IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

Cardiovascular Heart Disease Prediction Using Machine Learning

B Yadaiah

Asst. Professor, Dept. of CSE, MGIT, India

ABSTRACT: Machine Learning techniques have been extensively used for the identification of several disorders such as Heart-related diseases or cardiovascular diseases (CVDs). Cardiovascular Diseases are the main reason for a huge number of deaths in the world over the last few decades and have emerged as the most life-threatening disease, not only in India but in the whole world.
So, there is a need for a reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Cardiovascular disease is caused by correctable problems such as unhealthy diet, lack of exercise, being overweight, smoking and drinking. Most people die annually from cardiovascular diseases than any other disease. In India 1 out 4 die due to cardiovascular disease. Most Cardiovascular Diseases can be prevented by addressing behavioral risk factors. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. The model we are proposing uses a web interface that facilitates the hospital management to enter the details of the patient. These details serve as an input to the machine learning algorithm. Data set of both healthy people and people affected with cardiovascular diseases will be taken to compare with the given inputs and validate them by applying Support Vector Machines (SVM), Navies Bayes, Decision Trees (DT), Random Forest (RF) classifier, and Regression concepts of Machine Learning.

KEYWORDS: Machine Learning, cardiovascular, Navies Bayes, Decision Trees , Random Forest classifier.

I. INTRODUCTION

Heart disease describes a range of conditions that affect our heart. The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina), or stroke. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects that a person is born with (congenital heart defects), among others.

With the rampant increase in heart stroke rates at juvenile ages, we need to put a system in place to be able to detect the symptoms of a heart stroke at an early stage and thus prevent it. It is impractical for a common man to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of heart disease. Thus we propose to develop a machine learning model which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate, etc., and also give a planned diet to the user.

Symptoms can include:

- Chest pain, chest tightness, chest pressure, and chest discomfort (angina)
- Shortness of breath
- Pain, numbness, weakness, or coldness in the legs or arms if the blood vessels in those parts of your body are narrowed
- Pain in the neck, jaw, throat, upper abdomen, or back

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

II. LITERATURE SURVEY

There is number of works has been done related to disease prediction systems using different machine learning algorithms in medical Center's.

Senthil Kumar Mohan et al,[1] proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy that objective is to finding critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. We produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, Such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease.

Sonam Nikhar et al [2] has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms by This exploration plans to give a point by point portrayal of NaÃ¯ve Bayes and decision tree classifier that are applied in our examination especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the equivalent dataset, and the result uncovers that Decision Tree beats over Bayesian classification system

Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD), [3] Prediction of Heart Disease Using Machine Learning, In this paper proposed system they used the neural network algorithm multi-layer perceptron (MLP) to train and test the dataset. In this algorithm there will be multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in input layer is connected to output nodes through these hidden layers. This connection is assigned with some weights. There is another identity input called bias which is with weight b, which added to node to balance the perceptron. The connection between the nodes can be feedforwarded or feedback based on the requirement.

Abhay Kishore et al,[4] developed Heart Attack Prediction Using Deep Learning in which This paper proposes a heart attack prediction system using Deep learning procedures, explicitly Recurrent Neural System to predict the probable prospects of heart related infections of the patient. Recurrent Neural Network is a very ground-breaking characterization calculation that utilizes Deep Learning approach in Artificial Neural Network. The paper talks about in detail the significant modules of the framework alongside the related hypothesis. The proposed model deep learning and data mining to give the precise outcomes least blunders. This paper gives a bearing and point of reference for the advancement of another type of heart attack prediction platform. Prediction stage.

Lakshmana Rao et al,[5] Machine Learning Techniques for Heart Disease Prediction in which the contributing elements for heart disease are more (circulatory strain, diabetes, current smoker, high cholesterol, etc..). So, it is difficult to distinguish heart disease. Different systems in data mining and neural systems have been utilized to discover the seriousness of heart disease among people. The idea of CHD ailment is bewildering, in addition, in this manner, the disease must be dealt with warily. Not doing early identification, may impact the heart or cause sudden passing. The perspective of therapeutic science furthermore, data burrowing is used for finding various sorts of metabolic machine learning a procedure that causes the framework to gain from past information tests, models without being expressly customized. Machine learning makes rationale dependent on chronicled information.
Mr. Santhana Krishnan.J and Dr. Geetha.S, [6] Prediction of heart disease using machine learning algorithm This Paper predicts heart disease for Male Patient using Classification Techniques. The detailed information about Coronary Heart diseases such as its Facts, Common Types, and Risk Factors has been explained in this paper.

