Heart Attack Prediction Problem

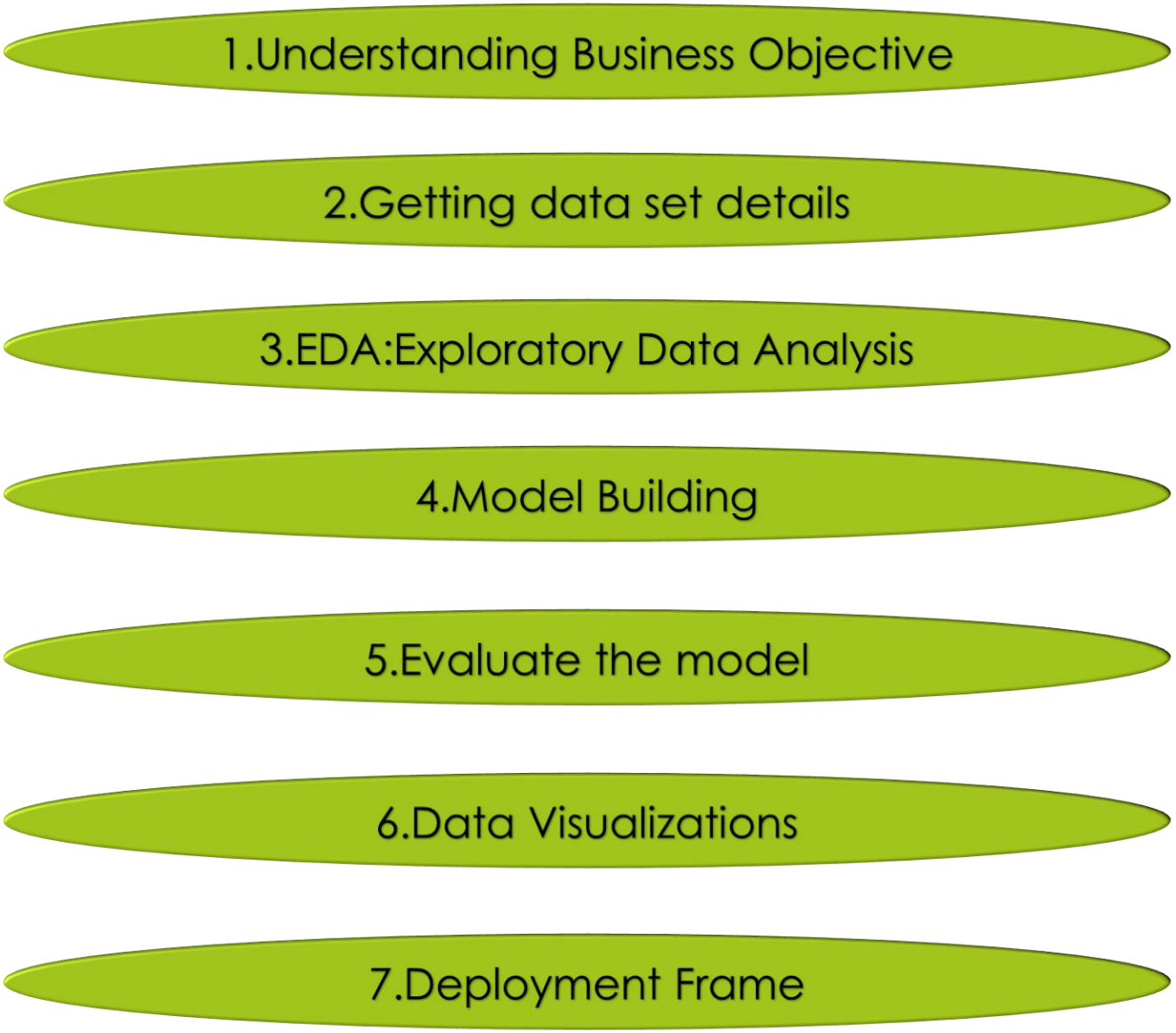
Business Problem:

INNODATATICS
Innovation • Data • Analytics

Business objective-

The main aim of the project is to check whether a person will get heart attack or not?

Project Architecture / Project Flow



Exploratory Data Analysis (EDA) and Feature Engineering

Data set details

The data set has 14 columns and 294 rows.
Following are the Features /Input variables and their description:

- 1)age: age in years
- 2) sex: sex (1 = male; 0 = female)
- 3) cp: chest pain type
  -- Value 1: typical angina
  -- Value 2: atypical angina
  -- Value 3: non-anginal pain
  -- Value 4: asymptomatic
- 4) trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- 5) chol: serum cholestoral in mg/dl
- 6) fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- 7) restecg: 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
- 8) thalach: maximum heart rate achieved
- 9) exang: exercise induced angina (1 = yes; 0 = no)
- 10) oldpeak = ST depression induced by exercise relative to rest
- 11) slope: the slope of the peak exercise ST segment
  -- Value 1: upsloping, Value 2: flat Value 3: downsloping
- 12) ca: number of major vessels (0-3) colored by flourosopy
- 13) thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Data set details

14)The output variable is ‘num’
num: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing(Person will not get heart attack) -- Value 1: > 50% diameter narrowing(Person will get heart attack)

