Dataset Name: Heart Disease

Group Name: Group-294

Contribution Table:

S. No	Name (as appears in Canvas)	ID EMAIL ID	Contribution
1	J PON DEEPAK ANTONY RAJ	2021sc04108@wilp.bits-pilani.ac.in	Equal (100%)
2	VEERAMREDDY SANJANA	2021sa04053@wilp.bits-pilani.ac.in	Equal (100%)
3	MOHIT DANG	2021sa04052@wilp.bits-pilani.ac.in	No (0%)

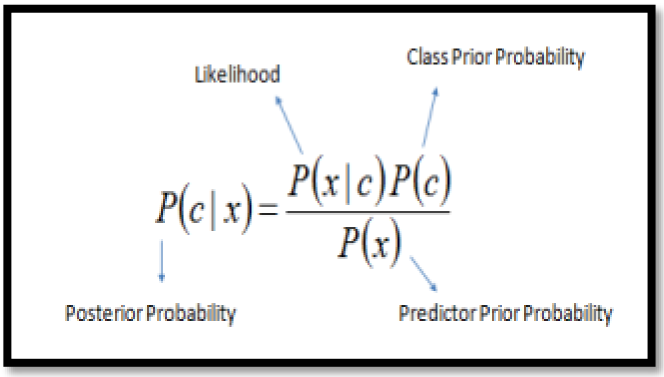
Abstract: This dataset can be used to predict heart disease and it has been collected at a hospital over a period.

Solution: Analyzed the data set, investigated, and evaluated the result and predicted the overall performance.

Introduction: Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases. Heart diseases have emerged as one of the most prominent causes of death all around the world. According to World Health Organisation, heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the leading cause of mortality. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15,2017. Heart related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organisation (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or cardiovascular diseases. Thus, feasible and accurate prediction of heart related diseases is very important. Medical organisations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately

Algorithms and Techniques Used: **Random Forest has good result as compare to other algorithms**

1. Naive Bayes (Scikit-learn): Naive Bayes is a simple but an effective classification technique which is based on the Bayes Theorem. It assumes independence among predictors, i.e., the attributes or features should be not correlated to one another or should not, in anyway, be related to each other. Even if there is dependency, still all these features or attributes independently contribute to the probability and that is why it is called Naïve. In, Naive Bayes has achieved an accuracy of 84.1584% with the 10 most significant features which are selected using SVMRFE (Recursive Feature Elimination) and gain ratio algorithms whereas in,Naive Bayes has achieved an accuracy of 83.49% when all 13 attributes of the Cleveland dataset are used.



2. Support Vector Machine (Linear) (Scikit-learn): Support Vector Machine is an extremely popular supervised machine learning technique (having a pre-defined target variable) which can be used as a classifier as well as a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall. Shan Xu et al. have used SVM to achieve an accuracy of 98.9% in People's Hospital dataset .In, SVM performs the best with 85.7655% of correctly classified instance and in SVM is used with boosting technique to give an accuracy of 84.81%. HoudaMezrigui et al. have used SVM to attain a f-measure value of 93.5617 . In SVM classifies the pixel variation with an accuracy of 92.1% helping to identify the affected region accurately.

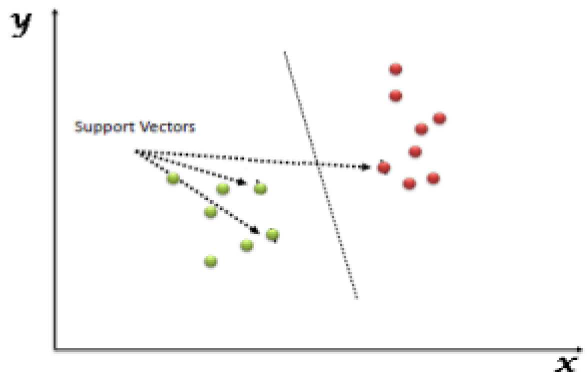
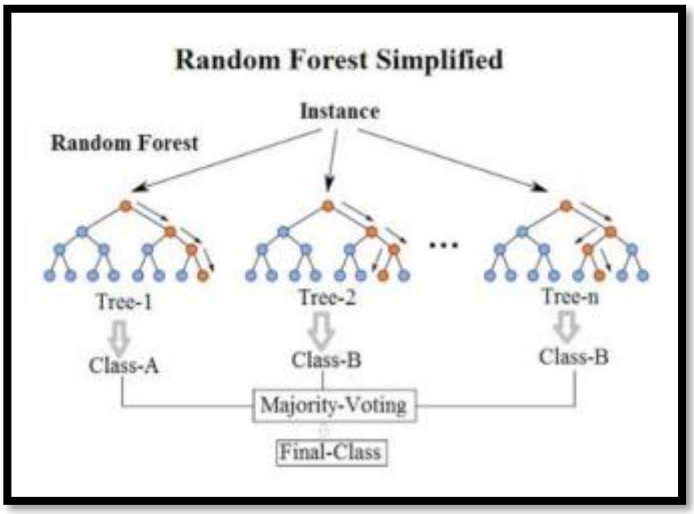


