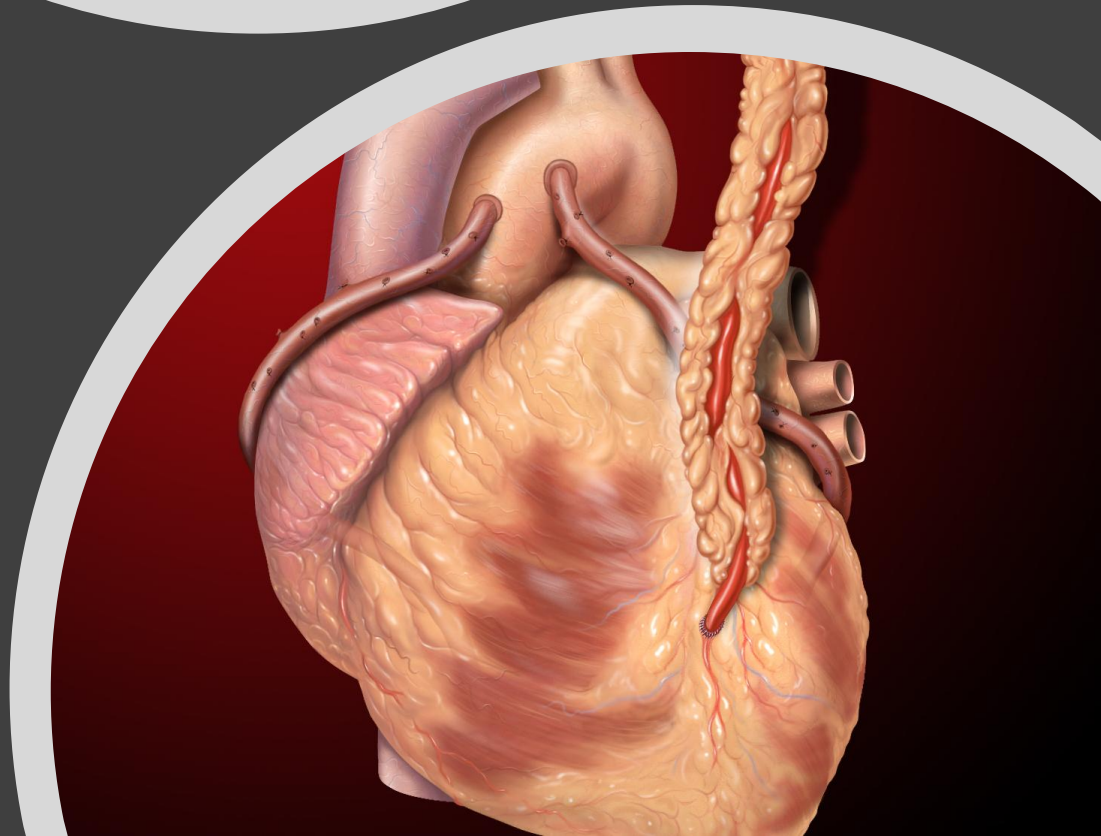
Heart Deaseas Predictation

TEAM MEMBER

- SNEHSIS ROY
- SUBHODIP ROY
- MOUMITA MAJI
- DRIPPTA DUTTA
- ARPITA MAJI



PROJECT MENTOR :

PROF. ARNAB CHAKRABORTY

Content

- Project Objective & Scope
- Data Description
- Methodology
- Data Preprocessing
- Models Used
- Accuracy Comparison
- Inference
- Future Scope of Improvements

PROJECT OBJECTIVE & SCOPE

Objective:

- **Given :** Framingham Heart disease dataset taken from Github (contains training and test data).
- **Goal :** To predict whether a patient will have Coronay Heart Disease in recent 10 years.
- **Finally :** Apply on the test dataset and compare the differences in the results

Scope:

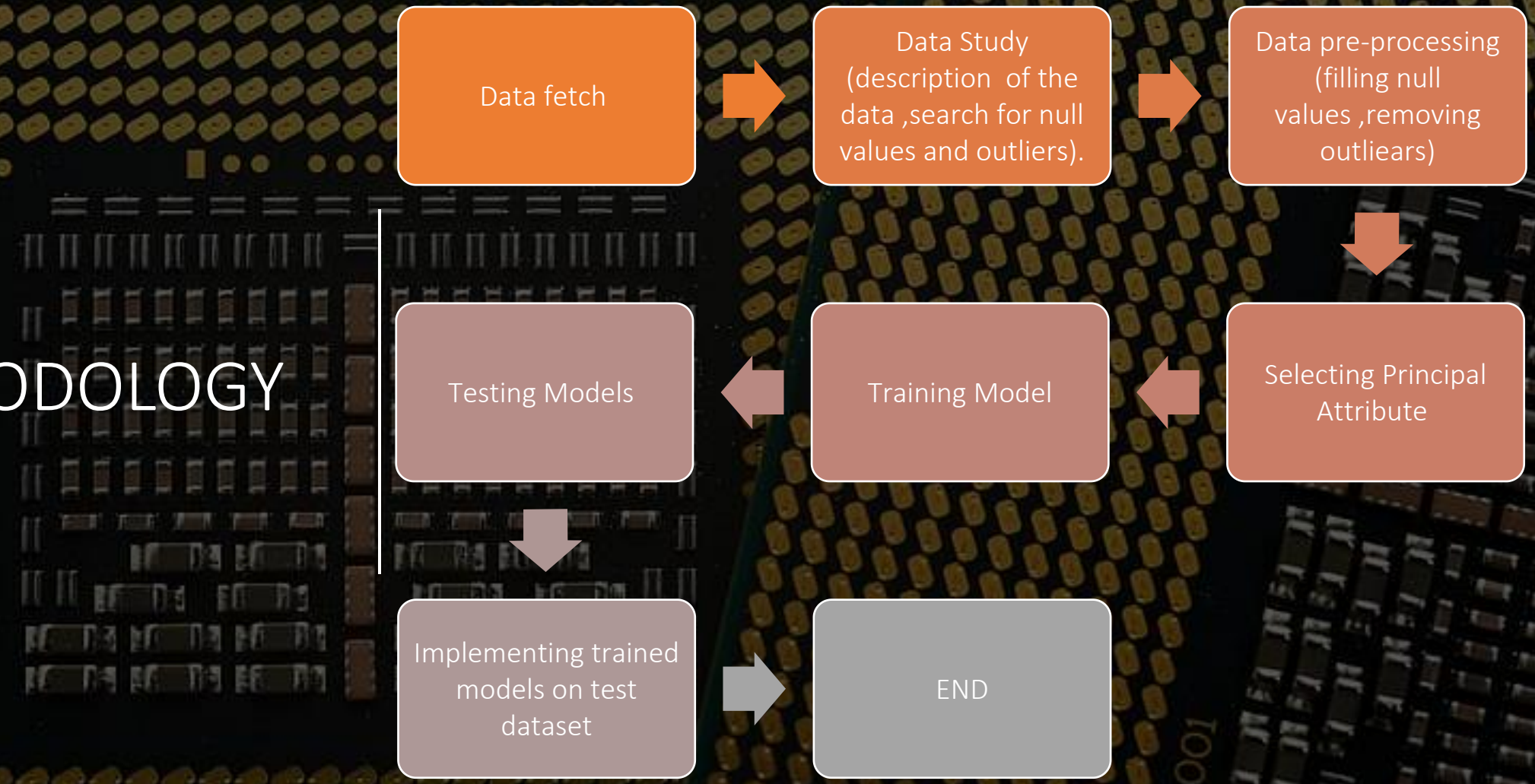
- It is a useful project as the Classifier models can be used to quickly determine wether the patient have chance to have coronary heart disease in recent 10 years.
- The results might have some mismatch with the real-world applications. But that can be avoided if the models are trained for small datasets.