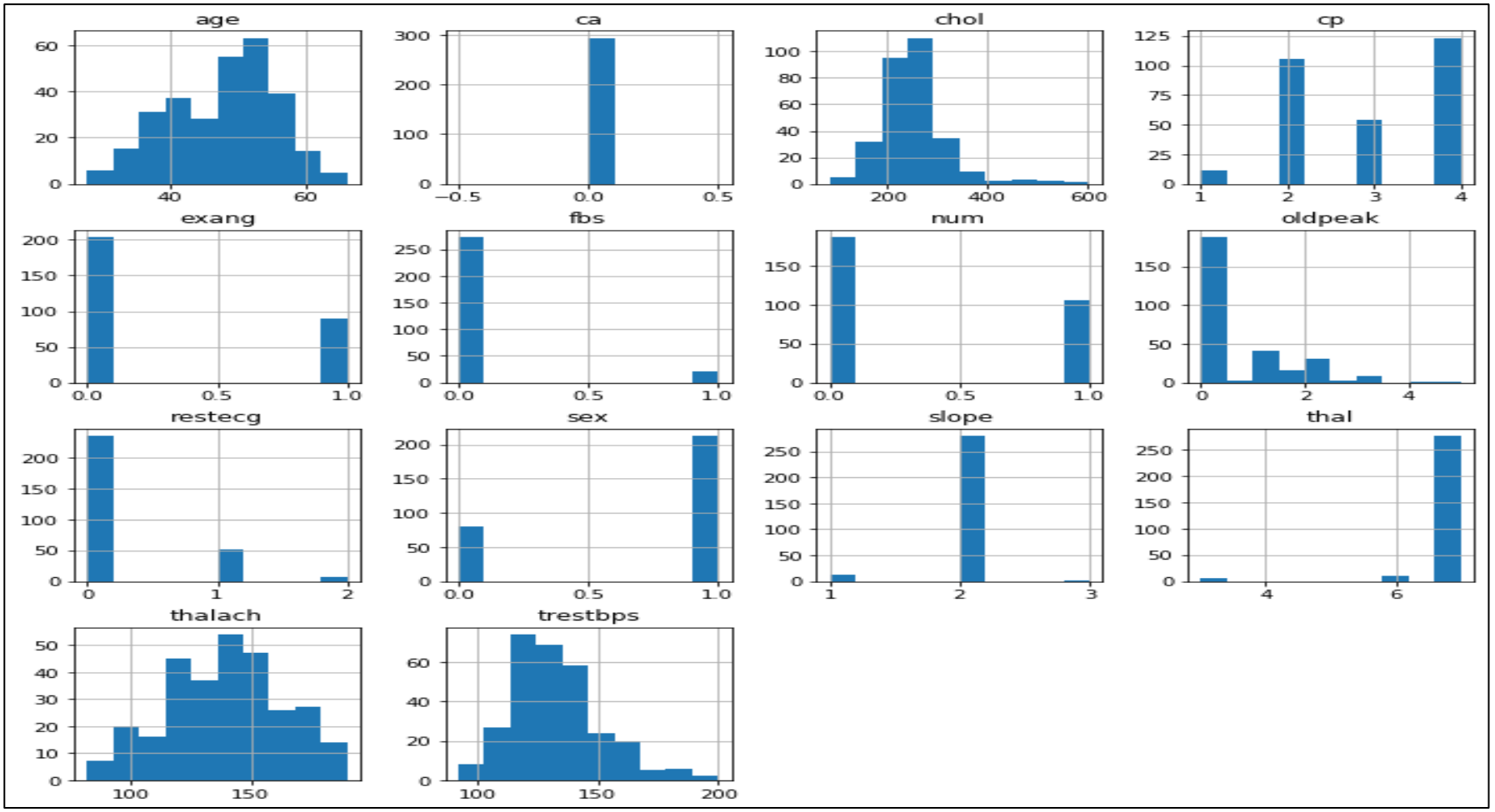
Below snap shot mentions the first 5 rows of the data set

Index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	28	1	2	130	132	0	2	185	0	0	nan	nan	nan	0
1	29	1	2	120	243	0	0	160	0	0	nan	nan	nan	0
2	29	1	2	140	nan	0	0	170	0	0	nan	nan	nan	0
3	30	0	1	170	237	0	1	170	0	0	nan	nan	6	0
4	31	0	2	100	219	0	1	150	0	0	nan	nan	nan	0

NA values are observed for all variables apart from age,sex,cp,oldpeak and num. Snapshot is attached on the right side. We are filling those NA values with the help of Imputation technique .Here mode imputation is used.

