Snehal Jain - BAP 220727

Heart Attack Prediction

Heart Failure Prediction: Supervised Machine Learning Problem

Data Mining II

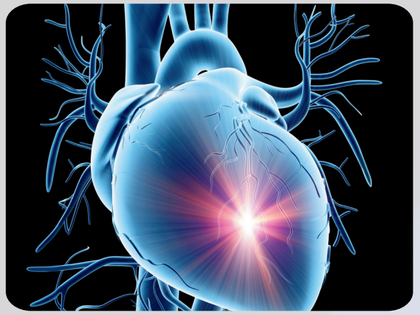
HEART ATTACK PREDICTION

Predictive Analysis for Cardiovascular Disease

PROBLEM STATEMENT

Statement

Develop a machine learning model capable of early detection and management of cardiovascular diseases in individuals with risk factors such as hypertension, diabetes, hyperlipidemia, or established disease.



DATA SET USED

A Heart attack prediction dataset containing **13 features** related to cardiovascular health, such as blood pressure, cholesterol levels, and other relevant medical indicators.

DATA MINING ALGORITHMS

Utilizing **decision tree, AdaBoost, XGBoost, and random forest** algorithms to predict the likelihood of cardiovascular disease based on the provided dataset.

SOLUTION

Provide a predictive model that can assist in early detection and management of CVD's, thus potentially reducing the number of premature deaths caused by heart attacks and strokes.

HEART ATTACK PREDICTION

Data Dictionary: Attributes in the Dataset

DATA DICTIONARY

Age
- Age of the patient

Sex
- Sex of the patient
  - Value 0: Male
  - Value 2: Female

Trtbps
- Resting blood pressure (in mm Hg)

Fbs
- (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

Thalachh
- maximum heart rate achieved

Old Peak
- Previous peak

Slp
- Slope

Caa
- Number of major vessels (0-3)

Thall
- Thalium Stress Test result ~ (0,3)

Exng
- Exercise induced angina (1 = yes; 0 = no)

Cp
- Chest Pain type chest pain type
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic

Restecg
- Resting electrocardiographic results
  - Value 0: normal
  - Value 1: having ST-T wave abnormality
  - Value 2: Ventricular hypertrophy by Estes' criteria

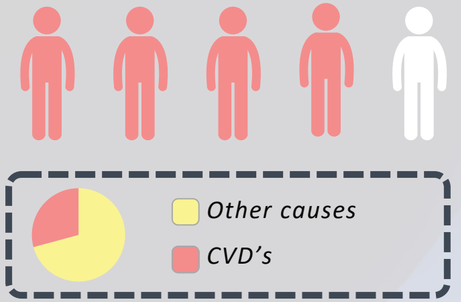
Output
- 0 = less chance of heart attack
- 1 = more chance of heart attack

Chol
- Cholesterol in mg/dl fetched via BMI sensor

DOMAIN KNOWLEDGE

Background

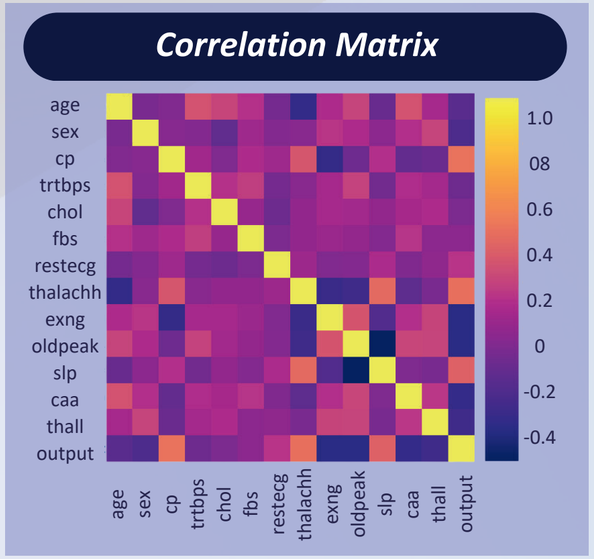
Cardiovascular diseases (CVDs) claim 17.9 million lives annually, representing 29% of global deaths, with 4 out of 5CVD deaths due to heart attacks and strokes, and one-third occurring prematurely in individuals under 70.