DATA DESCRIPTION

The description of the data with types and description of each of the attribute is given shown in this table

Columns	Attribute name	Type	Description	Target Attribute
Male	male	Categorical	Gender of the patient (0/1)	No
Age	age	Non-Categorical	Age of the patient	No
Education	Education	Categorical	Education status of the patient (1/2/3/4)	
Current Smoker	currentSmoker	Non-Categorical	The number of cigarettes that the person smoked on average in one day	No
BP Meds	BPMeds	Categorical	Whether or not the patient as on blood pressure medication.(0/1)	No
Prevalent Stroke	prevalentStroke	Categorical	Whether or not the patient had previously had a stroke.(0/1)	No
Prevalent Hypertensive	prevalentHyp	Categorical	Whether or not the patient was hypertensive.(0/1)	No
Diabetes	diabetes	Categorical	Whether or not the patient had diabetes.(0/1)	No
Total Cholestrol	totChol	Non-Categorical	Total cholesterol level	No
Systolic Blood Pressure	sysBP	Non-Categorical	Systolic blood pressure	No
Diastolic Blood Pressure	diaBP	Non-Categorical	Diastolic blood pressure	No
Body Mass Index	BMI	Non-Categorical	Body Mass Index	No
Heart Rate	heartRate	Non-Categorical	Heart Rate	No
Glucose	glucose	Non-Categorical	Glucose level	No
TenYearCHD	TenYearCHD	Non-Categorical	10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)	Yes

