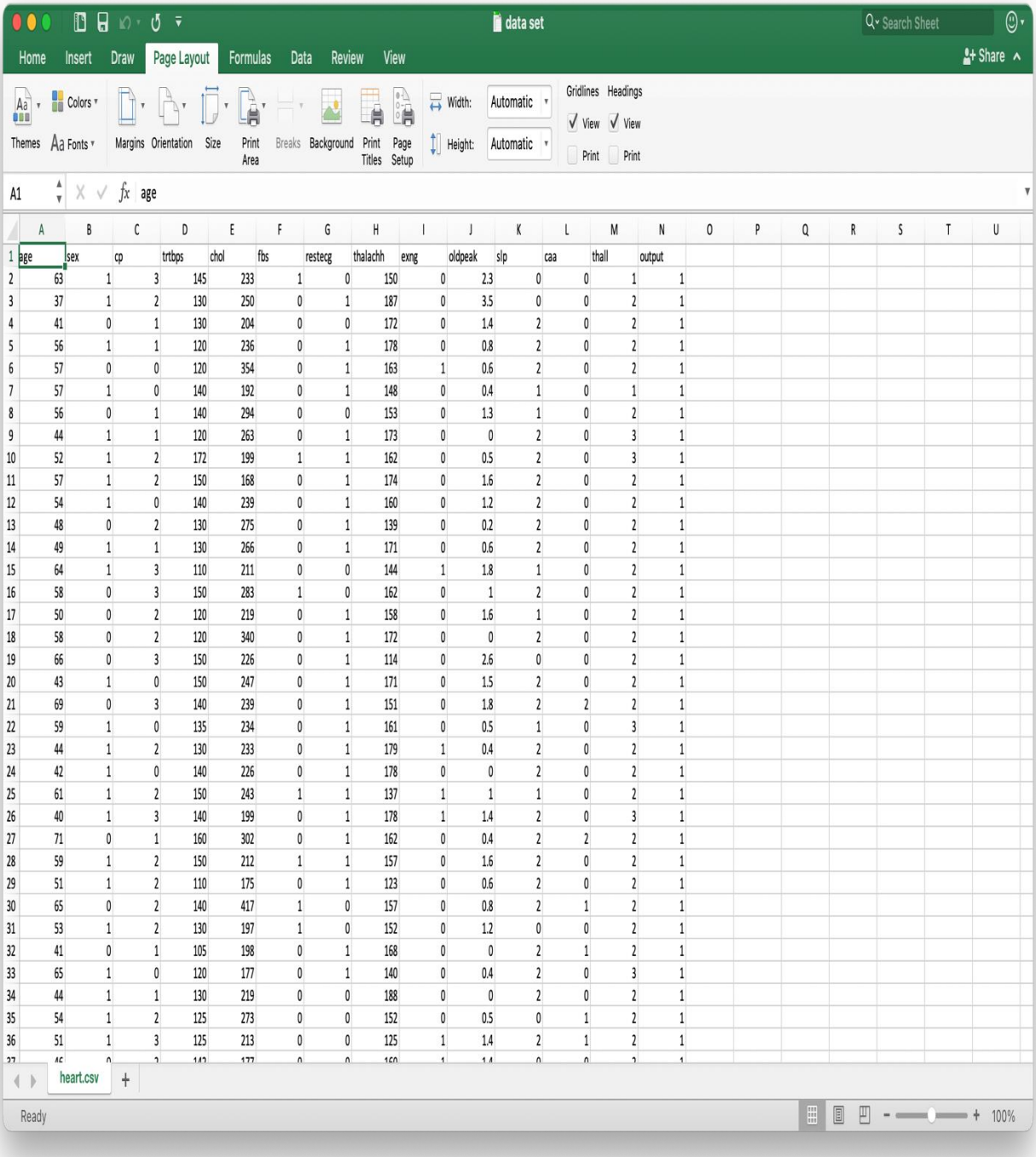
- chol: cholesterol in mg/dl fetched via BMI sensor • fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- rest_ecg: resting electrocardiographic results Value 0: normal

Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

- thalach: maximum heart rate achieved
- target: 0= less chance of heart attack 1= more chance of heart attack.

