ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 6, June 2022

HEALTH CARE - HEART ATTACK POSSIBILITY PREDICTION

Abirami S
Computer Science & Engineering
Salem College of Engineering and Technology,
Salem, India.
toabis.13@gmail.com

Lavanya S
Computer Science & Engineering
Salem College of Engineering and Technology,
Salem, India.
lavanya95252@gmail.com

Prathima S
Computer Science & Engineering
Salem College of Engineering and Technology,
Salem, India.
prathubabi@gmail.com

Archana M
Computer Science & Engineering
Salem College of Engineering and Technology, Salem,
India.
arch.rukmani2404@gmail.com

***Abstract*—** The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data. Some advanced data mining techniques are used. In this study, a Heart Attack Possibility Prediction is developed using Naïve Bayes and Decision Tree algorithms for predicting the risk level of heart disease. The system uses 13 parameters such as age, sex, blood pressure, cholesterol and obesity for prediction. This system predicts the likelihood of patients getting heart disease. It enables significant knowledge. Eg: relationships between medical factors related to heart disease and patterns, to be established. We have employed the multilayer perception neural network with back propagation as the training algorithm. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

.

IEEE Keywords— Prediction, records, dataset, decision.

I.INTRODUCTION

Heart disease describes a range of conditions that affect your heart. Today, cardiovascular diseases are the leading cause of death 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 (irregular heartbeat), swollen legs, and in some cases weight gain occurring quite fast; it can be 1-2 kg dailyz. 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 patient. But as time is passing, a lot of research data and patients records of hospitals are available. *ere 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.Nowadays 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 models to diagnose the disease and classify or predict the results. [1] emphasized that people who are visually impaired have a hard time navigating their surroundings, recognizing objects, and avoiding hazards on their own since they do not know what is going on in their immediate 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 which sensitivity is achieved as 93.3 percent and specificity is achieved as 63.5 percent. Diagnosis and prediction of Heart Disease and Blood Pressure along with other attributes using the aid neural networks was introduced by R. Subramanian et al.. A deep Neural Network was Built incorporating the given attributes related to the disease which were able to produce a output which was carried out by the output perceptron and almost included 120 hidden layers which is the basic and most variant technique of ensuring a accurate result of having heart disease if we use the model for Test Dataset. supervised network has been advised for diagnosis of heart diseases [16]. [3] discussed about an eye blinking sensor. Nowadays heart attack patients are increasing day by day."Though it is tough to save the heart attack patients, we can increase the statistics of saving the life of patients & the life of others whom they are responsible for. [5] discussed about a system, GSM based AMR has low infrastructure cost and it reduces man power. The system is fully automatic, hence the probability of error is reduced. The data is highly secured and it not only solve the problem of traditional meter reading system but also provides additional features such as power disconnection, reconnection and the concept of power management.

OVERVIEW OF THE PROJECT

It is implemented in Python and different classification algorithms are used. Below is a brief description of the general approach that I employed:

MODULES:-

- Data Cleaning and pre processing
- Exploratory Data Analysis

II. DATA SOURCE

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and in accordance with other medical conditions . Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to Cardiovascular diseases. We take a data source which is comprised of medical history of 303 different patient of different age groups. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 303 patients that helps us detecting if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 13 columns, where each row corresponds to a single record. All attributes are listed in 'Table 1'.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.0
mean	52.496970	0.563636	1.375758	129.303030	242.230303	0.139394	0.593939	150.466667	0.139394	0.963030	1.593939	0.363636	2.121212	1.0
std	9.650661	0.497444	0.952222	16.169613	53.952872	0.347412	0.504818	19.174276	0.347412	0.780883	0.593636	0.848894	0.465752	0.0
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	96.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.0
25%	44.000000	0.000000	1.000000	120.000000	208.000000	0.000000	0.000000	149.000000	0.000000	0.000000	1.000000	0.000000	2.000000	1.0
50%	52.000000	1.000000	2.000000	130.000000	234.000000	0.000000	1.000000	161.000000	0.000000	0.200000	2.000000	0.000000	2.000000	1.0
75%	59.000000	1.000000	2.000000	140.000000	267.000000	0.000000	1.000000	172.000000	0.000000	1.000000	2.000000	0.000000	2.000000	1.0
max	76.000000	1.000000	3.000000	180.000000	564.000000	1.000000	2.000000	202.000000	1.000000	4.200000	2.000000	4.000000	3.000000	1.0

