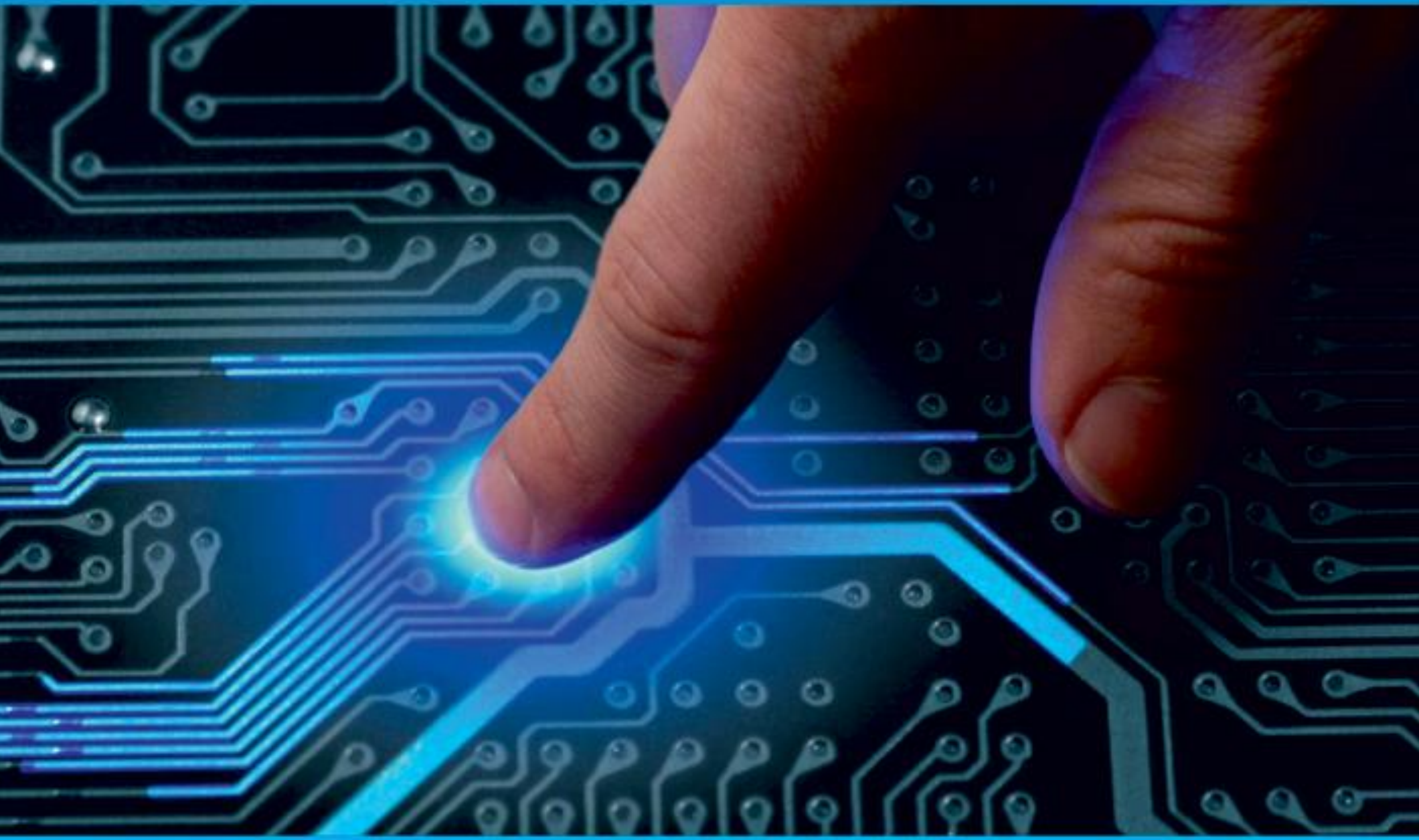
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

9940 572 462 6381 907 438 ijircce@gmail.com www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205176 |

Heart Disease Prediction Using Machine Learning

Priya Nagbhidkar, Prof Pushpa Tandekar, Prof Lowlesh Yadav

Department of Computer Science and Engineering, Shri Sai College of Engineering and Technology, Chandrapur, India

Department of Computer Science and Engineering, Shri Sai College of Engineering and Technology, Chandrapur, India

Department of Computer Science and Engineering, Shri Sai College of Engineering and Technology, Chandrapur, India

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.