Target audience: People with cardiovascular disease or who are at high cardiovascular risk

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1

A glimpse at the dataset

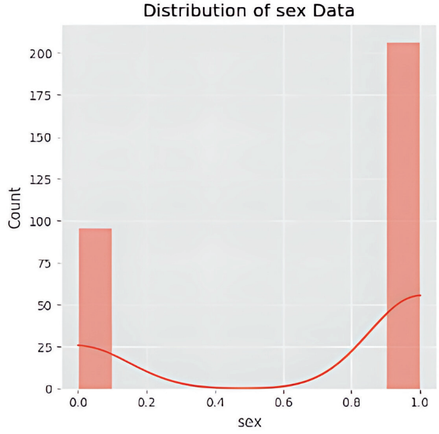


HEART ATTACK PREDICTION

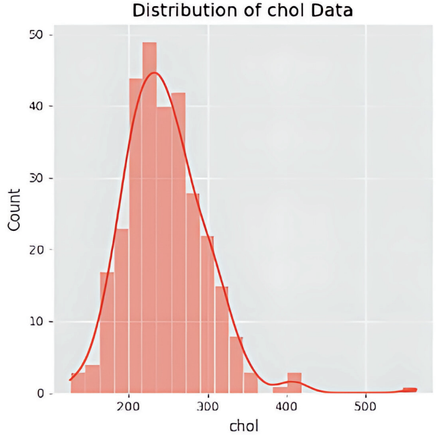
Kernel Density Estimate: Check the linearity of the variables

