Challenge of the Month February

Heart disease classification

Name:Saanvi Tayal

Challenge of the Month February

1.Problem Scoping

Step 3: Using the 4Ws problem canvas method, include the following in the document:

1. 4Ws - Who, What, Where, Why 2. Problem Statement

Who?	People above 40 years who have high cholesterol and blood pressure and feel chest pain.
What?	Symptoms of cardiovascular diseases are neglected but in future may cause death.
Where?	All over the world.
Why?	Unhealthy diet and lifestyle of people

Problem-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.

Goal- Find the presence of heart disease in the patient

2.Data Acquisition

Step 5: Collect / Acquire a dataset related to the project. Read, analyze and perform exploratory data analysis on the dataset.

Source -Kaggle

Details of columns

1. Age
2. sex (1 = male; 0 = female)
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by fluoroscopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

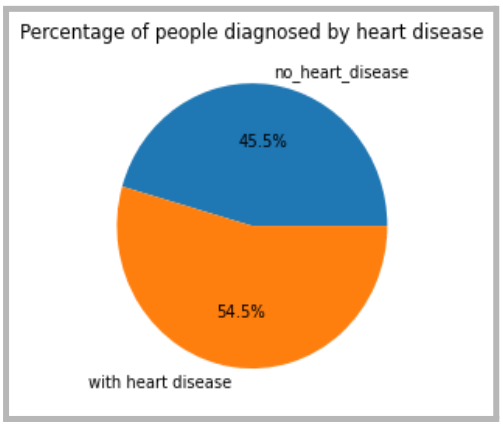
3.Data Exploration (Exploratory Data Analysis)

Step 6: Clean the data and visualise with graphs and charts to extract trends and patterns within the dataset.

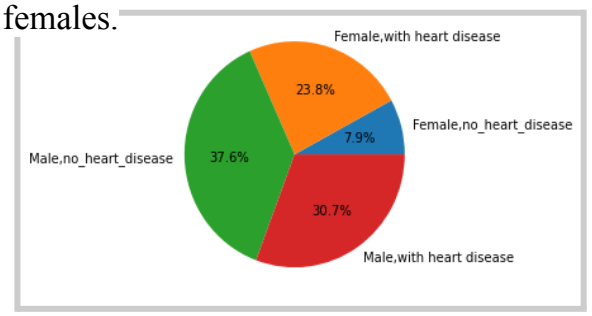
Dataset had 14 features (columns) and 303 rows.I found an already refined dataset so no anomalies were discovered.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

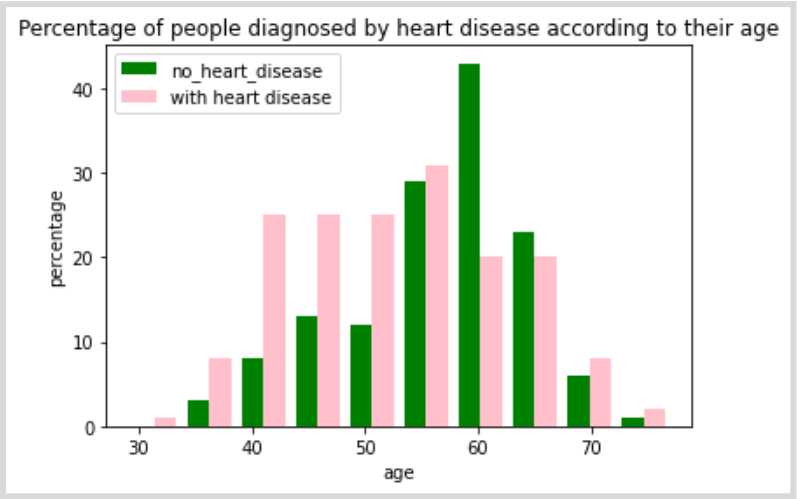
Percentage of people diagnosed by heart disease