I. INTRODUCTION

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.. There 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 model to 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.

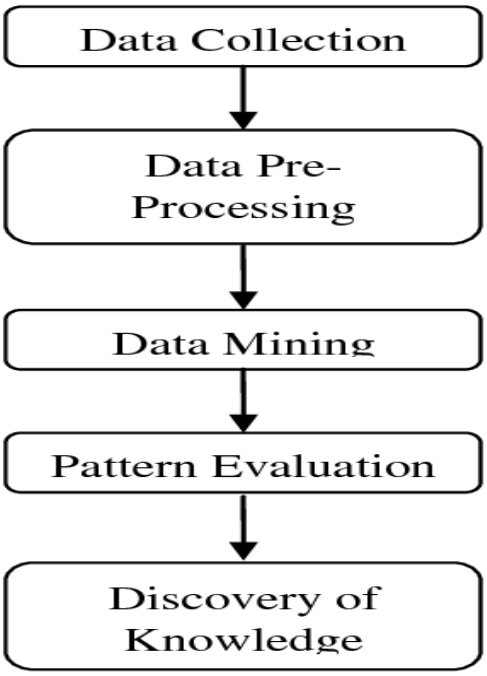
II. OBJECTIVES

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.

III. METHODOLOGY FLOW CHART

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. 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 of 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.

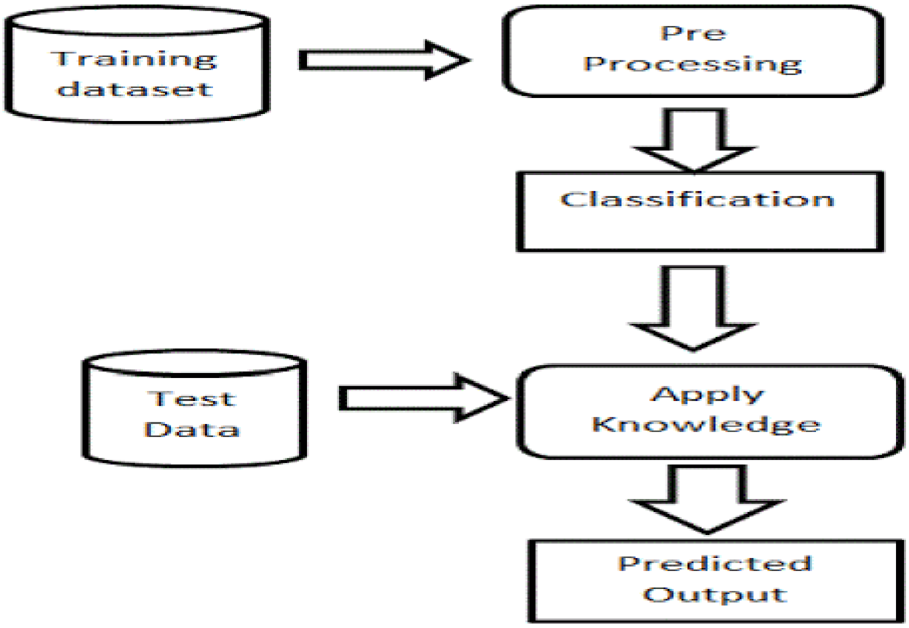
Attribute Description :

- 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.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205176 |



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 without 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.