III. METHODOLOGY

1. Architecure

We can define the machine learning workflow in 5 stages.

- Gathering data from the standard dataset and analyzing the patients details

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

- Data pre-processing is done on the dataset which consists of patients details and all the missing values are filled
- The data is split into training and test datasets and 75% data is considered as training data and remaining 25% is used as test data
- Training and testing the model
- Algorithm is applied on the data for classification and algorithms can be KNN , Naïve Bayes etc
- Finally the the persons who has the heart disease are predicted and diet is recommend to the patient accordingly

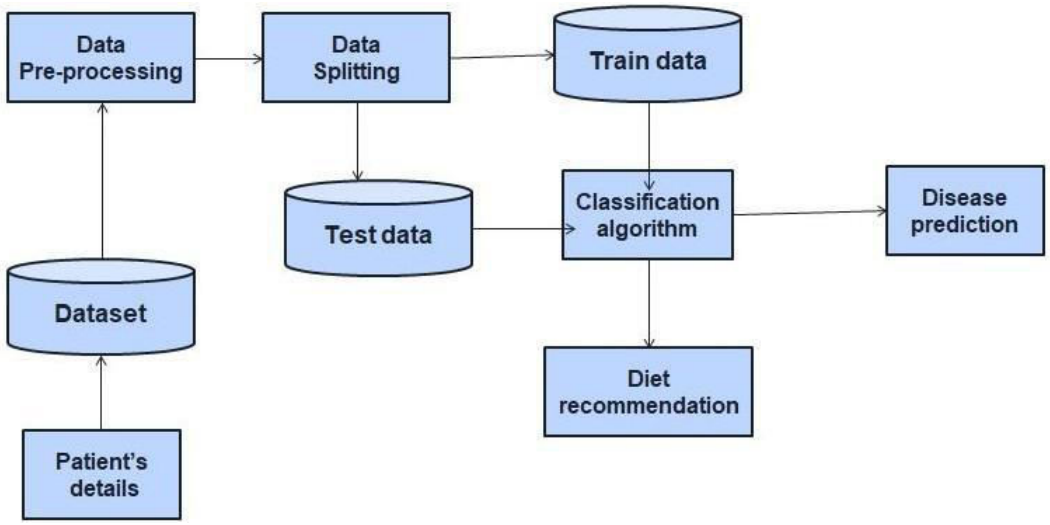

Figure.3.1 Architecure of Heart Disease Prediction using machine learning

IV. RESULTS

1. Results

The input to the machine learning model is dataset consisting of patient details it is considered as input and we get prediction as output and diet is recommended to the patient accordingly.

Dataset output screen

The dataset consists of 14 attributes which play a major role in predicting the heart disease.The attributes include Age,Sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal ,target.
Among these 14 attributes age, sex, resting blood pressure, serum cholesterol, resting ecg, fasting blood pressure, chest pain type are the main important attributes and these attributes play a major role in predicting the heart disease using machine learning.

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.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

	age	sex	cp	trestbps	chol	fbs	...	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	...	0	2.3	0	0	1	1
1	37	1	2	130	250	0	...	0	3.5	0	0	2	1
2	41	0	1	130	204	0	...	0	1.4	2	0	2	1
3	56	1	1	120	236	0	...	0	0.8	2	0	2	1
4	57	0	0	120	354	0	...	1	0.6	2	0	2	1
5	57	1	0	140	192	0	...	0	0.4	1	0	1	1
6	56	0	1	140	294	0	...	0	1.3	1	0	2	1
7	44	1	1	120	263	0	...	0	0.0	2	0	3	1
8	52	1	2	172	199	1	...	0	0.5	2	0	3	1
9	57	1	2	150	168	0	...	0	1.6	2	0	2	1
10	54	1	0	140	239	0	...	0	1.2	2	0	2	1
11	48	0	2	130	275	0	...	0	0.2	2	0	2	1
12	49	1	1	130	266	0	...	0	0.6	2	0	2	1
13	64	1	3	110	211	0	...	1	1.8	1	0	2	1
14	58	0	3	150	283	1	...	0	1.0	2	0	2	1
15	50	0	2	120	219	0	...	0	1.6	1	0	2	1
16	58	0	2	120	340	0	...	0	0.0	2	0	2	1
17	66	0	3	150	226	0	...	0	2.6	0	0	2	1
18	43	1	0	150	247	0	...	0	1.5	2	0	2	1
19	69	0	3	140	239	0	...	0	1.8	2	2	2	1