DISTRIBUTION GRAPH

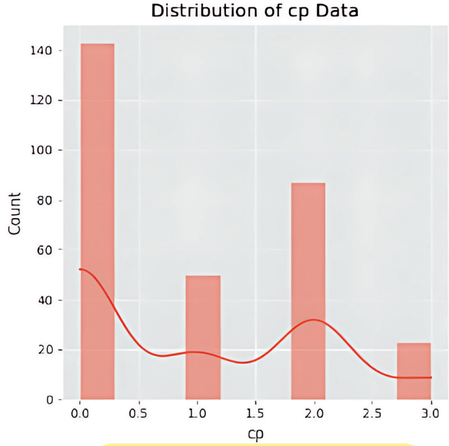
Age group: 50-65



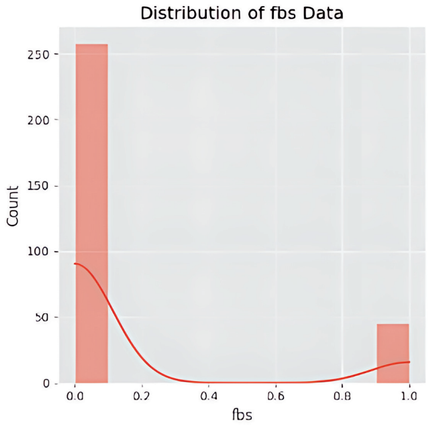
Males



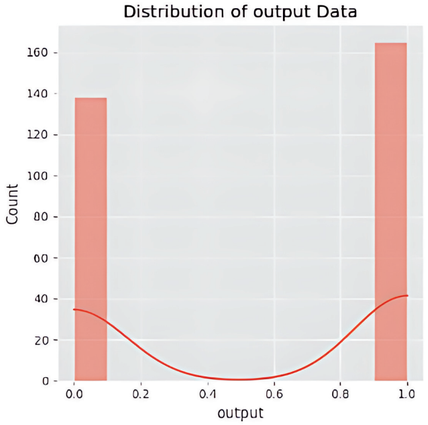
208-286mg/dL

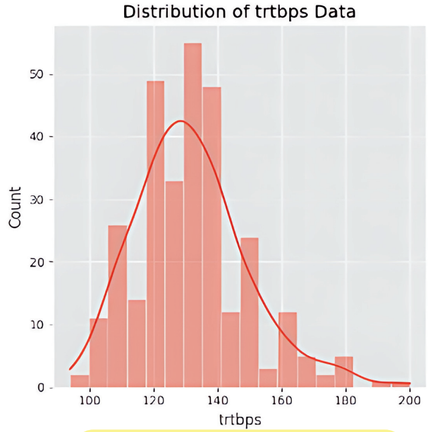


Typical angina



Greater than 120



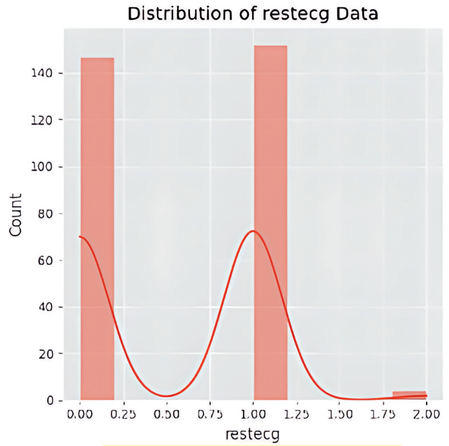


120-158mm HG

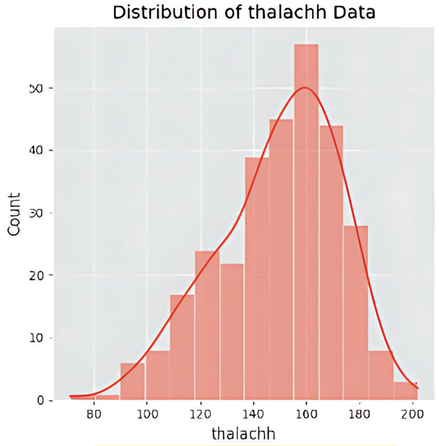
HEART ATTACK PREDICTION

Kernel Density Estimate: Check the linearity of the variables

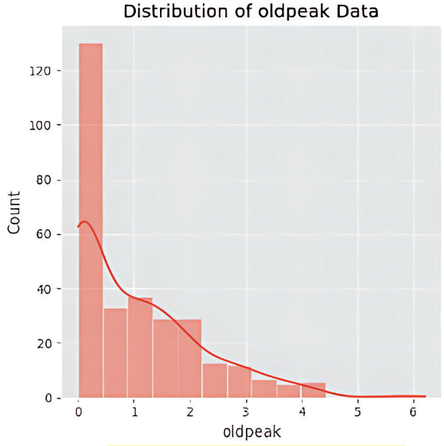
DISTRIBUTION GRAPH



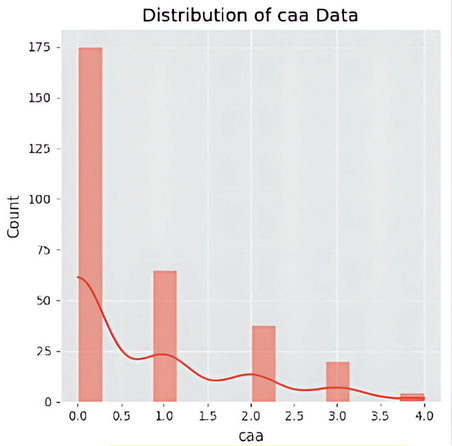
Having abnormality



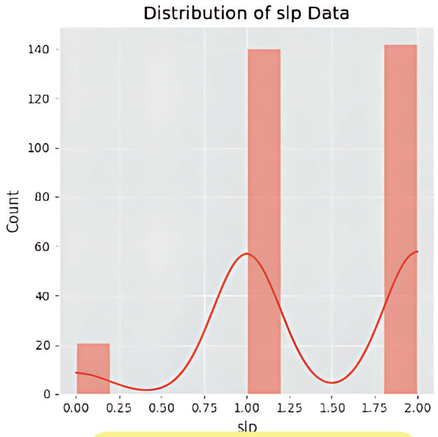
122-145mm Hg



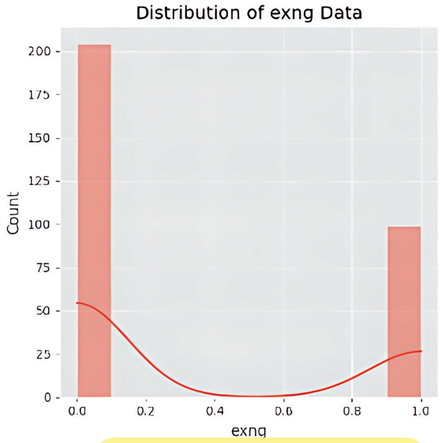
0-1.25



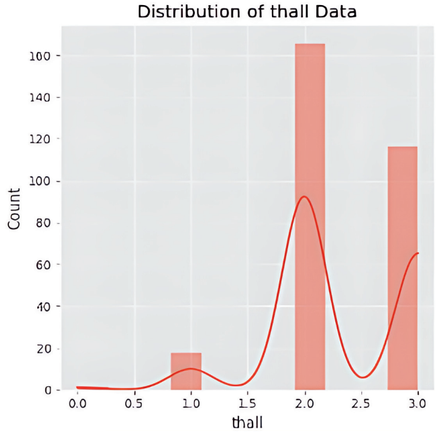
Negligible



Flat and downsloping



50% don't exercise

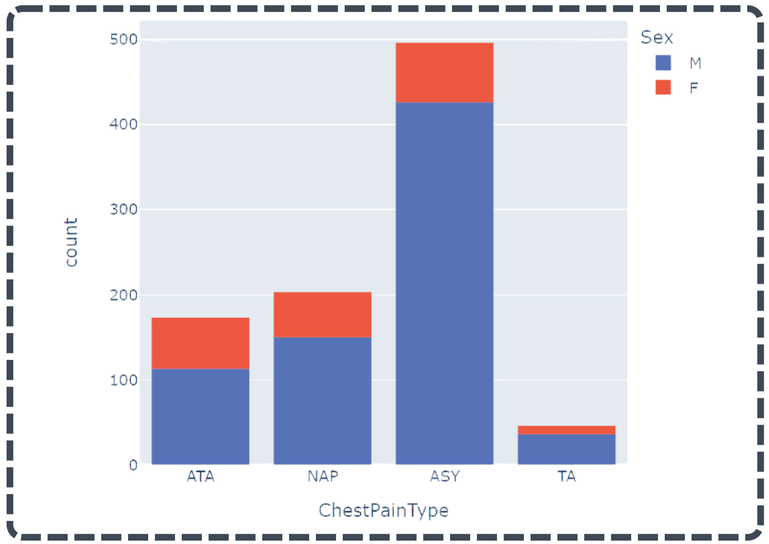


2 Thallium stress

HEART ATTACK PREDICTION