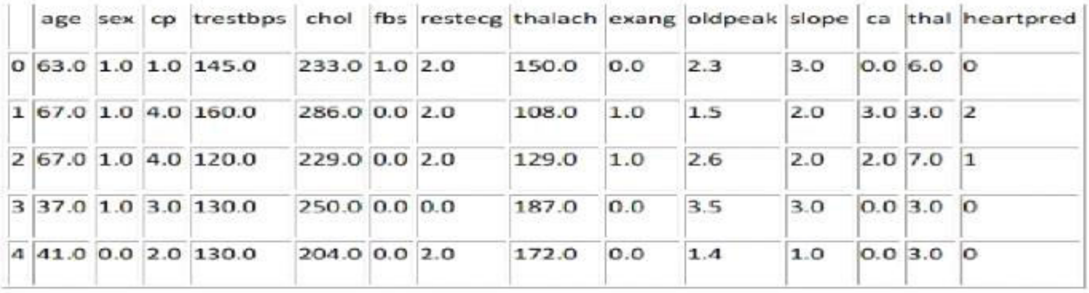
• We see that the heart disease occurred 54.46% of the time in the dataset, whilst 45.54% was the no heart disease.

• 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

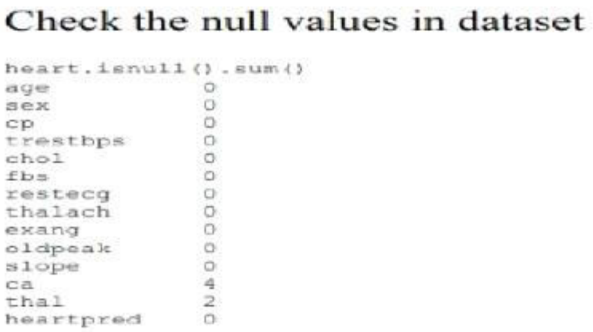
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205176 |

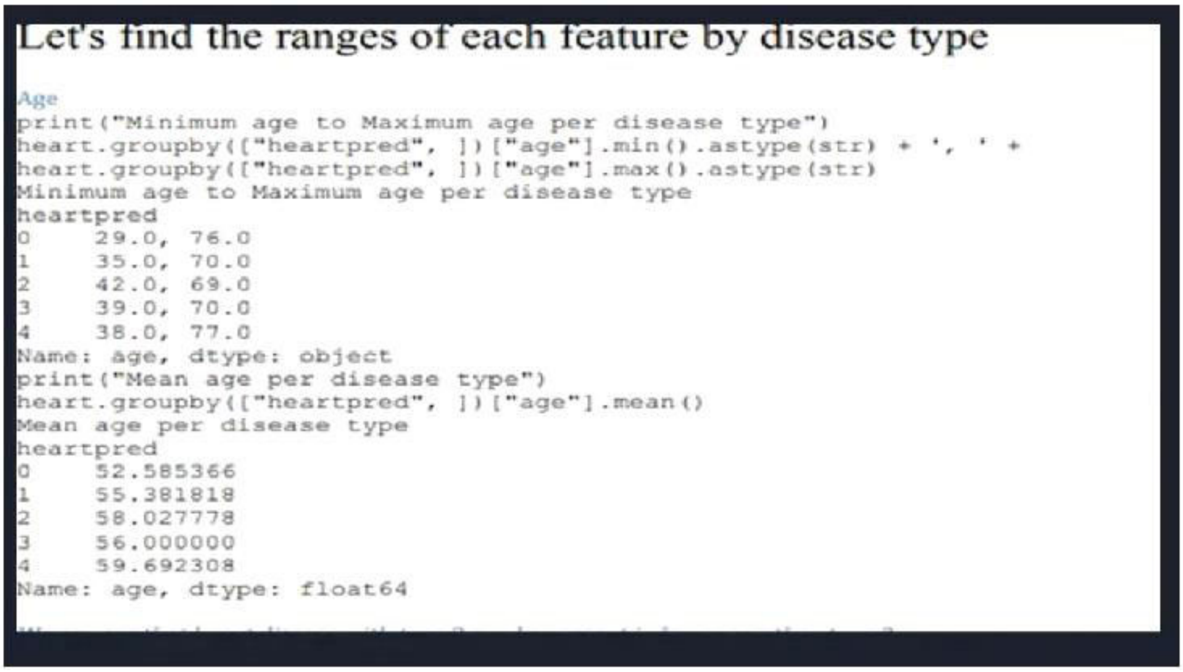


	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	heartpred
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0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