DATA PREVIEW



data set

Search Sheet

Home Insert Draw Page Layout Formulas Data Review View Share

Themes Colors Fonts Margins Orientation Size Print Area Breaks Background Print Titles Page Setup Width: Automatic Height: Automatic Gridlines View Print Headings View Print

A1 age

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output							
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1							
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1							
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1							
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1							
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1							
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1							
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1							
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1							
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1							
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1							
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1							
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1							
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1							
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1							
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1							
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1							
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1							
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1							
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1							
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1							
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1							
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1							
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1							
25	61	1	2	150	243	1	1	137	1	1	1	0	2	1							
26	40	1	3	140	199	0	1	178	1	1.4	2	0	3	1							
27	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1							
28	59	1	2	150	212	1	1	157	0	1.6	2	0	2	1							
29	51	1	2	110	175	0	1	123	0	0.6	2	0	2	1							
30	65	0	2	140	417	1	0	157	0	0.8	2	1	2	1							
31	53	1	2	130	197	1	0	152	0	1.2	0	0	2	1							
32	41	0	1	105	198	0	1	168	0	0	2	1	2	1							
33	65	1	0	120	177	0	1	140	0	0.4	2	0	3	1							
34	44	1	1	130	219	0	0	188	0	0	2	0	2	1							
35	54	1	2	125	273	0	0	152	0	0.5	0	1	2	1							
36	51	1	3	125	213	0	0	125	1	1.4	2	1	2	1							

heart.csv +

Ready 100%

2.2 METHODOLOGY

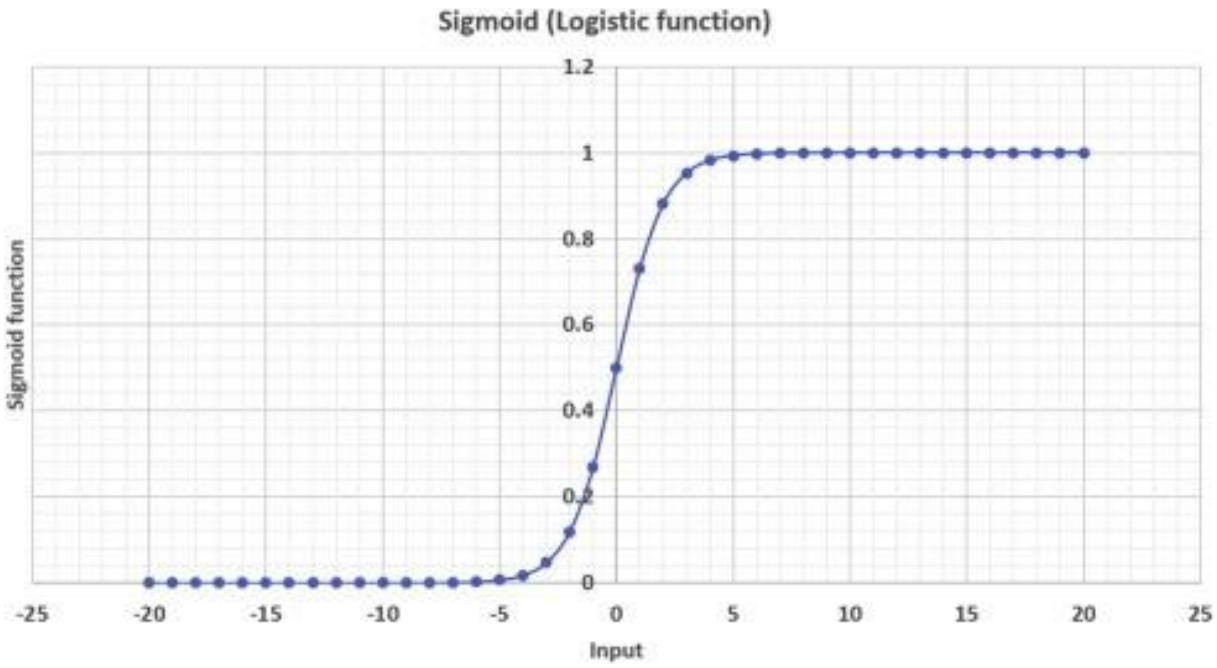
The above data is modelled as Logistic Regression as Risk of Heart disease is a categorical variable which is dependent and rest are independent variable.