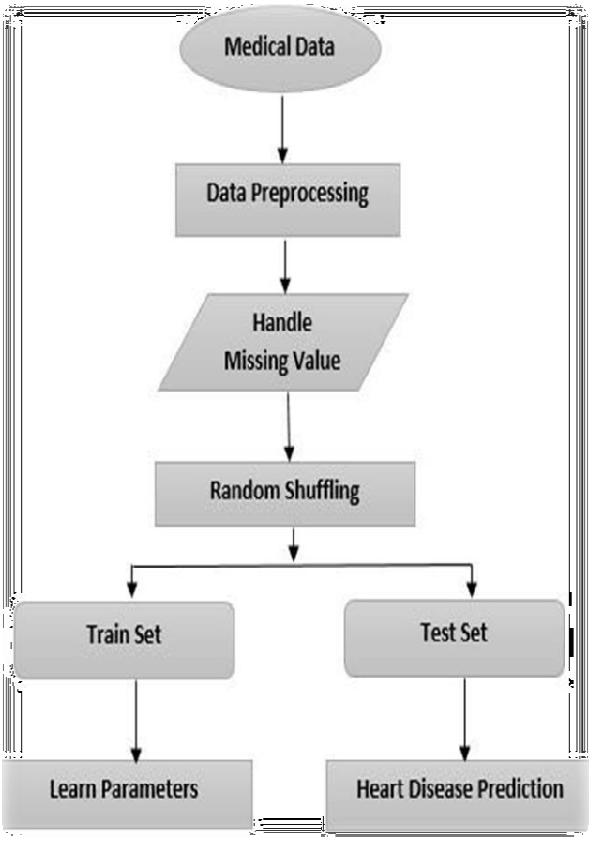


Fig: Proposed Model

III. METHODOLOGY

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease.This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model.. The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the preprocessing stage where we explore the data.Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier. Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective Heart Disease Prediction System.

DATACLEANING AND PROCESSING:

Here I checked and dealt with missing and duplicate variables from the data set as these can grossly affect the performance of different machine learning algorithms.

```
In [6]: # Display Missing Values

        print(ht.isna().sum())
```

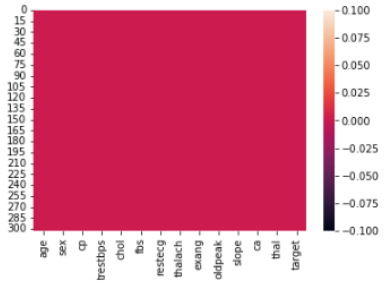
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

EXPLORATORY DATA ANALYSIS:

Here I wanted to gain important statistical insights from the data and the things that I checked for were the distributions of the different attributes, correlations of the attributes with each other and the target variable and I calculated important odds and proportions for the categorical attributes.

Exploratory Data Analysis (EDA)

```
In [7]: # Display Missing Values using HeatMap

        sns.heatmap(ht.isnull())
        plt.show()
```



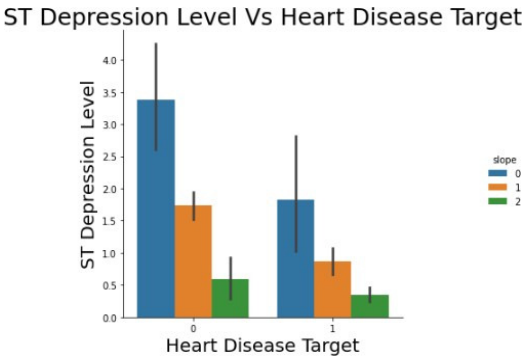
Filtering Data by Positive and Negative Heart Disease Patient