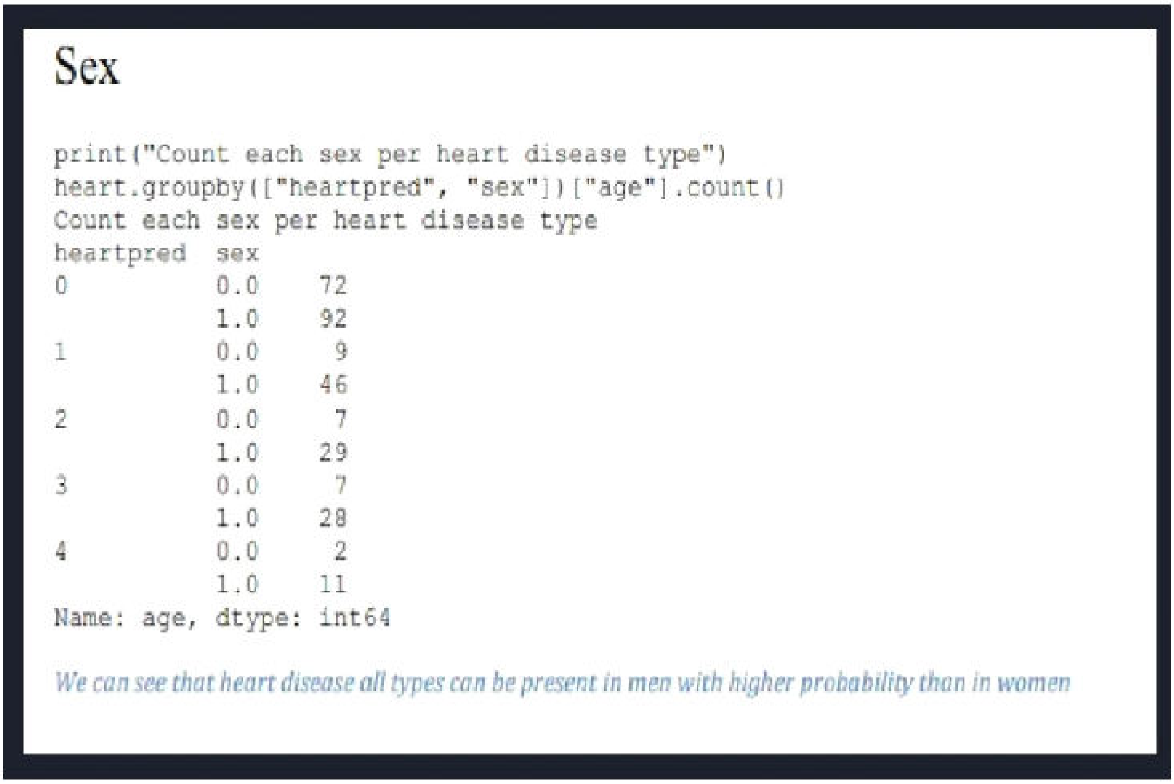
```
heart.isnull().sum()
age          0
sex          0
cp           0
trestbps     0
chol         0
fbs          0
restecg      0
thalach      0
exang        0
oldpeak      0
slope        0
ca           4
thal         2
heartpred    0
```



