Heart Disease Prediction

Heart failure is a life-threatening cardiovascular condition that affects millions of individuals worldwide. Timely detection and accurate prediction of heart failure risk are crucial for improving patient outcomes and reducing healthcare costs. This project aims to develop a predictive model for heart failure using machine learning techniques. By analyzing a diverse set of clinical and patient data, this model will assist in identifying individuals at risk of heart failure, allowing for early intervention and improved patient care.

The project will employ data from various medical sources, preprocess it, and train a machine learning model to predict the likelihood of heart failure in a patient. The goal is to create a valuable tool for healthcare professionals, empowering them to make informed decisions and provide better care to their patients. This introductory phase sets the stage for the comprehensive development and implementation of a heart failure prediction model, which holds the potential to save lives and improve the quality of healthcare.

Dataset: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

Data Structure:

Our data, from the Kaggle source has a shape of 918*12 I.e., total of 918 observations and 12 attributes. Out of the 12 attributes there are categorical and numerical attributes of 7,5. The description of the data is given, as,

SL. no	Attribute	Description	dType
1	Age	Age is in range 29-77 years.	int64
2	Sex	Male: 0 Female: 1	int64
3	Chest Pain	Different types of Chest pain, ATA, NAP, ASY, TA	int64
4	Resting BP	Resting Blood pressure in mm HG range: 94-200	int64
5	Cholesterol	Serum Cholesterol in mg/del, range: 126-564	int64
6	Fasting BS	Range < and > 120mg/dl True: 1 False:0	int64
7	Resting ECG	Resting Electrocardiogram Normal, ST, LVH	int64
8	Max HR	Max heart rate 71-202	int64
9	Exercise Agnima	Pain during exercise Yes: 1 or No :0	int64
10	Old peak	ST depression due exercise w.r.t rest 0 to 2	int64
11	ST Slope	Slope of peak exercise, Up, Flat, down	int64
12	Heart disease	Class label. Cardiac Disease yes: 1, no: 0	int64

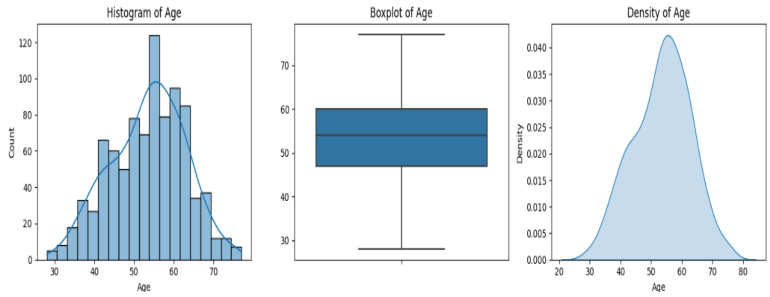
Let’s dive deeper into the attributes,

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Numerical Attributes: Attributes that have real value or integer valued domain.

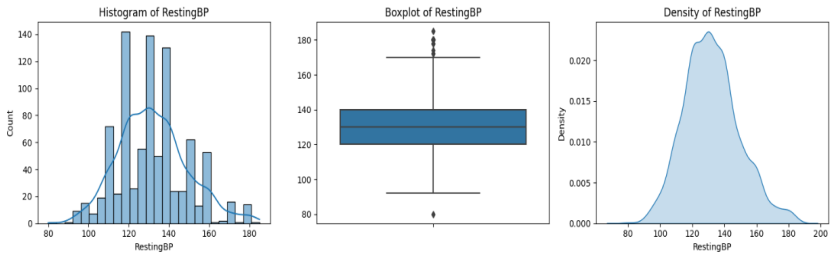
Age:

- Age is one of the key features in medical diagnosis.
- Age is a continuous variable.
- The Age is normally distributed, for this we’ve used Histogram plots, box plot and density plot.
- The average age lies around 53.5, with a maximum age of 77 and a minimum age of 28.0



Resting BP:

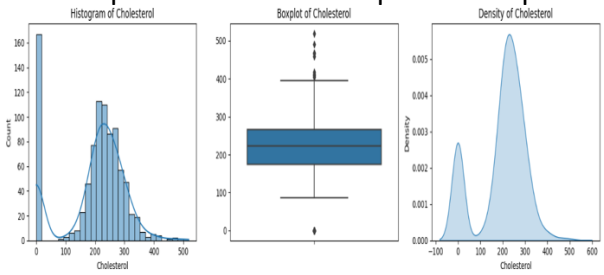
- Resting Blood pressure is a known parameter, where we get to know the blood pressure in our body, this is usually measured in mm HG.
- It is continuous.
- The mean of resting BP is 132.396514, and the maximum value is around 200.00



Cholesterol:

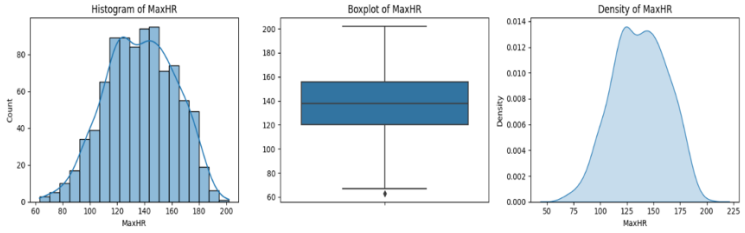
- Cholesterol levels which are measured in mm/dl. People with higher cholesterol are likely to suffer from heart disease.

- Cholesterol is discrete.
- The mean value is 198.799564, the maximum is around 603.00.
- As you can see, in the distribution, the cholesterol has more null values, we are using Mode imputation as mean imputation is prone to outliers.



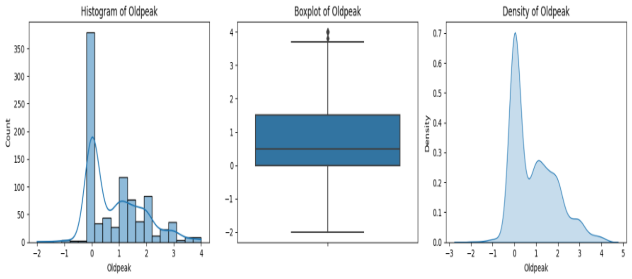
Max HR:

- Maximum heart rate achieved during observation or during test.
- This is a discrete feature.
- The mean value of Maximum heart rate is 136.809368 and the maximum value is of 202.0



Old Peak:

- Old peak is continuous variable.
- ST depression induced by exercise relative to rest.
- The mean value is around, 0.667 and the maximum value is 6.200

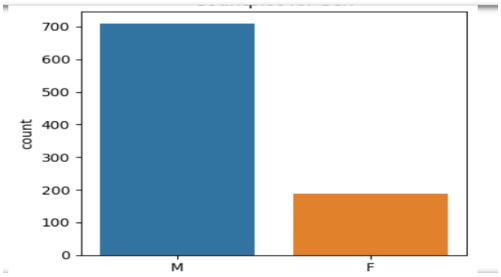


Categorical Attributes: These have a set value in a particular domain. These are generally Nominal or ordinal attributes.

Sex:

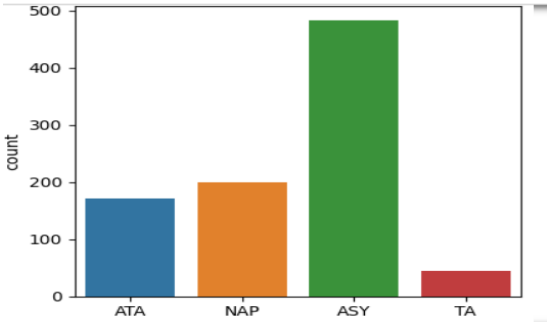
- Sex is a nominal attribute.
- The gender of patients is classified as males and females.

- Upon using count plots and understanding the distribution of the ‘sex’ feature, we observed that 79% of data has males and 21% as females, which means that the model will be biased if we consider this data.



