HEART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

Submitted in fulfillment of the summer internship in APSCHE-IIDT virtual internship program with industry partner-Black Bucks

By

CHALLAGUNDLA PRAVEEN (22JK1A4226)

***ABSTRACT*:**

Day by day the cases of heart diseases are increasing at a rapid rate and it's very Important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naïve bayes etc. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the.pynb format.

INTRODUCTION

Heart Disease Prediction:

Heart disease describes a range of conditions that affect your heart. Today, cardiovascular diseases are the leading cause ofdeath worldwide with 17.9 million deaths annually, as per the World Health Organization reports. Various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc.. *ere are certain signs which the American Heart Association lists like the persons having sleep issues, a certain increase and decrease in heart rate (ir regular heart beat), swollen legs, and in some cases weight gain occurring quite fast; it can be 1-2 kg daily. All these symptoms resemble different diseases also like it occurs in the aging persons, so it becomes a difficult task to get a correct diagnosis, which results in fatality in near future. But as time is passing, a lot of research data and patients records of hospitals are available. There are many open sources for accessing the patient's records and researches can be conducted so that various computer technologies could be used for doing the correct diagnosis of the patients and detect this disease to stop it from becoming fatal. Now a days it is well known that machine learning and artificial intelligence are playing a huge role in the medical industry. We can use different machine learning and deep learning modelsto diagnose the disease and classify or predict the results. A complete genomic data analysis can easily be done using machine learning models. Models can be trained for knowledge pandemic predictions and also medical records can be transformed and analyzed more deeply for better predictions. Many studies have been performed and various machine learning models are used for doing the classification and prediction for the diagnosis of heart disease. An automatic classifier for detecting congestive heart failure shows the patients at high risk and the patients at low risk by Melillo et al; they used machine learning algorithm as CART which stands for Classification and Regression in whichsensitivity is achieved as 93.3 percent and specificity is achieved as 63.5 percent. Then for improving the performance electrocardiogram (ECG) approach is suggested by Rahhal et al in

which deep neural networks are used for choosing the best features and then using them. Then, for detecting heart failures, a clinical decision support system is contributed by Guidi et al. for preventing it at an early stage. They tried to compare different machine learning models and deep learning models especially neural net works, as support vector machine, random forest, and CART algorithms. An 87.6 percent accuracy was achieved by random forest and CART, which outperformed everyone used in the classification. Combining the natural language processing with the rule-based approach, Zhang et al. achieved 93.37 percent accuracy when the NYHA HF class was found from the unstructured clinical notes. SVM techniques used for detecting patients who already have diabetes and then predicting heart disease by Parthiban and Srivatsa achieved a 94.60 percent accuracy rate, and the features taken were common like blood sugar level, age of the patient, and their blood pressure data. In machine learning, a common problem is the high dimensionality of the data; the datasets which we use contain huge data and sometimes we cannot view that data even in 3D, which is also called the curse of dimensionality. So, when we perform operations on this data, we require a huge amount of memory, and sometimes the data can also grow exponentially and overfitting can happen. The weighting features can be used, so the redundancy in the dataset can be decreased which in turn also helps in decreasing the processing time of the execution. For decreasing the dimensionality of the dataset, there are various feature engineering and feature selection techniques which can be used to remove that data not having that much importance in the dataset. In literature, when feature engineering and feature se lection are applied, the results improve, both for classification as well as predictions.Dun et al. tried variousmachine learning and deep learning techniques for detectingthe heart disease and also performed hyperparameters tuning for increasing the results accuracy. Neural networks achieved high accuracy of 78.3 percent, and the other models were logistic regression, SVM, and ensemble techniques like Random Forest, etc. For reducing the cardiovascular features, Singh et al. used generalized discriminant analysis for extracting nonlinear features; a binary classifier like an extreme learning machine for less overfitting and increasing the training speed and the ranking method used for all these was Fisher. The accuracy achieved was 100 percent for detecting coronary heart disease. Arrhythmias classification was done by Yaghouby et al. for heart rate variability. A

multilayer perceptron neural network was used for doing the classification and 100 percent accuracy is achieved by reducing the features or Gaussian Discriminant Analysis. Aslet al.used Gaussian discriminant analysis for reducingthe HRV signal features to 15 and 100 percent precision is achieved using the SVM classifier. For dealing with data that are of high variance or high dimensional data, by using appropriate dimensionality reduction techniques like PCA, we can store valuable in formation in new components. PCA is used by many researchers as the first preference while dealing with high dimensionality data. Rajagopal and Ranganathan used five different dimensionality reduction techniques which are unsupervised (linear and nonlinear), and neural network is used as a classifier for classifying cardiac arrhythmia. Fas tICA (used for independent component analysis) with a minimum of 10 components was able to achieve an F1 score of 99.83 percent. Zhang et al. used the AdaBoost al gorithm which is based on PCA for detecting breast cancer. Negi et al. combined uncorrelated discriminant analysis with PCA so that the best features that are used for con trolling the upper limb motions can be selected and the results were great. Avendaño-Valencia et al. tried to reduce heart sounds to increase performance by applying PCA techniques on time-frequency representations. Kamencay et al. tried a new method for different medical images reaching an accuracy of 83.6 percent when trained on 200 images by using PCA-KNN which is a scale invariant feature used in medical images for the scaling purpose. Ratnasari et al. used a gray-level threshold of 150 based on PCA and ROI, all of these used for reducing features of the X-ray images. The studies of the past are mainly based on a 13-feature dataset. The classification is common in every study to predict if a patient has heart disease or not, and also one most common pattern which can be seen is that the dataset commonly used is of Cleveland. The results obtained achieved great accuracies like random forest with 89.2 percent accuracy ; decision tree with 89.1 percent ac curacy ; ANN with 92.7 percent accuracy , 89 percent , and 89.7 percent accuracy ; and SVM accuracy with 88 percent . A hybrid model is created which achieved an accuracy of 94.2 percent by GA þ NN. PCA models achieved an accuracy of 92 and 95.2 percent as PCA regression and PCA1þNN . The dimensionality reduction was the main focus here for learning three things: (i) selection of the best features, (ii) validation of performance, and (iii) use of six different classifiers for calculating the 74 features which

are selected. Heart disease is very fatal and it should not be taken lightly. Heart disease happens more in males than females, which can be read further from Harvard Health Publishing. Researchers found that, throughout life, men were about twice as likely as women to have a heart attack. *at higher risk persisted even after they accounted for tradi tional risk factors of heart disease, including high choles terol, high blood pressure, diabetes, body mass index, and physical activity. The researchers are working on this dataset as it contains certain important parameters like dates from 1998, and it is considered as one of the benchmark datasets when someone is working on heart disease prediction. This is dataset dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V, and the results achieved are quite promising.