2.2.1 LOGISTIC REGRESSION

Logistic regression is another powerful supervised ML algorithm used for binary classification problems (when target is categorical). The best way to think about logistic regression is that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function defined below to model a binary output variable (Tolles & Meurer, 2016).

The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio.

$$Logistic \quad Function = \frac{1}{1 + e^{-x}}$$



Logistic regression is another fundamental method initially formulated by David Cox in 1958[32] that builds a logistic model (also known as the logit model). Its most significant advantage is that it can be used both for classification and class probability estimation, because it is tied with logistic data distribution. It takes a linear combination of features and applies to them a nonlinear sigmoidal function. In the basic version of logistic regression, the output variable is binary, however, it can be extended into multiple classes (then it is called multinomial logistic regression). The binary logistic model classifies specimen into two classes, whereas the multinomial logistic model extends this to an arbitrary number of classes without ordering them.

The mathematics of logistic regression rely on the concept of the “odds” of the event, which is a probability of an event occurring divided by the probability of an

event not occurring. Just as in linear regression, logistic regression has weights associated with dimensions of input data. In contrary to linear regression, the relationship between the weights and the output of the model (the “odds”) is exponential, not linear.

Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today .

Random Forest Classifier

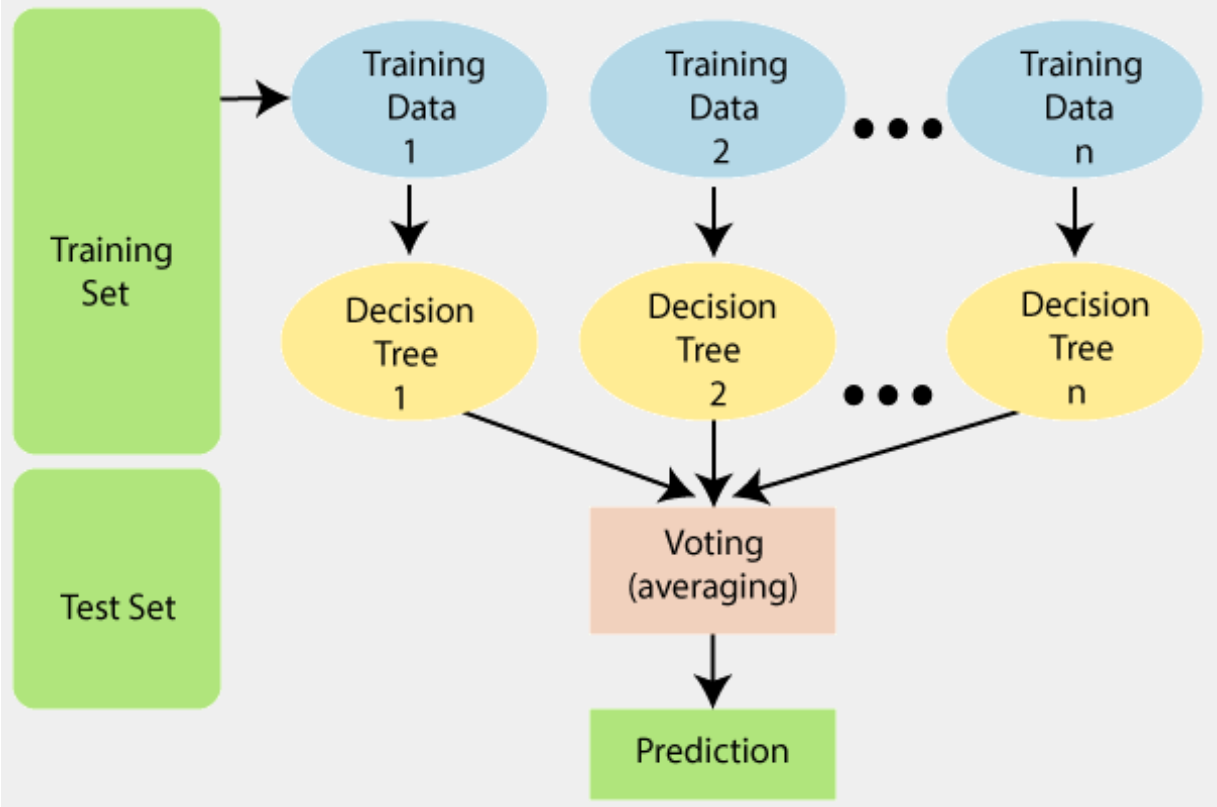
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the

concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:



Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Software Used

Python

Python is developed under an OSI-approved open source license, making it freely usable and distributable, even for commercial use. Python's license is administered by the Python Development Foundation.

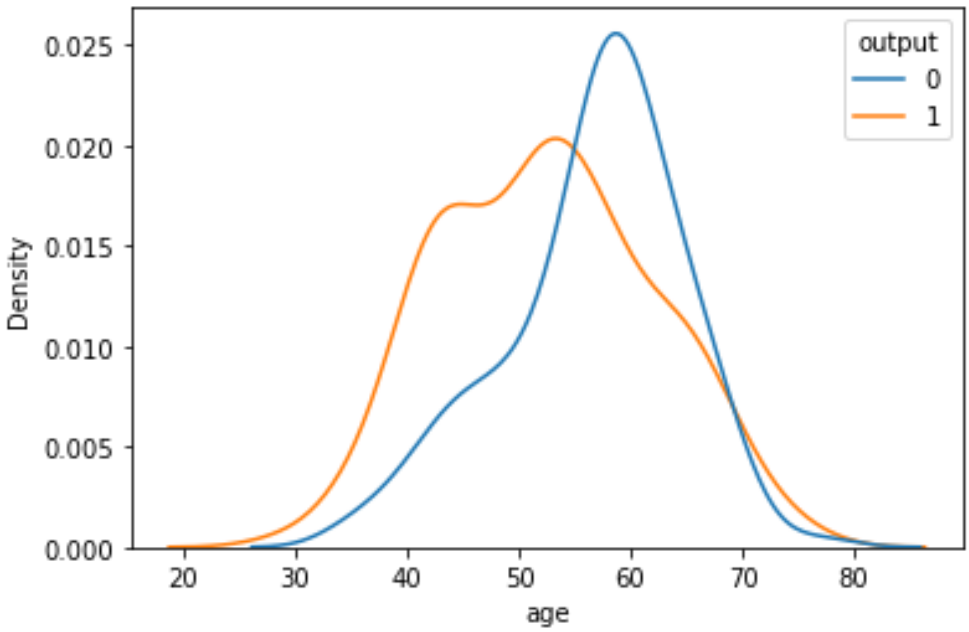
The Python Package Index(PyPI) hosts thousands of third-party modules for Python. Both Python's standard library and the community-contributed modules allow for endless possibilities.

Jupyter Notebook

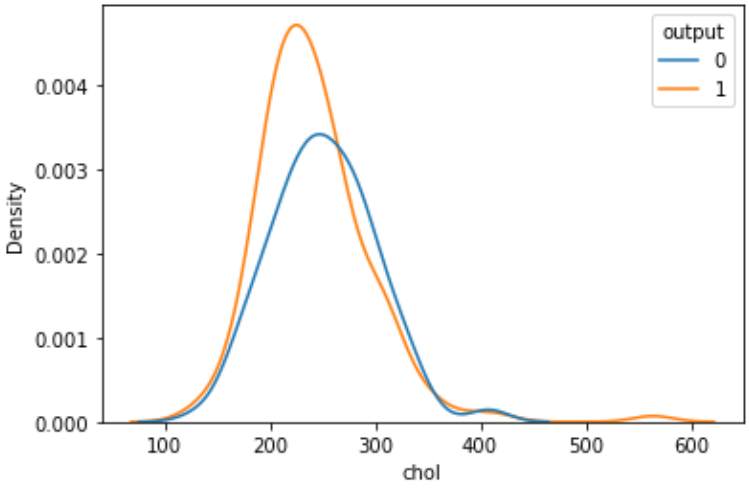
Jupyter Notebook is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