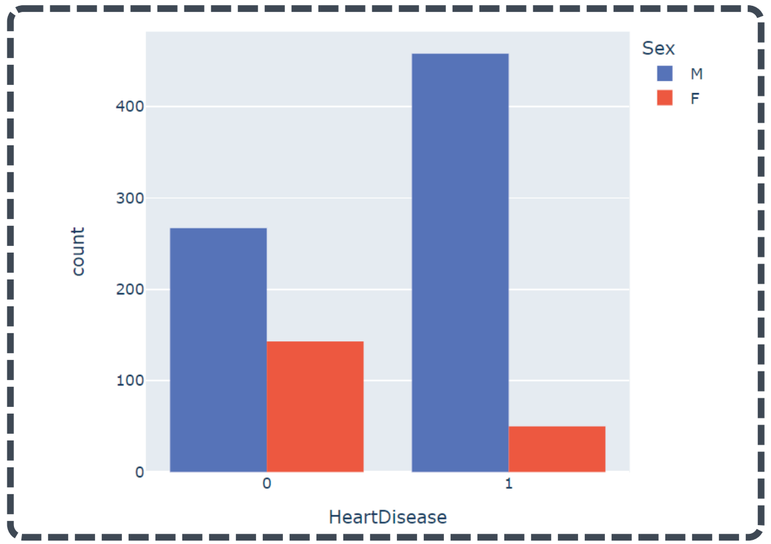
Data Distribution of Chest Pain Type and Heart Diseases in Males and Females

CHEST PAIN TYPE



In the distribution of types of chest pain, males outnumber females in all four categories

HEART DISEASES



In the distribution of heart disease , males outnumber females meaning they are more prone to such diseases

HEART ATTACK PREDICTION

Data preprocessing: Null values and feature scaling

DATA PREPROCESSING

NULL VALUES

age	0
sex	0
cp	0
trtbps	0
chol	0
fbs	0
restecg	0
thalachh	0
exng	0
oldpeak	0
slp	0
caa	0
thall	0
output	0

Null Values

Removing Nan Values

In any real-world dataset, there are always few null values. It doesn't really matter whether it is a regression, classification or any other kind of problem, no model can handle these NULL or NaN values on its own so we need to intervene.