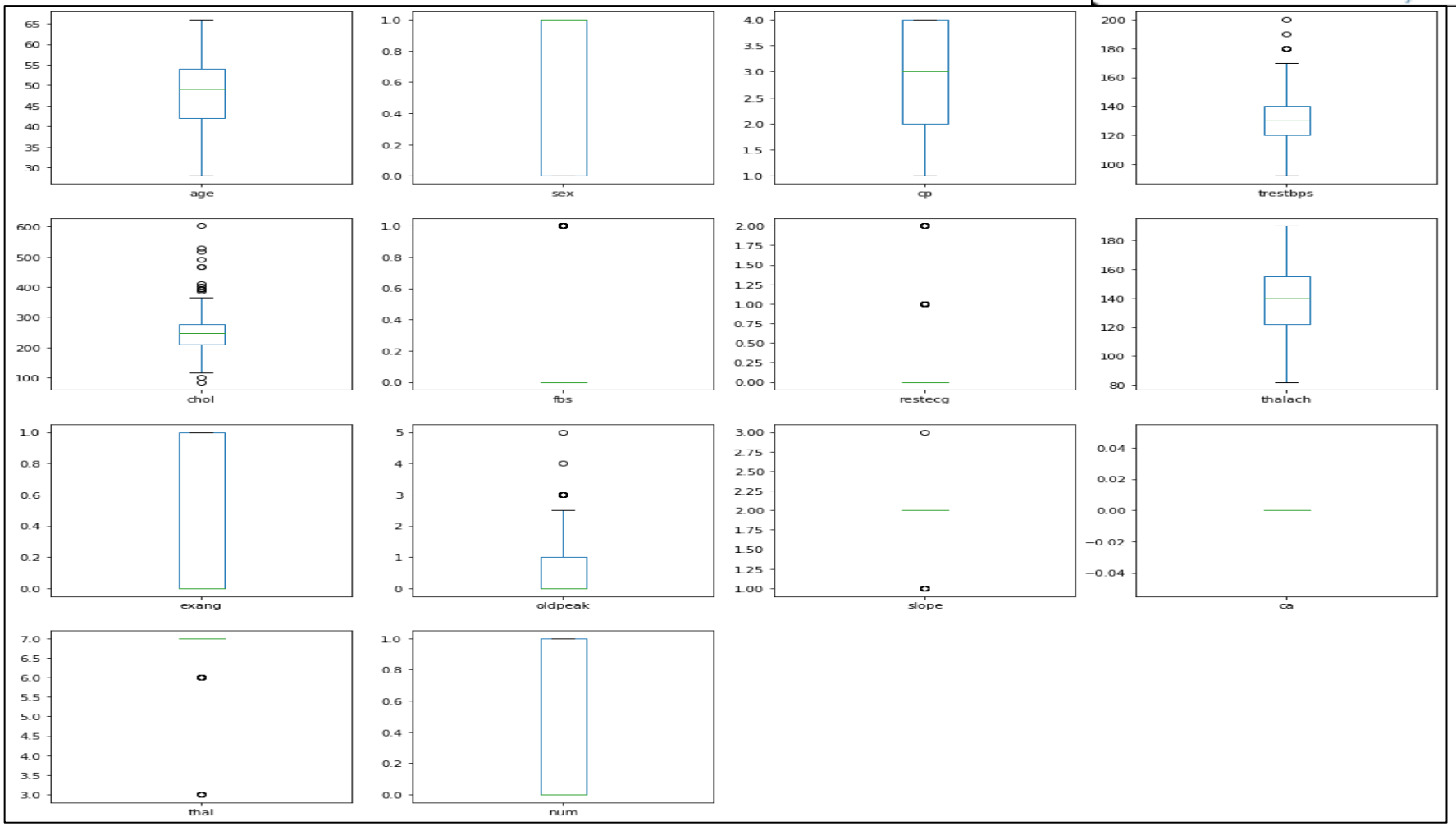
```
age          0
sex          0
cp           0
trestbps     1
chol        23
fbs          8
restecg      1
thalach      1
exang        1
oldpeak      0
slope      190
ca         291
thal       266
num          0
dtype: int64
```

Data visualization :



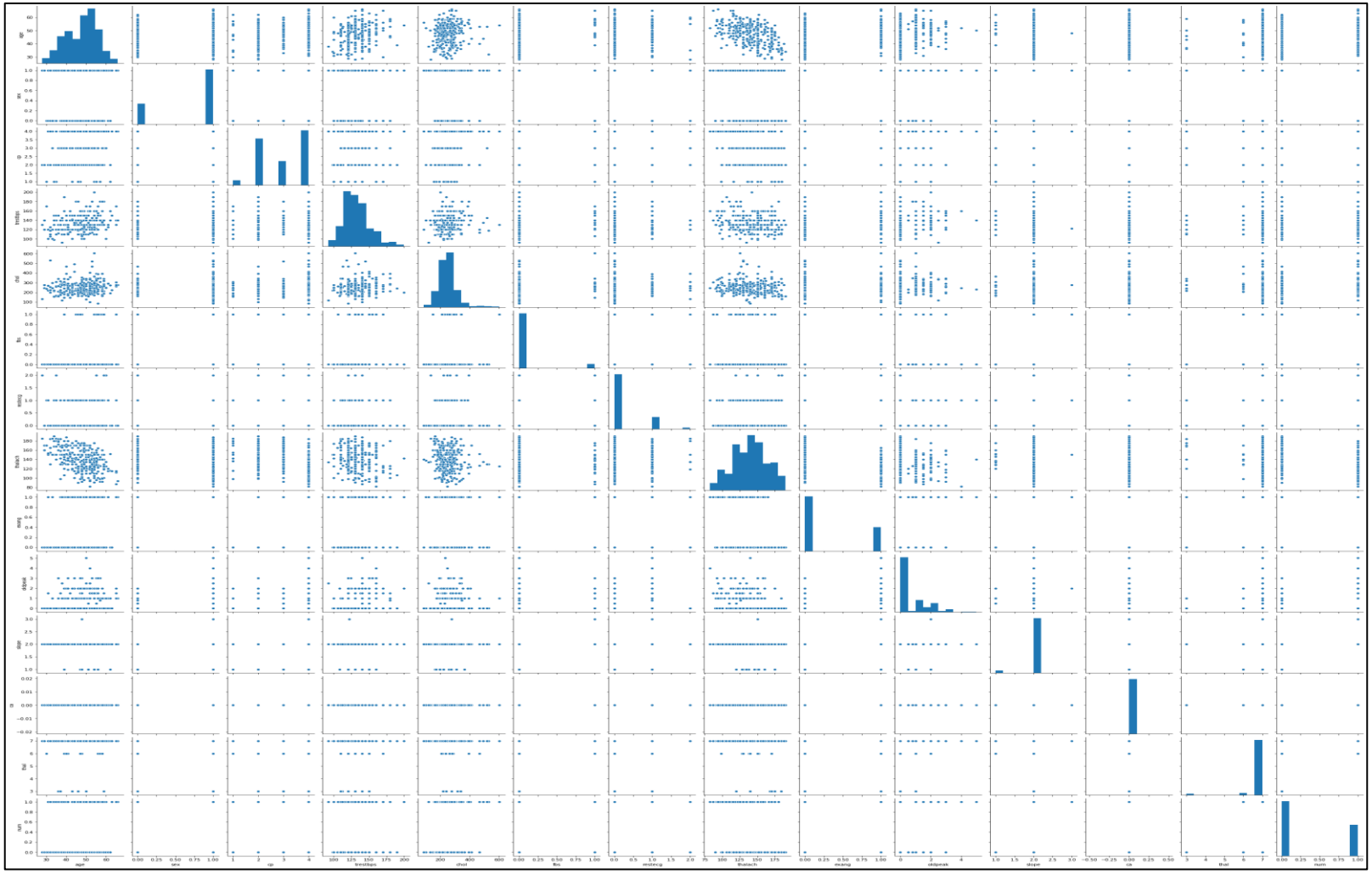
1)The above plot is the histogram plot for all the features in the data set. Since Since we have sex','cp','fbs','restecg','exang','slope','ca','thal' these are in categorical values we need to convert some variable to dummy variable. A **histogram** makes it easier to identify different data, the frequency of its occurrence and categories

Data visualization :



1)The above plot is the box plot for all the features in the data set. The main advantage of boxplot is it shows **outliers.** An **outlier** is a data point that differs significantly from other observations.