Background of study:

Heart disease predictor is an offline platform designed and developed to explore the path of machine learning. The goal is to predict the health of the patient from collective data to be able to detect configurations at risk for the patient, and therefore, in cases requiring emergency medical assistance, alert the appropriate medical staff of the situation of the latter. We initially have a dataset collecting information of many patients with which we can conclude the results into a complete form and can predict data precisely. The results of the predictions, derived from the predictive models generated by machine learning, will be presented through several distinct graphical interfaces according to the datasets considered. We will then bring criticism as to the scope of our results.

Data has been collected from Kaggle. Data collection is the process of gathering and measuring information from countless different sources to use the data.

***Methodology*:**

Description of the Dataset

The dataset used for this research purpose was the Public Health Dataset and it is dating from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The “target” field refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. The first four rows and all the dataset features are shown in Table 1 without any preprocessing. Now the attributes which are used in this research purpose are described as follows and for what they are used or resemble:

- Age—age of patient in years, sex—(1 = male; 0 = female).
- Cp—chest pain type.
- Trestbps—resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if you have a normal blood pressure reading, it is fine, but if it is a little higher than it should be, you should try to lower it. Make healthy changes to your lifestyle).
- Chol—serum cholesterol shows the amount of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).
- Fbs—fasting blood sugar larger than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
- Restecg—resting electrocardiographic results.
- Thalach—maximum heart rate achieved. The maximum heart rate is 220 minus your age.
- Exang—exercise-induced angina (1 yes). Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.
- Oldpeak—ST depression induced by exercise relative to rest.

- Slope—the slope of the peak exercise ST segment.
- Ca—number of major vessels (0–3) colored by fluoroscopy.
- Thal—no explanation provided, but probably thalassemia (3 normal; 6 fixed defects; 7 reversible defects).
- Target (T)—no disease = 0 and disease = 1, (angiographic disease status).

Machine Learning :

Machine learning is used to provide the good learning to the machines and analyze some pattern for handling the data in extra efficient manner. Sometimes, it may happens that after viewing the data, we even unable to predict the actual pattern or acquire the valuable information from the data. In this condition, we have to go for machine learning. The motive of machine learning is to grasp some knowledge from the data by themselves. Even, many studies has been terminated which highlights the purpose of machine learning that how do machines learn by its.

Machine Learning Techniques :

The main ML techniques can be classified as follows.

Supervised Learning :

The supervised machine learning alg0rithms are those which demand some external assistance. The input dataset splits into training and test dataset. The trained dataset composed of output variable which is t0 be predicted or classified. Each algorithm get to know a specific pattern from the training dataset and just apply them to the test dataset for prediction or classification purposes. This algorithm is named as supervised learning in view of the fact that the process of algorithm learning from the training dataset can be thought 0f as a teacher supervising the learning process. Three most prominent supervised learning algorithms are considered below.

Generic Model Predicting Heart Disease:

Data Collection and Preprocessing

The dataset used was the Heart disease Dataset which is a combination f 4 different database, but only the UCI Cleveland dataset was used. This database consists of a total

of 76 attributes but all published experiments refer to using a subset of only 14 features . Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis. The complete description of the 14 attributes used in the proposed work is mentioned in Table 1 shown below.

Attribute Description :

- Distinct Values of Attribute
- Age- represent the age of a person
- Multiple values between 29 & 71
- Sex- describe the gender of person (0- Feamle, 1-Male)-0,1
- CP- represents the severity of chest pain patient is suffering.-0,1,2,3
- RestBP-It represents the patients BP.
- Multiple values between 94& 200
- Chol-It shows the cholesterol level of the patient.
- Multiple values between 126 & 564
- FBS-It represent the fasting blood sugar in the patient-0,1
- Resting ECG-It shows the result of ECG-0,1,2
- Heartbeat- shows the max heart beat of patient
- Multiple values from 71 to 202.
- Exang- used to identify if there is an exercise induced angina. If yes=1 or else no=0-0,1.

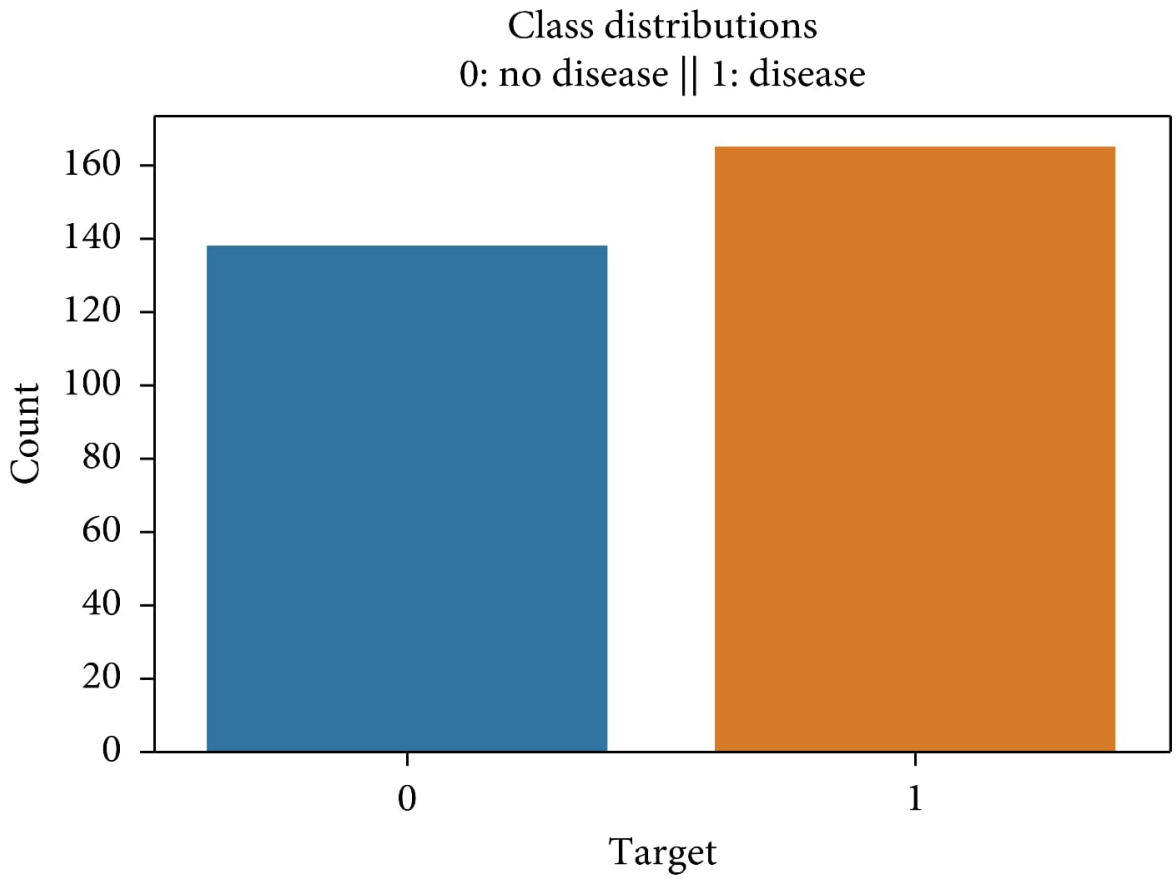
Preprocessing of the Dataset :

- The dataset does not have any null values. But many outliers needed to be handled properly, and also the dataset is not properly distributed. Two approaches were used.

