HEART FAILURE PREDICTION USING MACHINE LEARNING

1. Introduction:

Heart disease is one of the major cause of mortality in the world today. Prediction of Heart Failure disease is a critical challenge in the field of clinical data analysis. With the advanced development in machine learning (ML), artificial intelligence (AI) and data science has been shown to be effective in assisting in decision making and predictions from the large quantity of data produced by the healthcare industry. ML approaches has brought lot of improvements and broadens the study in medical field which recognizes patterns in the human body by using various algorithms and correlation techniques. One such reality is coronary heart disease, various studies gives impression into predicting heart disease with ML techniques. Initially ML was used to find degree of heart failure, but also used to identify significant features that affects the heart disease by using correlation techniques. There are many features/factors that lead to heart disease like age, blood pressure, sodium creatinine, ejection fraction etc. In this paper we propose a method to finding important features by applying machine learning techniques. The work is to design and develop prediction of heart disease by feature ranking machine learning. Hence ML has huge impact in saving lives and helping the doctors, widening the scope of research in actionable insights, drive complex decisions and to create innovative products for businesses to achieve key goals.

In today's contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension [2]. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection of that disease. The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences.

Records of large set of medical data created by medical experts are available for analyzing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Mostly the medical database consists of discrete information. Hence, decision making using discrete data becomes complex and tough task. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage [5].

2. Related Work

Lot of work has been carried out to predict heart disease using UCI Machine Learning dataset. Different levels of accuracy have been attained using various data mining techniques which are explained as follows. Avinash Golande and et, al. studies various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared [1]. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

T.Nagamani, et al. have proposed a system [2] which deployed data mining techniques along with the Map Reduce algorithm. The accuracy obtained according to this paper for the 45 instances of testing set, was greater than the accuracy obtained using conventional fuzzy artificial neural network. Here, the accuracy of algorithm used was improved due to use of dynamic schema and linear scaling. Fahd Saleh Alotaibi has designed a ML model comparing five different algorithms [3]. Rapid Miner tool was

used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy. Anjan Nikhil Repaka, et al. proposed a system in [4] that uses NB (Naïve Bayesian) techniques for classification of dataset and AES (Advanced Encryption Standard) algorithm for secure data transfer for prediction of disease. Theresa Princy. R, et al, executed a survey including different classification algorithm used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (KNearest Neighbour), Decision tree and accuracy of the classifiers was analyzed for different number of attributes [5].

Problem Formulation:

According to the World Health Organization more than 10 million die due to Heart diseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis [1]. This dataset gives a number of variables along with a target condition of having or not having heart disease. This dataset contain 12 attributes. The "goal" field refers to the presence of heart disease in the patient. The goal is to predict the patient having heart disease or not.

Preliminaries:

This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyze the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This research presents performance analysis of various ML techniques such as XGBoost, Decision Tree, Logistic Regression, CatBoost, Support Vector Classifier and Random Forest for predicting heart disease at an early stage [3].

3. Exponential Data Analysis(EDA) :

It's a clean, easy to understand set of data. However, the meaning of some of the column headers are not obvious. Here's what they mean,

Data Description:

- Age : The person's age in years
- Sex : The person's sex (M = male, F = female)
- Chest Pain Type : The chest pain experienced(TA :typical angina, ATA: atypical angina, NAP: non-anginal pain, ASY: asymptomatic)
- RestingBP: The person's resting blood pressure
- Cholesterol: The person's cholesterol
- FastingBS: The person's fasting blood sugar (1 = true; 0 = false)
- RestingECG: Resting electrocardiographic measurement (normal, having ST-T wave abnormality, left ventricular hypertrophy)
- MaxHR: The person's maximum heart rate achieved
- ExerciseAngina: Exercise induced angina (Y = yes; N = no)
- oldpeak: ST depression induced by exercise relative to rest
- ST_slope: the slope of the peak exercise ST segment (UP, flat, downsloping)
- Heart disease: (0 = no, 1 = yes)

Data Information and Describe:

The info() function is used to count the values and show us either dataset contain missing values or not and also tell us the data type of columns and describe() function is used to measure the data description like count, mean, Standard Deviation, Minimum, 25%, 50%, 75% and Maximum values in the dataset.