Data set details



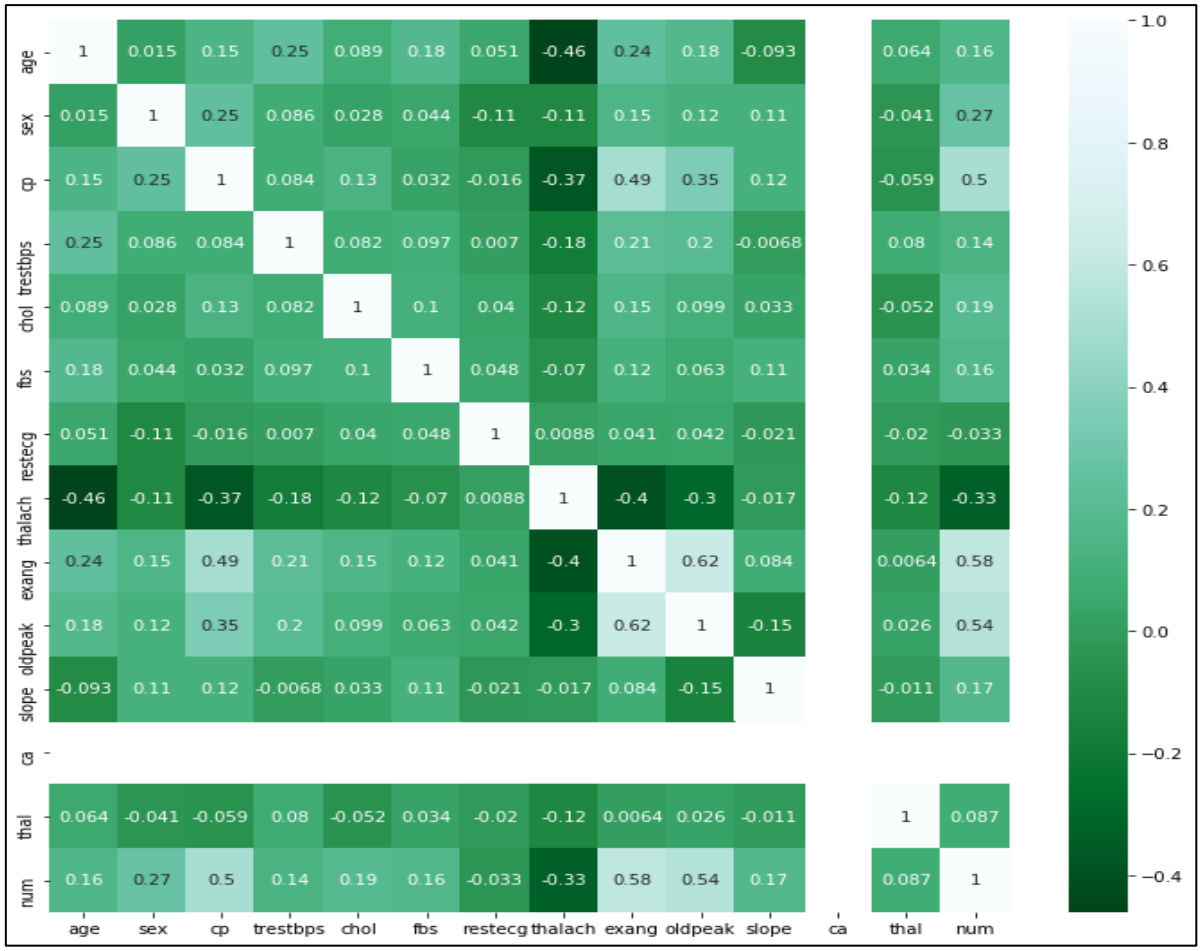
1)The above plot represents the scatter plot for all the features in the data set. A **pairplot** plot a pairwise relationships in a dataset. A “**pairplot**” is also known as a scatterplot, in which one variable in the same data row is matched with another variable's value,.