- One witout outliers and feature selection process and directly applying the data to the machine learning algorithms, and the results which were achieved were not promising.
- But after using the normal distribution of dataset for overcoming the overfitting problem and then applying Isolation Forest for the outlier's detection, the results achieved are quite promising.
- Various plotting techniques were used for checking the skewness of the data, outlier detection, and the distribution of the data. All these preprocessing techniques play an important role when passing the data for classification or prediction purposes.

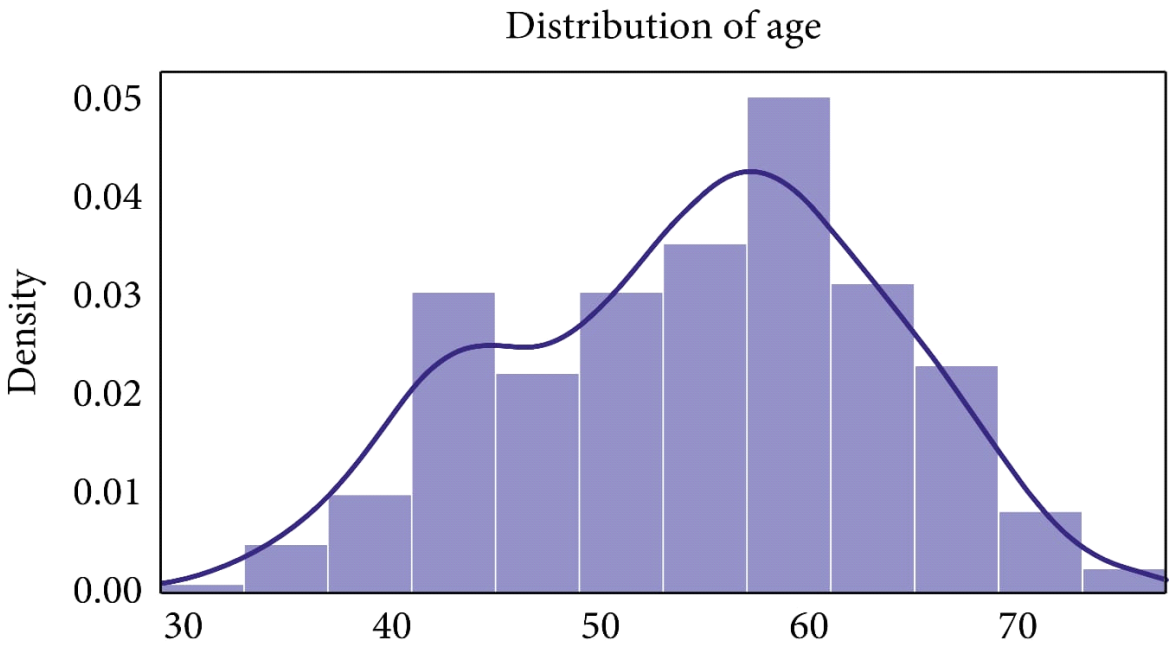
Checking the Distribution of the Data :

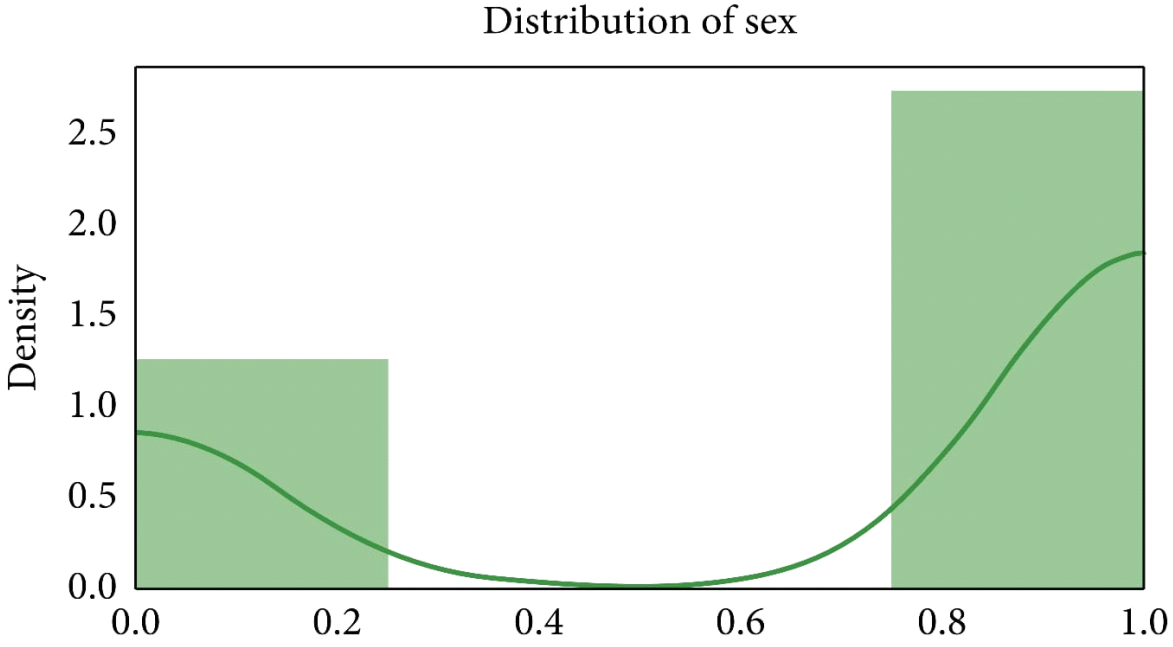
- The distribution of the data plays an important role when the prediction or classification of a problem is to be done.
- We see that the heart disease occurred 54.46% of the time in the dataset, whilst 45.54% was the no heart disease.
- So, we need to balance the dataset or otherwise it might get overfit. This will help the model to find a pattern in the dataset that contributes to heart disease