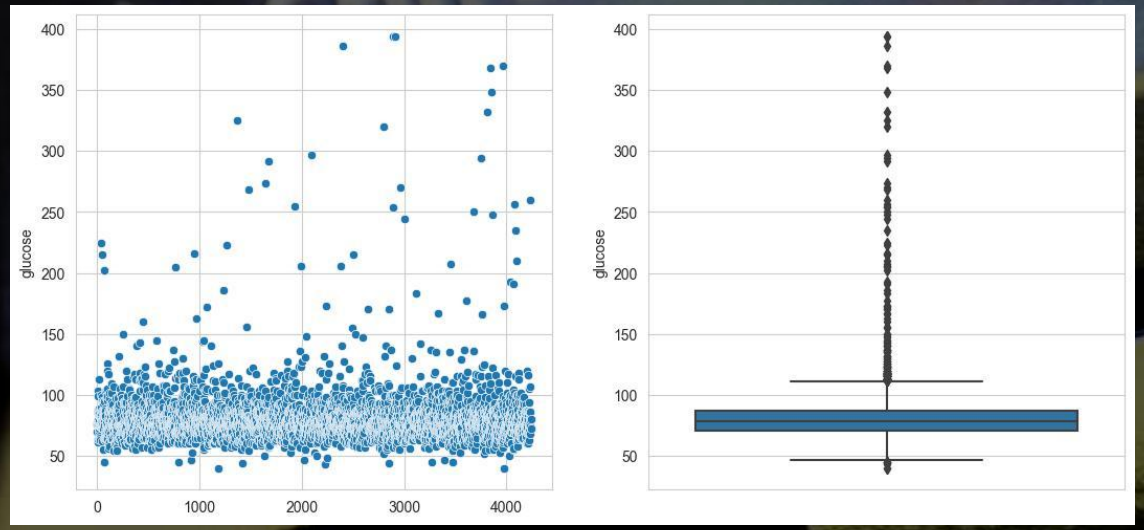
METHODOLOGY



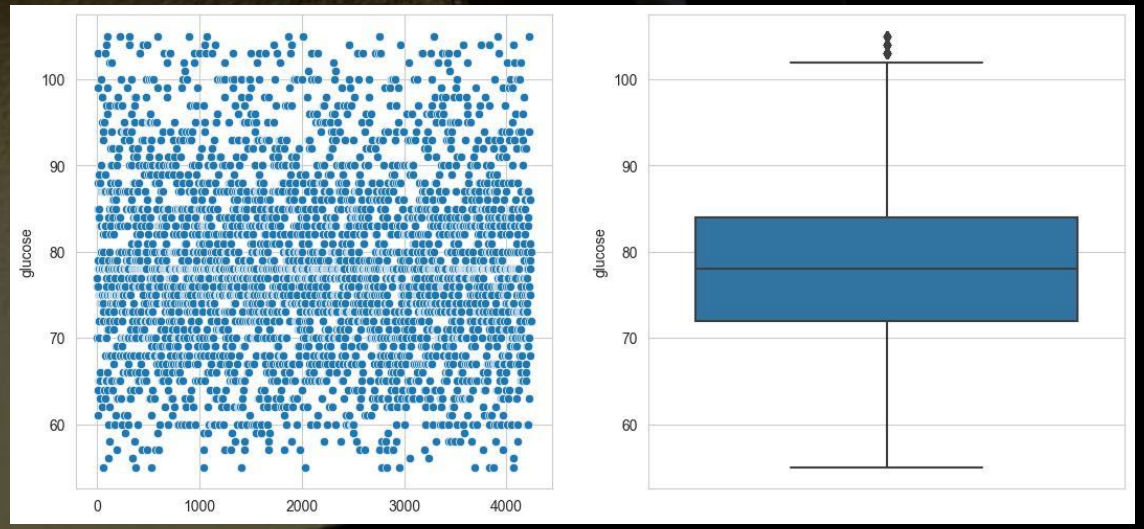
DATA PROCESSING

Removing outliers from Glucose column

Before



After



MODEL USED

- Logistic Regression
- Random Forest
- K – NN Classification
- Decision Tree

LOGISTIC REGRESSION

- **Logistic Regression** is a type of regression analysis.Regression analysis is a type of predictive modelling technique which is used to find the relationship between a dependent variable (usually known as the “Y” variable) and either one independent variable (the “X” variable) or a series of independent variables. When two or more independent variables are used to predict or explain the outcome of the dependent variable, this is known as multiple regression.Logistic Regression models the data using sigmoid function.