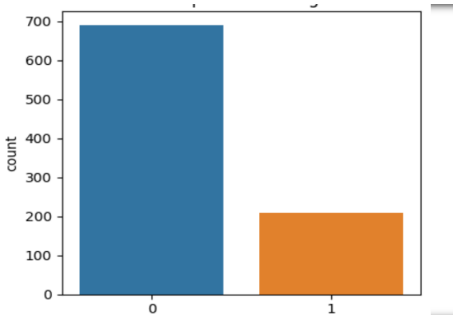
Chest Pain Type:

- This categorical feature explains the type of pain the patient is experiencing.
- This has four unique features, TA, ATA, NAP, ASY
- This is a nominal feature



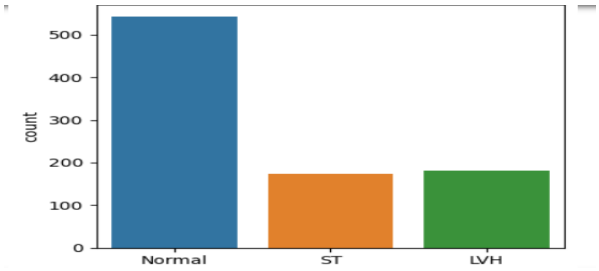
Fasting BS:

- This feature is Binary which is a special case of Discrete.
- The Blood Sugar is greater than 120 and has value of 1 and 0 otherwise.
-



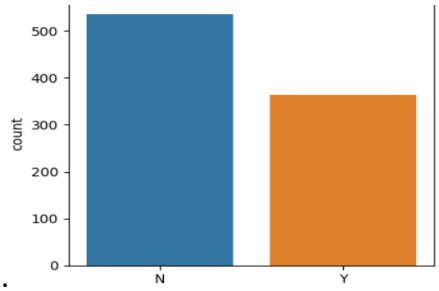
Resting ECG:

- This is a nominal feature.
- Normal: When the ECG has no abnormalities or spikes in the graph.
- ST: When the ECG shows us a T wave inversions or ST elevations of depressions.
- LVH: When the ECG shows some hypertrophy on the left ventricular part of the heart.



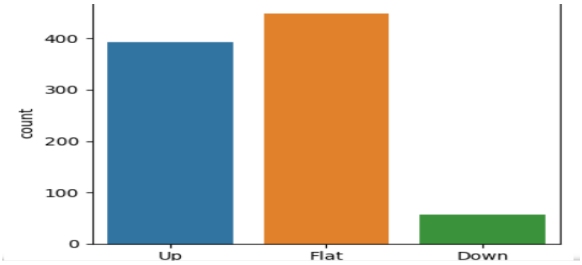
Exercise Angina:

- Whether or not the patient is experiencing pain while doing exercises. It's Y for yes and N for no.
- This is also a nominal feature.



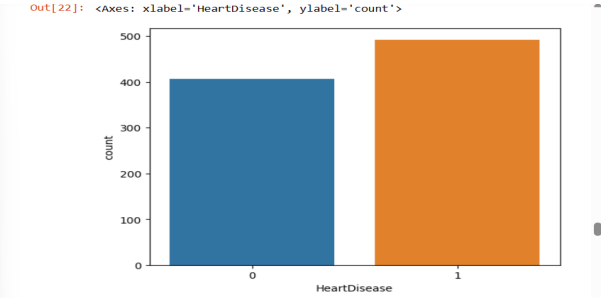
ST Slope:

- The slope is up, flat or down.
- Slope is a nominal feature.



Heart disease:

- This is the class label, which indicates 0 for no heart disease and 1 for the risk of getting heart disease.
- This is also a binary feature.
- The data is balanced, as we saw the labels proportion is in safe range.



Null & Duplicate Values:

The entire data has no null or duplicate values.

Outliers:

- For the outlier- identification, the Z-score method, with a threshold of 3, identified almost 19 outliers.

- Our plan is to build the models using outliers and by removing outliers. As some outliers might be true values but are not in the range of the data and some outliers are the values entered by mistake.