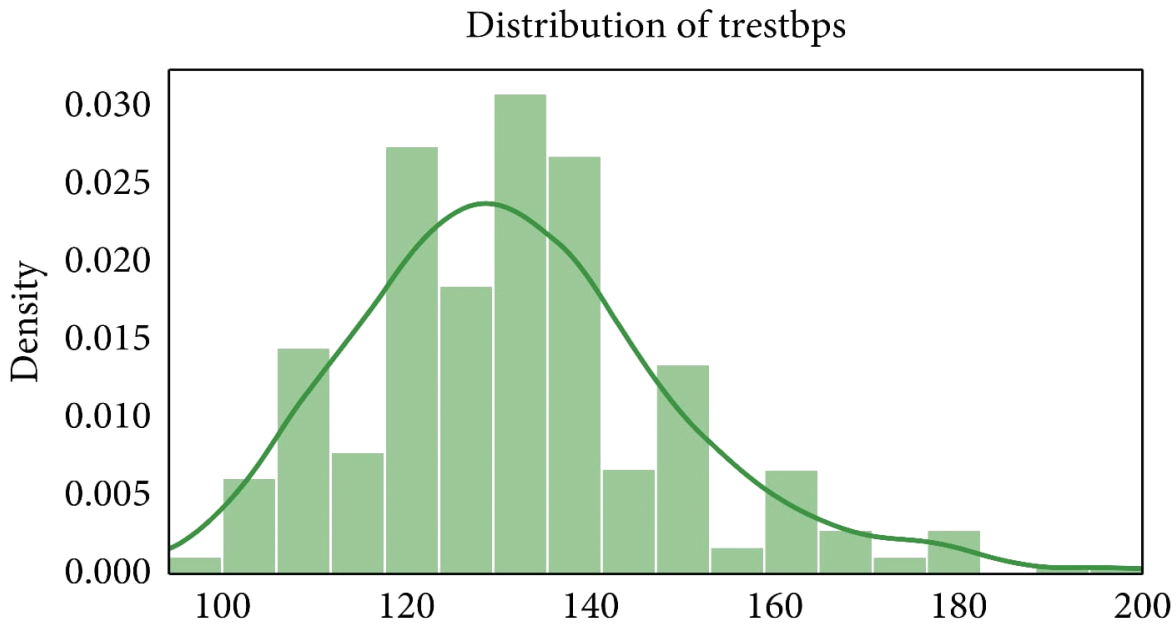


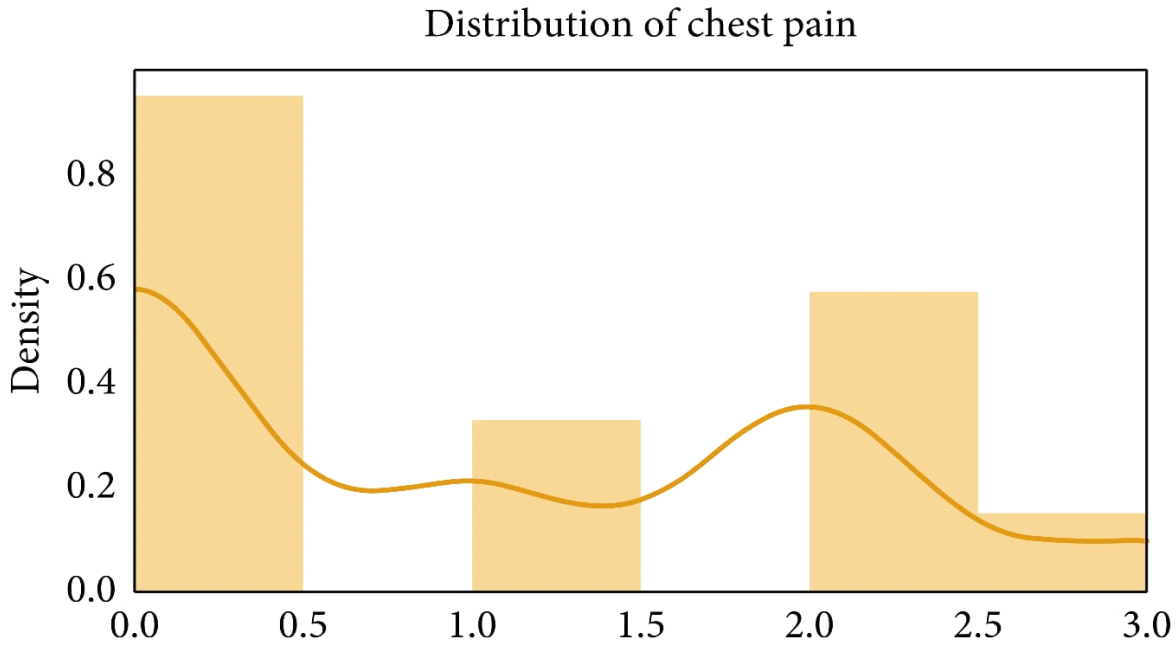
Checking the Skewness of the Data

For checking the attribute values and determining the skewness of the data (the asymmetry of a distribution), many distribution plots are plotted so that some interpretation of the data can be seen. Different plots are shown, so an overview of the data could be analyzed. The distribution of age and sex, the distribution of chest pain and trestbps, the distribution of cholesterol and fasting blood, the distribution of ecg resting electrode and thalach, the distribution of exang and oldpeak, the distribution of slope and ca, and the distribution of thal and target all are analyzed and the conclusion.


Distribution of age
Density
0.05
0.04
0.03
0.02
0.01
0.00
30
40
50
60
70


Distribution of sex
Density
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

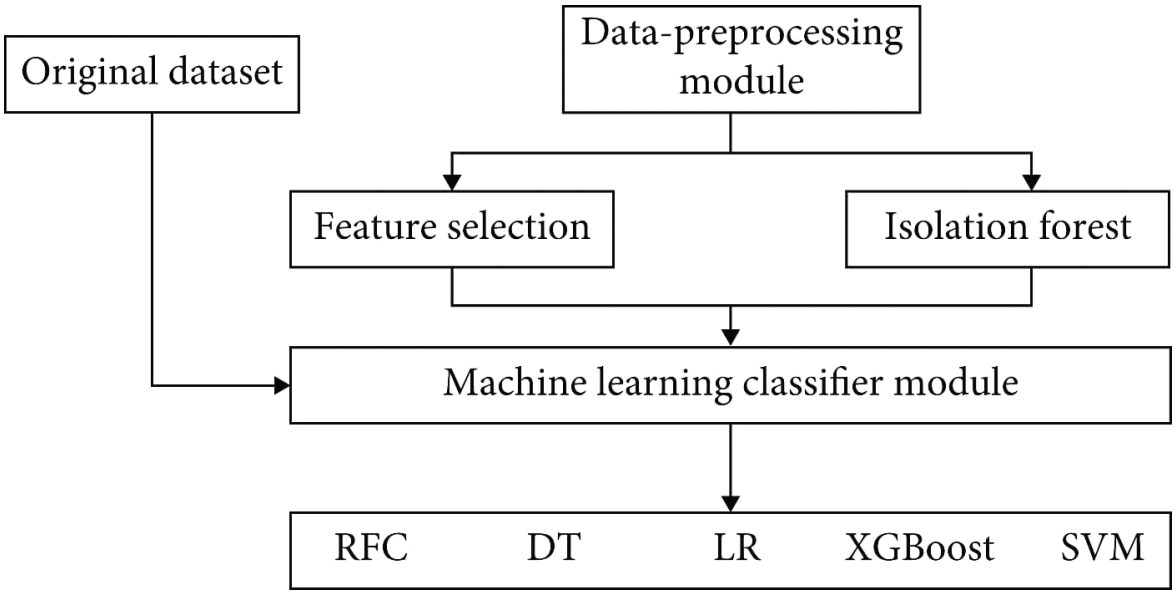




By analyzing the distribution plots, it is visible that thal and fasting blood sugar is not uniformly distributed and they needed to be handled; otherwise, it will result in overfitting or underfitting of the data.