Data set details

Index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
age	1	0.0145162	0.146616	0.246296	0.0885991	0.18113	0.0506718	-0.460308	0.239223	0.178172	-0.092651	nan	0.0644588	0.15986
sex	0.0145162	1	0.245769	0.0861142	0.0275826	0.0443724	-0.108656	-0.109523	0.154925	0.115959	0.110755	nan	-0.0409074	0.270099
cp	0.146616	0.245769	1	0.0838079	0.131841	0.0319303	-0.0163724	-0.369317	0.494674	0.351735	0.117079	nan	-0.058952	0.504631
trestbps	0.246296	0.0861142	0.0838079	1	0.0815384	0.0968206	0.00695436	-0.182737	0.212989	0.199892	-0.00675529	nan	0.0803875	0.138525
chol	0.0885991	0.0275826	0.131841	0.0815384	1	0.104676	0.0401976	-0.12146	0.152966	0.0987178	0.0333787	nan	-0.0519455	0.193823
fbs	0.18113	0.0443724	0.0319303	0.0968206	0.104676	1	0.0479877	-0.070128	0.115503	0.0631788	0.114321	nan	0.0341241	0.162353
restecg	0.0506718	-0.108656	-0.0163724	0.00695436	0.0401976	0.0479877	1	0.00879219	0.0412905	0.0421931	-0.0213767	nan	-0.0199466	-0.0332471
thalach	-0.460308	-0.109523	-0.369317	-0.182737	-0.12146	-0.070128	0.00879219	1	-0.401415	-0.297866	-0.0174622	nan	-0.118966	-0.330625
exang	0.239223	0.154925	0.494674	0.212989	0.152966	0.115503	0.0412905	-0.401415	1	0.624965	0.0838337	nan	0.00637129	0.583847
oldpeak	0.178172	0.115959	0.351735	0.199892	0.0987178	0.0631788	0.0421931	-0.297866	0.624965	1	-0.14565	nan	0.0258435	0.544957
slope	-0.092651	0.110755	0.117079	-0.00675529	0.0333787	0.114321	-0.0213767	-0.0174622	0.0838337	-0.14565	1	nan	-0.0111136	0.170642
ca	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
thal	0.0644588	-0.0409074	-0.058952	0.0803875	-0.0519455	0.0341241	-0.0199466	-0.118966	0.00637129	0.0258435	-0.0111136	nan	1	0.0870547
num	0.15986	0.270099	0.504631	0.138525	0.193823	0.162353	-0.0332471	-0.330625	0.583847	0.544957	0.170642	nan	0.0870547	1

1)The above plot represents correlation plot for all the features in the data set. A **correlation matrix** is a table showing **correlation** coefficients between variables. Each cell in the table shows the **correlation** between two variables. A **correlation matrix** is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

Data set details



1) The above diagram shows the heat map for all the features in the data set.Heatmaps provide a visual approach to understanding numeric values. It is a representation of data in the form of a map or diagram in which data values are represented as colours. A heat map is data analysis software that uses color the way a bar graph uses height and width: as a data visualization tool.

Model Building

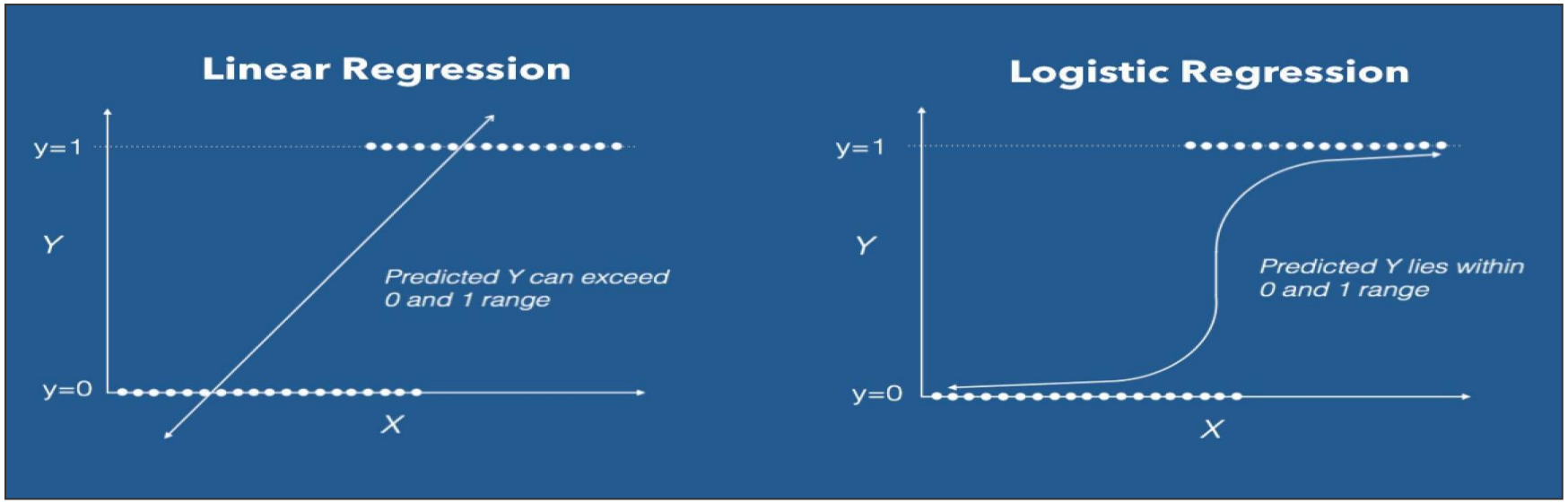
Model Building

Following are the models used for Model building in this project:
1)Logistic regression.
2)Decision tree classifier.
3) Random forest classifier.
4) Extra tree classifier.
5) Support Vector Machine(SVM) Classifier.
6)Neural Networks
7)Bagging classifier method.

Logistic regression:

1)Logistic regression is often referred as logit model is a technique used to predict the probability associated with each dependent variable category.
2)Logistic Regression Model is a generalized form of Linear Regression Model. It is a very good Discrimination Tool.
3)Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using it's underlying logistic function.
4)The probability in logistic regression curve can be given by :

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



Logistic regression:

Advantages of Logistic Regression:

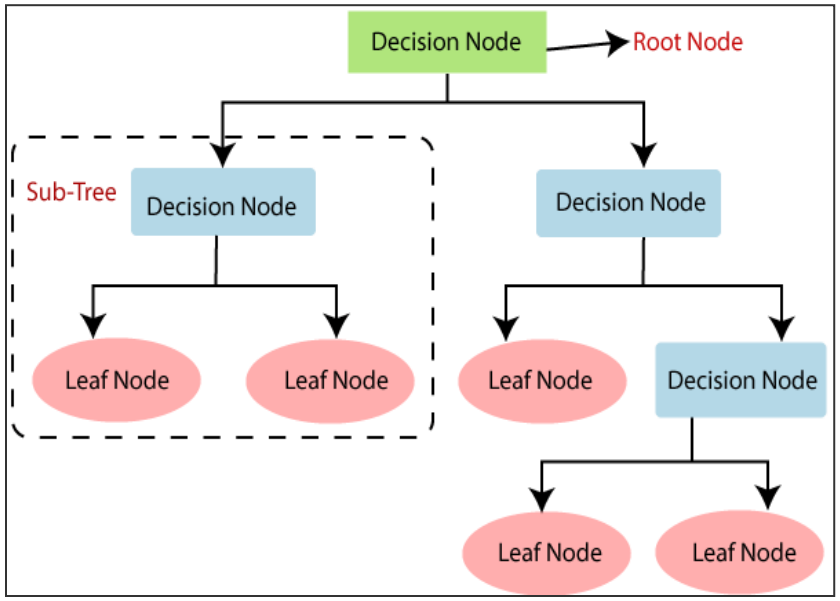
1. Logistic Regression performs well when the dataset is linearly separable.
2. Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.
3. Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).
4. Logistic regression is easier to implement, interpret and very efficient to train.