Fig. 1: Support Vector Machine

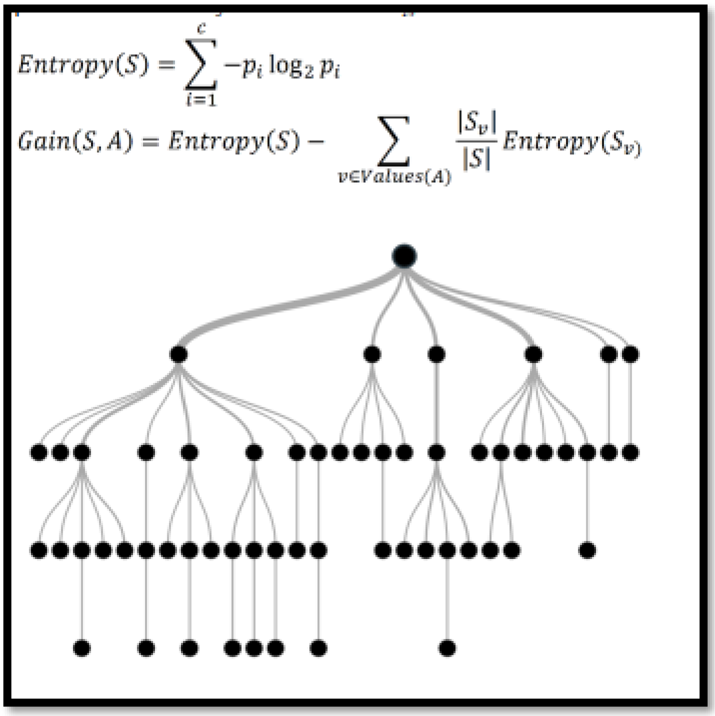
3. Random Forest: Random Forest is also a popularly supervised machine learning algorithm. This technique can be used for both regression and classification tasks but generally performs better in classification tasks. As the name suggests, Random Forest technique considers multiple decision trees before giving an output. So, it is basically an ensemble of decision trees. This technique is based on the belief that more number of trees would converge to the right decision. For classification, it uses a voting system and then decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees. It works well with large datasets with high dimensionality. In , random forest performs exceptionally well. In Cleveland dataset, random forest has a significantly higher accuracy of 91.6% than all the other methods. In People's Hospital dataset, it achieves an accuracy of 97%. In random forest has achieved an f-measure of 0.86. In , random forest is used to predict coronary heart disease and it obtains an accuracy of 97.7%.



4. K – Nearest Neighbour: In 1951, Hodges et al. introduced a nonparametric technique for pattern classification which is popularly known the K-Nearest Neighbour rule. K-Nearest Neighbour technique is one of the most elementary but very effective classification techniques. It makes no assumptions about the data and is generally be used for classification tasks when there is very less or no prior knowledge about the

data distribution. This algorithm involves finding the k nearest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the found data points to it. In KNN gives an accuracy of 83.16% when the value of k is equal to 9 while using 10-cross validation technique. In KNN with Ant Colony Optimization performs better than other techniques with an accuracy of 70.26% and the error rates is 0.526.Ridhi Saini et al. have obtained a efficiency of 87.5%, which is very good.