RESULTS & DISCUSSION

3.1 EXPLORATORY DATA ANALYSIS

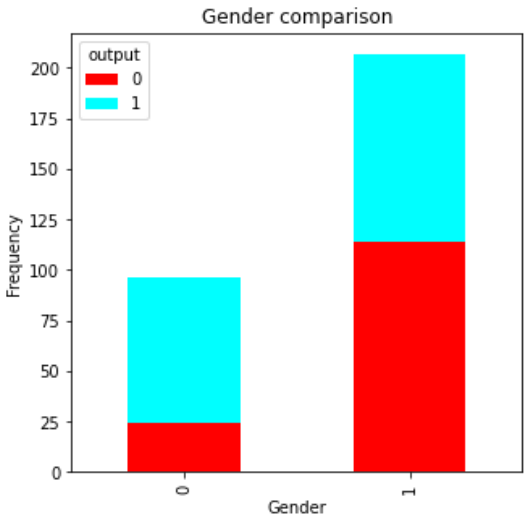


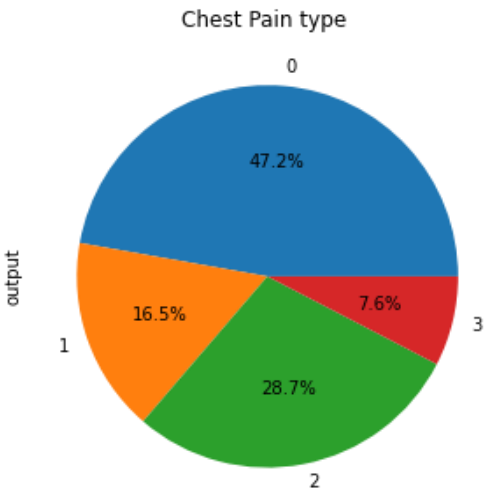
The above graph is density curve according to age we can conclude that Males tend to have heart attack more than females .It is likely that wouldn't have a heart attack if age passes mid-fifties.