Machine Learning Classifiers Proposed

The proposed approach was applied to the dataset in which firstly the dataset was properly analyzed and then different machine learning algorithms consisting of linear model selection in which Logistic Regression was used. For focusing on neighbor selection technique KNeighbors Classifier was used, then tree-based technique like DecisionTree Classifier was used, and then a very popular and most popular technique of ensemble methods RandomForest Classifier was used. Also for checking the high dimensionality of the data and handling it, Support Vector Machine was used. Another approach which also works on ensemble method and Decision Tree method combination is XGBoost classifier



Deep Learning Pseudocode:

- Dataset of training
- Dataset of testing
- Checking the shape/features of the input
- The procedure of initiating the sequential layer
- Adding dense layers with dropout layers and ReLU activation functions
- Adding a last dense layer with one output and binary activation function
- End repeat

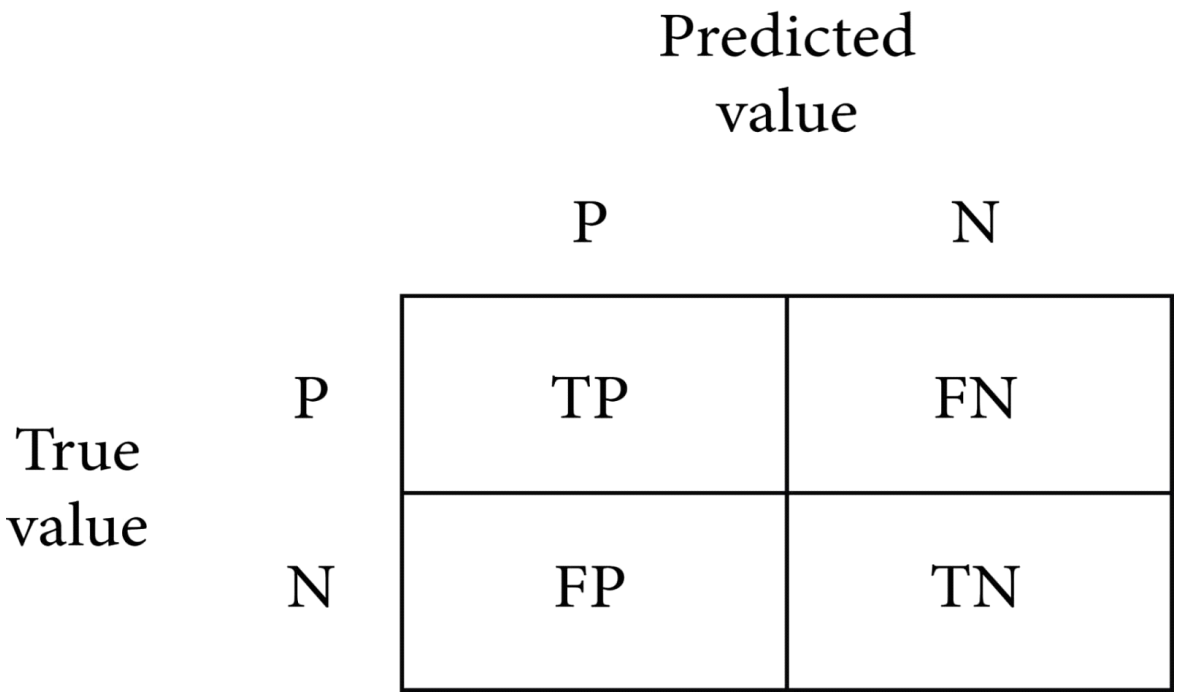
- L (output)
- End procedure

Deep Learning Proposed:

There are two ways a deep learning approach can be applied. One is using a sequential model and another is a functional deep learning approach. In this particular research, the first one is used. A sequential model with a fully connected dense layer is used, with the flatten and dropout layers to prevent the overfitting and the results are compared of the machine learning and deep learning and variations in the learning including computational time and accuracy can be analyzed and can be seen in the figures further discussed in the Results section.

Evaluation Process Used:

For the evaluation process, confusion matrix, accuracy score, precision, recall, sensitivity, and F1 score are used. A confusion matrix is a table-like structure in which there are true values and predicted values, called true positive and true negative. It is defined in four parts: the first one is true positive (TP) in which the values are identified as true and, in reality, it was true also. The second one is false positive (FP) in which the values identified are false but are identified as true. The third one is false negative (FN) in which the value was true but was identified as negative. The fourth one is true negative (TN) in which the value was negative and was truly identified as negative.



P=positive, N=negative, TP=true positive, FN=false negative, FP=false positive, TN=true negative.

Then for checking how well a model is performing, an accuracy score is used. It is defined as the true positive values plus true negative values divided by true positive plus true negative plus false positive plus false negative. The formula is

accuracy = TP+TN/TP+TN+FP+FN.

After accuracy there is specificity which is the proportion of true negative cases that were classified as negative; thus, it is a measure of how well a classifier identifies negative cases. It is also known as the true negative rate. The formula is

Specificity = TN/TN+FP.

Then there is sensitivity in which the proportion of actual positive cases got predicted as positive (or true positive). Sensitivity is also termed as recall. In other words, an unhealthy person got predicted as unhealthy. The formula is

Sensitivity = TP/TP+FN

Results:

By applying different machine learning algorithms and then using deep learning to see what difference comes when it is applied to the data, three approaches were used. In the first approach, normal dataset which is acquired is directly used for classification, and in the second approach, the data with feature selection are taken care of and there is no outliers detection. The results which are achieved are quite promising and then in the third approach the dataset was normalized taking care of the outliers and feature selection; the results achieved are much better than the previous techniques, and when compared with other research accuracies, our results are quite promising.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	heartpred
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

Check the null values in dataset

```
heart.isnull().sum()
age          0
sex          0
cp           0
trestbps     0
chol         0
fbs          0
restecg      0
thalach      0
exang        0
oldpeak      0
slope        0
ca           4
thal         2
heartpred    0
```

Preprocess Data and handle missing values using simple mean imputation methods

```
heart["thal"] = heart["thal"].fillna(heart["thal"].median())
heart["ca"] = heart["ca"].fillna(heart["ca"].median())
print(heart.describe())
```

```
              age         sex          cp    trestbps        chol
fbs  \
count  303.000000  303.000000  303.000000  303.000000  303.000000
303.000000
mean    54.438944    0.679868    3.158416  131.689769  246.693069
0.148515
std      9.038662    0.467299    0.960126   17.599748   51.776918
0.356198
min     29.000000    0.000000    1.000000   94.000000  126.000000
0.000000
25%     48.000000    0.000000    3.000000  120.000000  211.000000
0.000000
50%     56.000000    1.000000    3.000000  130.000000  241.000000
0.000000
75%     61.000000    1.000000    4.000000  140.000000  275.000000
0.000000
max     77.000000    1.000000    4.000000  200.000000  564.000000
1.000000
```

```
           restecg     thalach       exang     oldpeak       slope
ca  \
count  303.000000  303.000000  303.000000  303.000000  303.000000
303.000000
mean     0.990099  149.607261    0.326733    1.039604    1.600660
0.663366
std      0.994971   22.875003    0.469794    1.161075    0.616226
0.934375
min      0.000000   71.000000    0.000000    0.000000    1.000000
0.000000
25%      0.000000  133.500000    0.000000    0.000000    1.000000
0.000000
50%      1.000000  153.000000    0.000000    0.800000    2.000000
0.000000
75%      2.000000  166.000000    1.000000    1.600000    2.000000
1.000000
max      2.000000  202.000000    1.000000    6.200000    3.000000
3.000000

             thal   heartpred
count  303.000000  303.000000
mean     4.722772    0.937294
std      1.938383    1.228536
min      3.000000    0.000000
25%      3.000000    0.000000
50%      3.000000    0.000000

 75%      7.000000    2.000000
 max      7.000000    4.000000
```