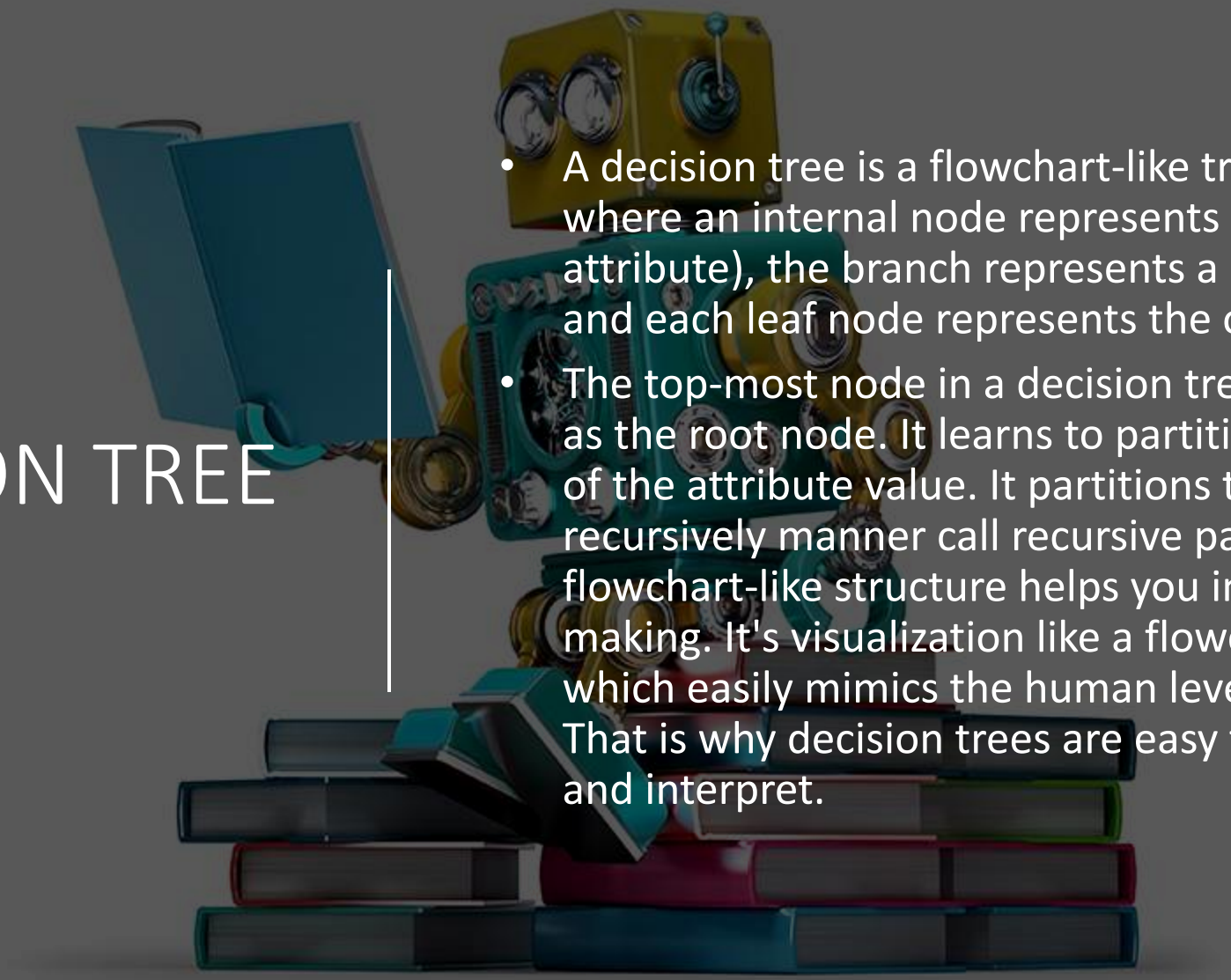
RANDOM FOREST

- 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 Machine Learning. 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.

K-NN CLASSIFIER

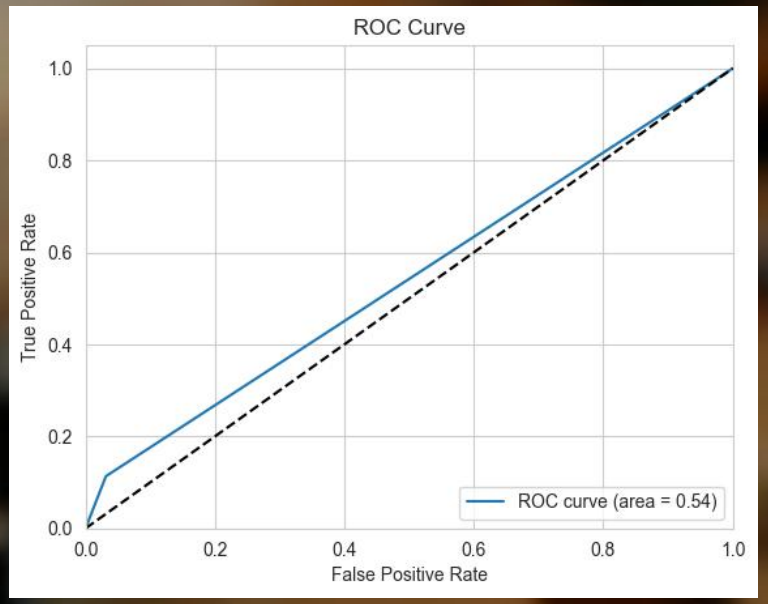
- K-nearest neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closet to the test data. The KNN algorithm calculates the probability of the test data belonging to the classes of ‘K’ training data and class holds the highest probability will be selected. In the case of regression, the value is the mean of the ‘K’ selected training points.