5. Decision Tree: Decision tree is a of supervised learning algorithm.This technique is mostly used in classification problems. It performs effortlessly withcontinuous and categorical attributes. This algorithm dividesthe population into two or more similar sets based on the most significantpredictors.Decision Treealgorithm, first calculates the entropy of each and every attribute. Then the dataset is split with the help of thevariables or predictors with maximum information gain or minimum entropy. These two steps are performed recursively with the remaining attributes. In decision tree has the worst performance with an accuracy of 77.55% but when decision tree is used with boosting technique it performs better with an accuracy of 82.17%.In decision tree performs very poorly with a correctly classified instance percentage of 42.8954% whereas in also uses the same dataset but used the J48 algorithm for implementing Decision Trees and the accuracy thus obtained is 67.7% which is less but still an improvement on the former. Renu Chauhan et al. have obtained an accuracy of 71.43%. M.A. Jabbar et al. have used alternating decision trees with principle component analysis to obtain an accuracy 92.2%.Kamran Farooq et al. have achieved the best results on using decision tree-based classifier combined with forward selection which achieves a weighted accuracy of 78.4604%. 686 Internat.



$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Step by step breakdown of code:

1)Importing libraries

2)Loading data set

3)Displaying sample entries

4)Data cleaning &Pre-processing

a)Renaming features to proper name

b)Converting features to categorical features

c)Checking top 5 entries

d)Checking missing entries

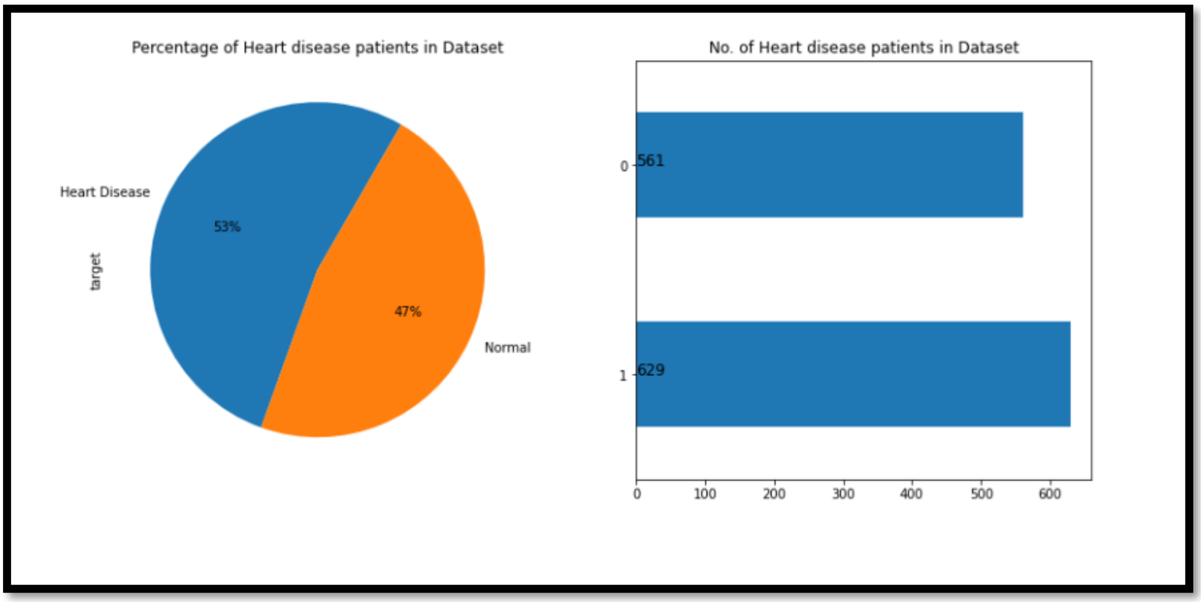
5)Exploratory data analysis(EDA)

a)checking shape of data set

b)summary statistics of numerical columns

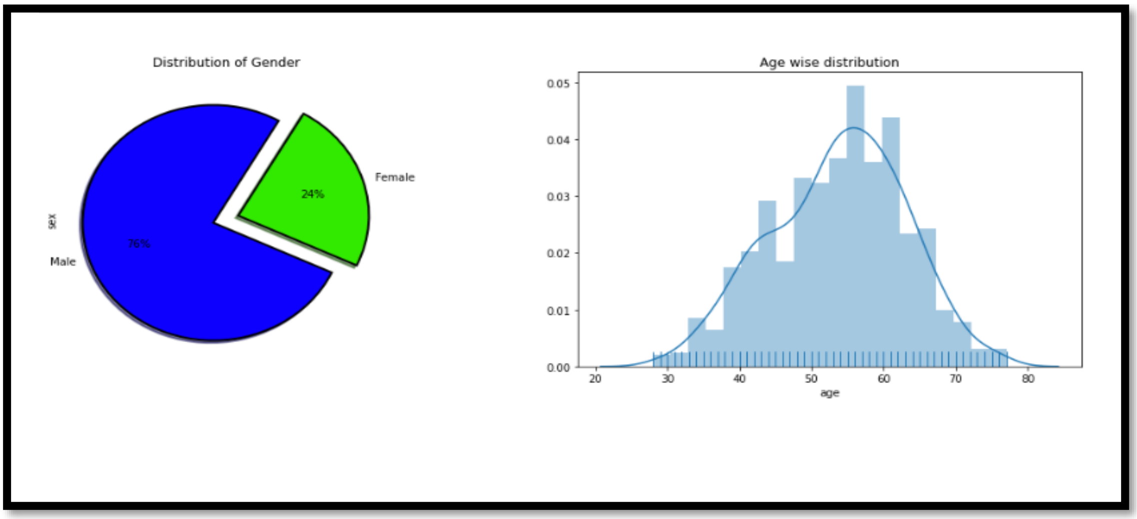
c)summary statistics of categorical columns