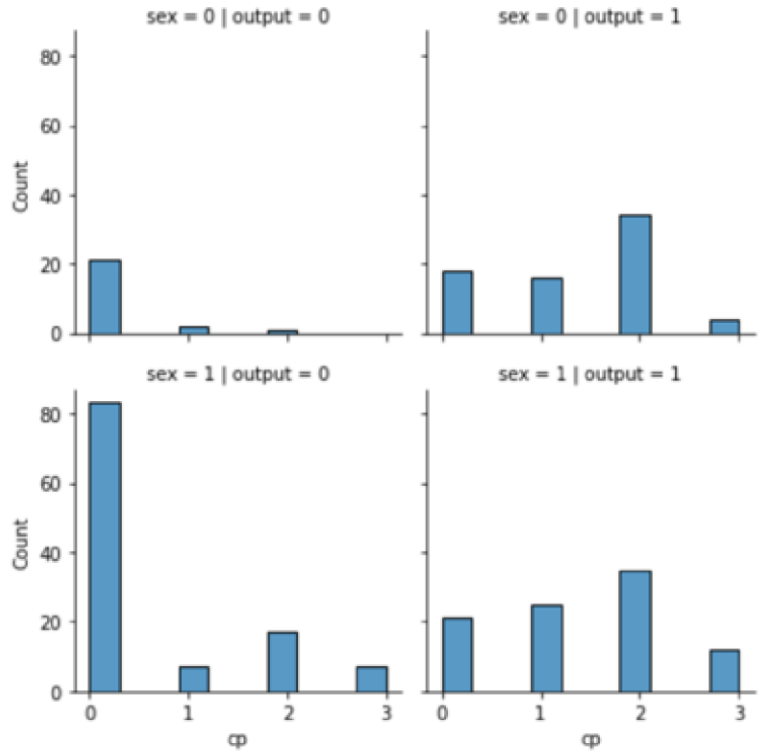


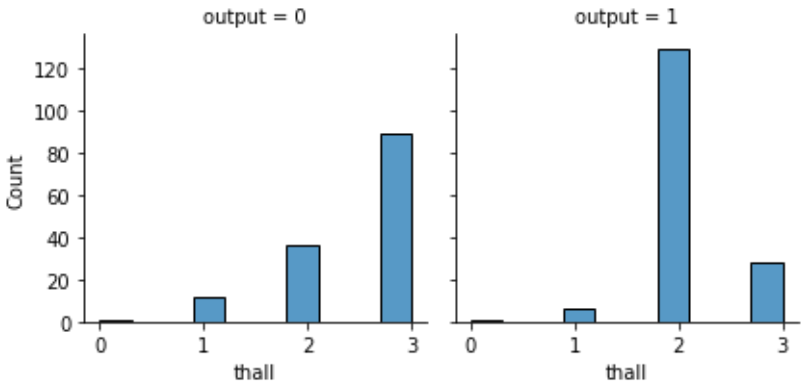
Higher Cholestrol level increases chances of getting heart attack.

