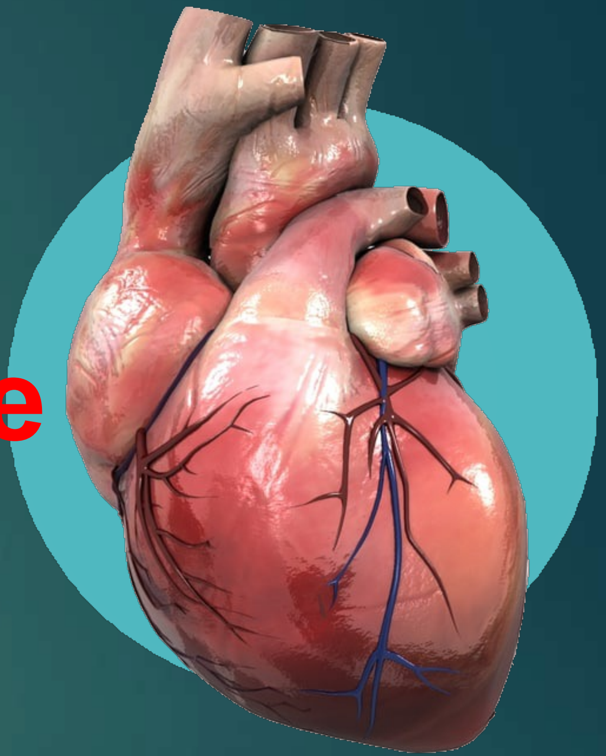
Classification of Heart Disease



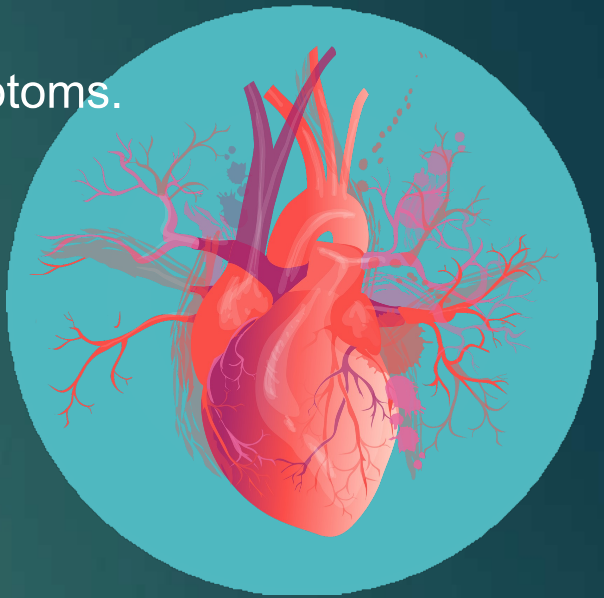
JOE ANSON AQUINO
May 10, 2023

Springboard

Springboard Data Science

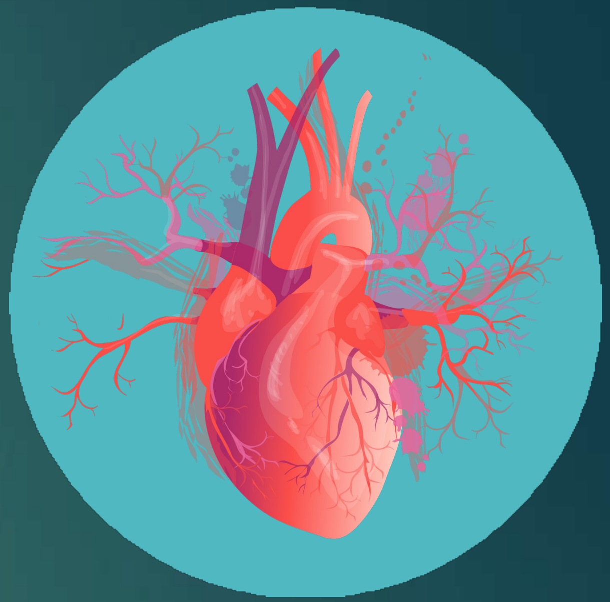
Introduction

- US No.1 silent killer that leads to the person's death without apparent symptoms.
- Early diagnosis of heart disease reduces high-risk patient complications.
- Assistance of Machine learning decisions and predictions using algorithms.



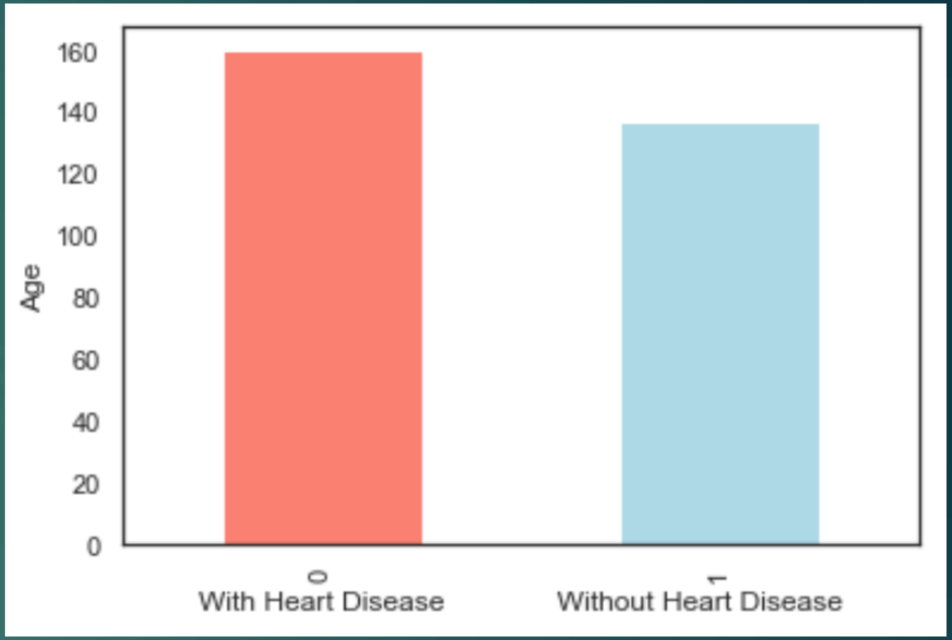
Approach

- Gather data from the UCI machine learning repository
- Organize it to ensure it is well-defined.
- Clean and explore the data before utilizing a machine-learning algorithm
- Predict heart disease using the Cleveland Dataset.
- Study and analyze using Logistic Regression Classifier, Random Forest Classifier, Light GBM Classifier, xgBoost Classifier, and Decision Tree.