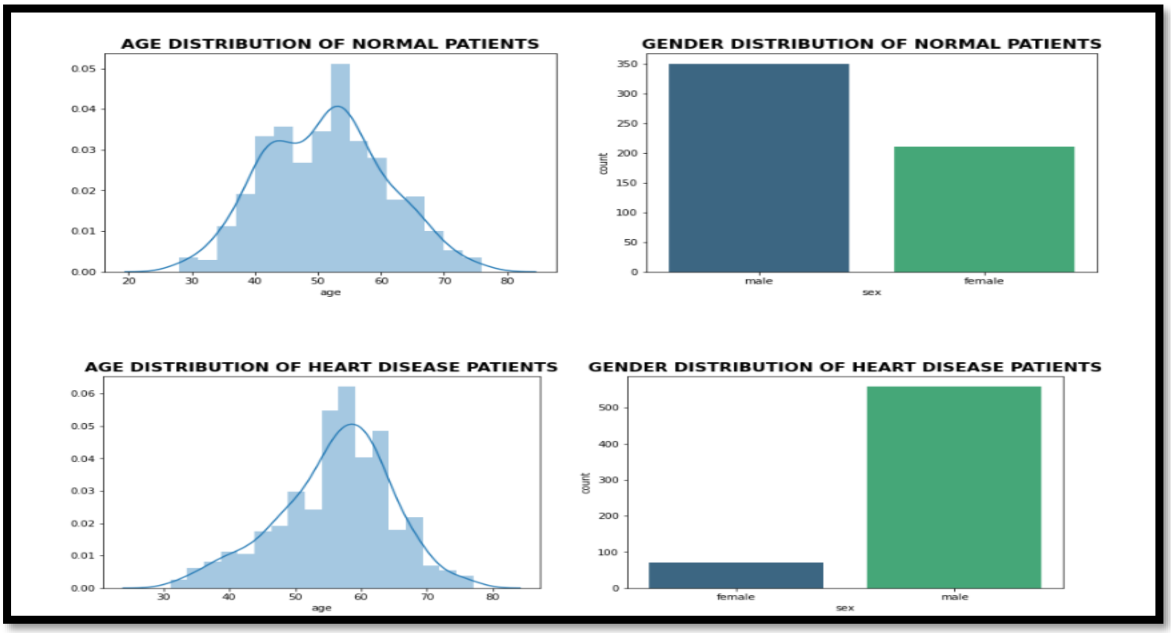
6)Distribution of heart disease



a)Plotting attribution of employees

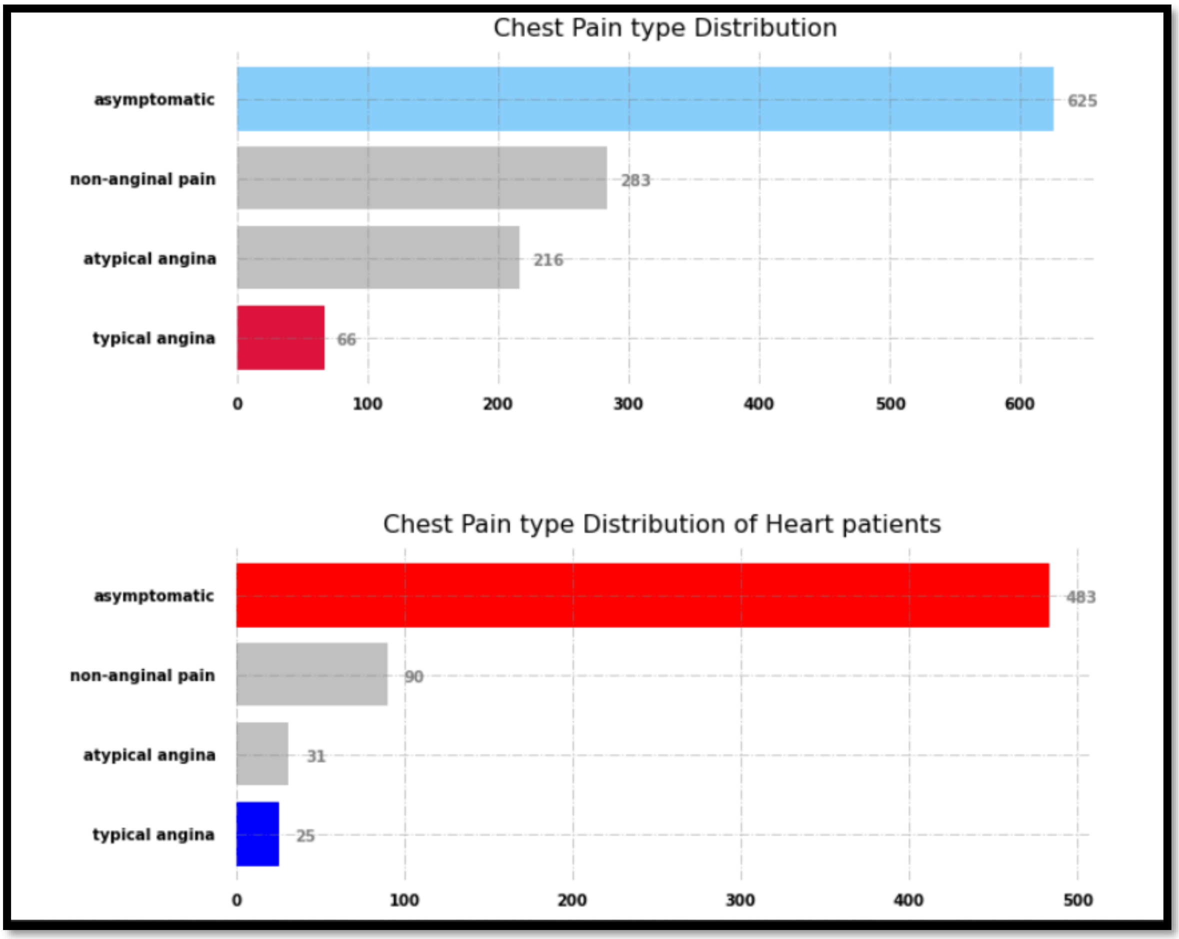
7)Checking gender & Age wise Distribution