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

Age

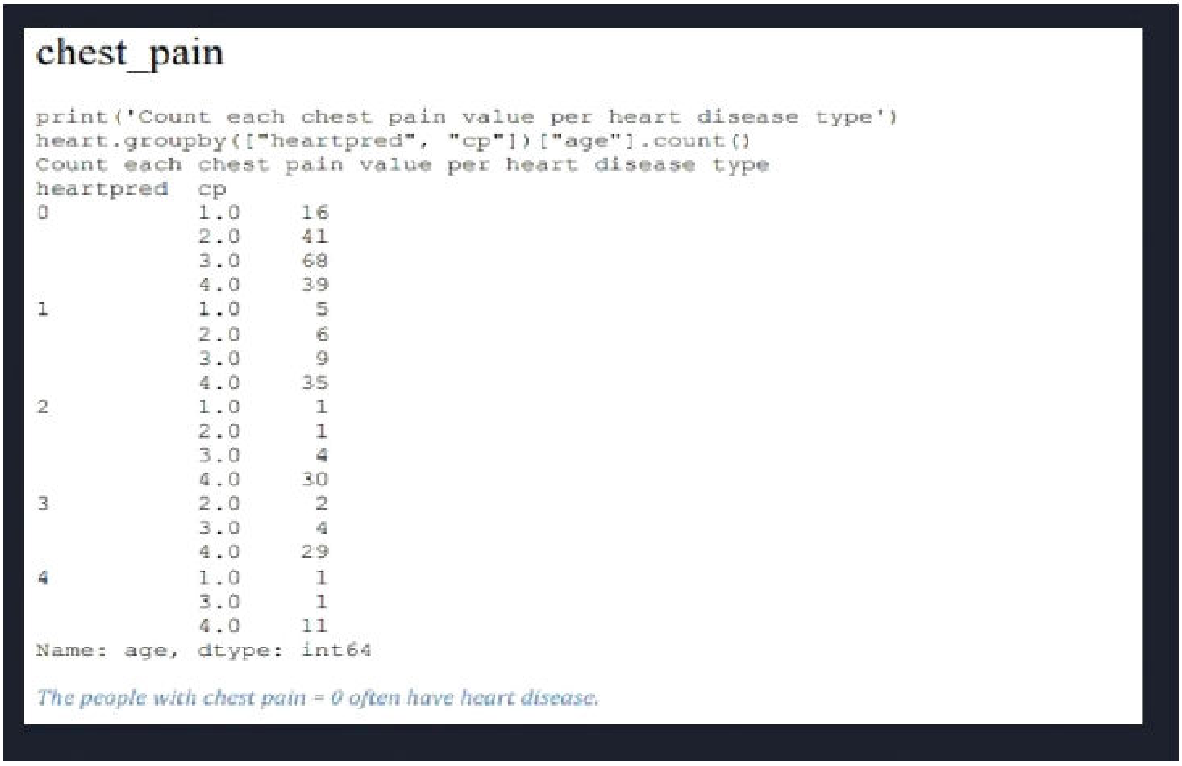
```
print("Minimum age to Maximum age per disease type")
heart.groupby(["heartpred", ])["age"].min().astype(str) + ', ' +
heart.groupby(["heartpred", ])["age"].max().astype(str)
Minimum age to Maximum age per disease type
heartpred
0     29.0, 76.0
1     35.0, 70.0
2     42.0, 69.0
3     39.0, 70.0
4     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
```