Dataset

(age) age in years
(sex) sex (1 = male; 0 = female)
(cp) chest pain type –
Value 1: typical angina
Value 2: atypical angina
Value 3: non-anginal pain
Value 4: asymptomatic
(trestbps) resting blood pressure (in mm Hg on admission to the hospital)
(chol) serum cholesterol in mg/dl
(fbs) (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
(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
(thalach) maximum heart rate achieved
(exang) exercise induced angina (1 = yes; 0 = no)
(oldpeak) ST depression induced by exercise relative to rest
(slope) Value 1: upsloping
Value 2: flat
Value 3: downsloping
(ca) number of major vessels (0-3) colored by fluoroscopy
(thal) 3 = normal; 6 = fixed defect; 7 = reversable defect
(num) (the predicted attribute) diagnosis of heart disease
Value 0: < 50% diameter narrowing
Value 1: > 50% diameter narrowing (in any major vessel: attributes 59 through 68 are vessels)



Heart disease prediction

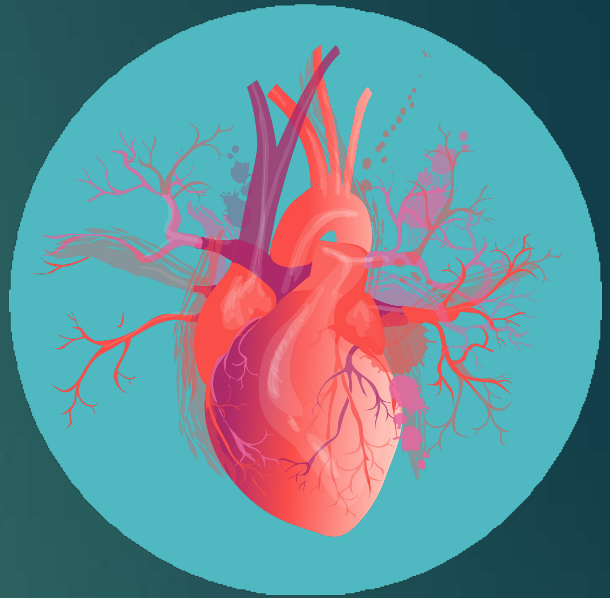
- Firstly, the user sends a feature input and the heart disease dataset which may contain a number of instances and characteristics.

- Following with the algorithm we have taken for classification

- After giving the complete inputs to the machine by using machine learning algorithms like decision tree, random forest regression etc…

- The machine performs wrangling the dataset from the algorithm

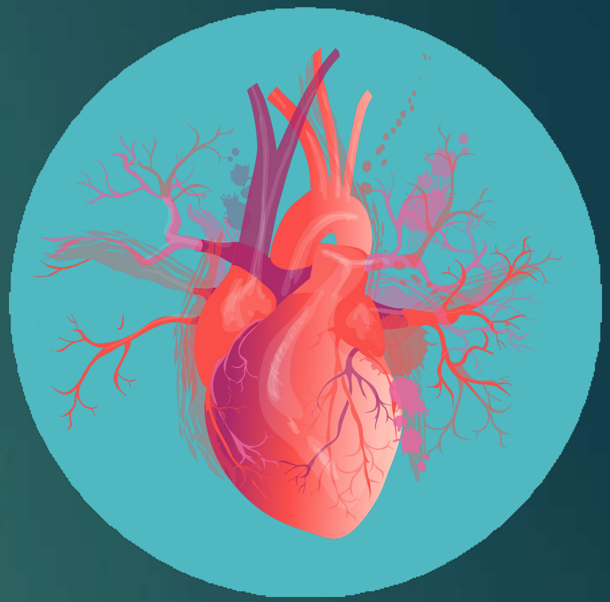
- Finally, it gives a predictable output to the user.



Machine learning algorithm models

- Logistic regression: One of the very popular algorithms is considered as logistic regression which is a supervised learning model. It performs categorical predictions which can be ‘true’ or ‘false’. This model provides probabilistic values instead of exact ones. This algorithm works on both continuous and discrete values. A simple S-Shaped curve can elaborate the logistic regression very precisely.

- Random forest is a flexible, easy-to-use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most-used algorithms, due to its simplicity and diversity (it can be used for both classification and regression tasks).



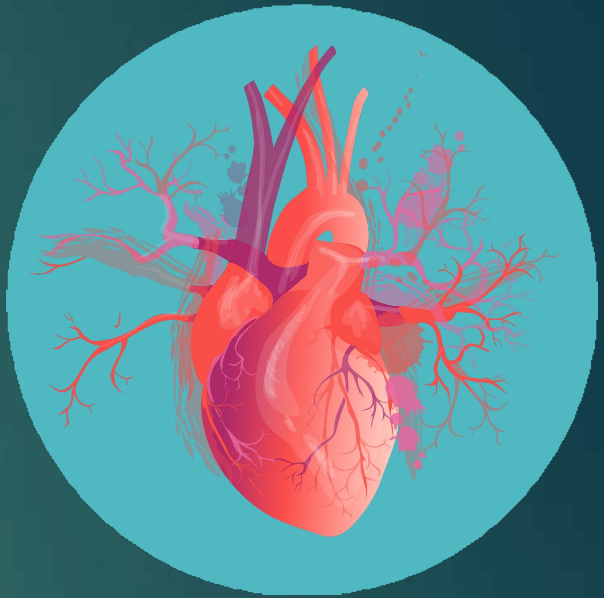
Machine learning algorithm models

- LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks.

- XGBoost: It is a decision tree classifier which has been implemented on gradient boosting framework. This model works on the principle that weak learners should be combined to produce best predictions. Ensembling is performed in sequential manner.

- Decision Tree creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values of attributes.

Note: All applied with hyper-parameter tuning.

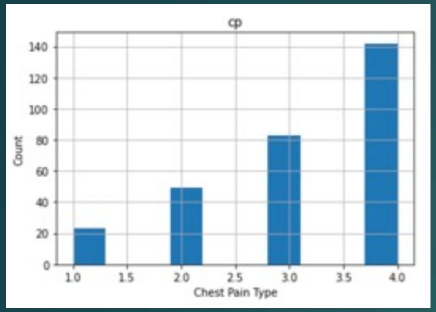


Results

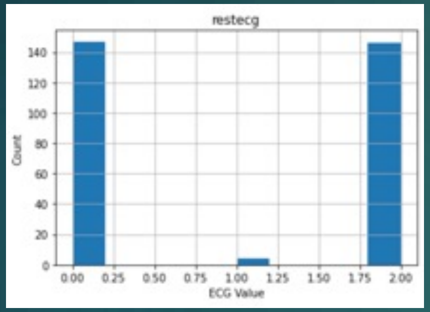
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
count	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00	297.00
mean	54.54	0.68	3.16	131.69	247.35	0.14	1.00	149.60	0.33	1.06	1.60	0.68	4.73	0.46
std	9.05	0.47	0.96	17.76	52.00	0.35	0.99	22.94	0.47	1.17	0.62	0.94	1.94	0.50
min	29.00	0.00	1.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	1.00	0.00	3.00	0.00
25%	48.00	0.00	3.00	120.00	211.00	0.00	0.00	133.00	0.00	0.00	1.00	0.00	3.00	0.00
50%	56.00	1.00	3.00	130.00	243.00	0.00	1.00	153.00	0.00	0.80	2.00	0.00	3.00	0.00
75%	61.00	1.00	4.00	140.00	276.00	0.00	2.00	166.00	1.00	1.60	2.00	1.00	7.00	1.00
max	77.00	1.00	4.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	3.00	3.00	7.00	1.00

Mean, std, 25% and 75% on the continuous variables.