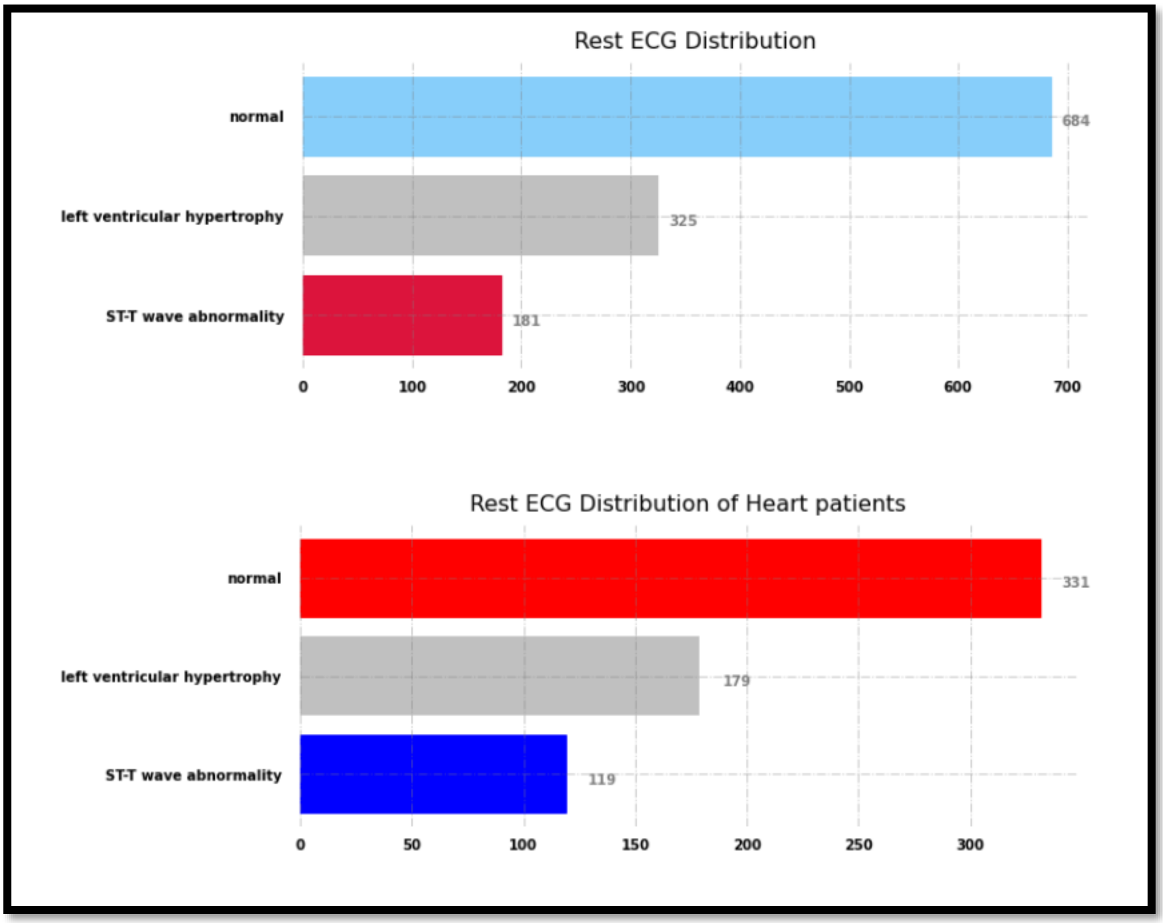
8)Distribution of chest type

a)Exploring the heart disease patients based on chest pain type

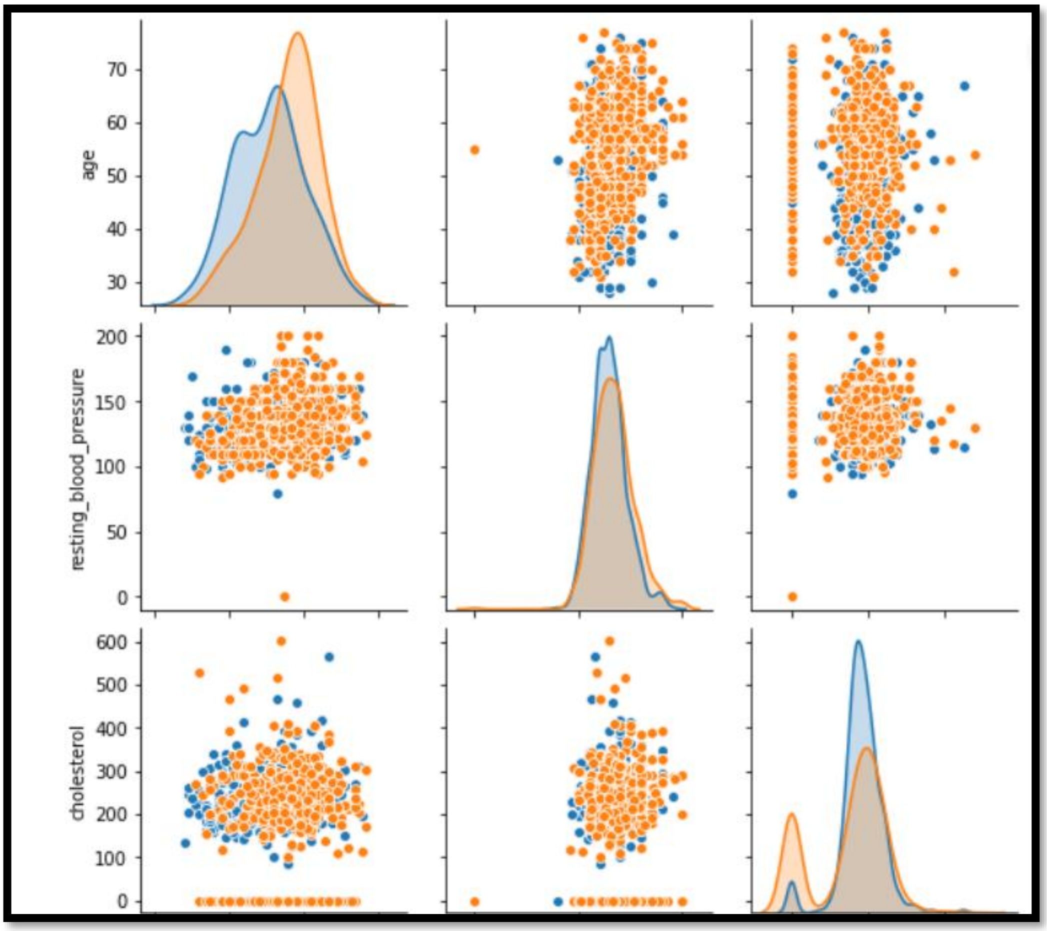


9)Distribution of rest ECG

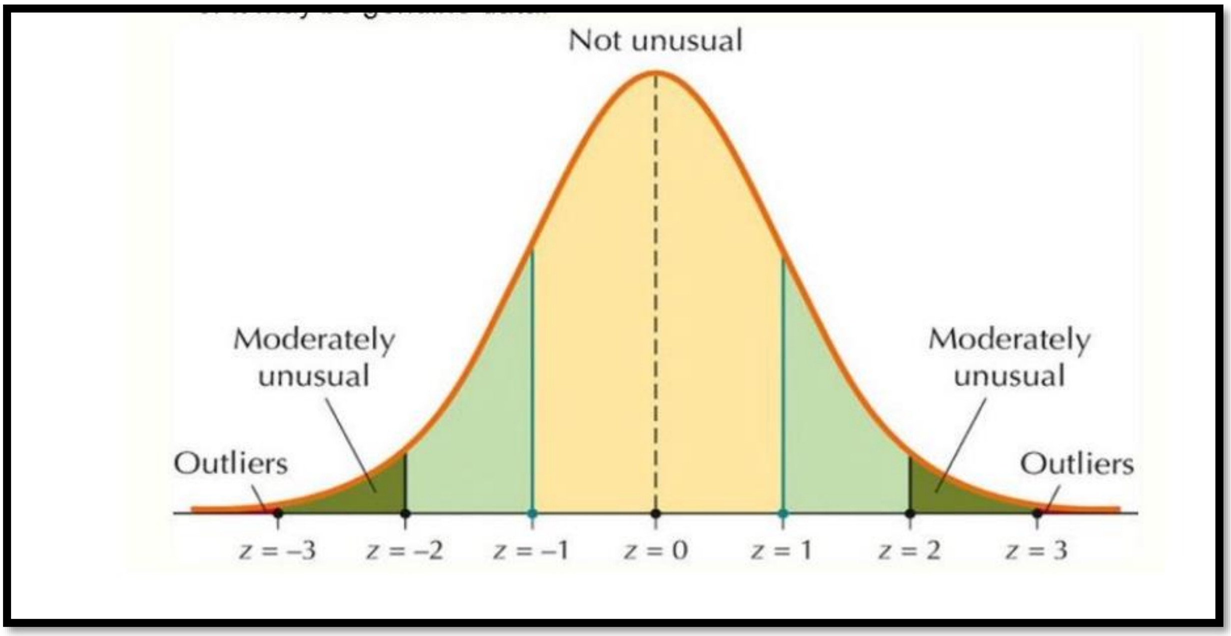
a)Exploring the heart disease patients based on REST ECG