One-hot encoding:

One-hot encoding is the conversion of categorical information into a format that may be fed into machine learning algorithms to improve prediction accuracy. One-hot encoding is a common method for dealing with categorical data in machine learning.

- For the Chest pain Type attribute, applying One-hot encoding results in four new attributes of ATA, NAP, ASY, TA
- For the Resting ECG feature, the one-hot encoding results in, Normal, ST, LVH.

Label-encoder:

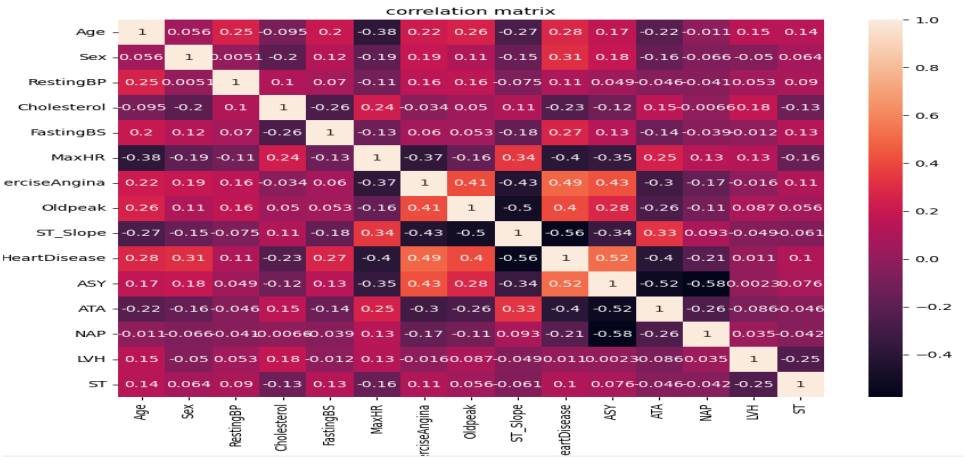
- Label encoding is an effective way to convert Categorical variables into numerical form.
- The attribute- Sex, Male is encoded as 1 and Female is encoded as 0.

Fit transform:

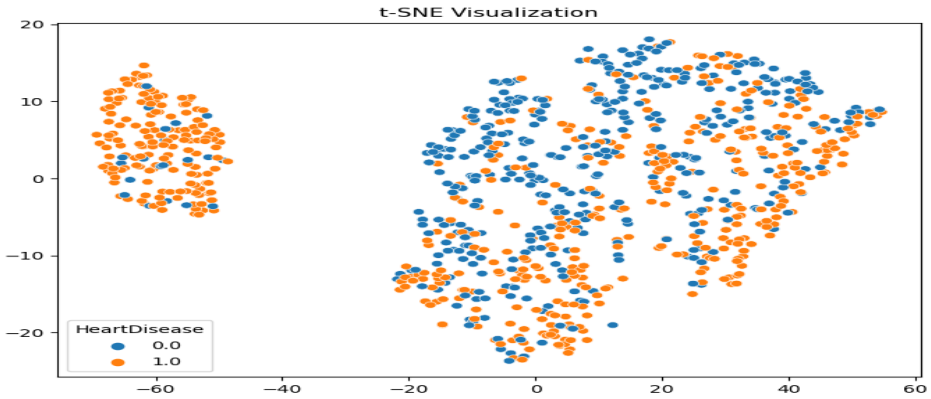
- By applying fit-transform method, the Exercise_Anigma which has Y, N is now transformed into 0 ,1.
- The ST_slope which has up, flat. Down is now transformed into 0, 1, 2.

Correlation:

- The correlation between the attributes, after complete pre-processing, is shown in the below, heat map.
- There are few attributes, that are negatively correlated with the heart disease class label, the Cholesterol, MaxHR, ST_Slope, ATA, NAP of Chest pain.