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
```

```
1 data.describe()
```

	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

Data Shape:

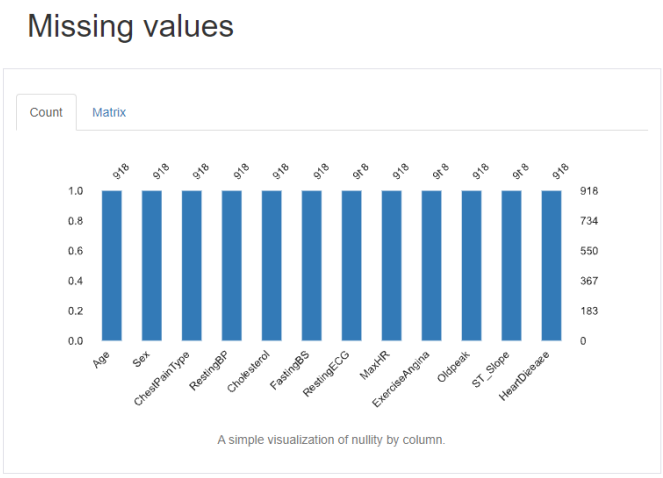
The Heart Failure Dataset contain 918 row and 12 columns

```
1 data.shape
```

```
(918, 12)
```

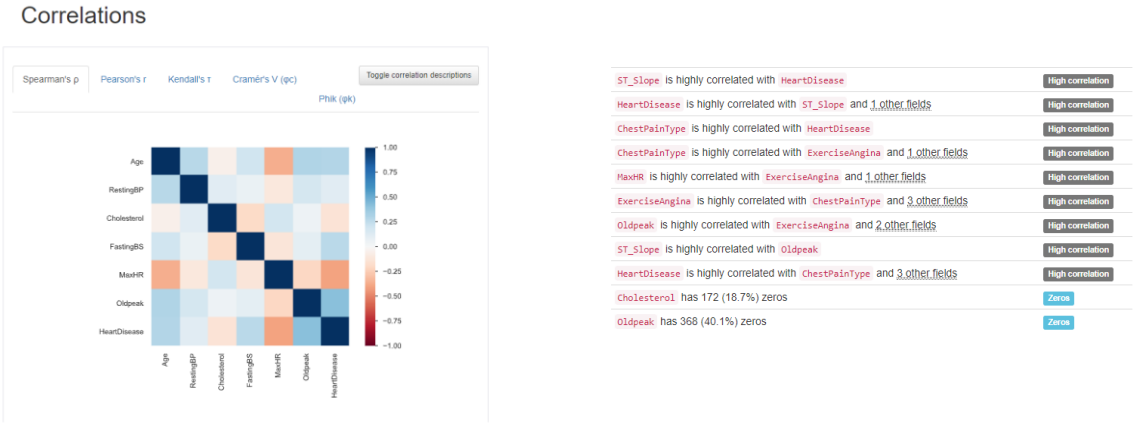
Missing Values:

There is no missing values in this dataset, every column have 918 rows.



Correlation:

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable. Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables.



Comparing Sex Column with Other columns:

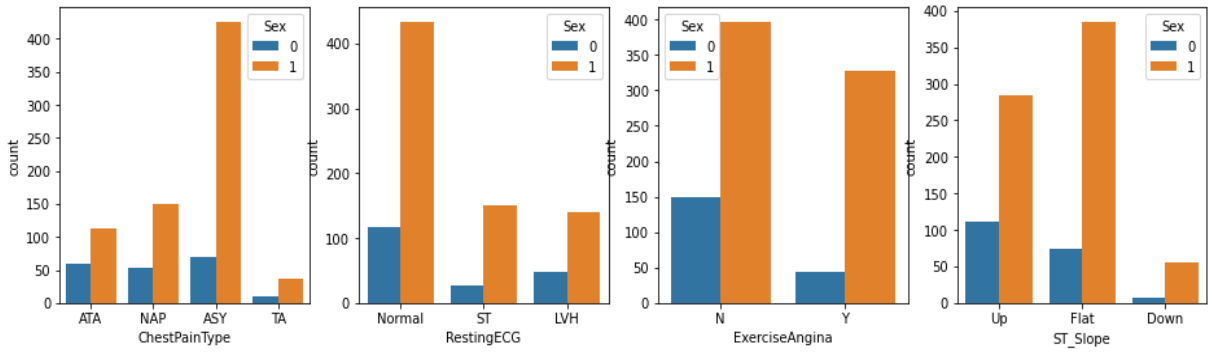
We have ChestPainType, RestingECG, ExerciseAngina and ST_Slope columns to comparison with Sex column. In this figure we can see how many Male and Female having these problems. In this Sex columns we mapped the values 1 : Male and 0 : Female.