Disadvantages of Logistic Regression:

1. Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.
2. If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit.
3. Logistic Regression can only be used to predict discrete functions. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set. This restriction itself is problematic, as it is prohibitive to the prediction of continuous data

Decision tree classifier :

1)It is called as Greedy algorithm and Supervised classification model.
2)Decision tree algorithm is like a tree like structure where it is the combination of root node ,branch node and leaf nodes.
3)With the help of Entropy and Information gain we have to choose Root node.
4)Outcomes here are leaf nodes.
5)Overfitting is the main problem in decision tree classifier. As the model overfits we call it as greedy. We can go with pruning method for removing the branches without removing the information.



Root Node: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
Leaf Node: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
Splitting: Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
Branch/Sub Tree: A tree formed by splitting the tree.

Decision tree classifier :

Advantages of the Decision Tree:

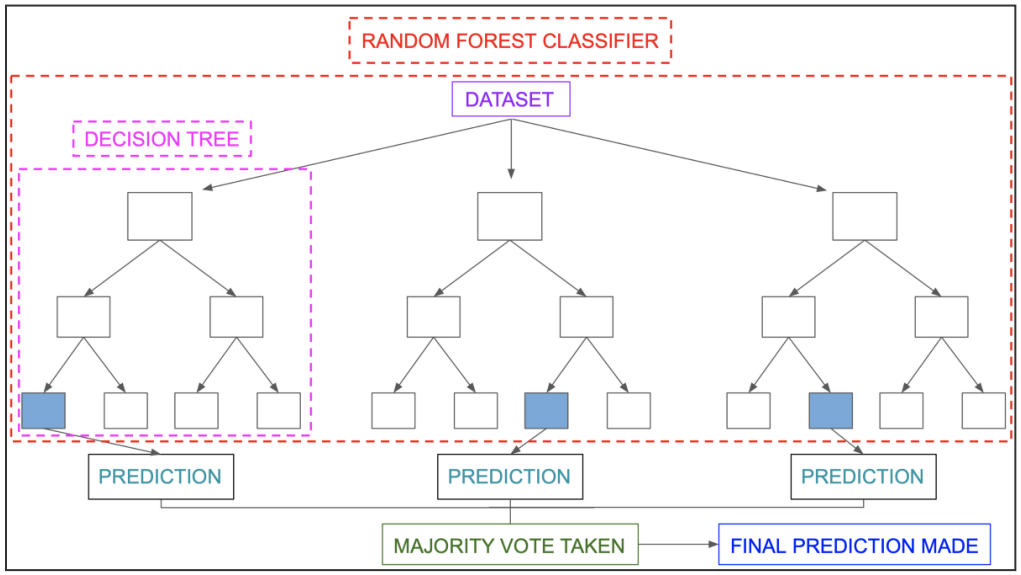
1)It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
2)It can be very useful for solving decision-related problems.
3)It helps to think about all the possible outcomes for a problem.
4)There is less requirement of data cleaning compared to other algorithms.

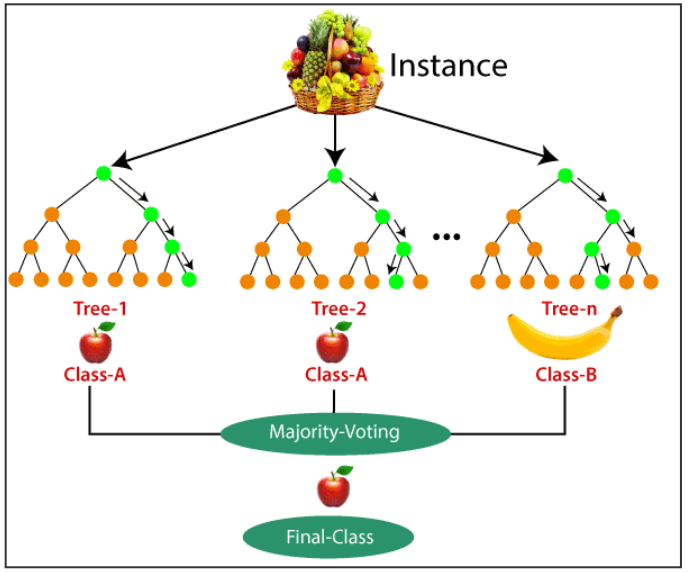
Disadvantages of the Decision Tree:

1)The decision tree contains lots of layers, which makes it complex.
2)It may have an overfitting issue, which can be resolved using the **Random Forest algorithm**.
3) For more class labels, the computational complexity of the decision tree may increase.

Random Forest classifier :

1) Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It is combination of multiple decision tress. Its is an **ensemble** method.
2) 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.
3) The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.





Random Forest classifier :

Advantages of Random Forest:

1)Random Forest is capable of performing both Classification and Regression tasks.
2)It is capable of handling large datasets with high dimensionality.
3)t enhances the accuracy of the model and prevents the overfitting issue.
4)It can handle null values in the data set.