Figure 4.1 Dataset of 967 individual patients taken from kaggle

Preprocessing screen

The data is preprocessed from the dataset. The missing values are filled with the help of imputers and the feature scaling is done in this phase. The values are normalized so that there is no difference between values and all the values are under same range.The categorical values are converted into dummy variables for preprocessing.

```
#pre processing
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values = np.nan, strategy = 'mean', verbose = 0)

imputer = imputer.fit(X[:, 0:13])
X[:, 0:13] = imputer.transform(X[:, 0:13])

#for diet withhout scaling
D_train,D_test, DE_train, DE_test = train_test_split(X, Y, test_size=0.3, random_state=0)

# Feature Scaling -> Normalising
sc = StandardScaler()
X = sc.fit_transform(X)

#data splitting
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=0)
```

Figure 4.2 Preprocessing is done in the machine learning model

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

KNN Algorithm output

The KNN Model is used to predict the heart disease prediction using machine learning.

```
plt.plot([k for k in range(1, 21)], knn_scores, color = 'red')

for i in range(1,21):
    plt.text(i, knn_scores[i-1], (i, knn_scores[i-1]))
plt.xticks([i for i in range(1, 21)])
plt.xlabel('Number of Neighbors (K)')
plt.ylabel('Scores')
plt.title('K Neighbors Classifier scores for different K values')
```

```
Text(0.5, 1.0, 'K Neighbors Classifier scores for different K values')
```

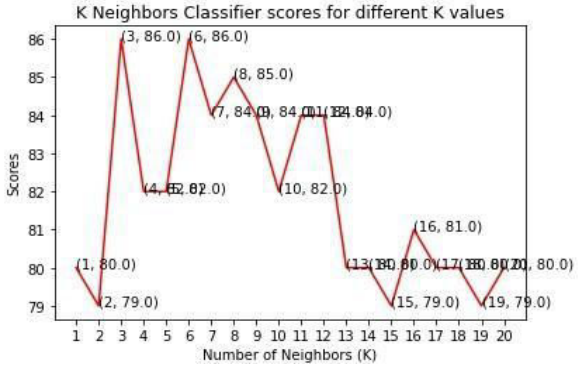

Figure 4.3 KNN Algorithm

```
#KNN
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 7)

classifier.fit(X_train, Y_train)

Y_pred = classifier.predict(X_test)
result = accuracy_score(Y_test,Y_pred)*100
print("Accuracy:",result)
```

```
Accuracy: 83.51648351648352
```

Figure 4.4 Output for K Nearest neighbour algorithm

Naïve Bayes Algorithm

```
from sklearn.ensemble import RandomForestClassifier

rf=RandomForestClassifier(n_estimators=1000,random_state=1)
rf.fit(X_train,Y_train)

rf_pred=rf.predict(X_test)
print('Test accuracy{:.2f}%'.format(rf.score(X_test,Y_test)*100))
score_rf=accuracy_score(rf_pred,Y_test)
score_rf=score_rf*100
```

```
Test accuracy82.42%
```

Figure 4.5 Naive bayes algorithm

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

Decision Tree Algorithm

```
dt = DecisionTreeClassifier(random_state=best_x)
dt.fit(X_train,Y_train)

Y_pred_dt = dt.predict(X_test)
score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)

print("The accuracy score achieved using Decision Tree is: "+str(score_dt)+" %")
```

```
The accuracy score achieved using Decision Tree is: 75.82 %
```

Figure 4.6 Decision Tree Algorithm

Random forest

```
from sklearn.ensemble import RandomForestClassifier

rf=RandomForestClassifier(n_estimators=1000,random_state=1)
rf.fit(X_train,Y_train)

rf_pred=rf.predict(X_test)
print('Test accuracy{:.2f}%'.format(rf.score(X_test,Y_test)*100))
score_rf=accuracy_score(rf_pred,Y_test)
score_rf=score_rf*100
```

```
Test accuracy82.42%
```

Figure 4.7 Random Forest Algorithm

MLP classifier

```
idx=0
true=0
false=0
for i in x_test:
    if predicted[idx]==Y_test[idx]:
        true+=1
    else:
        false+=1
    idx+=1

print("accuracy of mlpclassifierr:")
score_mlp=(true/(true+false))*100
print(score_mlp)
```

```
accuracy of mlpclassifierr:
68.13186813186813
```

