Heart Disease Prediction using Machine Learning

J. Gowri, R. Kamini, G. Vaishnavi, S. Thasvin, C. Vaishna

Abstract*: Heart is one most important organ in our body. The prediction of heart disease is most complicated task in today world. There are number of instruments available in today's worlds. These instruments are so expensive some of them can afford that instrumentals some of them cannot afford the instruments. Early prediction of heart disease will reduce the death rate. we can tell the patients before the hand. In todays world we all have the good amount of data using that good amount of data we can predict the heart disease using various machine learning techniques. The proposed method will tell to patients probabilities of heart diseases. In this paper using the UCI dataset performed various machine learning techniques like Logistic Regression, Decision tree, KNN, Naïve Bayes, Random Forest, XGBoost, Support vector machine . In this paper we used proposed methodology from PHASE I to PHASE VII Using the evaluation metrics we can check the performance of the machine learning which gives more accuracy from the above seven machine learning algorithm..

***Keywords*: Logistic Regression, Decision tree, KNN, Naïve Bayes, Random Forest, XG Boost, Support vector machine, Accuracy, Machine learning, Prediction, heart.**

I. INTRODUCTION

The most important subject matter is prediction the usage of system mastering techniques. Machine mastering is extensively used now a day in lots of commercial enterprise packages like e trade and lots of greater. Heart disorder prediction is one a number of the main complex duties in clinical discipline. Because coronary heart is the subsequent fundamental organ evaluating to mind which has greater precedence in every member of human race body. It transport the blood and materials to all organs of the entire body. It is one of the most heinous disorders, specifically the silent coronary heart assault, which assaults someone so all at once that there's no time to get it dealt with and such disorder could be very tough to be diagnosed. Various clinical records mining and system mastering strategies are being applied to extract the precious records concerning the coronary heart disorder prediction. Yet, the accuracy of the favored consequences isn't always satisfactory. Prediction of occurrences of coronary heart illnesses in clinical discipline is full-size paintings. This Model proposes a coronary heart assault prediction device the usage of Machine mastering strategies. Health care discipline has an extensive quantity of records, for processing the ones records sure strategies are used. In the era, about one-character dies in line with minute way to coronary heart situation. This System predicts the bobbing up opportunities of heart disease. As coronary heart situation prediction can be complicated task, In today's world there's demand to automate the prediction method to keep away from dangers associated with it and alert the affected person properly beforehand. Data analytics is beneficial for prediction from greater records and it facilitates clinical middle to expect of numerous disorders. Huge quantity of affected person associated records is maintained on month-to-month basis. The saved records may be beneficial for supply of predicting the incidence of destiny disorder. The proposed paintings predict the possibilities of coronary heart situation and classifies affected person's threat degree via way of means of imposing exclusive records processing strategies like KNN, Naive Bayes, Decision Tree, Logistic Regression, Random Forest, SVM, XGB oost.

Manuscript received on 19 June 2022 | Revised Manuscript received on 25 June 2022 | Manuscript Accepted on 15 July 2022 Manuscript published on 30 July 2022.
* Correspondence Author

J. Gowri, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.
R. Kamini*, PG Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.
G. Vaishnavi, PG Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.
S. Thasvin, PG Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.
C. Vaishna, PG Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.

II. LITERATURE SURVEY

[1] Senthilkumar Mohan recommend mine to get invisible information for effective deciding. Finding of hidden relationships along with patterns frequently goes untapped. He used Naïve Bayes and Decision Tree algorithm techniques used for predicting the heart disease. Decision Tree gives higher accuracy than naïve bayes.

[2] Aditi Gavhane superscription dataset is training and testing is performed using the neural networks. In that RNN gives good accuracy in deep learning.

[3] Santhana Krishnan foretell the originating chance of heart condition. The upshot of this method the probabilities of occurring cardiovascular disease in terms of percent. The datasets are processed using four main Machine Learning Algorithm they are Logistic Regression, Decision Tree, Support Vector Machine and Naive Bayes Algorithm. From the above ML algorithm Decision Tree gives good accuracy.

[4] Tsien et al in their study indicated that classification trees, which have certain advantage over logistic regression models, with patients having myocardial infarct (MI). The results shown that the occurrence of MI has been noticed in male than the feminine. Age, Systolic blood pressure, smoking has been found to be the important risk consider the patients with ML.

Retrieval Number: 100.1/ijitee.H91480711822
DOI: 10.35940/ijitee.H9148.0711822
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

[5] Aakash Chauhan foretell that the data is directly received from the electronic machine and it also reduce manual work. The trail gives more accuracy in prediction of heart disease.

III. METHODOLOGY

A. Data collection

Data series is defined due to the fact the system of gathering , gauging and examining precise in sagacity for probe. A probe can compare their speculation on the basis of accumulated information. In maximum cases, information series is that the number one and maximum great step for probe, irrespective of the arena of probe. In our study, we use a dependent information set of UCI with the size of 303 rows and thirteen columns are provided in Table - I.