Disadvantages of Random Forest:

1)Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

Bagging Classifier :

1) Bagging is a technique that can be used to improve the accuracy of Classification & Regression Trees (CART).They're known as 'ensemble' methods.
2) Bagging reduces overfitting (variance) by averaging or voting, however, this leads to an increase in bias, which is compensated by the reduction in variance though.
3) A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

Extra tree classifier:

1) The **Extra Trees algorithm** works by creating a large number of unpruned decision **trees** from the training dataset. Predictions are made by averaging the prediction of the decision **trees** in the case of regression or using majority voting in the case of classification

2) ExtraTreesClassifier is an ensemble learning method fundamentally based on decision trees. ExtraTreesClassifier, like RandomForest, randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting.

3) Extra Trees is like Random Forest, in that it builds multiple trees and splits nodes using random subsets of features, but with two key differences: it does not bootstrap observations (meaning it samples without replacement), and nodes are split on random splits, not best splits. So, in summary,

ExtraTrees:

1) builds multiple trees with **bootstrap** = **False** by default, which means it samples without replacement
2)nodes are split based on **random** splits among a **random subset** of the features selected at every node
4)In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.Also called as **Extremely Randomized Trees Classifier(Extra Trees Classifier)**

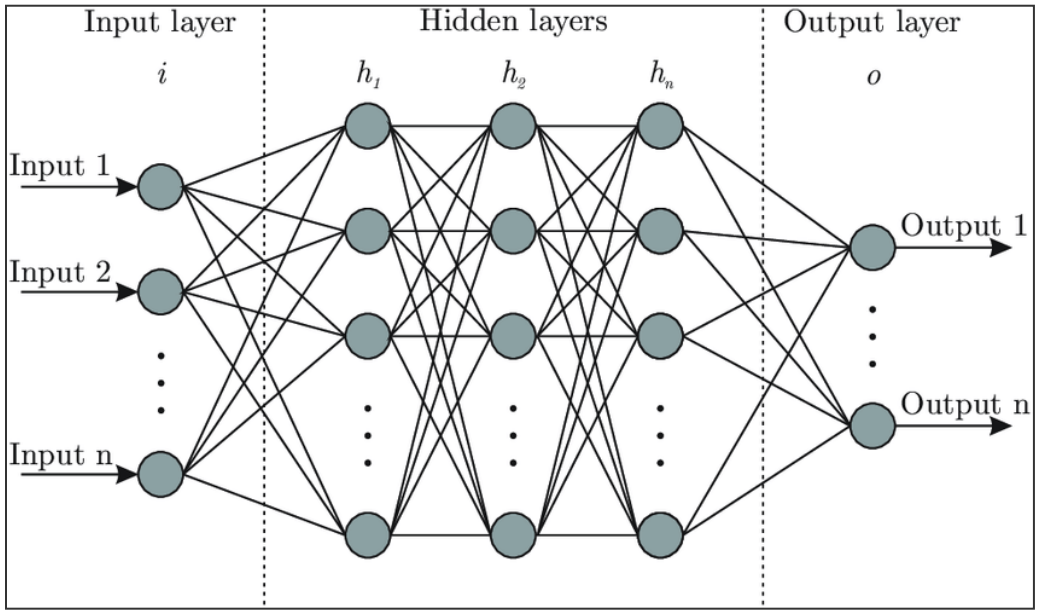
Extra tree classifier:

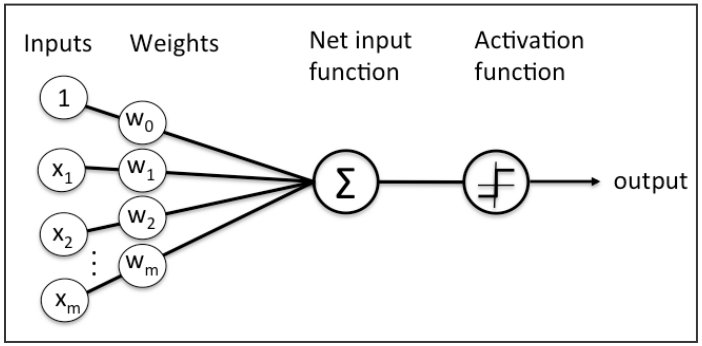
	Decision Tree	Random Forest	Extra Trees
Number of trees	1	Many	Many
No of features considered for split at each decision node	All Features	Random subset of Features	Random subset of Features
Boostrapping(Drawing Sampling without replacement)	Not applied	Yes	No
How split is made	Best Split	Best Split	Random Split

1) Major difference between Decision tree ,Radom forest and Extra trees classifier.

Neural Networks :

1) A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes. Thus a neural network is either a biological neural network, made up of real biological neurons, or an artificial neural network, for solving artificial intelligence(AI) problems.
2) Neural network architecture :



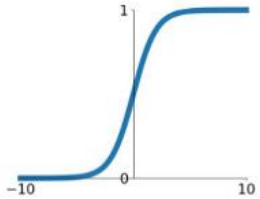


Neural Networks :

Different types of Activation functions:

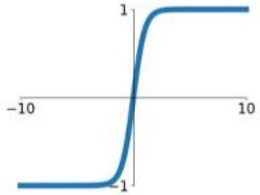
Sigmoid

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



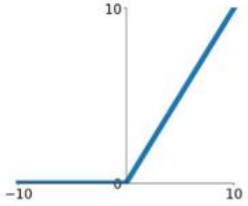
tanh

$$\tanh(x)$$



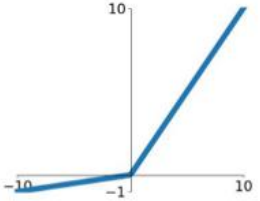
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