Let's find the ranges of each feature by disease type

Age

```
print("Minimum age to Maximum age per disease type")
heart.groupby(["heartpred", ])["age"].min().astype(str) + ', ' +
heart.groupby(["heartpred", ])["age"].max().astype(str)
```

```
Minimum age to Maximum age per disease type
heartpred
0    29.0, 76.0
1    35.0, 70.0
2    42.0, 69.0
3    39.0, 70.0
4    38.0, 77.0
Name: age, dtype: object
```

```
print("Mean age per disease type")
heart.groupby(["heartpred", ])["age"].mean()
```

```
Mean age per disease type
heartpred
0    52.585366
1    55.381818
2    58.027778
3    56.000000
4    59.692308
Name: age, dtype: float64
```

Sex

```
print("Count each sex per heart disease type")
heart.groupby(["heartpred", "sex"])["age"].count()
```

```
Count each sex per heart disease type
heartpred  sex
0          0.0     72
           1.0     92
1          0.0      9
           1.0     46
2          0.0      7
           1.0     29
3          0.0      7
           1.0     28
4          0.0      2
           1.0     11
Name: age, dtype: int64
```

We can see that heart disease all types can be present in men with higher probability than in women

chest_pain

```
print('Count each chest pain value per heart disease type')
heart.groupby(["heartpred", "cp"])["age"].count()
Count each chest pain value per heart disease type
heartpred  cp
0          1.0    16
           2.0    41
           3.0    68
           4.0    39
1          1.0     5
           2.0     6
           3.0     9
           4.0    35
2          1.0     1
           2.0     1
           3.0     4
           4.0    30
3          2.0     2
           3.0     4
           4.0    29
4          1.0     1
           3.0     1
           4.0    11
Name: age, dtype: int64
```

The people with chest pain = 0 often have heart disease.

blood pressure

```
print("Minimum blood pressure to Maximum  blood pressure per disease type")
heart.groupby(["heartpred"])["trestbps"].min().astype(str) + ', ' +
heart.groupby(["heartpred"])["trestbps"].max().astype(str)
Minimum blood pressure to Maximum  blood pressure per disease type
heartpred
0     94.0, 180.0
1    108.0, 192.0
2    100.0, 180.0
3    100.0, 200.0
4    112.0, 165.0
Name: trestbps, dtype: object
print("Mean resting blood pressure per disease type")
heart.groupby(["heartpred", ])["trestbps"].mean()
Mean blood pressure per disease type
heartpred
0    129.250000
1    133.254545
2    134.194444
3    135.457143
4    138.769231
Name: trestbps, dtype: float64
```

As bigger is mean blood pressure as higher is type of heart disease

```
print("Minimum serum_cholestoral to Maximum serum_cholestoral per disease
type")
heart.groupby(["heartpred"])["chol"].min().astype(str) + ', ' +
heart.groupby(["heartpred"])["chol"].max().astype(str)
Minimum serum_cholestoral to Maximum serum_cholestoral per disease type
heartpred
0    126.0, 564.0
1    149.0, 335.0
2    169.0, 409.0
3    131.0, 353.0
4    166.0, 407.0
Name: chol, dtype: object
```

serum_cholestoral

```
print("Mean serum_cholestoral per disease type")
heart.groupby(["heartpred", ])["chol"].mean()
Mean serum_cholestoral per disease type
heartpred
0    242.640244
1    249.109091
2    259.277778
3    246.457143
4    253.384615
Name: chol, dtype: float64
```

fasting_blood_sugar

```
print("Count each fasting_blood_sugar per heart disease type")
heart.groupby(["heartpred", "fbs"])["age"].count()
Count each fasting_blood_sugar per heart disease type
heartpred  fbs
0          0.0    141
           1.0     23
1          0.0     51
           1.0      4
2          0.0     27
           1.0      9
3          0.0     27
           1.0      8
4          0.0     12
           1.0      1
Name: age, dtype: int64
```

electrocardiographic results

```
print("Count each electrocardiographic per heart disease type")
heart.groupby(["heartpred", "restecg"])["age"].count()
Count each electrocardiographic per heart disease type
heartpred  restecg
0          0.0        95
           1.0         1
           2.0        68
1          0.0        23
           2.0        32
2          0.0        19
           1.0         1
           2.0        16
3          0.0        12
           1.0         1
           2.0        22
4          0.0         2
           1.0         1
           2.0        10
Name: age, dtype: int64
```

max_heart_rate

```
print("Minimum max_heart_rate to Maximum max_heart_rate per disease type")
heart.groupby(["heartpred"])["thalach"].min().astype(str) + ', ' +
heart.groupby(["heartpred"])["thalach"].max().astype(str)
Minimum max_heart_rate to Maximum max_heart_rate per disease type
heartpred
0      96.0, 202.0
1      88.0, 195.0
2      71.0, 170.0
3      90.0, 173.0
4     114.0, 182.0
Name: thalach, dtype: object
print("Mean max_heart_rate per disease type")
heart.groupby(["heartpred", ])["thalach"].mean()
Mean max_heart_rate per disease type
heartpred
0    158.378049
1    145.927273
2    135.583333
3    132.057143
4    140.615385
Name: thalach, dtype: float64
```

induced_angina

```
print("Count  induced_angina per heart disease type")
heart.groupby(["heartpred", "exang"])["age"].count()
Count  induced_angina per heart disease type
heartpred  exang
0          0.0       141
           1.0        23
1          0.0        30
           1.0        25
2          0.0        14
           1.0        22
3          0.0        12
           1.0        23
4          0.0         7
           1.0         6
Name: age, dtype: int64
```