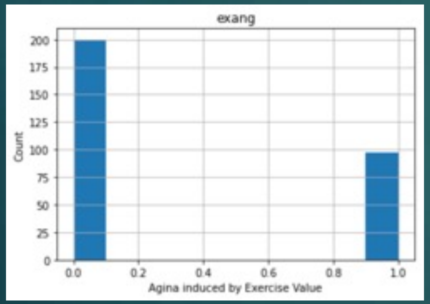
Results



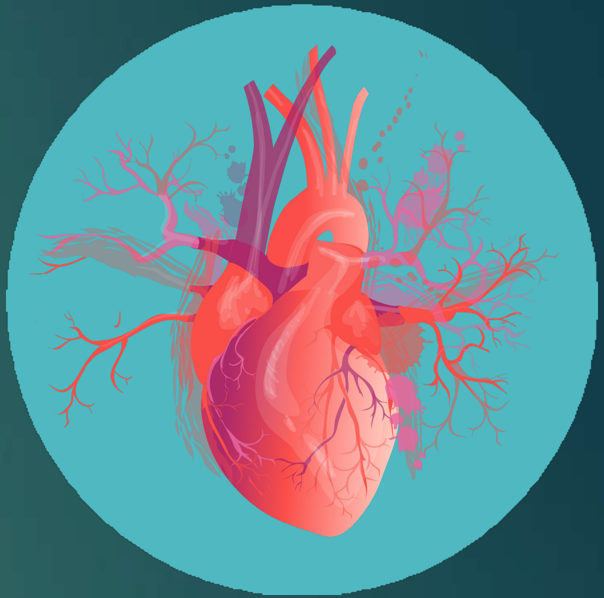
'cp'
Chest pain People with cp 2, 3, 4 are more likely to have heart disease than people with cp 1.



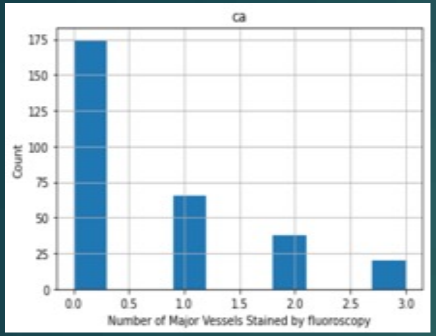
'restecg'
resting EKG results People with a value of 1 having ST-T wave abnormality and with value 2 showing probable or definite left ventricular hypertrophy by Estes' criteria , reporting an abnormal heart rhythm, which can range from mild symptoms to severe problems are more likely to have heart disease.



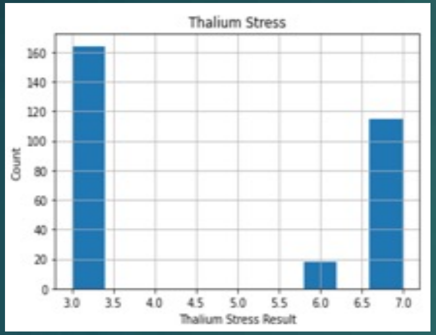
'slope'
{the slope of the ST segment of peak exercise}: People with a slope value of 3 (Downslopins: signs of an unhealthy heart) are more likely to have heart disease than people with a slope value of 1 slope (Upsloping: best heart rate with exercise) or 2 (Flat sloping: minimal change (typical healthy heart)).



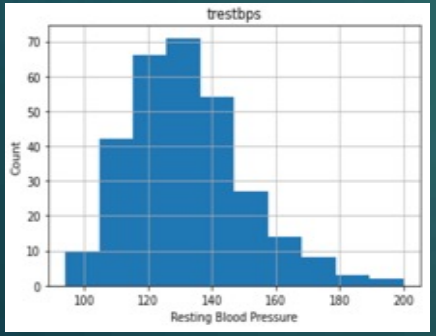
Results



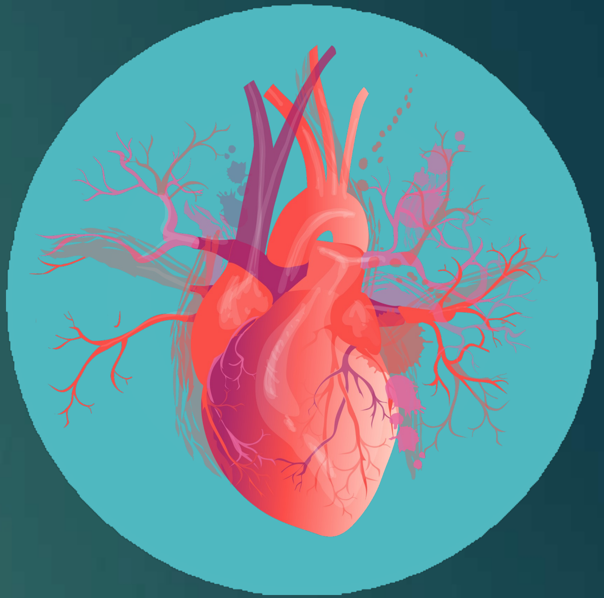
'ca'
number of major vessels (0-3) stained by fluoroscopy: the more blood movement the better, so people with 'ca' equal to 1 are more likely to have heart disease.



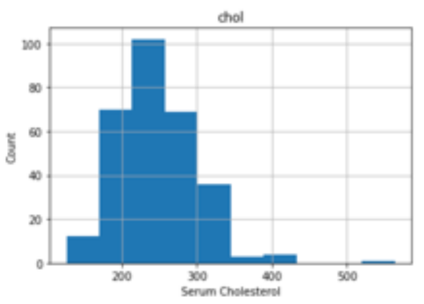
'thal'
thalium stress result: People with a thal value of 3 with no blood flow in some part of the heart are more likely to have heart disease.



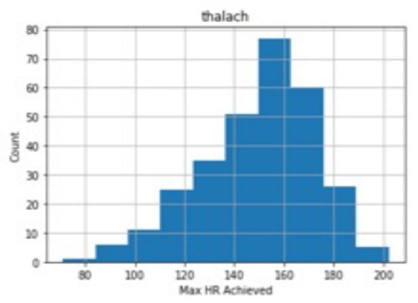
'trestbps': resting blood pressure anything above 130-140 is generally of concern



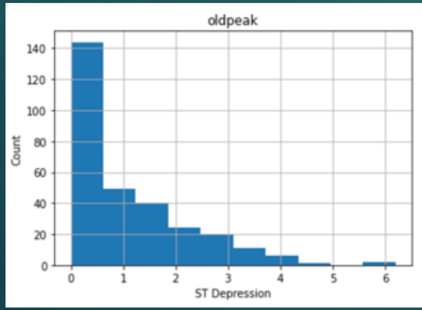
Results



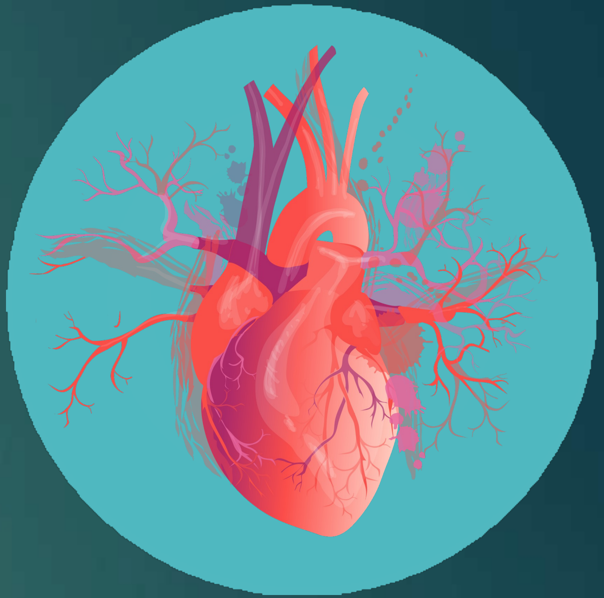
'chol':
greater than 200 is of concern.