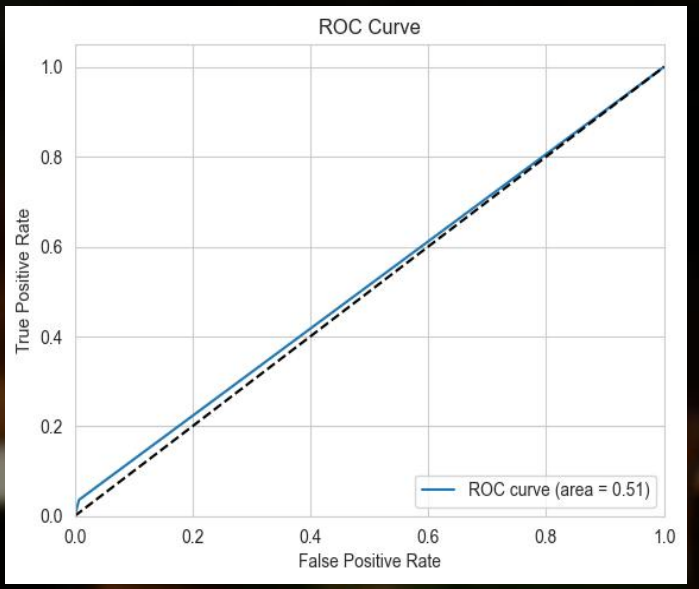
DECISION TREE

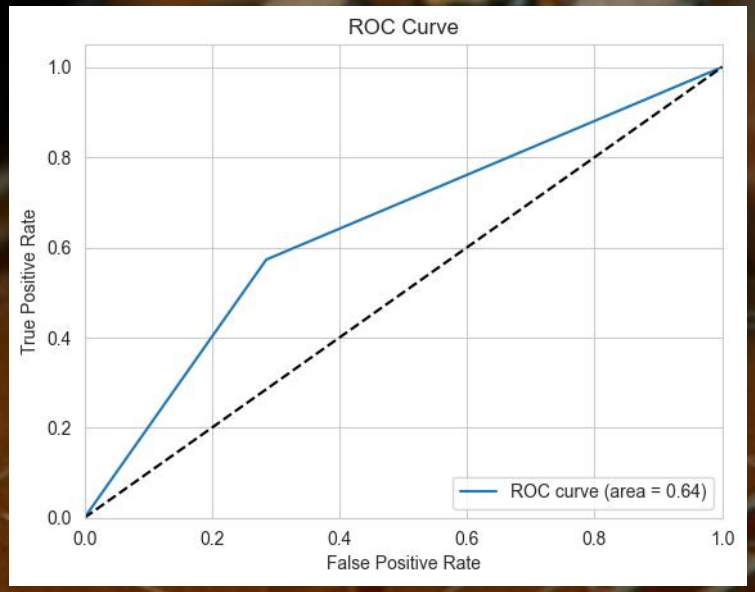


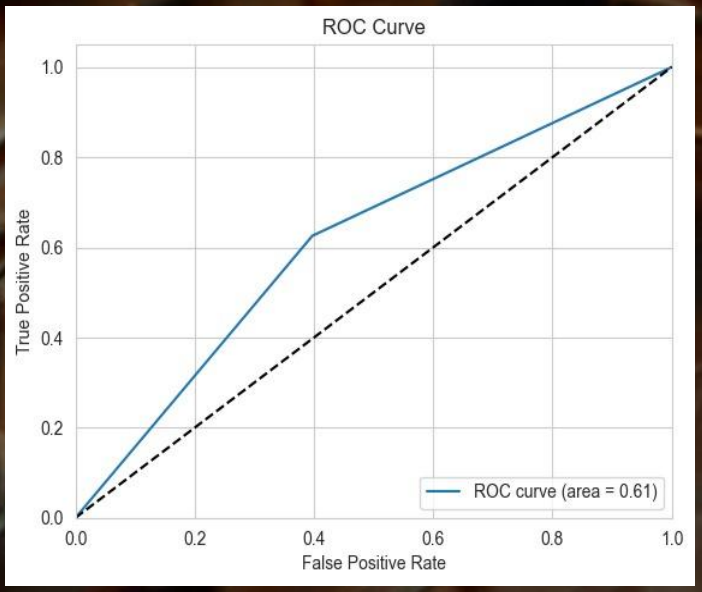
- A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.
- The top-most node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