B. Data-Preprocessing

Before beginning the software of Machine Learning algorithms, we put together the information to be implemented; this segment is completed in steps: Features choice This step is primarily based totally at the correlation matrix. we've got thirteen attributes in Table – I which are associated and structured each other. The information of the chosen functions are defined in

Table- I: UCI Dataset Attributes

Attributes	Description	
	Description	Values
Age	age	Age 29 to 62 years
Sex	Sex	0-male 1-female
CP	Chest pain type	0-typical angina pectoris 1-atypical angina 2-non-anginal pain 3-symptomatic
trestbps	Resting blood pressure in mm/Hg	Numeric value : example: 140mm/Hg
Chol	Serum cholesterol in mg/dl	Numeric value : example: 289mg/hg
Fbs	Fasting blood pressure>120mg/dl	1-Yes 0-No
Restecg	Resting electrocardiographic results	0-normal, 1-have the ST-T 2-hypertrophy
Thalach	Maximum heart rate achieved	Numeric value : Example: 140,173
Exang	Exercise induced angina	1-Yes 0-No
Oldpeak	ST depression induced by exercise relative to rest	Numeric Value
Slope	The slope of the peak exercise ST segment	1-on the rise 2-flat 3-the downhill slope
Ca	Number of major vessels colored by flourosopy	0-3 vessels
Thal	Thalassemia	1-normal, 2-defect repaired 3-reversible defect

C. Manual Exploration

Data exploration or Manual Exploration is that the preliminary step in information analysis, in which customers discover an outsized information set in an unstructured way to find preliminary patterns, characteristics, and factors of interest. This system isn't predestined to reveal similarly of statistics an statistics set holds, how ever as an alternative to contribution create a complete photograph of great traits and important factors to test in more detail. In our study, we upload a column in our information set (Target) which incorporates zero or one (0 = unhealthy, 1 = healthy).

D. Selection of attributes

In the selection of features we have to select the dependent features of the give dataset. The independent features of the dataset will be dropped from the dataset for performing the various machine learning techniques,

IV. HISTOGRAM OF FEATURES

A. Histogram of attributes

The histogram of features repulicate the variety of dataset code and features which is practiced to generate it.

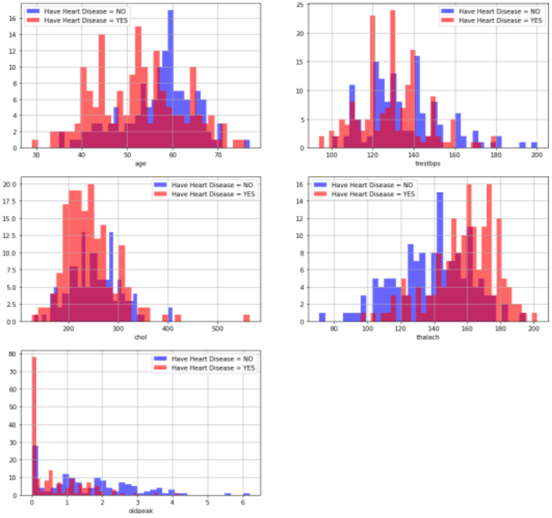


Fig. 1. Histogram of continuous values in the UCI dataset

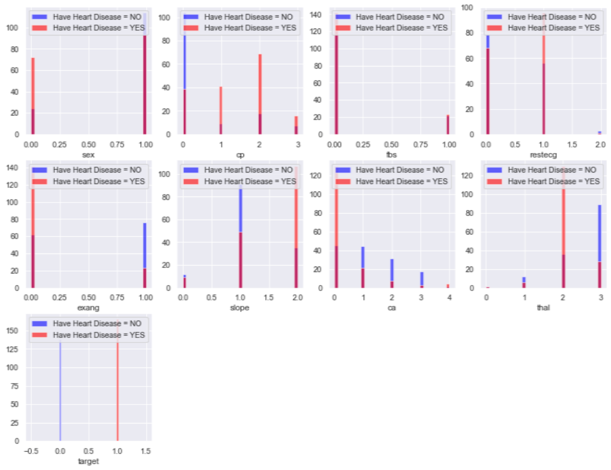


Fig. 2. Histogram of categorical values of UCI dataset

V. PROPOSED SYSTEM

In this paper we are implementing the heart disease prediction using the ML algorithm. First we have to give the input in csv file After that perform data processing and check it out any missing values or null values. if null values or missing values are found means it can be replaced by using mean median and mode. After that perform machine learning algorithm from Phase I to Phase VII.