'thalach':
People with a maximum of over 140 are more likely to have heart disease.

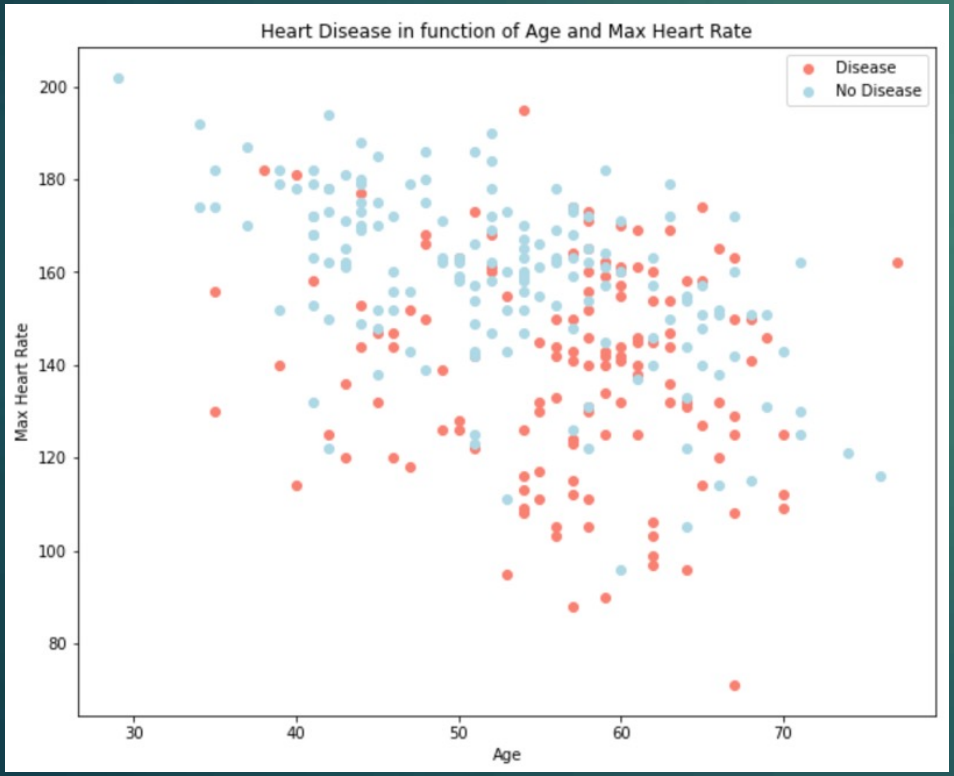


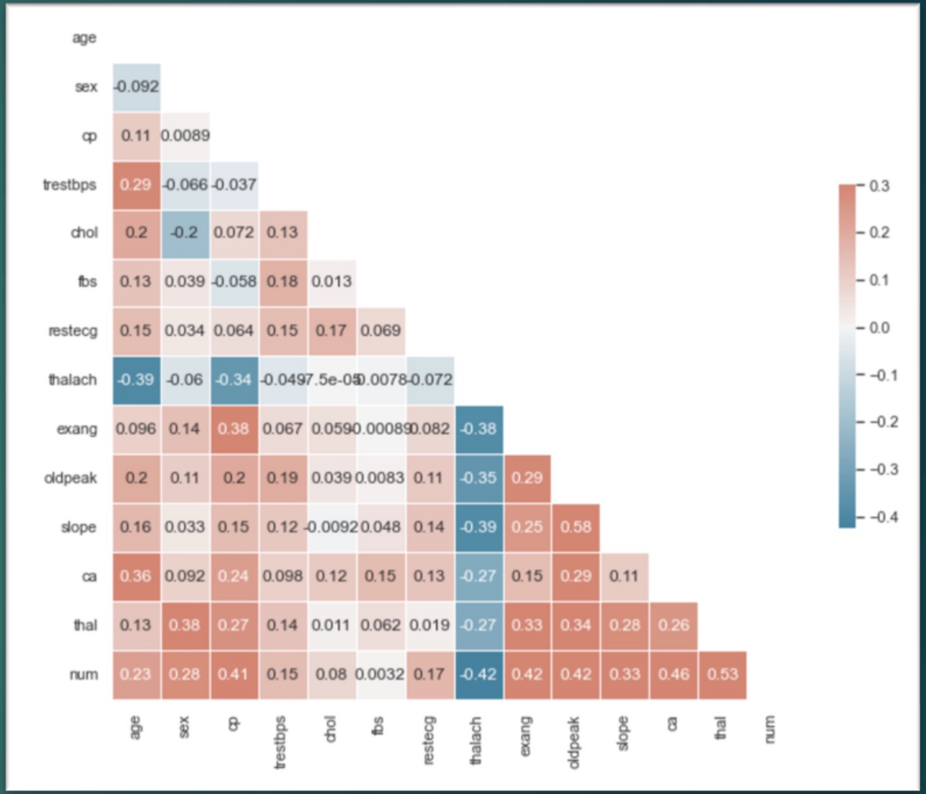
'oldpeak'
of exercise-induced ST depression vs. rest looks at heart stress during exercise an unhealthy heart will stress more.



Results

'Maximum Heart Rate' versus 'Age' showed 50 years of age and higher mostly have heart disease.