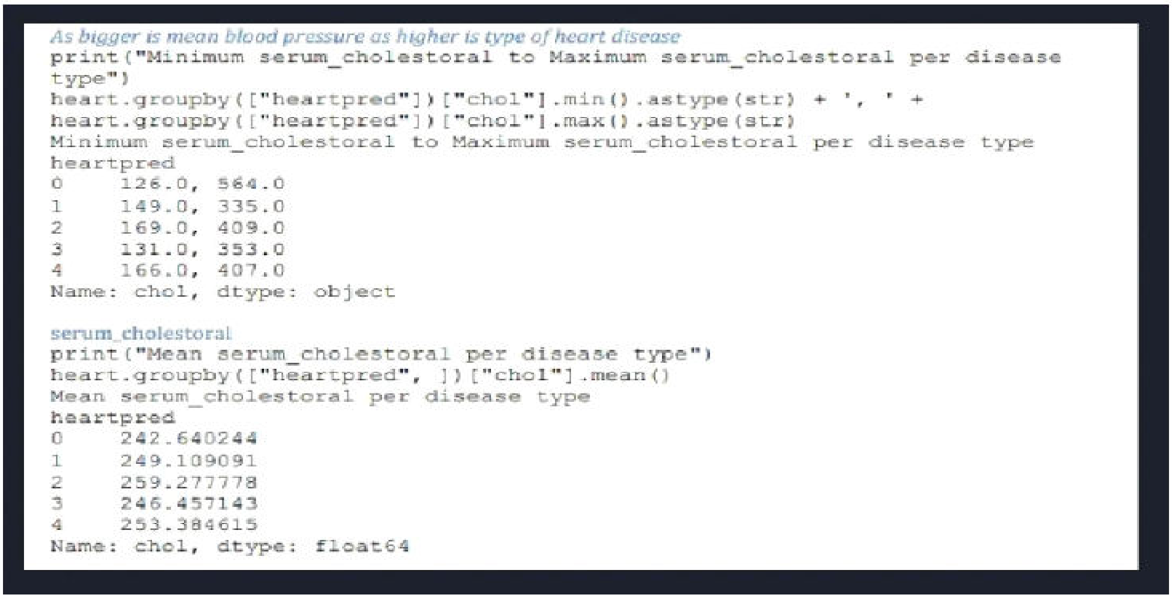
We can see that heart disease all types can be present in men with higher probability than in women



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.