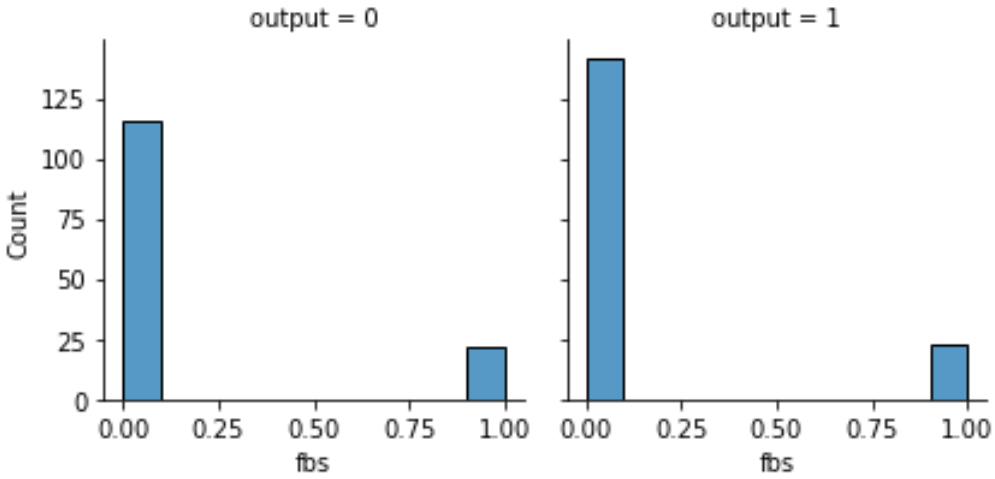
Variance – Covariance Matrix

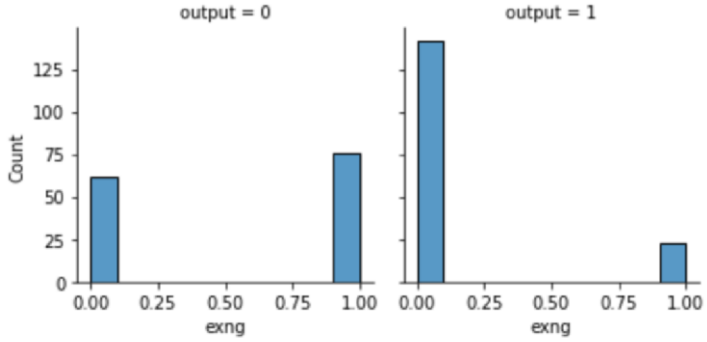

Gender comparison
output
0
1
Frequency
Gender

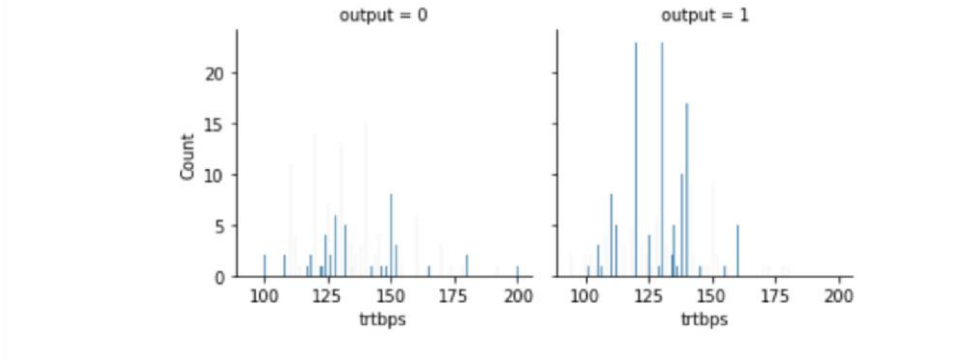

Chest Pain type
output
47.2%
16.5%
7.6%
28.7%
0
1
2
3

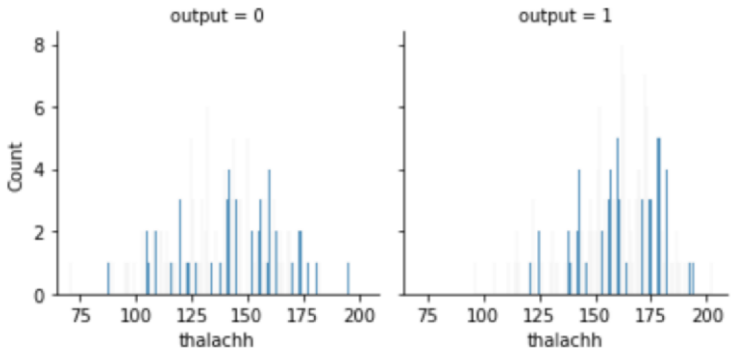

sex = 0 | output = 0
sex = 0 | output = 1
sex = 1 | output = 0
sex = 1 | output = 1
Count
cp

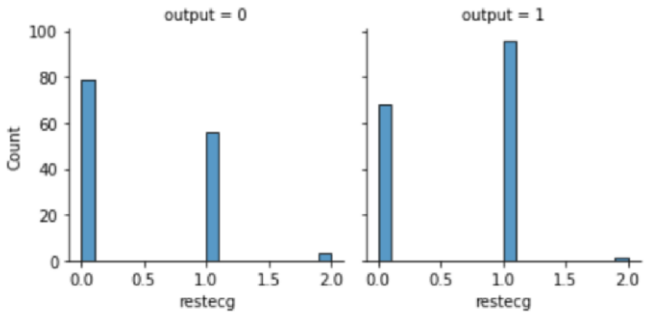

output = 0
output = 1
Count
thall


output = 0
output = 1
Count
125
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
fbs


output = 0
output = 1
Count
125
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
exng


output = 0
output = 1
Count
20
15
10
5
0
100
125
150
175
200
trtbps


output = 0
output = 1
Count
8
6
4
2
0
75
100
125
150
175
200
thalachh




```
              precision    recall  f1-score   support

           0       0.94      0.80      0.86        40
           1       0.81      0.94      0.87        36

    accuracy                           0.87        76
   macro avg       0.88      0.87      0.87        76 weighted
   avg       0.88      0.87      0.87        76

[[32  8]
 [ 2 34]]
LR accuracy : 86.84%


              precision    recall  f1-score   support

           0       0.97      0.80      0.88        40
           1       0.81      0.97      0.89        36

    accuracy                           0.88        76
   macro avg       0.89      0.89      0.88        76 weighted
   avg       0.90      0.88      0.88        76

RF accuracy :88.16%
```

	Model	Accuracy
1	Random Forest	88.157895
0	Logistic Regression	86.842105

Summary

- Males tends to have Heart Attack more than females.
- Chest Pain of Type 1 is more likely to occur.
- The probability of men getting chest pain type 1 is 4 times higher than women getting it.

- No high correlation between chest pain type 1

and actually getting heart attack.

- for “thall “ type 2 indicates a higher probability of getting heart attack.
- There is no significant indication that higher blood sugar indicates a heart attack.
- People who survived a previous stroke before a high chance of 50% to get heart attack.
- People with high amount of cholesterol in their blood are more likely to get heart attack.
- For blood pressure it’s weak correlation - at normal there is no correlation between getting a heart attack .
- Above 120 bp which is normal it’s more likely to get a heart attack.

- Above 160 bp - there is no certainty that a person would get heart attack.

- There is a strong correlation between achieving a heart rate higher than 140 and getting heart attack.
.
- Accuracy of Logistic Regression - 86.84%
.
- Accuracy of Random Forest - 88.16%
.
- Random Forest is the best Classifier in this case.