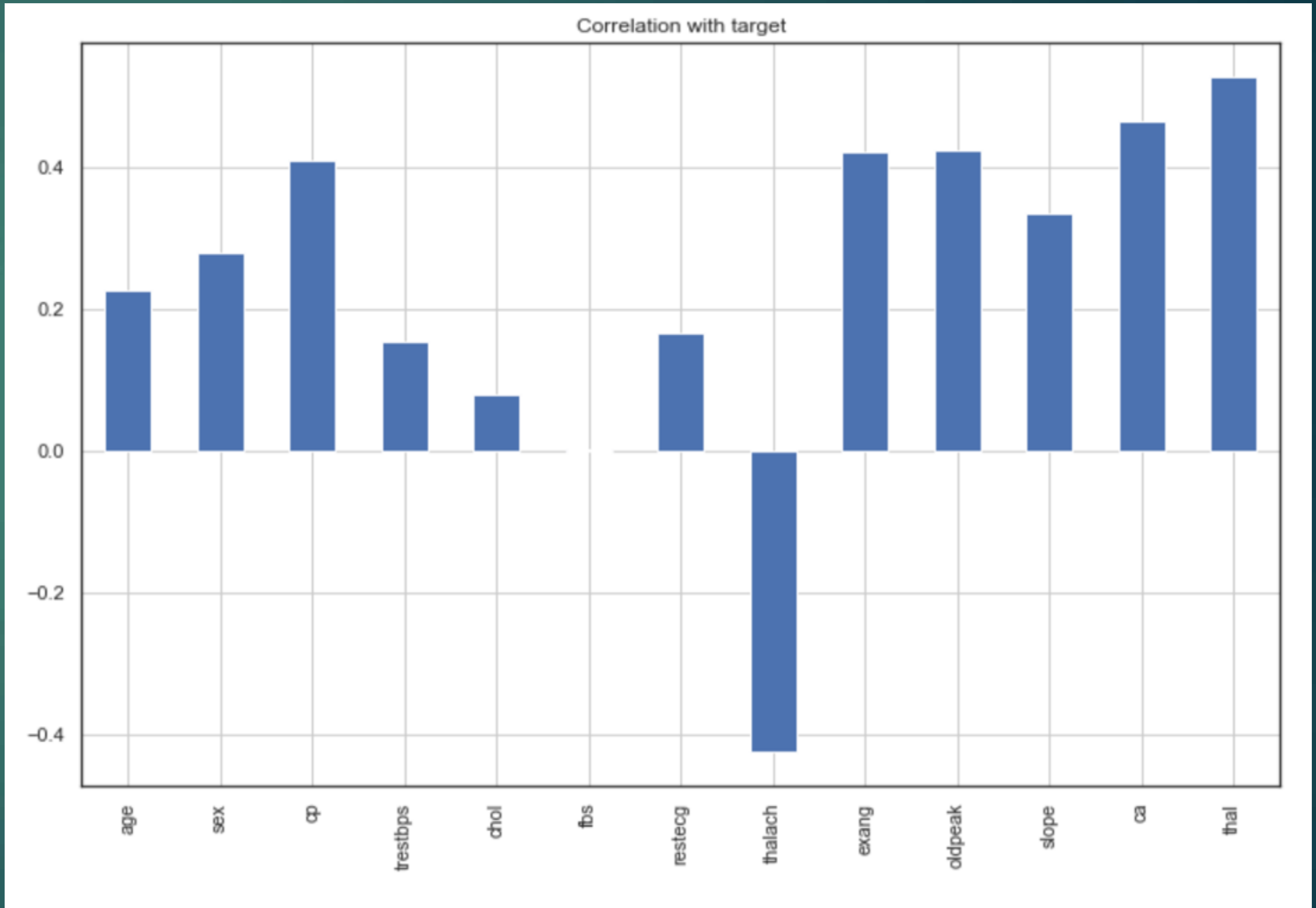




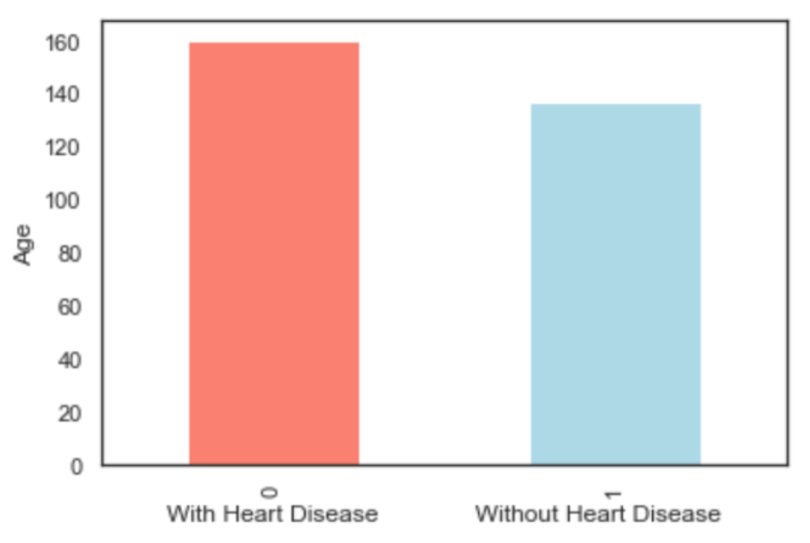
Heatmap shows between 'oldpeak' and 'slope has are highly positively correlated. Our target 'num' is mostly correlated to all our features except 'num' and 'thalach' with a negatively correlation.

Results

This Bar Graph show the Correlation with the target



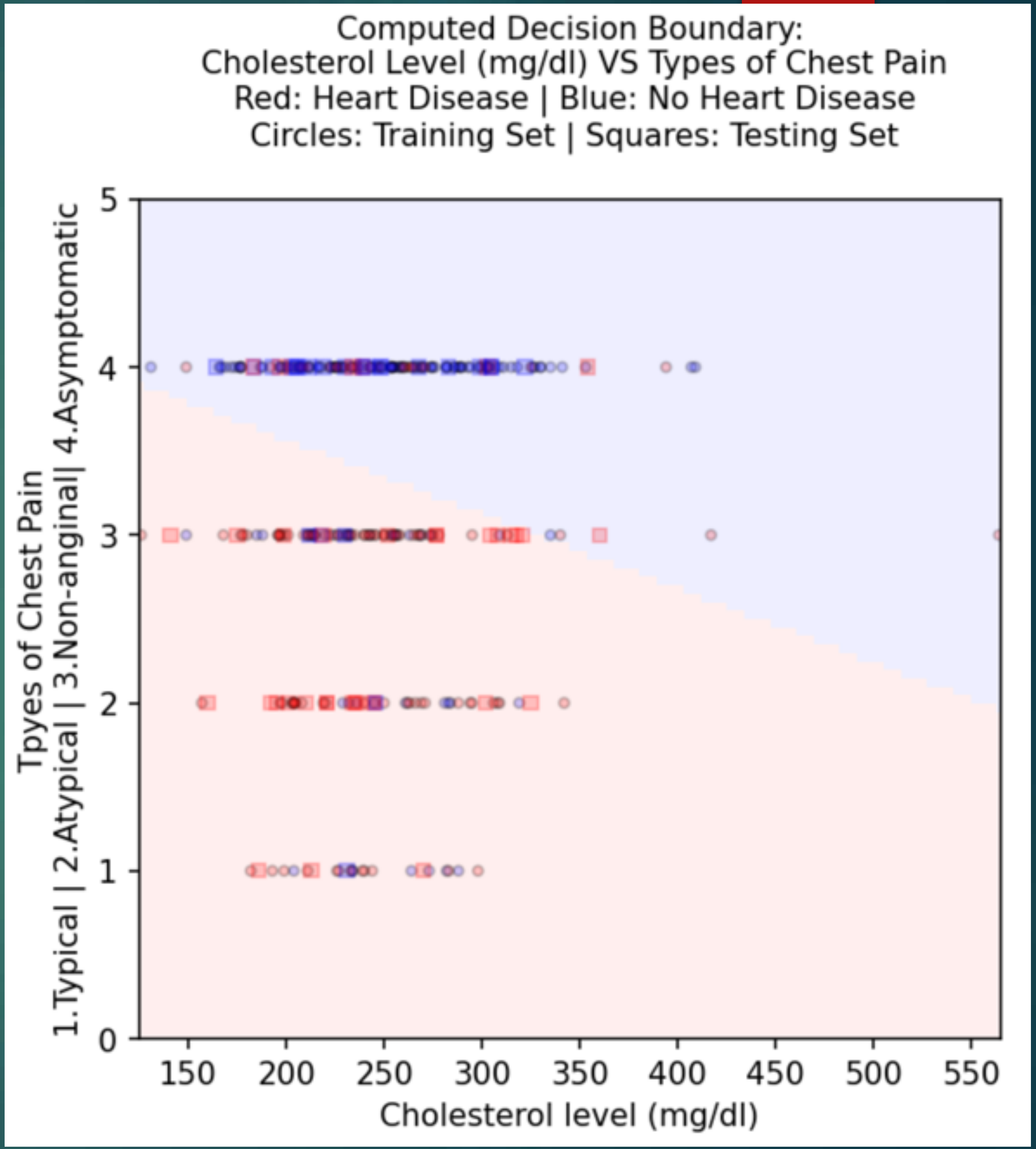
This Bar Graph show people with or without heart disease