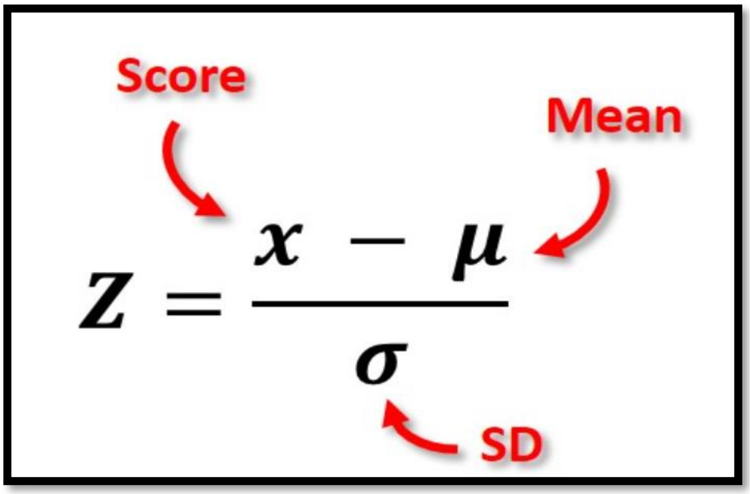


10)Distribution of numerical features



11)Outlier Detection & Removal





$$Z = \frac{x - \mu}{\sigma}$$

a) filtering numeric features as age , resting bp, cholestrol and max heart rate achieved has outliers as per EDA
b) calculating zscore of numeric columns in the dataset
c) Defining threshold for filtering outliers
d) filtering outliers retaining only those data points which are below threshold
e) checking shape of dataset after outlier removal
f) encoding categorical variables
g) segregating dataset into features i.e., X and target variables i.e., ythe shape of dataset