REFERENCES

- Tolles,Juliana,Meure,William J(2016) -"Logistic Regression Relating to Patient Characteristics to Outcome" – 533
- Hosmer , David W:Lemeshow -"Applied Logistic Regression " -2000
- Thomas W. Edgar, David O. Manz, in Research Methods for Cyber Security, 2017

Website Links :

https://www.wikipedia.com

https://www.mayoclinic.org

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

APPENDIX-(Python code)

```
# import data science basic libraries
import numpy as np
import pandas as pd
```

- data = pd.read_excel("/Users/faizulabbas/Desktop/dataset.xls")
- data
- import seaborn as sns
- import matplotlib.pyplot as plt
- x= data.drop('output' , axis=1)
- y= data['output']
- x
- y
- from sklearn.feature_selection import mutual_info_classif
- from sklearn.preprocessing import LabelEncoder , MinMaxScaler
- from sklearn.model_selection import train_test_split
- from sklearn.metrics import accuracy_score , confusion_matrix ,recall_score , precision_score
- from sklearn.metrics import classification_report
- data.describe()
- data.info()
- age_prob = sns.kdeplot(data=data,hue='output' , x='age')
- chol_prob = sns.kdeplot(data=data , hue = 'output' , x = 'chol')
- plt.figure(figsize=(15,12))
- sns.heatmap(data.corr( ) , annot= True)

- pd.crosstab(data.sex , data.output).plot(kind="bar", stacked=True , figsize=(5,5),color=['red','cyan'])
- plt.title('Gender comparison')
- plt.xlabel('Gender')
- plt.ylabel('Frequency')
- plt.show()
- pie= data.groupby('cp')["output"].count().plot(kind="pie ",autopct='%1.1f%%',figsize=(5,5)
- , title="Chest Pain type ")
- g= sns.FacetGrid(data , col = 'output' , row= 'sex' , height=3)
- g.map(sns.histplot, "cp" , binwidth= 0.3)
- g= sns.FacetGrid(data , col = 'output' , row= 'sex' , height=3)
- g.map(sns.histplot, "cp" , binwidth= 0.3)
- g=sns.FacetGrid(data,col='output',height=3)
- g.map(sns.histplot, "fbs" ,binwidth=0.1)
- g=sns.FacetGrid(data,col='output',height=3)
- g.map(sns.histplot,"exng",binwidth=0.1)
- g=sns.FacetGrid(data,col='output',height=3)
- g.map(sns.histplot,"trtbps",binwidth=0.3)
- g=sns.FacetGrid(data,col='output',height=3)
- g.map(sns.histplot,"thalachh",binwidth=0.3)
- g=sns.FacetGrid(data,col='output',height=3)
- g.map(sns.histplot,"restecg",binwidth=0.1)
- x_train , x_test , y_train , y_test = train_test_split(x,y,test_size=0.25,

- random_state=100)
- from sklearn.linear_model import LogisticRegression
- LRclassifier = LogisticRegression()
- LRclassifier.fit(x_train , y_train)
- y_predict= LRclassifier.predict(x_test)
- print(classification_report(y_test,y_predict))
- print(confusion_matrix(y_test,y_predict))
- from sklearn.metrics import accuracy_score
- LRAcc = accuracy_score(y_predict,y_test)
- print('LR accuracy : {:.2F}%'.format(LRAcc*100))
- from sklearn.ensemble import RandomForestClassifier
- RFClassifier = RandomForestClassifier(n_estimators =10 , max_depth=3)
- RFClassifier.fit(x_train,y_train)
- y_predict1=RFClassifier.predict(x_test)
- print(classification_report(y_test,y_predict1))
- from sklearn.metrics import accuracy_score
- RFAcc = accuracy_score(y_predict1,y_test)
- print('RF accuracy :{:.2f}%'.format(RFAcc*100))

- compare= pd.DataFrame({'Model':['Logistic Regression','Random Forest']
    - ,'Accuracy': [LRAcc*100,RFAcc*100]})
- compare.sort_values(by='Accuracy',ascending=False)