RECEIVER OPERATING CHARACTERISTIC CURVES

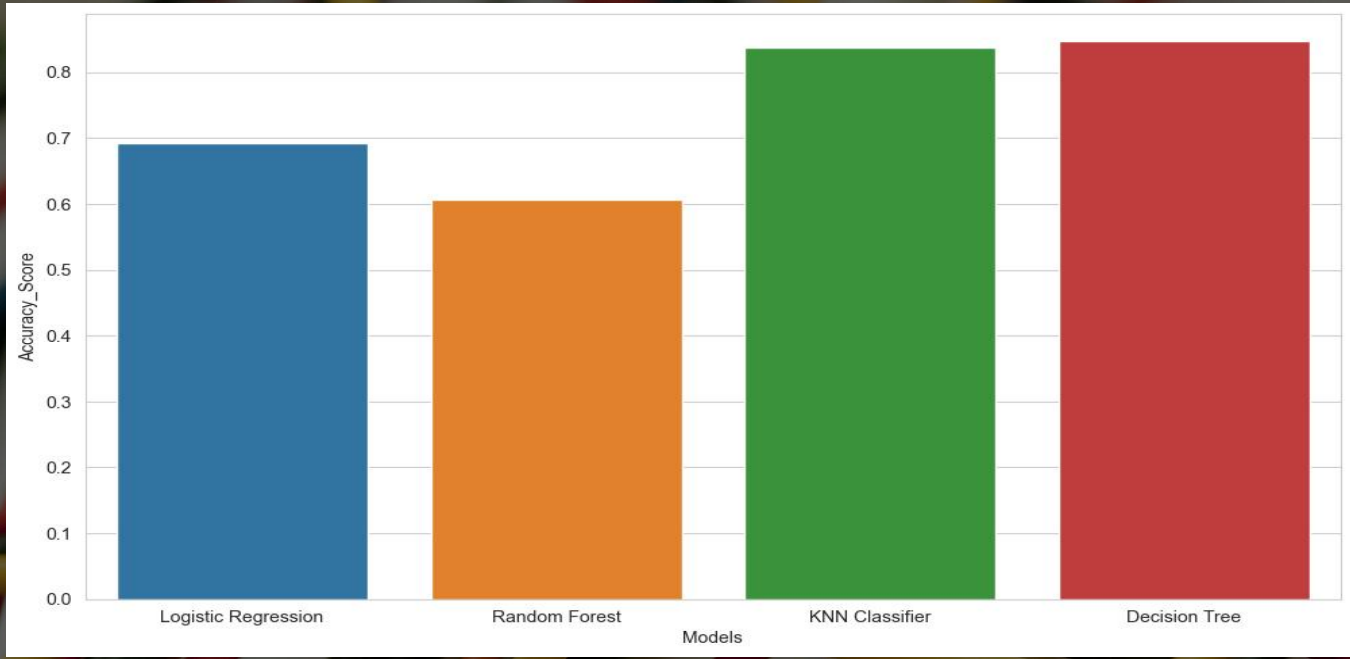








ACCURACY COMPARISON GRAPH



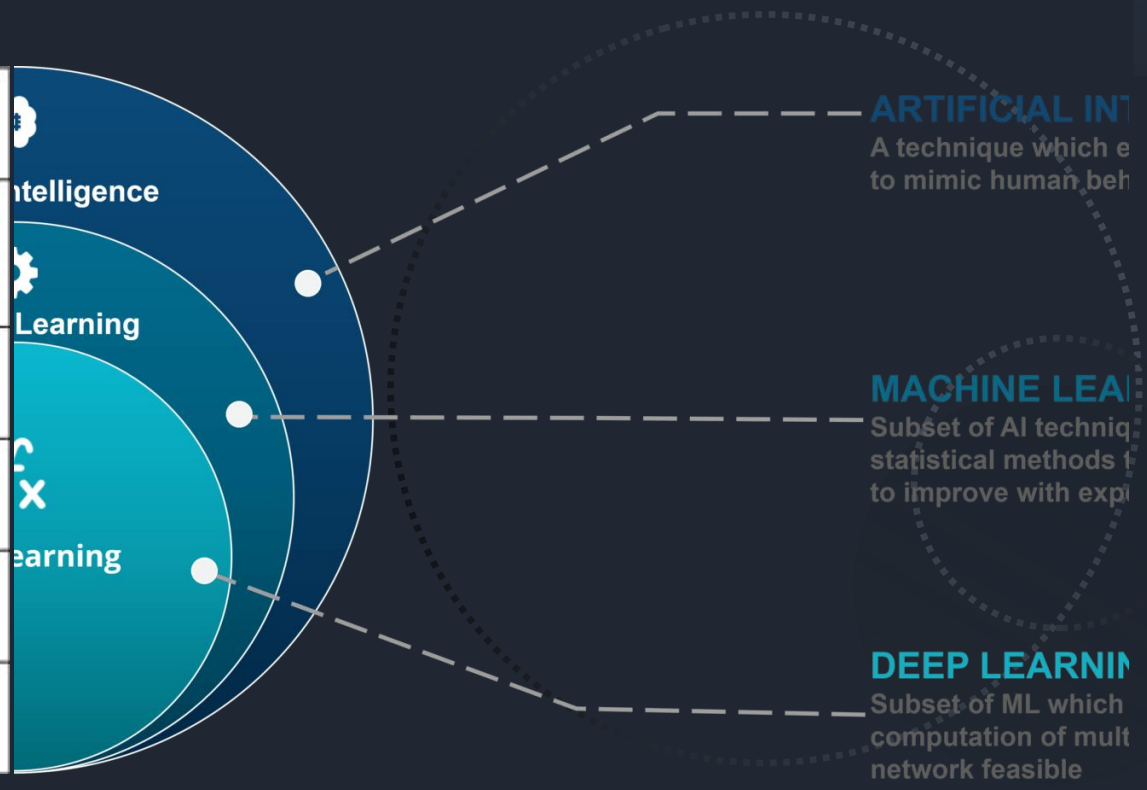
Models Used	Accuracy %
Logistic Regression	69.5755
Random forest	64.7799
K – Nearest Neighbor	81.6824
Decision Tree	82.9403

- The data shown above is an accuracy score of our models.
- We see that the highest accuracy for the train dataset is in Decision Tree.
- But we select Logistic Regression Model to predict the test dataset , because it is giving more balanced result.

INFERENCE

- The output result of the test dataset obtained by model namemodel is inferred to be accepted.

Classifier Model	Correct Outcomes	
	0 (No)	1(Yes)
Logistic Regression	785	112
Random Forest	706	118
K-Nearest Neighbour	1021	18
Decision Tree	1061	12



FUTURE SCOPE OF IMPROVEMENTS

- User who are intented to check their chances of heart disease in recent 10 years can use our trained model to check wether they will have a coronary heart disease or not. This trained models can be implemented on graphical user interface to make it easy to use.

- Various hospital can use this model and modify them according to there needs.To predict the heart health of the patient

THANK YOU