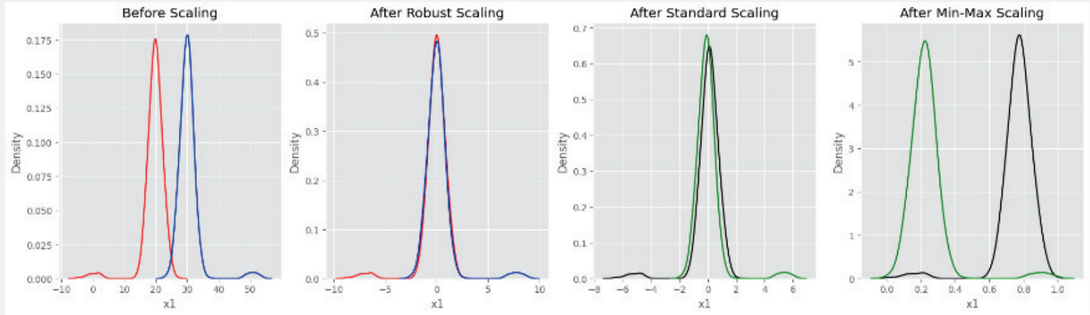
Feature Scaling

Standardization & Normalization

Tree-based algorithms are fairly insensitive to the scale of the features. A decision tree is only splitting a node based on a single feature. The decision tree splits a node on a feature that increases the homogeneity of the node. This split on a feature is not influenced by other features.

Graphical Representaion

Robust Scaler & Standard Scaler

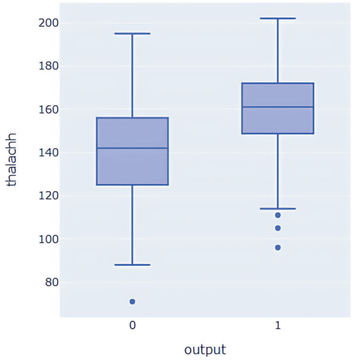


HEART ATTACK PREDICTION

Understanding Cardiovascular Disease Indicators

CVD INDICATORS: REMOVING OUTLIERS

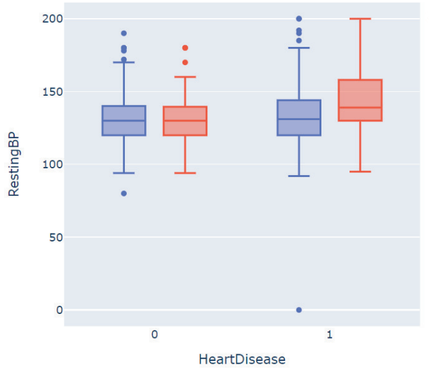
01



Max Heart Rate

Range: 122 - 145mm Hg
The above maximum heart rate achieved is prone to a heart attack

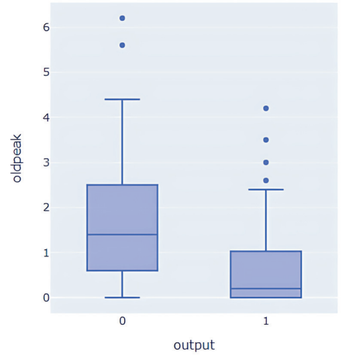
02



Resting BP

Males: 120 - 144mm Hg
Females: 130 - 158mm Hg
The above BP rate is vulnerable to a heart attack

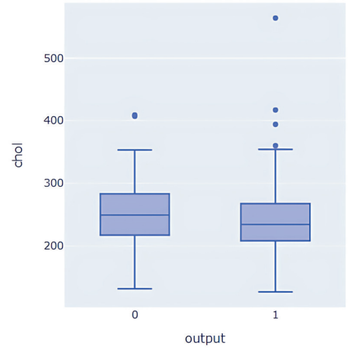
03



Old Peak

Range: 0-1.5
A higher Old Peak value might point towards a greater risk of heart complications, underscoring its clinical significance

04



Cholestrol

Range: 208-265mg/dl
Extremely high cholesterol are levels at fair risk or at high risk for cardiovascular disease