```
ChestPainType :> No of Unique Values :  ['ATA' 'NAP' 'ASY' 'TA']

RestingECG :> No of Unique Values :  ['Normal' 'ST' 'LVH']

ExerciseAngina :> No of Unique Values :  ['N' 'Y']

ST_Slope :> No of Unique Values :  ['Up' 'Flat' 'Down']
```

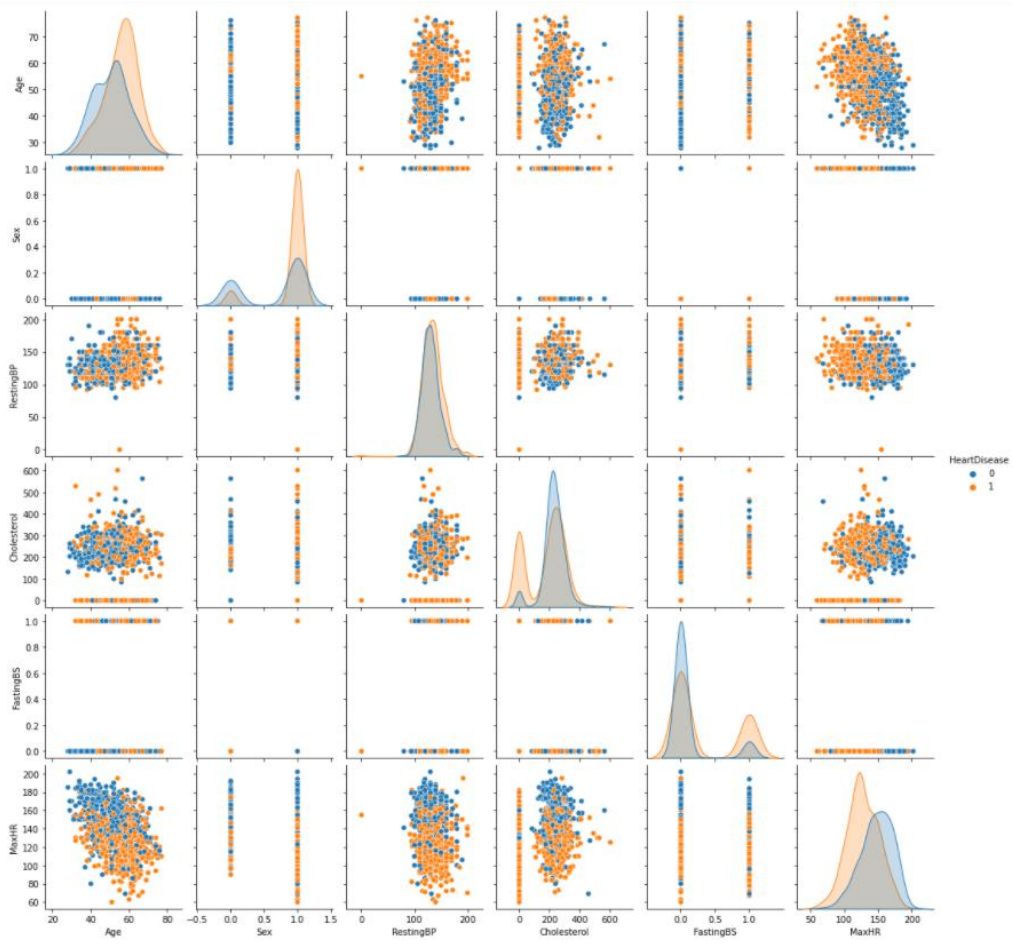


- Cardiovascular disease is the leading cause of death in Man in Australia with 90% of women having one risk factor.
- The causes including high blood pressure, high cholesterol, smoking, diabetes, weight and family history are discussed.

- Men's hearts are affected by stress and depression more than women's. Depression makes it difficult to maintain a healthy lifestyle

Pair Plot:

A pair-plot plot a pairwise relationships in a dataset. The pair-plot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column.



4. Experimental Results and Analyses:

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not

balanced. In this modeling process we use 80% of data for training and 20% of data from testing. We have used a number of machine learning algorithms, these are following:

- **1. Logistic Regression**

  The meaning of the term regression is very simple: any process that attempts to find relationships between variables is called regression. Logistic regression is regression because it finds relationships between variables. It is logistic because it uses logistic function as a link function.

- **2. Support vector machine (SVM)**

  A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. So you're working on a text classification problem.

- **3. K nearest neighborhood (KNN)**

  K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition

- **4. CATBoost**

  CatBoost is an algorithm for gradient boosting on decision trees. It is developed by Yandex researchers and engineers, and is used for search, recommendation systems, personal assistant, self-driving cars, weather prediction and many other tasks at Yandex and in other companies, including CERN, Cloudflare, Careem taxi. It is in open-source and can be used by anyone.

- **5. Random Forest Classifier**

  Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees

- **6. XGBoost**

  XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

- **7. Decision Tree Classifier algorithms**

  A decision tree classifier is a tree in which internal nodes are labeled by features. ... The classifier categorizes an object xi by recursively testing for the weights that the features labeling the internal nodes have in vector xi, until a leaf node is reached. The label of this node is then assigned to xi.

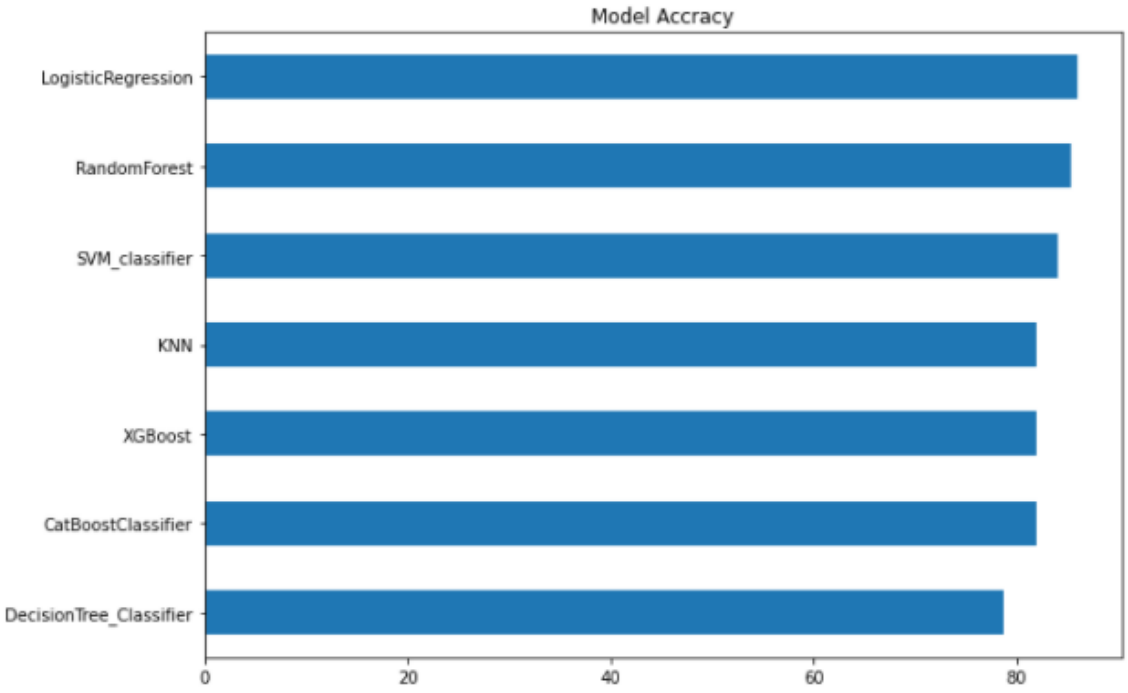
Performance of Machine Learning Models:

In this session, I'll discuss common metrics used to evaluate models. When performing classification predictions, there's four types of outcomes that could occur. **True positives** are when you predict an observation belongs to a class and it actually does belong to that class. **True negatives** are when you predict an observation does not belong to a class and it actually does not belong to that class. **False positives** occur when you predict an observation belongs to a class when in reality it does not. **False negatives** occur when you predict an observation does not belong to a class when in fact it does.

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

Precision is defined as the fraction of relevant examples (true positives) among all of the examples which were predicted to belong in a certain class.

Recall is defined as the fraction of examples which were predicted to belong to a class with respect to all of the examples that truly belong in the class.



These are the results of different algorithms which are applied on the Dataset, The best performance model is Logistic Regression with 86% of accuracy on test dataset.

5. Conclusion

Heart is the most essential organ of the human body and day by day the loss of Human Life is increasing exponentially due to heart failure. Hence there is an urgent need for research to focus into the causes for heart failure and to design a robust prediction system to detect at early stage so that loss of life can be avoided. Even though there were many heart diseases prediction systems available at

present but each one has its own limitations. The main objective of this research work is to overcome the difficulty faced by other researchers and to build a robust system which works efficiently and will able to predict accurately the possibility of heart attack at very early stage. By using the Logistic Regression this model could able to predict with an accuracy of about 86.0% which is highest as compared to other algorithms. In future work, we'll be able to work with Deep Learning model which is very flexible and efficient. We'll use LSTM and CNN neural network to measure the accuracy and performance.

6. **Reference**

[1] Avinash Golande, Pavan Kumar T,"Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[6] Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.

[7] UCI, ―Heart Disease Data Set.[Online]. Available (Accessed on May 1 2020): https://www.kaggle.com/hishaamarmghan/machine-learning-for-heart-failure-prediction/data

[8] Sayali Ambekar, Rashmi Phalnikar,"Disease Risk Prediction by Using Convolutional Neural Network",2018 Fourth International Conference on Computing Communication Control and Automation.

[9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[10] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018