12)Train Test split
13)Cross Validation
14)Model Building

a) Multi Layer Perceptron
b) K nearest neighbour
c) Extra Tree Classifier
d) XGBoost
e) Support Vector Classifier
f) Stochastic Gradient Descent
g) Adaboost Classifier
h) decision Tree Classifier
i) gradient boosting machine

10)Model Selection
11)Stacked Ensemble
12)Model Interpretation

Exploratory Data Analysis (EDA) is **an approach to analyze the data using visual techniques**. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

Heart-Disease-Prediction-using-Machine-Learning:

Thus preventing Heart diseases has become more than necessary. Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives. This is where Machine Learning comes into play. Machine Learning helps in predicting the heart diseases, and the predictions made are quite accurate.

The project involved analysis of the heart disease patient dataset with proper data processing. Then, different models were trained and predictions are made with different algorithms KNN, Decision Tree, Random Forest, SVM, Logistic Regression etc This is the jupyter notebook code and dataset I've used for my Kaggle kernel 'Binary Classification with Sklearn and Keras'

I've used a variety of Machine Learning algorithms, implemented in Python, to predict the presence of heart disease in a patient. This is a classification problem, with input features as a variety of parameters, and the target variable as a binary variable, predicting whether heart disease is present or not.

Machine Learning algorithms used:

1. Logistic Regression (Scikit-learn)
2. Naive Bayes (Scikit-learn)
3. Support Vector Machine (Linear) (Scikit-learn)
4. K-Nearest Neighbours (Scikit-learn)

5. Decision Tree (Scikit-learn)
6. Random Forest (Scikit-learn)
7. XGBoost (Scikit-learn)
8. Artificial Neural Network with 1 Hidden layer (Keras)

Accuracy achieved: 95% (Random Forest)

GITHUB LINKS:

Jypyter Note	https://github.com/pondeepak25/dm/blob/main/bits-dm-hear-assignment-group-294.ipynb
Python Code	https://github.com/pondeepak25/dm/blob/main/bits-dm-hear-assignment-group-294.py
Dataset used	https://github.com/pondeepak25/dm/blob/main/heart.csv

Evaluation Matrix Checklist Verified:

S No	Criteria	
1	Data Understanding and Preparation along with EDA	• Data quality issues are identified and addressed • Appropriate data pre-processing measures are applied wherever applicable • Any notable exceptions are reported in form of comments, wherever appropriate • Attempt in right direction to find out contributing factors • Right set of visuals are used for univariate and bivariate data analysis

		• Meaningful insights are derived and presented in effective manner
2	Model building and evaluation	• Right data mining task is identified • Train and test data derived and used properly • Appropriate data mining technique is used for the model building • Model parameters are fine tuned to improve the model accuracy • Appropriate technique is used to identify the factors contributing to the model accuracy • Model evaluation is done based on the appropriate measures and criteria
3	Effective Story telling through the report/presentation	• The presentation has proper structure, not too big, not too small. Elaborate the important points in more precise manner. • Has focus on the problem to be solved • Talks about the factors contributing to the issue along with right kind of proofs • Explaining the observations visually where visuals are showcasing the facts • The recommendations / suggestions are spelt out clearly. • Assumptions are specified at right places.
4	Code readability and organization	• Code is executing, no syntax errors

		• No customizations is needed to execute the code • Code is simple and augmented with proper comments wherever required • Built-in functions / libraries are used wherever possible • Repeated code is moved into functions and used appropriately when required • Long code snippets are broken down into small parts and made available as functions to increase the modularity of the code • Appropriate variable names are used to improve the readability of the code
5	Overall utilization of the concepts learnt in the course	• Appropriate steps are carried out in the data preparation stage which involves the concepts learnt in the class for the same • Various concepts from the data mining process are paid enough attention while developing the model