HEART ATTACK PREDICTION

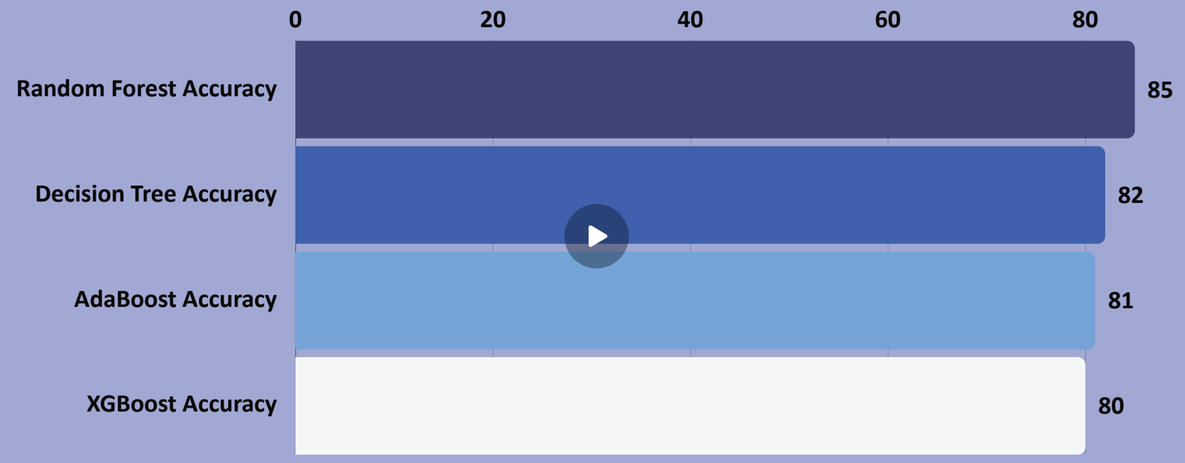
Test Accuracy: Random Forest, Adaboost, XGboost, Decision Tree

MACHINE LEARNING ALGORITHMS

Definition

To evaluate the performance of 4 machine learning models - Decision Tree, Random Forest, AdaBoost and XGBoost - in predicting heart disease to understand how well each model can generalize its predictions on unseen data, ensuring it's not just memorizing the training set (overfitting).

GRAPHICAL COMPARISION OF ACCURACIES



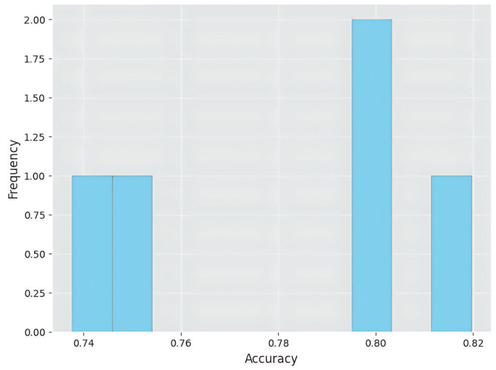
This metric is more indicative of a model's real-world potential. It shows how effectively a model can generalize its learnt "rules" to predict outcomes on new, unseen data.

RANDOM FOREST ACCURACY SCORE - 85%

HEART ATTACK PREDICTION

Fold-Wise Accuracy: Random Forest, Adaboost, XGboost, Decision Tree

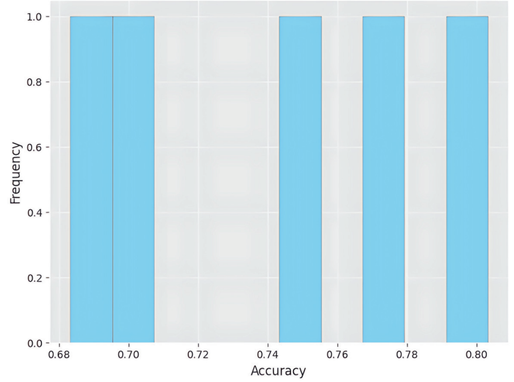
XGBOOST FOLD-WISE ACCURACY



Insight

Its performance is commendable, coming close to Random Forest, but highly overfiitted for our dataset, a red flag.