Figure 4.8 MLP classifier

Diet recommendation

Figure 4.9 Diet recommendation output

```
No Risk for heart disease
age: 60 bp:  125 gender:  1 dia:  0
no. of calories per day: 2000
Breakfast: Cooked meat(100gm) + Cooked meat(100gm) + Berries(100gm)
Lunch: Cooked meat(100gm) + whole egg or 2 egg whites  + Any vegetable(80g)  + Any vegetable(80g) + Leafy Greens +1 corn tortilla +
Dinner:  Cooked meat(100gm) + Any vegetable(80g) + Leafy Greens +1 corn tortilla
Snack: Fruit Juice(125ml)

--------------------------------------------------------------------------------------------

No Risk for heart disease
age: 62 bp:  130 gender:  1 dia:  0
no. of calories per day: 2000
Breakfast: Tofu(200gm) + Cooked fish(200gm) + Banana
Lunch: Tofu(200gm) + Cooked fish(200gm)  + Any vegetable(80g)  + Any vegetable(80g) + Leafy Greens +Half Large Potato(100g) + 1 TBS
Dinner:  Cooked fish(200gm) + Any vegetable(80g) + Leafy Greens +1 corn tortilla
Snack: Dried Fruits(Handfull)

--------------------------------------------------------------------------------------------
```

Figure 4.10 Output for diet recommendation

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

Correlation Matrix

Correlation is an indication about the changes between two variables. We can plot correlation matrix to show which variable is having a high or low correlation in respect to another variable. Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together**.** In simple terms, it tells us how much does one variable changes for a slight change in another variable. It may take positive, negative and zero values depending on the direction of the change. A high correlation value between a dependent variable and an independent variable indicates that the independent variable is of very high significance in determining the output. In a multiple regression setup where there are many factors, it is imperative to find the correlation between the dependent and all the independent variables to build a more viable model with higher accuracy. One must always remember that more number of features does not imply better accuracy. More features may lead to a decline in the accuracy if they contain any irrelevant features creating unrequired noise in our model The correlation matrix is a symmetrical matrix with all diagonal elements equal to +1. We would like to emphasise that a correlation matrix only provides insight to a data scientist about correlation, and it is NOT a reliable tool to study causation. Indeed, the correlation values should be carefully interpreted by an expert to avoid nonsense statements.

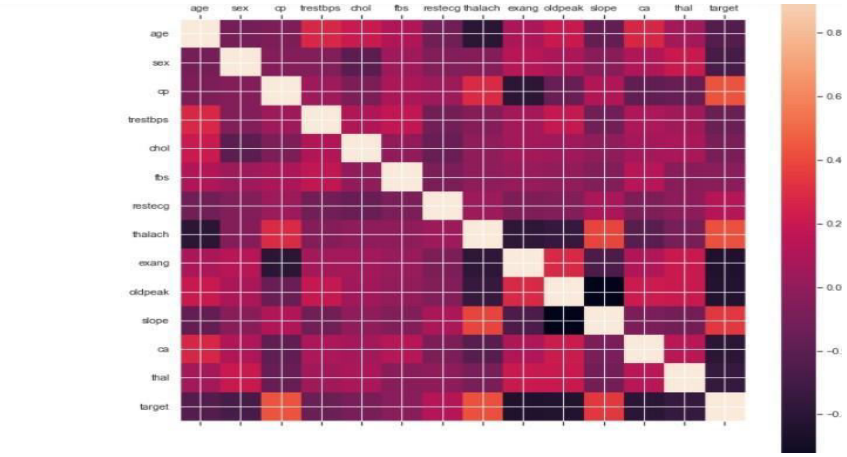

Figure 4.11 Correlation Matrix

Comparision of different Models

The different models are being applied to predict the heart disease using machine learning and comparision is being done on different models.

Accuracy Score

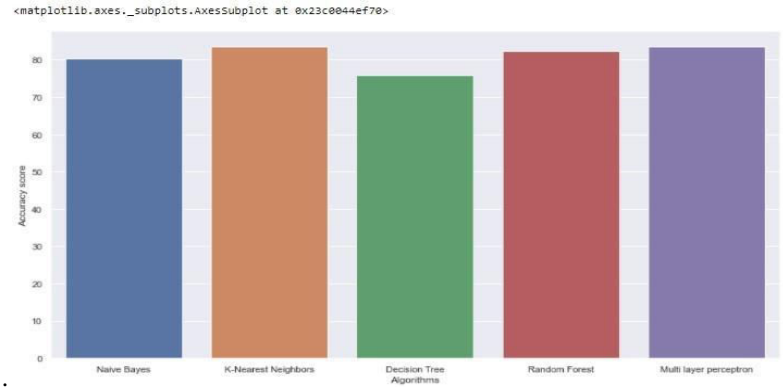

Figure 4.12 Accuracy score comparison

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

From the above figure, it is evident that the KNN algorithm has given maximum accuracy. More the accuracy more is the chance for the correct prediction of the disease.