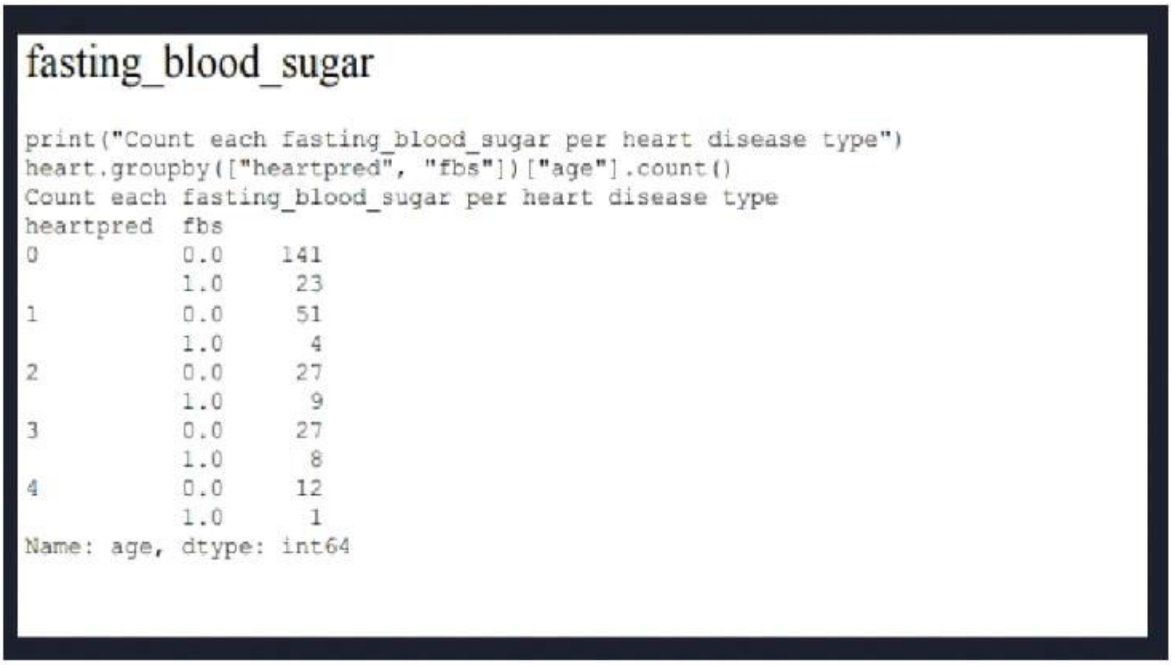


As bigger is mean blood pressure as higher is type of heart disease

```
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

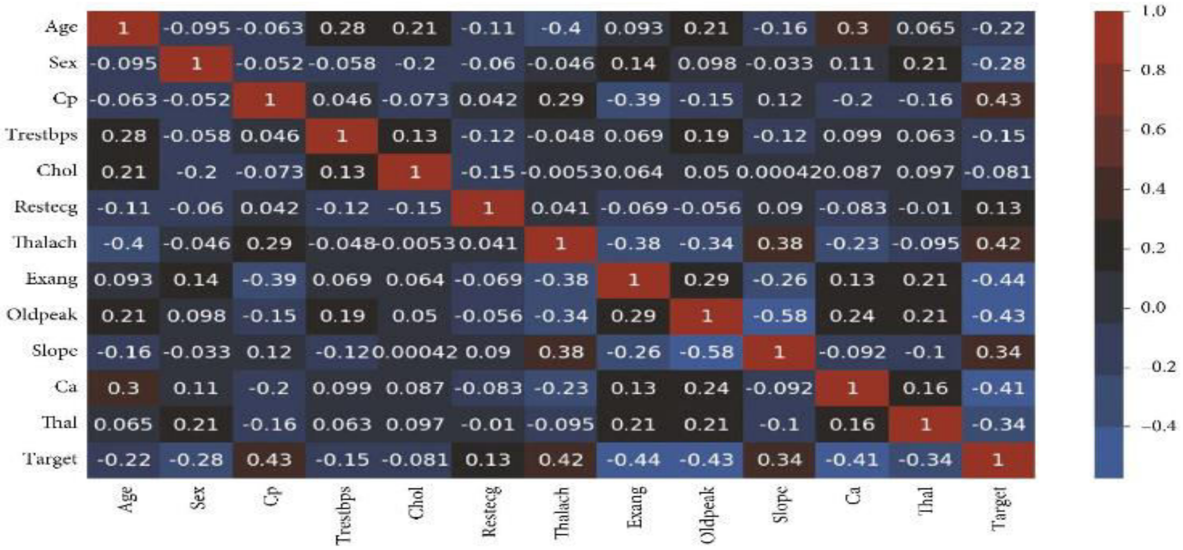
```
print("Mean serum_cholestoral per disease type")
heart.groupby(["heartpred", ])["chol"].mean()
Mean serum_cholestoral per disease type
heartpred
0     242.640244
1     249.109091
2     259.277778
3     246.457143
4     253.384615
Name: chol, dtype: float64
```

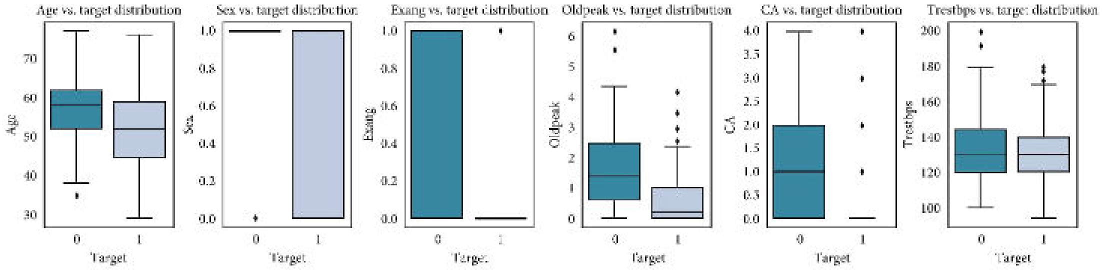


fasting_blood_sugar