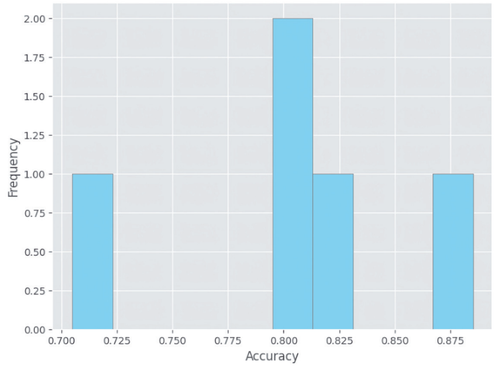
DECISION TREE FOLD-WISE ACCURACY



Insight

They can be prone to overfitting, especially if not pruned properly. The difference between test and cross-validation accuracy suggests a modest overfit

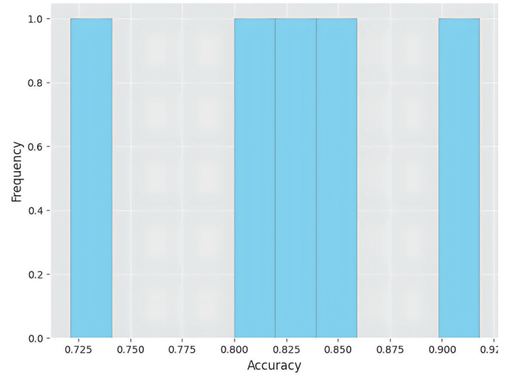
ADABOOST FOLD-WISE ACCURACY



Insight

Performance is fairly good in comparison to XGBoost however careful tuning is needed

RANDOM FOREST FOLD-WISE ACCCURACY

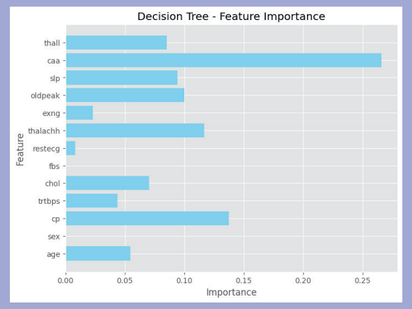


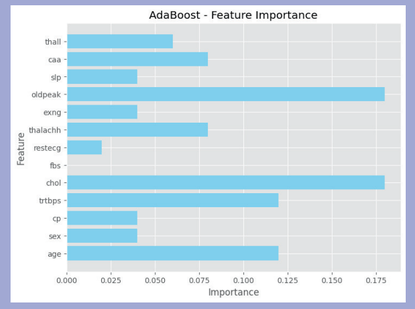
Insight

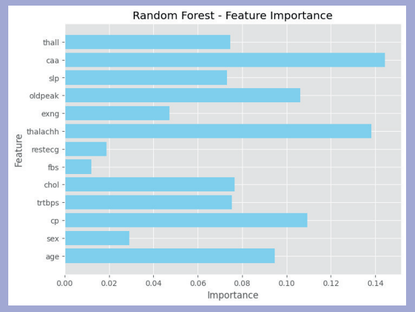
By virtue of its ensemble nature, it diversifies risk and often achieves higher accuracy

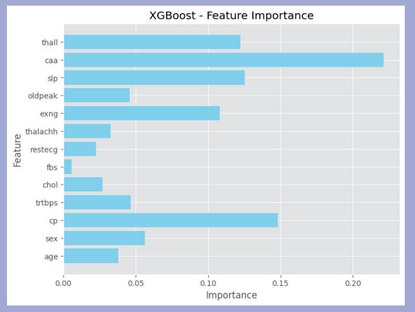
WHY RANDOM FOREST CLASSIFIER WORKS BEST HERE?

- *Random Forest operates by constructing **multiple decision trees** during training and outputs the majority class of the individual trees for classification. This **ensemble approach** inherently minimizes the risk of errors posed by individual trees.*
- *Random Forest is adept at handling **large datasets with higher dimensionality**. It can handle input variables without variable deletion, providing a comprehensive insight into which features matter most.*
- *By using multiple trees, Random Forests tend to **avoid overfitting** that single trees might succumb to.*
- *Random Forest can handle imbalanced datasets by balancing error in the class population through "class_weight" parameters or by creating a **balanced bootstrap sample** for each tree.*









LINK TO GOOGLE COLAB FILE

Thank you!

Details

Name: Snehal Jain

Course: B.A.P (CA + Maths)

Roll no. 220727