False Negative Rate

In medical testing, and more generally in binary classification, a false positive is an error in data reporting in which a test result improperly indicates the presence of a condition, such as a disease (the result is *positive*), when in reality it is not present, while a false negative is an error in which a test result improperly indicates no presence of a condition (the result is *negative*), when in reality it is present.

An example for a false negative is a test indicating that a woman is not pregnant whereas she is actually pregnant. Another example is a truly guilty prisoner who is acquitted of a crime.

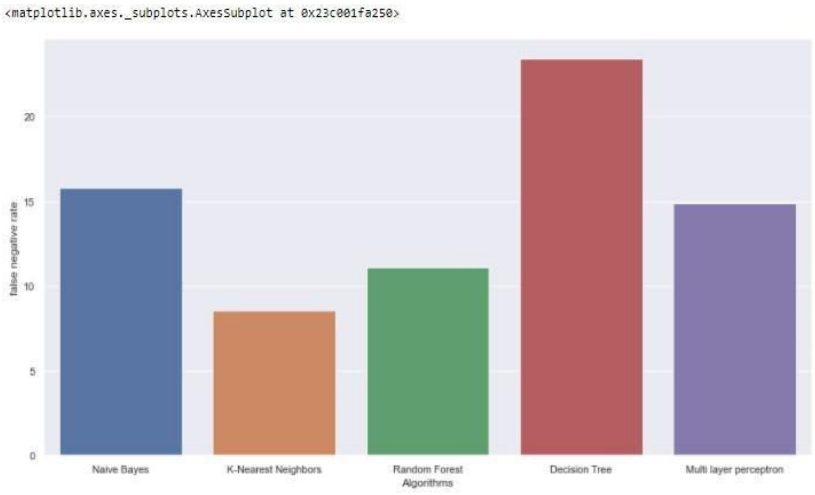

Figure 4.13 False-negative rate comparison

Values of confusion matrix

Algorithm	True positive	False positive	False Negative	True Negative
KNN	22	5	4	30
Naïve Bayes	21	6	3	31
Random Forest	22	5	6	28
Decision Tree	25	2	4	30

Table 4.1 Values obtained for confusion matrix

Analysis of machine learning

Algorithm	Precision	Recall	F-measure	Accuracy
Decision tree	0.845	0.823	0.835	81.97%
Regression	0.857	0.882	0.869	85.25%

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106225 |

Random Forest	0.937	0.882	0.909	90.16%
Naïve Bayes	0.837	0.911	0.873	85.25%

Table 4.2 Analysis of machine learning

V. CONCLUSION AND FUTURE SCOPE

1. Conclusion

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

2. Future Scope

One of the major drawbacks of these works is that the main focus has been on the application of classification techniques for heart disease prediction, rather than studying various data cleaning and pruning techniques that prepare and make a dataset suitable for mining. It has been observed that a properly cleaned and pruned dataset provides much better accuracy than an unclean one with missing values. Selection of suitable techniques for data cleaning along with proper classification algorithms will lead to the development of prediction systems that give enhanced accuracy

In the future, we would like to enhance the application by making it customer-specific and deploying this model on a hospital website, and help the doctors detect any early signs of the disease.

REFERENCES

1. https://en.wikipedia.org/wiki/Cardiovascular_disease.
2. http://www.heart.org/HEARTORG/Conditions/HeartAttack/WarningSignsofa HeartAttack/Warning-Signs-of-aHeartAttack_UCM_002039_Article.jsp#.WNpKgPl97IU.
3. www.who.int/cardiovascular_diseases/en/.
4. http://food.ndtv.com/health/world-heart-day-2015-heart-disease-in-india-is-agrowing- concern-ansari-1224160
5. Shaikh Abdul Hannan, A.V. Mane, R. R. Manza,and R. J. Ramteke, Dec 2010,"Prediction of Heart Disease Medical Prescription using Radial Basis Function",IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2010.5705900 ,28-29 .
6. AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "HDPS: Heart Disease Prediction System",Computing in Cardiology, ISSN: 0276-6574, pp.557- 560.
7. Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010,"Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network",International Conference on Computer and Communication Technology (ICCCT),DOI:10.1109/ICCCT.2010.5640377, 17-19.
8. Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map",IEEE International Conference on Fuzzy Systems (FZZ),pp. 1377-1382

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details