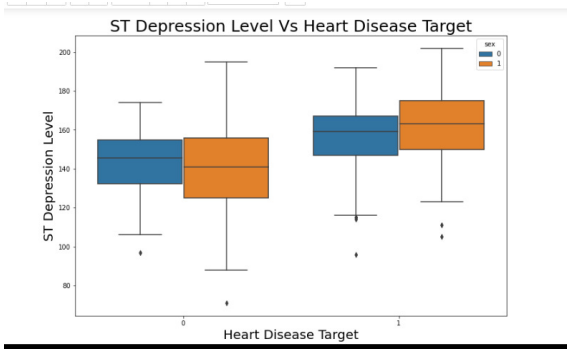
```
In [14]: # Filtering Data by Positive Heart Disease Patient

         pos_ht=ht[ht["target"]==1]
         pos_ht.describe()
```

Out[14]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
count	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.000000	165.0
mean	52.496970	0.563636	1.375758	129.303030	242.230303	0.139394	0.593939	158.466667	0.139394	0.583030	1.593939	0.363636	2.1
std	9.550651	0.497444	0.952222	16.169813	53.552872	0.347412	0.504818	19.174276	0.347412	0.780683	0.593635	0.848894	0.4
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	96.000000	0.000000	0.000000	0.000000	0.000000	0.0
25%	44.000000	0.000000	1.000000	120.000000	208.000000	0.000000	0.000000	149.000000	0.000000	0.000000	1.000000	0.000000	2.0
50%	52.000000	1.000000	2.000000	130.000000	234.000000	0.000000	1.000000	161.000000	0.000000	0.200000	2.000000	0.000000	2.0
75%	59.000000	1.000000	2.000000	140.000000	267.000000	0.000000	1.000000	172.000000	0.000000	1.000000	2.000000	0.000000	2.0
max	76.000000	1.000000	3.000000	180.000000	564.000000	1.000000	2.000000	202.000000	1.000000	4.200000	2.000000	4.000000	3.0

FEATURE SELECTION:
Since having irrelevant features in a data set can decrease the accuracy of the models applied, I used the Boruta Feature Selection technique to select the most important features which were later used to build different models.





MODEL DEVELOPMENT AND COMPARISON:
I used six classification models, i.e., Logistic Regression, K-Nearest Neighbors, Decision Trees and Support Vector Machine, After which I compared the performance of the models using their accuracy and F1 scores. I then settled with the best performing model.

.

IMPLEMENTATION OF PROJECT:
The proposed system taken input from the CSV file (Comma, Separated, Values). The CSV file contains medical data about patient's information for predicting the customer's information based on what we need. That dataset contains,

Model 1 : Logistic Regression

```
In [21]: model1=LogisticRegression(random_state=1)
         model1.fit(x_train,y_train)

         y_pred1=model1.predict(x_test)
         print(classification_report(y_test,y_pred1))
```

```
              precision    recall  f1-score   support

           0       0.77      0.67      0.71        30
           1       0.71      0.81      0.76        31

    accuracy                           0.74        61
   macro avg       0.74      0.74      0.74        61
weighted avg       0.74      0.74      0.74        61
```

Model 2 : K-NN (K-Nearest Neighbors)

```
In [22]: model2=KNeighborsClassifier()
         model2.fit(x_train,y_train)

         y_pred2=model2.predict(x_test)
         print(classification_report(y_test,y_pred2))
```

```
              precision    recall  f1-score   support

           0       0.78      0.70      0.74        30
           1       0.74      0.81      0.77        31

    accuracy                           0.75        61
   macro avg       0.76      0.75      0.75        61
weighted avg       0.76      0.75      0.75        61
```

Model 3 : SVM (Support Vector Machine)

```
In [23]: model3=SVC(random_state=1)
         model3.fit(x_train,y_train)

         y_pred3=model3.predict(x_test)
         print(classification_report(y_test,y_pred3))
```

	precision	recall	f1-score	support
0	0.80	0.67	0.73	30
1	0.72	0.84	0.78	31
accuracy			0.75	61
macro avg	0.76	0.75	0.75	61
weighted avg	0.76	0.75	0.75	61

Model 4 : Naives Bayes Classifier

```
In [24]: model4=GaussianNB()
         model4.fit(x_train,y_train)

         y_pred4=model4.predict(x_test)
         print(classification_report(y_test,y_pred4))
```

	precision	recall	f1-score	support
0	0.79	0.73	0.76	30
1	0.76	0.81	0.78	31
accuracy			0.77	61
macro avg	0.77	0.77	0.77	61
weighted avg	0.77	0.77	0.77	61