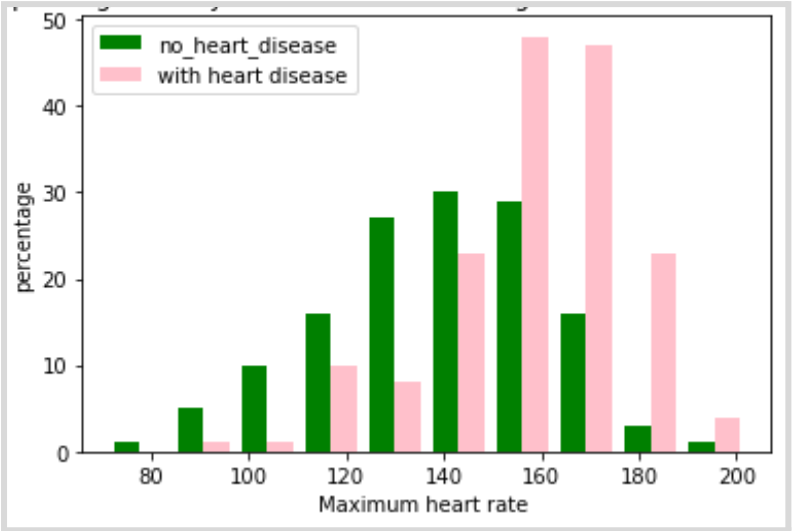
- Over 50% of people were diagnosed with heart disease.
- The ratio of male has heart disease is 30.7%, a little bit higher than females.



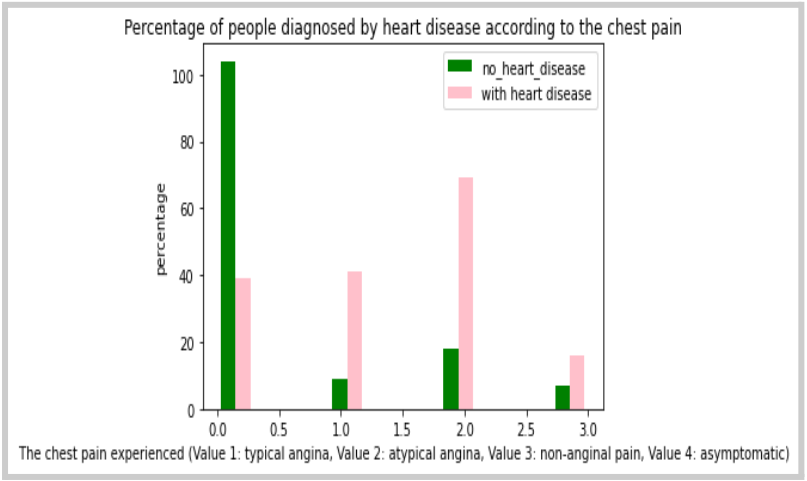
Relation between age, cholesterol , resting blood pressure, minimum blood pressure and occurrence of heart disease.



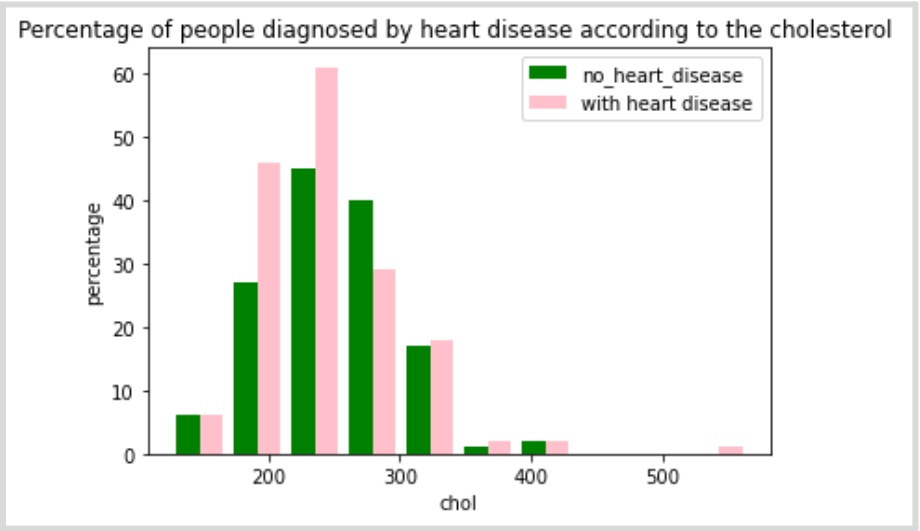
- This graph shows that people above 40 years are more prone to heart diseases .