```
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
```

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.





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 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%. 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. So our algorithm produced greater accuracy and more promising than other approaches. The comparison of different classifiers of ML and DL

IV. 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.

V. 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.

VI. FUTURE SCOPE

The proposed model requires an efficient processor with good memory configuration to implement it in real time. The proposed model has wide area of application like grid computing, cloud computing, robotic modeling, etc. To increase the performance of our classifier in future, we will work on ensembling two algorithms called Random Forest and Adaboost. By ensembling these two algorithms we will achieve high performance.

REFERENCES

[1] F. Przerwa, A. Kukowka, K. Kotrych, and I. Uzar, "Probiotics in prevention and treatment of cardiovascular diseases," Herba Polonica, vol. 67,no. 4, pp. 77–85, 2021.

[2] J. Rehm, C. T. Sempos, and M. Trevisan, "Average volume of alcohol consumption, patterns of drinking and risk of coronary heart disease-a review," Journal of Cardiovascular Risk, vol. 10, no. 1, pp. 15–20, 2003.

[3] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives bayesian," in 2019 3rd International conference on trends in electronics and informatics (ICOEI). IEEE, 2019, pp. 292–297.

[4] R. Sun, M. Liu, L. Lu, Y. Zheng, and P. Zhang, "Congenital heart disease: causes, diagnosis, symptoms, and treatments," Cell biochemistry and biophysics, vol. 72, no. 3, pp. 857–860, 2015.

[5] W. P. Castelli, R. D. Abbott, and P. M. McNamara, "Summary estimates of cholesterol used to predict coronary heart disease." Circulation, vol. 67, no. 4, pp. 730–734, 1983.

[6] S. Anitha and N. Sridevi, "Heart disease prediction using data mining techniques," Journal of analysis and Computation, 2019.

[7] J.-K. Kim, J.-S. Lee, D.-K. Park, Y.-S. Lim, Y.-H. Lee, and E.-Y. Jung, "Adaptive mining prediction model for content recommendation to coronary heart disease patients," Cluster computing, vol. 17, no. 3, pp. 881–891, 2014.

[8] H. Sharma and M. Rizvi, "Prediction of heart disease using machine learning algorithms: A survey," International Journal on Recent andInnovation Trends in Computing and Communication, vol. 5, no. 8, pp. 99–104, 2017

[9]. Lowlesh Yadav and Asha Ambhaikar, "IOHT based Tele-Healthcare Support System for Feasibility and perfor-mance analysis," Journal of Electrical Systems, vol. 20, no. 3s, pp. 844–850, Apr. 2024, doi: 10.52783/jes.1382.

[10] L. Yadav and A. Ambhaikar, "Feasibility and Deployment Challenges of Data Analysis in Tele-Healthcare System," 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), Raipur, India, 2023, pp. 1-5, doi: 10.1109/ICAIIHI57871.2023.10489389.

[11] L. Yadav and A. Ambhaikar, "Approach Towards Development of Portable Multi-Model Tele-Healthcare System," 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), Raipur, India, 2023, pp. 1-6, doi: 10.1109/ICAIIHI57871.2023.10489468.

[12] Lowlesh Yadav and Asha Ambhaikar, Exploring Portable Multi-Modal Telehealth Solutions: A Development Approach. International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), vol. 11, no. 10, pp. 873–879, Mar. 2024.11(10), 873–879, DOI: 10.13140/RG.2.2.15400.99846.

[13] Lowlesh Yadav, Predictive Acknowledgement using TRE System to reduce cost and Bandwidth, March 2019. International Journal of Research in Electronics and Computer Engineering (IJRECE), VOL. 7 ISSUE 1 (JANUARY-MARCH 2019) ISSN: 2393-9028 (PRINT) | ISSN: 2348-2281 (ONLINE).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details