Results

In this figure, class 0 **(NO heart disease)** is shaded **RED**, and class 1 **(WITH heart disease)** is shaded **BLUE**. The train labels are plotted as circles, using the same color scheme, while the test data are plotted as squares.

The classifier tends to suggest heart disease either with Chest Pain “cp” or cholesterol “chol” increase. This seems possibly correct.



Results

Training Classification Report

Classification Report for Training Data

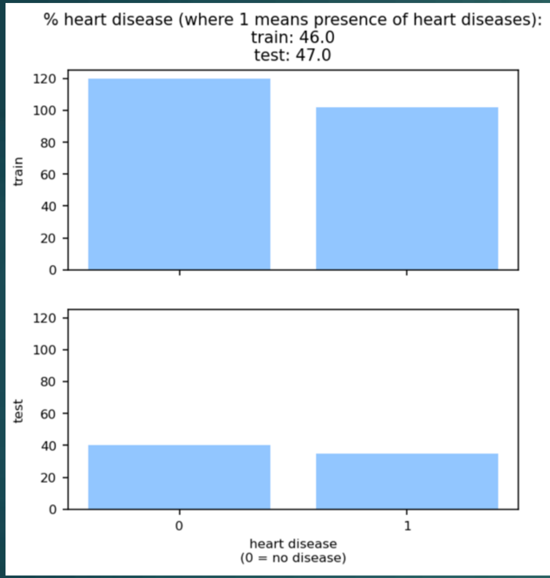
	precision	recall	f1-score	support
0	0.76	0.71	0.73	124
1	0.70	0.75	0.73	113
accuracy			0.73	237
macro avg	0.73	0.73	0.73	237
weighted avg	0.73	0.73	0.73	237

Classification Report for Test Data

	precision	recall	f1-score	support
0	0.85	0.81	0.83	36
1	0.73	0.79	0.76	24
accuracy			0.80	60
macro avg	0.79	0.80	0.79	60
weighted avg	0.80	0.80	0.80	60

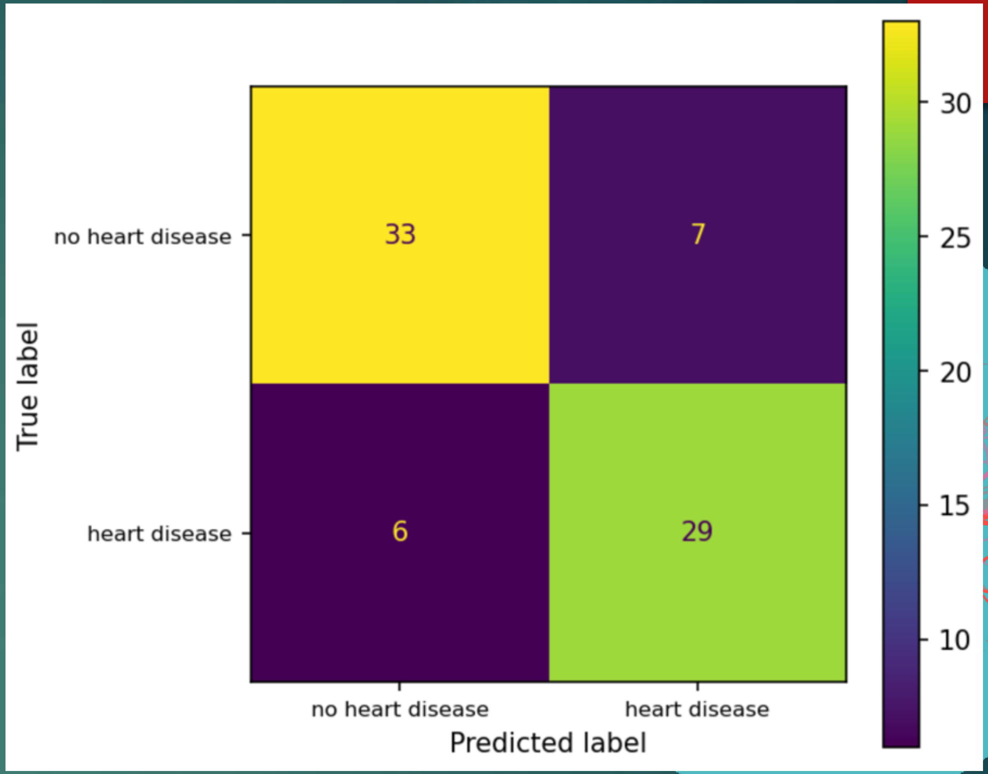
Test Classification Report

Results



Comparison of TRAIN and TEST response data

Class 1 (heart disease) has 29 samples we expected to be positive came back positive, these are my True Positive. 6 samples that we expected to be positive came back negative, these are my False Negative and for class 0 (No heart disease) there are 7 samples that we expected to be negative came back positive, these are my False Positive. 33 samples that we expected to negative came back negative, these are my True Negative.