- This graph shows there are more chances of heart disease if the blood pressure is high. (above 140 mmHg).In addition, everyone should always keep an eye on the resting blood pressure. The ideal resting blood pressure is lower than 120mmHg, but if blood pressure is much lower than the 120mmHg, it means that you are under high risk of heart disease.



- The most obvious symptom is chest pain. There are three types of chest pain, atypical angina is strongly related to heart disease.(Atypical angina is represented by 2.0 on x-axis)



- Also, amounts of people having heart disease are over 200mg/dl of chol. According to the research, the normal value of chol should be lower than 200mg/dl.

4.Modelling

Step 7: Choose a suitable algorithm and train the model with the data.

I. Split dataset between training and testing

Library/Function used :sklearn.model_selection,train_test_split

II. Scaling

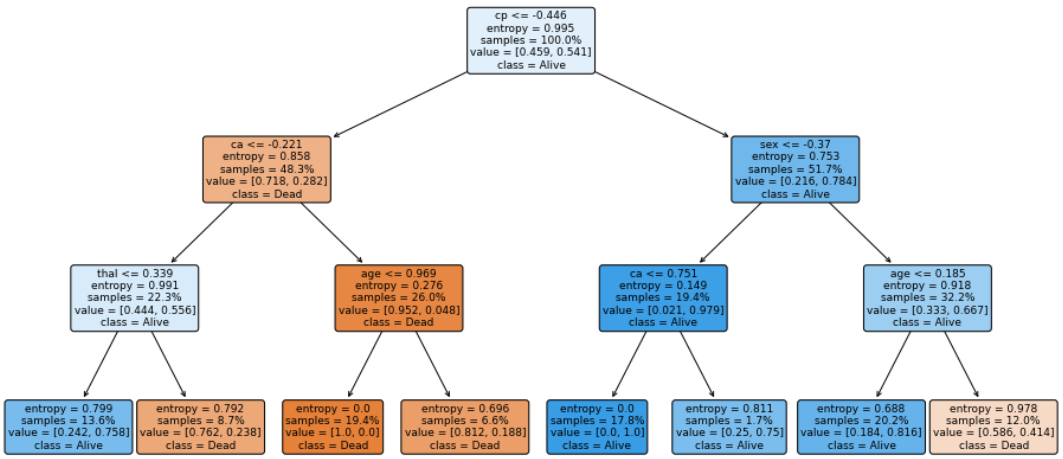
If the data in any condition has data points far from each other, scaling is a technique to make them closer to each other.

Library/Function used :sklearn.preprocessing,StandardScaler(fit_transform)

III. Training different models to choose the best one.

A. Decision trees

Library/Function used: sklearn.tree,DecisionTreeClassifier



B. Logistic Regression
Library/Function used: sklearn.linear_model ,LogisticRegression

C. Random Forest Classifier
Library/Function used: sklearn.ensemble ,RandomForestClassifier

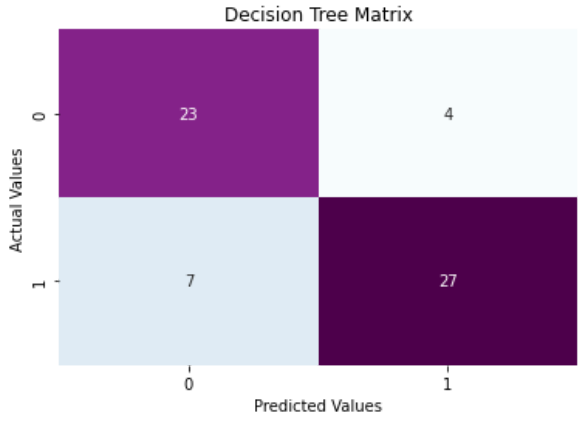
D. Support Vector Classifier
Library/Function used: sklearn.svm ,SVC