ST_depression

```
print("Count  mean ST_depression per heart disease type")
heart.groupby(["heartpred"])["oldpeak"].mean()
Count  mean ST_depression per heart disease type
heartpred
0    0.586585
1    1.005455
2    1.780556
3    1.962857
4    2.361538
Name: oldpeak, dtype: float64
```

slope

```
print("Count slope per heart disease type")
heart.groupby(["heartpred", "slope"])["age"].count()
Count slope per heart disease type
heartpred  slope
0          1.0      106
           2.0       49
           3.0        9
1          1.0       22
           2.0       31
           3.0        2
2          1.0        7
           2.0       26
           3.0        3
3          1.0        6
           2.0       24
           3.0        5
4          1.0        1
           2.0       10
           3.0        2
Name: age, dtype: int64
```

vessels

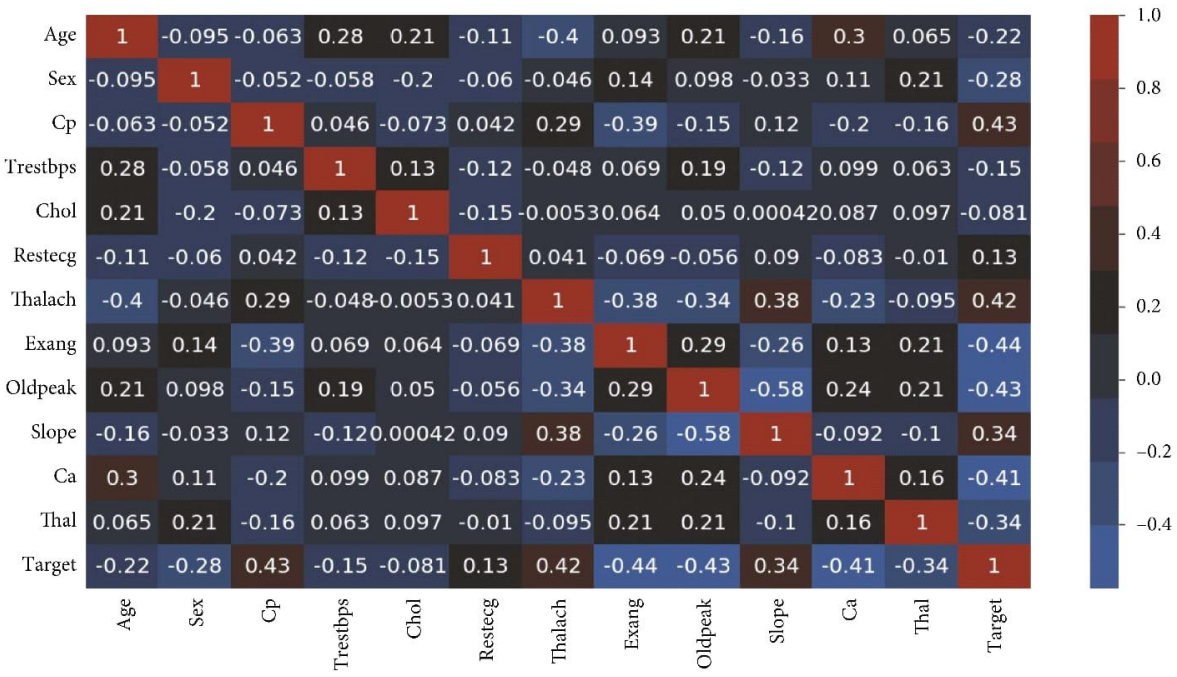
```
print("Count  mean vessels per heart disease type")
heart.groupby(["heartpred"])["ca"].mean()
Count  mean vessels per heart disease type
heartpred
0    0.268293
1    0.727273
2    1.222222
3    1.457143
4    1.692308
Name: ca, dtype: float64
```

Thal

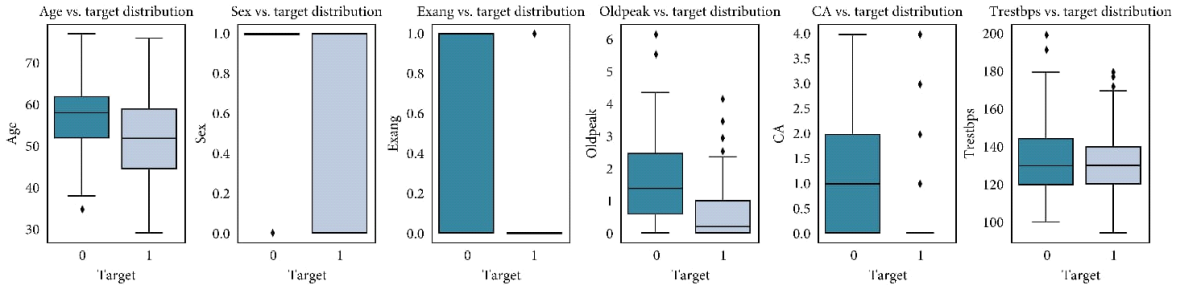
```
print("Count  mean thal per heart disease type")
heart.groupby(["heartpred"])["thal"].mean()
Count  mean thal per heart disease type
heartpred
0    3.792683
1    5.345455
2    5.944444
3    6.285714
4    6.230769
Name: thal, dtype: float64
```

Using the First Approach (without Doing Feature Selection and Outliers Detection):

As can be seen in the dataset is not normalized, there is no equal distribution of the target class, it can further be seen when a correlation heatmap is plotted, and there are so many negative values.



So, even if the feature selection is done, still, we have outliers.



By applying the first approach, the accuracy achieved by the Random Forest is 76.7%, Logistic Regression is 83.64%, KNeighbors is 82.27%, Support Vector Machine is 84.09%, Decision Tree is 75.0%, and XGBoost is 70.0%. SVM is having the highest accuracy here which is achieved by using the cross-validation and grid search for finding the best parameters or in other words doing the hyperparameter tuning. Then after machine learning, deep learning is applied by using the sequential model approach. In the model, 128 neurons are used and the activation function used is ReLU, and in the output layer which is a single class prediction problem, the sigmoid activation function is

used, with loss as binary cross-entropy and gradient descent optimizer as Adam. The accuracy achieved is 76.7%.

Using the Second Approach (Doing Feature Selection and No Outliers Detection) :

After selecting the features (feature selection) and scaling the data as there are outliers, the robust standard scalar is used; it is used when the dataset is having certain outliers. In the second approach, the accuracy achieved by Random Forest is 88%, the Logistic Regression is 85.9%, KNeighbors is 79.69%, Support Vector Machine is 84.26%, the Decision Tree is 76.35%, and XGBoost is 71.1%. Here the Random Forest is the clear winner with a precision of 88.4% and an F1 score of 86.5%. Then deep learning is applied with the same parameters before and the accuracy achieved is 86.8%, and the evaluation accuracy is 81.9%, which is better than the first approach.