T-SNE plot:



Model Selection

The machine learning task for the heart failure prediction project is a supervised binary classification problem. We aim to predict whether a patient is at risk of heart failure (positive class) or not (negative class) based on their medical and clinical data. This problem is supervised because we have access to labeled data where each instance is associated with a binary outcome (heart disease or normal).
In the given dataset, there is one column called “HeartDisease” which has values 0 or 1 indicating normal condition and heart disease respectively.

Model Selection:

Upon successful completion of the data preprocessing and feature selection, we will experiment with a variety of machine learning algorithms, including **logistic regression**, **Decision trees**, **random forest**, **gradient boosting**, **support vector machines**, and **neural networks**. We will also incorporate dimensionality reduction techniques such as **Linear discriminant analysis** to check the model performance with the reduced dimensions. The model selection will be based on their performance metrics, interpretability, and ability to handle the specific characteristics of the dataset.

Model Training:

We will split the dataset into **training**, **validation**, and **testing** sets, ensuring that they are stratified to maintain class balance. Models will be trained on the training set with hyperparameter tuning to optimize their performance.

Hyper parameter tuning:

After selecting the best model, we will perform hyperparameter tuning to optimize its performance. Hyperparameters will be adjusted using techniques like grid search or random search.

Grid search is a systematic and exhaustive method for hyperparameter tuning. It involves defining a set of possible values for each hyperparameter and then evaluating the model's performance for all possible combinations of hyperparameters.

Random search is an alternative hyperparameter tuning method that, as the name suggests, explores hyperparameter values randomly within defined ranges. This approach is more computationally efficient compared to grid search, and it can often find good hyperparameter combinations with fewer evaluations.

We will explore the random search to obtain the best hyperparameters.

Model evaluation and testing:

Model evaluation involves validation and testing on the validation set and test dataset respectively. We will assess model performance on the validation dataset using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

Model validation can be performed by **K-Fold cross validation**. We will split the dataset into k folds of approximately equal size. The model is trained and evaluated k times, where each time it is trained on k-1 of the folds and evaluated on the remaining fold. This process ensures that the model has been exposed to different subsets of the training data in each iteration.

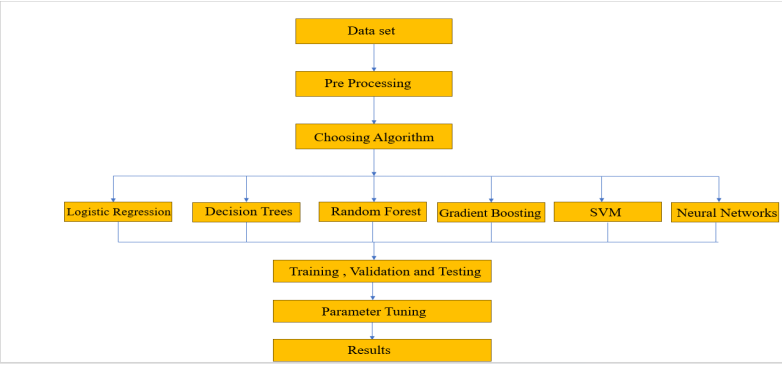
We will test the model by making predictions on the test dataset to evaluate the model on an independent dataset to measure its generalizability. We will iteratively fine tune the model based on the validation results.

Expected Results:

We expect the model to achieve high accuracy, precision, and recall, with a focus on minimizing false negatives (missed heart failure cases) due to the critical nature of the problem.We aim to

develop a model that will significantly contribute to the early detection of heart failure and ultimately improve patient care.

Flow Chart of Model:



References:

https://iopscience.iop.org/article/10.1088/1742-6596/2161/1/012013/pdf
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8740989
https://ieeexplore.ieee.org/abstract/document/9112443?casa_token=PM-CxIAklUcAAAAA:T6UXQ4FPkQb_ApxDevrmcL4hdBuQTpsL3FGujucb2G0lFsCyCgqi9AkdbwqFjS3_AK387voe

Team Members:

1. BalaSwamy Rusum
2. Satya Uma Praneetha Parupudi
3. Sai Mokshith Reddy Peta