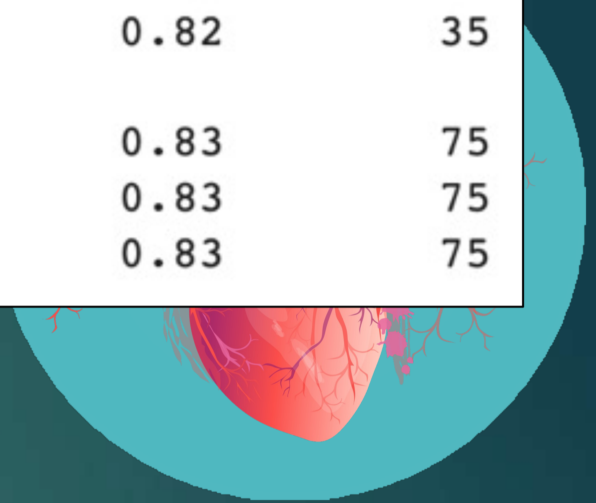


Results

Training Classification Report

The model's **TRAINING ACCURACY** (0.87) is pretty good

	precision	recall	f1-score	support
0	0.85	0.82	0.84	40
1	0.81	0.83	0.82	35
accuracy			0.83	75
macro avg	0.83	0.83	0.83	75
weighted avg	0.83	0.83	0.83	75



	precision	recall	f1-score	support
0	0.87	0.91	0.89	120
1	0.89	0.83	0.86	102
accuracy			0.87	222
macro avg	0.88	0.87	0.87	222
weighted avg	0.87	0.87	0.87	222

Test Classification Report

The model's **TEST ACCURACY** (0.83)

Results

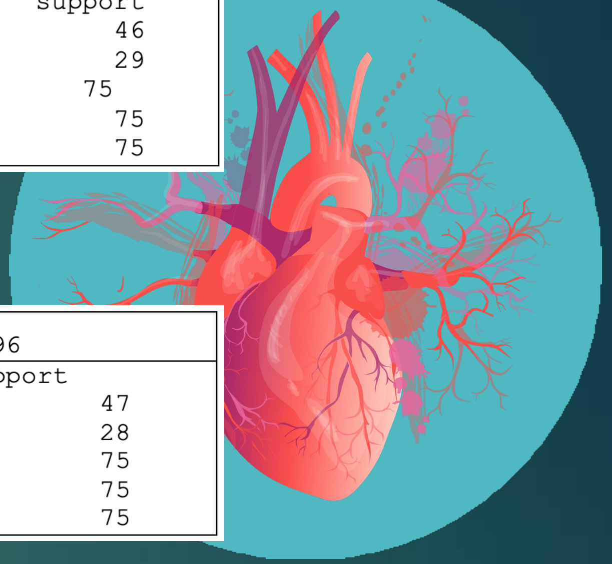
LOGISTIC REGRESSION CLASSIFIER

Logistic Regression Accuracy: 0.96

	precision	recall	f1-score	support
0	0.93	1.00	0.97	43
1	1.00	0.91	0.95	32
accuracy			0.96	75
macro avg	0.97	0.95	0.96	75
weighted avg	0.96	0.96	0.96	75

Logistic Regression with GridSearchCV Accuracy: 0.96

	precision	recall	f1-score	support
0	1.00	1.00	1.00	46
1	1.00	1.00	1.00	29
accuracy			1.00	75
macro avg	1.00	1.00	1.00	75
weighted avg	1.00	1.00	1.00	75



RANDOM FOREST CLASSIFIER

RandomForestClassifier Accuracy: 0.96

	precision	recall	f1-score	support
0	0.98	1.00	0.99	45
1	1.00	0.97	0.98	30
accuracy			0.99	75
macro avg	0.99	0.98	0.99	75
weighted avg	0.99	0.99	0.99	75

Random Forest with GridSearchCV Accuracy: 0.96

	precision	recall	f1-score	support
0	1.00	0.98	0.99	47
1	0.97	1.00	0.98	28
accuracy			0.99	75
macro avg	0.98	0.99	0.99	75
weighted avg	0.99	0.99	0.99	75

LIGHT GBM CLASSIFIER

LGBM Accuracy: 0.8667

	precision	recall	f1-score	support
0	0.87	0.91	0.89	44
1	0.86	0.81	0.83	31
accuracy			0.87	75
macro avg	0.87	0.86	0.86	75
weighted avg	0.87	0.87	0.87	75

LGBM with GridSearchCV Accuracy: 1.0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	46
1	1.00	1.00	1.00	29
accuracy			1.00	75
macro avg	1.00	1.00	1.00	75
weighted avg	1.00	1.00	1.00	75

Results

XGBOOST CLASSIFIER

XGBoost Classifier Accuracy: 1.0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	46
1	1.00	1.00	1.00	29
accuracy			1.00	75
macro avg	1.00	1.00	1.00	75
weighted avg	1.00	1.00	1.00	75

XGBoost Classifier with GridSerchCV Accuracy: 0.92

	precision	recall	f1-score	support
0	0.87	1.00	0.93	40
1	1.00	0.83	0.91	35
accuracy			0.92	75
macro avg	0.93	0.91	0.92	75
weighted avg	0.93	0.92	0.92	75

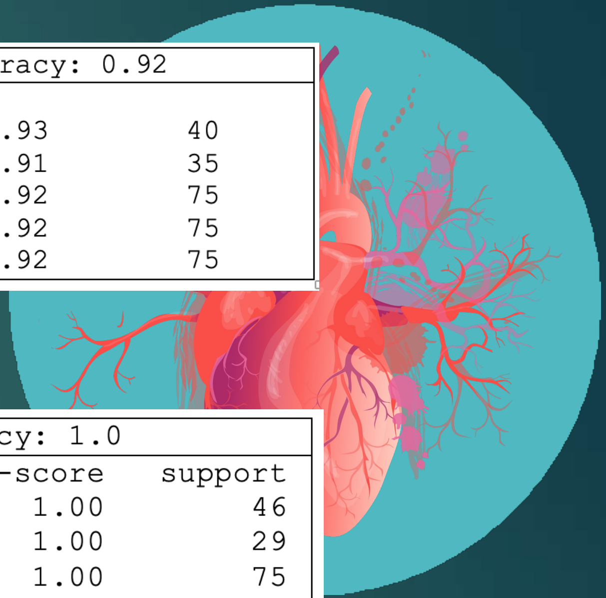
DECISION TREE

Decision Tree Accuracy: 1.0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	46
1	1.00	1.00	1.00	29
accuracy			1.00	75
macro avg	1.00	1.00	1.00	75
weighted avg	1.00	1.00	1.00	75

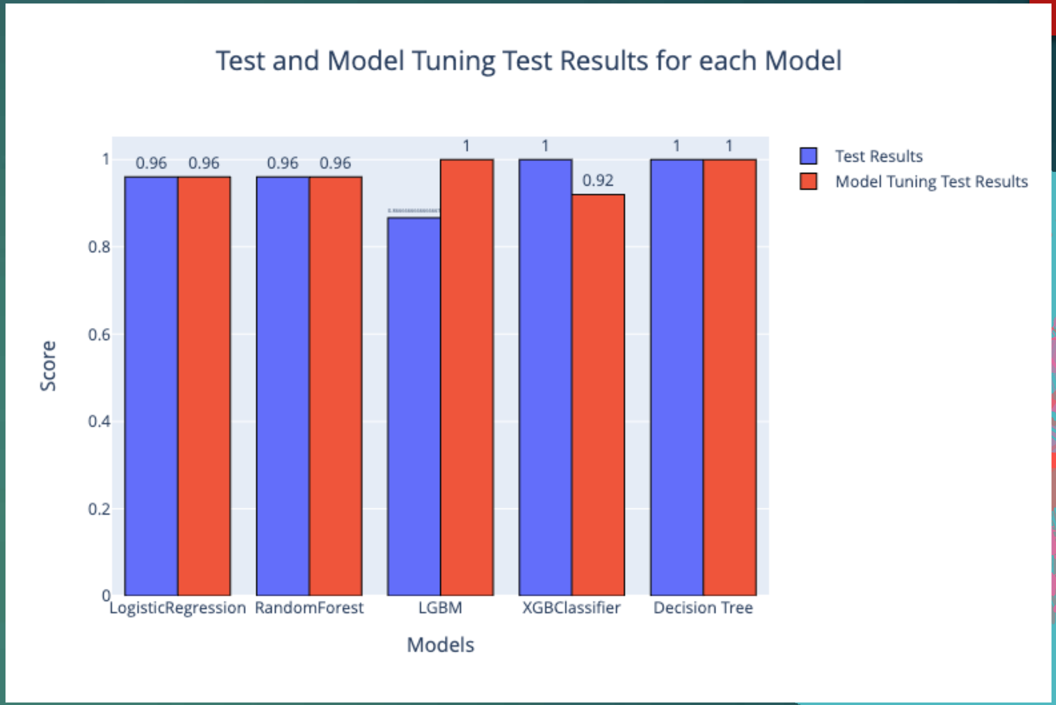
Decision Tree with GridSerchCV Accuracy: 1.0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	46
1	1.00	1.00	1.00	29
accuracy			1.00	75
macro avg	1.00	1.00	1.00	75
weighted avg	1.00	1.00	1.00	75