Using the Third Approach (by Doing Feature Selection and Also Outliers Detection):

In this approach, the dataset is normalized and the feature selection is done and also the outliers are handled using the Isolation Forest. The correlation comparison can be seen in Figure 10. The accuracy of the Random Forest is 80.3%, Logistic Regression is 83.31%, KNeighbors is 84.86%, Support Vector Machine is 83.29%, Decision Tree is 82.33%, and XGBoost is 71.4%. Here the winner is KNeighbors with a precision of 77.7% and a specificity of 80%. A lot of tips and tricks for selecting different algorithms are shown by Garate-Escamila et al. [38]. Using deep learning in the third approach, the accuracy achieved is 94.2%. So, the maximum accuracy achieved by the machine learning model is KNeighbors (83.29%) in the third approach, and, for deep learning, the maximum accuracy achieved is 81.9%. Thus, the conclusion can be drawn here that, for this dataset, the deep learning algorithm achieved 94.2 percent accuracy which is greater than the machine learning models. We also made a comparison with another research of the deep learning by Ramprakash et al. [39] in which they achieved 84% accuracy and Das et al. [33] achieved 92.7 percent accuracy. So our algorithm produced greater accuracy and more promising than other approaches. The comparison of different classifiers of ML and DL

Common Symptoms of Heart prediction:

- Machine learning allows building models to quickly analyze data and deliver results, leveraging the historical and real-time data, with machine learning that will help healthcare service providers to make better decisions on patient's disease diagnosis
- By analyzing the data we can predict the occurrence of the disease in our project.
- This intelligent system for disease prediction plays a major role in controlling the disease and maintaining the good health status of people by predicting accurate disease risk.
- Machine learning algorithms can also help provide vital statistics, real-time data and advanced analytics in terms of the patient's disease, lab test results, blood pressure, family history, clinical trial data, etc., to doctors.
- Heart disease predictor is an offline platform designed and developed to explore the path of machine learning.
- The goal is to predict the health of the patient from collective data to be able to detect configurations at risk for the patient, and therefore, in cases requiring emergency medical assistance, alert the appropriate medical staff of the situation of the latter.
- We initially have a dataset collecting information of many patients with which we can conclude the results into a complete form and can predict data precisely.
- The results of the predictions, derived from the predictive models generated by machine learning, will be presented through several distinct graphical interfaces according to the datasets considered. We will then bring criticism as to the scope of our results.

Types of Heart Disease:

- **Healthy Heart *:***

- Atrium.
- Plural atria.

Unhealthy heart:

- Coronary artery Disease.
- Heart Arrhythmias.
- Heart Failure.
- Heart Valve Disease.
- Cardiomyopathy

Eating a diet high in saturated fats, trans fat, and cholesterol has been linked to heart disease and related conditions, such as atherosclerosis. Also, too much salt (sodium) in the diet can raise blood pressure. Not getting enough physical activity can lead to heart disease.

Coronary artery disease

- Also called: CAD, atherosclerotic heart disease
- OverviewSymptomsTreatmentsNewsSpecialists
- Damage or disease in the heart's major blood vessels.
- The usual cause is the build-up of plaque. This causes coronary arteries to narrow, limiting blood flow to the heart.

High blood pressure

- Also called: HBP, hypertension
- OverviewSymptomsTreatmentsNewsSpecialists
- A condition in which the force of the blood against the artery walls is too high.
- Usually hypertension is defined as blood pressure above 140/90, and is considered severe if the pressure is above 180/120***.***
- ***Cardiac arrest***
- OverviewSymptomsTreatmentsNewsSpecialists
- Sudden, unexpected loss of heart function, breathing and consciousness.
- In cardiac arrest, the heart abruptly stops beating. Without prompt intervention, it can result in the person's death***.***

Heart Disease Prediction System Using Machine Learning :

Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

Supervised Learning :

This study, an effective heart disease prediction system (EHDPS) is developed using neural network for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction

Data insight: As mentioned here we will be working with the heart disease detection dataset and we will be putting out interesting inferences from the data to derive some meaningful results.

EDA: Exploratory data analysis is the key step for getting meaningful results.

Feature engineering: After getting the insights from the data we have to alter the features so that they can move forward for the model building phase.

Model building: In this phase, we will be building our Machine learning model for heart disease detection.

Conclusion:

The conclusion which we found is that machine learning algorithms performed better in this analysis. Many researchers have previously suggested that we should use ML where the dataset is not that large, which is proved in this work. In this paper, we proposed three methods in which comparative analysis was done and promising results were achieved. The conclusion which we found is that machine learning algorithms performed better in this analysis. Many researchers have previously suggested that we should use ML where the dataset is not that large, which is proved in this paper. The methods which are used for comparison are confusion matrix, precision, specificity, sensitivity, and F1 score. For the 13 features which were in the dataset, KNeighbors classifier performed better in the ML approach when data preprocessing is applied.

The computational time was also reduced which is helpful when deploying a model. It was also found out that the dataset should be normalized; otherwise, the training model gets overfitted sometimes and the accuracy achieved is not sufficient when a model is evaluated for real-world data problems which can vary drastically to the dataset on which the model was trained. It was also found out that the statistical analysis is also important when a dataset is analyzed and it should have a Gaussian distribution, and then the outlier's detection is also important and a technique known as Isolation Forest is used for handling this. The difficulty which came here is that the sample size of the dataset is not large. If a large dataset is present, the results can increase very much in deep learning and ML as well. The algorithm applied by us in ANN architecture increased the accuracy which we compared with the different researchers. The dataset size can be increased and then deep learning with various other optimizations can be used and more promising results can be achieved. Machine learning and various other optimization techniques can also be used so that the evaluation results can again be increased. More different ways of normalizing the data can be used and the results can be compared. And more ways could be found where we could integrate heart-disease-trained ML and DL models with certain multimedia for the ease of patients and doctors.

REFERENCE :

"Cardiovascular Diseases (Cvds)". Who.Int, 2020, https://www.who.int/zh/news-room/fact-

sheets/detail/cardiovascular-diseases-(cvds).

"Logistic Regression". En.Wikipedia.Org, 2020,

https://en.wikipedia.org/wiki/Logistic_regression.

"Understanding Random Forest". Medium, 2020,

https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

"Explanation Of The Decision Tree Model". Webfocusinfocenter.Informationbuilders.Com,

2020,

https://webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL_rstat/source/DecisionTree

47.htm[5] "Xgboost Algorithm: Long May She Reign!". Medium, 2020,

https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-

may-rein-edd9f99be63d.

"Neural Network Definition". Investopedia, 2020,

https://www.investopedia.com/terms/n/neuralnetwork.asp.

Perret-Guillaume C, et al. "Heart Rate As A Risk Factor For Cardiovascular Disease. -

Pubmed - NCBI". Ncbi.Nlm.Nih.Gov, 2020, https://www.ncbi.nlm.nih.gov/pubmed/19615487.

"Both Blood Pressure Numbers May Predict Heart Disease". Medicalnewstoday.Com, 2020,

https://www.medicalnewstoday.com/articles/325861.

"Angina (Chest Pain)". Www.Heart.Org, 2020, https://www.heart.org/en/health-

topics/heart-attack/angina-chest-pain.

2020, http://cooleysanemia.org/updates/Cardiac.pdf. Accessed 14 Mar 2020.

Saeed, M., Hetts, S., English, J., & Wilson, M. (2012, January). MR fluoroscopy in

vascular and cardiac interventions (review). Retrieved March 14, 2020, from

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3275732/

"What Can I Do To Avoid A Heart Attack Or A Stroke?". World Health Organization,

2020, https://www.who.int/features/qa/27/en/.