5.Evaluation

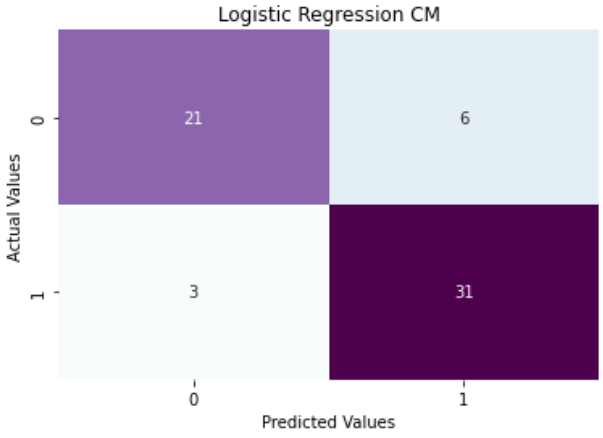
To evaluate , the library /function used : sklearn.metrics-classification_report, confusion_matrix, accuracy_score.

Step 8: Evaluate the performance of the model with relevant evaluation metrics such as accuracy, precision and so on.

I. Decision Trees
Accuracy :Decision Tree accuracy score: 81.97%
Confusion Matrix:



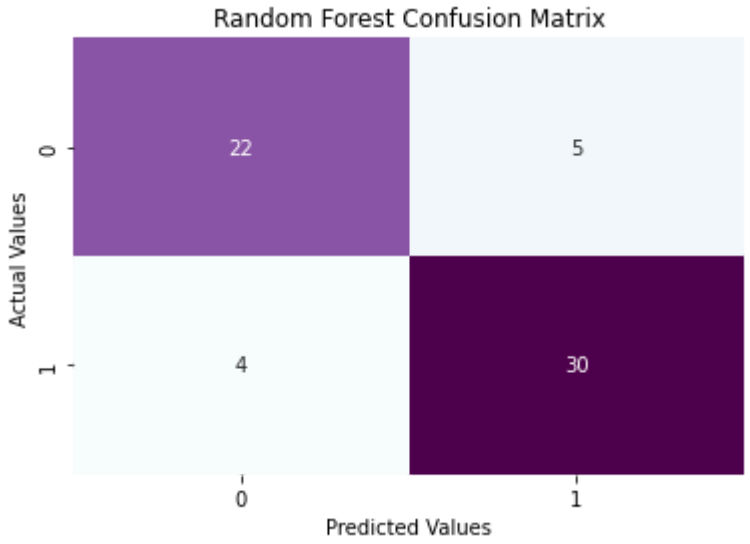
II. Logistic Regression
Accuracy :Logistic Regression accuracy score: 85.25%
Confusion Matrix:



III. Random Forest Classifier

Accuracy :Random Forest accuracy score: 85.25%

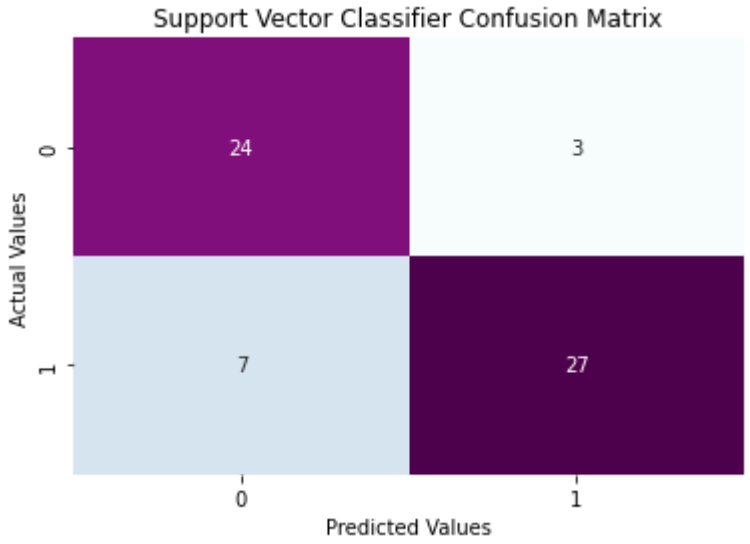
Confusion Matrix:



IV. Support Vector Classifier

Accuracy :SVC accuracy score: 83.61%

Confusion Matrix:



<u>Accuracy Comparison of models :</u>

Decision Trees <Support Vector Classifier <Random Forest Classifier = Logistic Regression

Thus the model chosen or which is best to classify this dataset is either random forest classifier or logistic regression.