Results

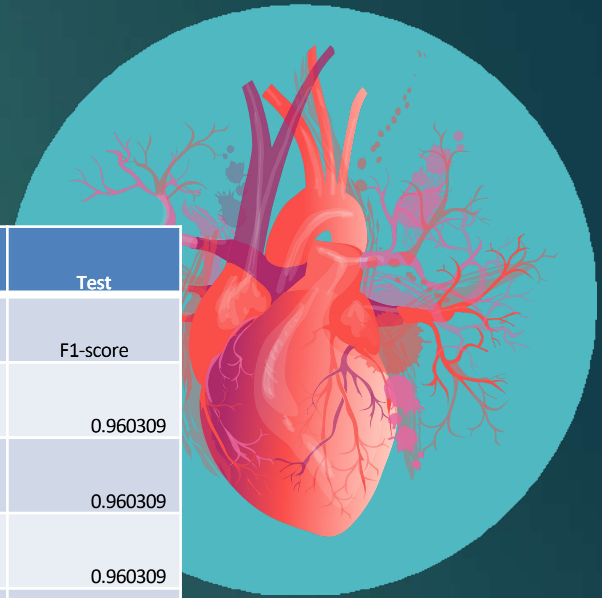
TEST and MODEL TUNING
Test Results for each Model



MODEL	Default	with GridSearchCV
LogisticRegression	0.960000	0.96
RandomForest	0.960000	0.96
LGBM	0.866667	1.00
XGBClassifier	1.000000	0.92
Decision Tree	1.000000	1.00

Results

Results achieved are discussed below presents the interface for taking input from users and predicting using machine learning.



	Train	Test	Train	Test	Train	Test	Train	Test
	Accuracy	Accuracy	Precision	Precision	Recall	Recall	F1-score	F1-score
Logistic_Regression	0.995495	0.96	0.988506	0.90625	1	1	0.9955	0.960309
Random_Forest	1	0.96	1	0.90625	1	1	1	0.960309
LGBM	0.995495	0.96	0.988506	0.90625	1	1	0.9955	0.960309
GBoost	0.995495	0.96	0.988506	0.90625	1	1	0.9955	0.960309
Decision_Tree	1	1	1	1	1	1	1	1

Conclusion

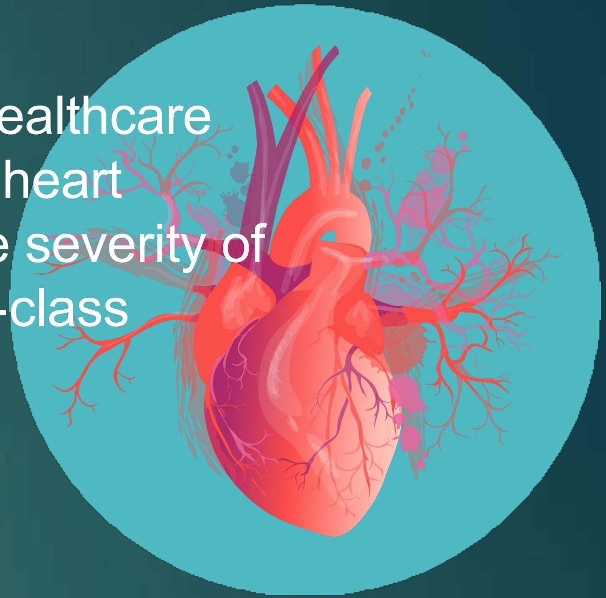
The comparative evaluation of five machine learning algorithms for the heart disease prediction was carried out in this study, with promising outcomes. In this investigation, the performance of ML approaches has been better. When data pre-processing was used, LGBM and Decision Tree performed better in the ML technique for the 13 features in the dataset. Deep learning algorithms are essential in application for the healthcare industry. Therefore, using deep learning techniques to forecast heat disease may produce superior results. In order to determine the severity of the sickness, we are also interested in category it as a multiclass problem.

Recommendations

- It is recommended to have **Additional data** from many sources could be taken so that the models would be able to predict for **different conditions** for the patients.
- **More features** that help determine whether a person would suffer from heart disease could be considered.
- Use **Decision Tree model**, which had the best performance, could be deployed in real-time to provide doctors with faster inference results. This could aid in the diagnosis of whether a person is suffering from heart disease or not.

FUTURE WORK

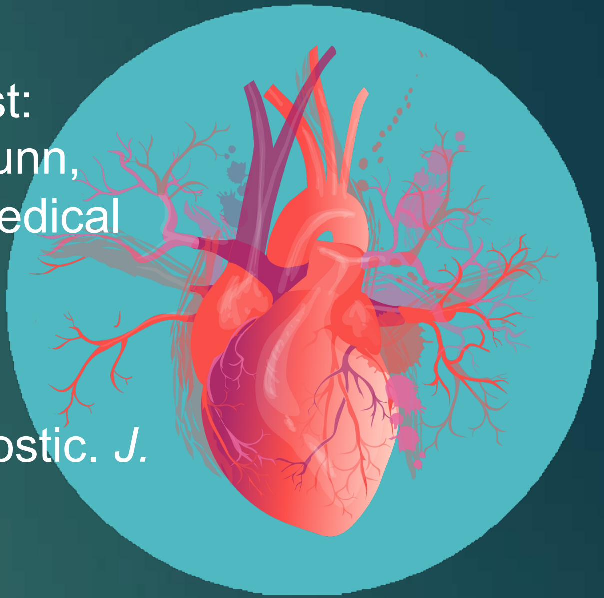
- Deep learning algorithms are essential in applications for the healthcare industry. Therefore, using deep learning techniques to forecast heart disease may produce superior results. In order to determine the severity of the sickness, we are also interested in categorizing it as a multi-class problem.



References

Heart Disease Data Set. Creators: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D., V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Fatima M, Pasha M: Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.* 2017; 09: 1–16. Publisher Full Text



Thank you