Model 5 : Decision Trees

```
In [25]: model5=DecisionTreeClassifier(random_state=1)
         model5.fit(x_train,y_train)

         y_pred5=model5.predict(x_test)
         print(classification_report(y_test,y_pred5))
```

	precision	recall	f1-score	support
0	0.68	0.70	0.69	30
1	0.70	0.68	0.69	31
accuracy			0.69	61
macro avg	0.69	0.69	0.69	61
weighted avg	0.69	0.69	0.69	61

Model 6 : Random Forest

```
In [26]: model6=RandomForestClassifier(random_state=1)
         model6.fit(x_train,y_train)

         y_pred6=model6.predict(x_test)
         print(classification_report(y_test,y_pred6))
```

	precision	recall	f1-score	support
0	0.88	0.70	0.78	30
1	0.76	0.90	0.82	31
accuracy			0.80	61
macro avg	0.82	0.80	0.80	61
weighted avg	0.81	0.80	0.80	61

IV. RESULT AND DISCUSSION

From these results we can see that although most of the researchers are using different algorithms such as SVC, Decision tree for the detection of patients diagnosed with Heart disease, Support vector, Decision tree KNN, Naïve bayes, Random Forest Classifier and Logistic regression yield a better result to out rule them.The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by Random Forest is 80% which is greater accuracies obtained from other algorithms. So, we summarize that our accuracy is improved due to the increased medical attributes that we used from the dataset we took.Our project also tells us that Logistic Regression, KNN, Support vector, Naïve bayes Classifier, Decision tree and Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease. This proves that Random Forest Classifier is better in diagnosis of a heart disease. The following ‘figure 1’, ‘figure 2’, ‘figure 3’, ‘figure 4’, ‘figure 5’ , ‘figure 6’ shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain:

V. CONCLUSION

cardiovascular disease detection model has been developed using six ML classification modeling techniques. This project predicts people with cardiovascular disease by extracting the patient medical history leads to a fatal heart disease from a dataset that includes patients’ medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic Regression, KNN, Support vector, Naïve bayes Classifier, Decision tree and Random Forest Classifier. The accuracy of our model is 80%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not. By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical databases that we can work on as these Machine learning techniques are better and they can predict better a human being which helps the patient as well as the doctors. Therefore, in conclusion this project helps predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying Random Forest and SVC to get an accuracy of an average of 80% on our model which is better than the models having an accuracy of 75%. Also, it is concluded that accuracy of Random Forest is highest between six algorithms that we have used i.e. 80%. ‘Figure 6’ shows 61% of people that are listed in the dataset are suffering from Heart Disease.

REFRENCES

[1] Christo Ananth, Stalin Jacob, Jenifer Darling Rosita, MS Muthuraman, T Ananth Kumar, “Low Cost Visual Support System for Challenged People”, 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), 978-1-6654-2111-9/22, IEEE, 10.1109/ICSTSN53084.2022.9761312, March 2022,pp. 1-4.

[2] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis:

overview of heart disease prediction. International Journal of Computer Applications,
17(8), 43-8

[3] Christo Ananth, S.Shafiqa Shalaysha, M.Vaishnavi, J.Sasi Rabiyathul Sabena, A.P.L.Sangeetha, M.Santhi, "Realtime Monitoring Of Cardiac Patients At Distance Using Tarang Communication", International Journal of Innovative Research in Engineering & Science (IJIRES), Volume 9, Issue 3,September 2014,pp-15-20.

[4] Dangare C S & Apte S S (2012).Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications,
47(10), 44-8.

[5] Christo Ananth, G.Poncelina, M.Poolammal, S.Priyanka, M.Rakshana, Praghash.K., "GSM Based AMR", International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST), Volume 1,Issue 4,July 2015, pp:26-28

Web Reference

- https:// www.w3schools.com
- https://www.tutorialspoint.com
- https://www.geeksforgeeks.com
- https://towardsdatascienc.com