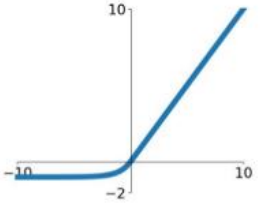


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

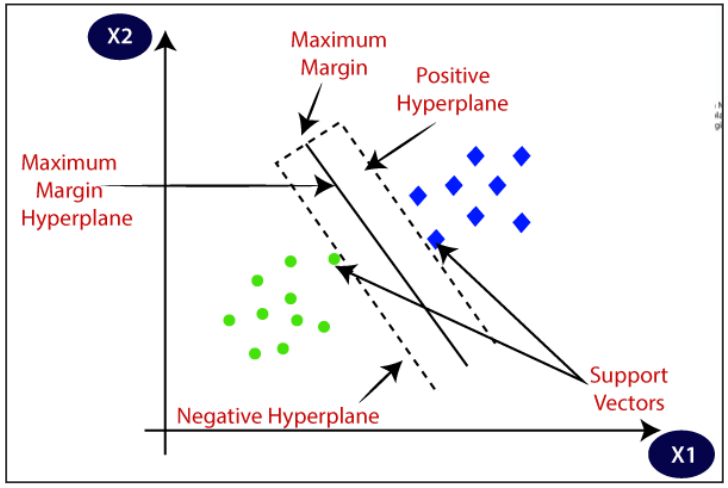
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

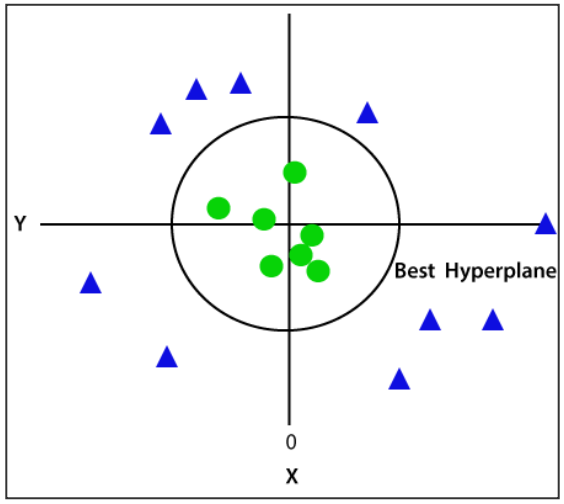


Support Vector Machine(SVM) Classifier:

1) Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.
2) The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
3) SVM is of two types .**a)Linear SVM :** It is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

b**) Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.





Model building Results:

Model Used	num	Precision	Recall	f1-score	Number of errors	Accuracy
Logistic Regression	0	0.86	0.91	0.88	8	0.86
	1	0.88	0.81	0.84		
Decision Tree Classifier	0	0.85	0.85	0.85	10	0.83
	1	0.81	0.81	0.81		
Random Forest Classifier	0	0.81	0.88	0.84	11	0.81
	1	0.83	0.73	0.78		
Extra tree Classifier	0	0.78	0.88	0.83	12	0.79
	1	0.82	0.69	0.75		
SVM Classifier	0	0.78	0.88	0.83	12	0.8
	1	0.82	0.69	0.75		
Neural Networks	0	0.83	0.88	0.85	10	0.83
	1	0.83	0.77	0.8		
Bagging Classifier	0	0.76	0.85	0.8	14	0.76
	1	0.77	0.65	0.71		

Model Used	cohen_kappa_score
Logistic Regression	0.72
Decision Tree Classifier	0.65
Random Forest Classifier	0.61
Extra tree Classifier	0.58
SVM Classifier	0.58
Neural Networks	0.65
Bagging Classifier	0.51

Evaluation step:

1)From all model building results we can conclude that cohen_kappa_Score for Logistic Regression model is 0.72 which is good compared to other models ,Number of predicting errors =8 which is less compared to other models. Accuracy is high 86 % which is good in Logistic regression. So the final model is Logistic regression
2)Precision = Quality of prediction .
3)Recall = Measurement of coverage.

Cohen_kappa_Score:
1)Cohen's Kappa statistic is a very useful, but under-utilised, metric. Sometimes in machine learning we are faced with a multi-class classification problem. In those cases, measures such as the accuracy, or precision/recall do not provide the complete picture of the performance of our classifier. In some other cases we might face a problem with imbalanced classes(Majority class and minority class problems)
2) Cohen's kappa statistic is a very good measure that can handle very well both multi-class and imbalanced class problems.
3) Cohen's kappa is given by the formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

Where : where po is the observed agreement, and pe is the expected agreement.

4) Cohen's kappa is always less than or equal to 1. Values of 0 or less, indicate that the classifier is useless. There is no standardized way to interpret its values. Landis and Koch (1977) provide a way to characterize values. According to their scheme a value < 0 is indicating no agreement , 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.

Model Deployment using

Model Deployment

- The Model Deployed is **Logistic Regression**.
- For deployment we have created files like app.py and Model.pkl.
- **Regression**(Higher Accuracy Model) saved as to **Model.pkl.**
- And **app.py** is the Flask deployment code.

Deployment Using flask

Flask File Creation:

- Import Flask from flask module
- Create an instance of the Flask class
- Using @app.route('/') to execute home function which gives home page for this use **home.html**
- Using @app.route('/predict', methods=[POST]) to execute predict function which gives results page for this use **result.html**
- After execute whole deployment code it gives link like **http://127.0.0:5000** run this link to get results.

Results:


127.0.0.1:5000/predict
Machine Learning App with Flask
Heart Attack prediction
Age of the person:
28
Sex:
1
cp-Chest pain type:
2
trestbps-Resting blood pressure:
112
chol-serum cholestoral in mg/dl:
132
fbs-fasting blood sugar :
0
restecg-resting electrocardiographic results:
0
thalach-maximum heart rate achieved:
165
exang-exercise induced angina:
0
oldpeak - ST depression induced by exercise relative to rest:
0
slope-Resting blood pressure:
1
ca number of major vessels (0-3) colored by flourosopy:
0
thal :
0
Predict
Person will not get heart attack

Thank you