Here we have used 7 ML algorithms they are

- Logistic regression
- KNN
- SVM
- Naïve Bayes
- Decision Tree
- Random forest
- XGBoost

Using the evaluation metrics we can calculate the accuracy of the ML algorithm

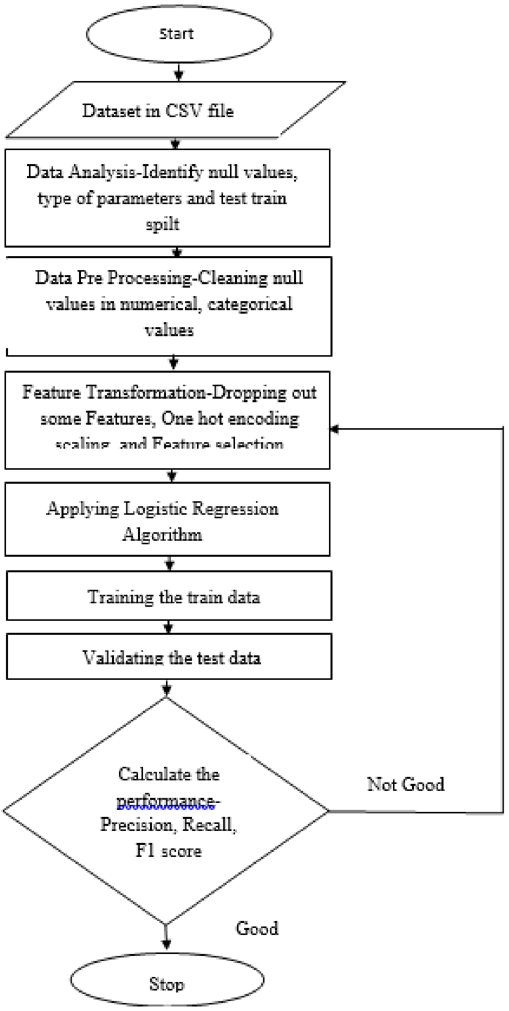

Start
Dataset in CSV file
Data Analysis-Identify null values, type of parameters and test train spilt
Data Pre Processing-Cleaning null values in numerical, categorical values
Feature Transformation-Dropping out some Features, One hot encoding scaling and Feature selection
Applying Logistic Regression Algorithm
Training the train data
Validating the test data
Calculate the performance-Precision, Recall, F1 score
Not Good
Good
Stop

Fig. 3. System flow diagram of proposed System

VI. PERFORMANCE ANALYSIS

A. Correlation of matrix

The matrix in machine learning it shows how one or two variables related to each other.

Fig. 4. Correlation matrix of UCI heart dataset

B. Confusion matrix

False positive (FP): the amount of unhealthy patients expected to be healthy.

False negative (FN): the amount of sick patients they are properly classified as sick.

True negative (TN): the amount of wellness patients they are incorrectly classified as sick.

True positive (TP): The amount of wellness patients were classified as healthy

The potency of the classifiers in distinguishing viscus sickness may be measured from the confusion matrix evaluation.

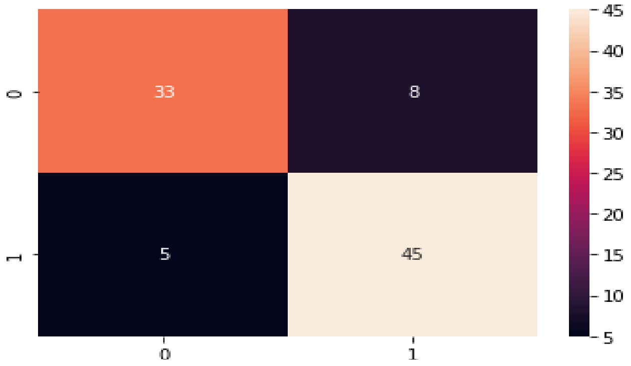


Fig. 5. Confusion matrix of SVM

C. Precision

It's the relation between the amount of positive predictions and therefore the total number of positive prediction category values. It will live the exactitude of the classifier.

Precision = (True Positive) / (True Positive + False Positive)

D. Recall

It's the measure of true prediction numbers divided by the number of true class values in testing data. It's the completeness of the classifiers.

Recall = (True Positive) / (True Positive + False Negative)

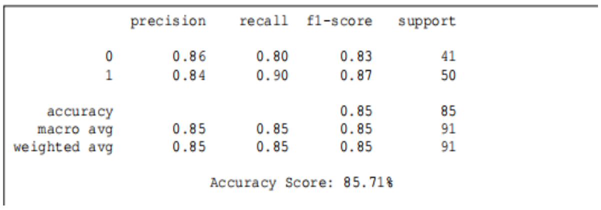
E. F-Measure

It expresses the balance between the recall and exactness. it's the mean value of each precision and recall.

F– Measure =2(True positive)/2(True positive) + False positive + False negative

F. Accuracy

Accuracy: it'sthe categoryifier"s ability to properly predict that the class of the instances were be labeled for all the instances.



```
              precision    recall  f1-score   support

           0       0.86      0.80      0.83        41
           1       0.84      0.90      0.87        50

    accuracy                           0.85        85
   macro avg       0.85      0.85      0.85        91
weighted avg       0.85      0.85      0.85        91

              Accuracy Score: 85.71%
```

Fig. 6. Evaluation metrics of SVM Algorithm

Accuracy=(True positive +True negative)/(True Positive +TN+ False positive+ False Negative)

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

VII. RESULT

While performing & ml algorithms svm gives more accuracy compared with other 6 ml algorithms. accuracy is calculated with the support of the confusion matrix of every algorithm, here the amount count of true positive, true negative, false positive, false negative is given and victimization the equation of accuracy, worth has been calculated and it's ended that svm is best with 85.7%.

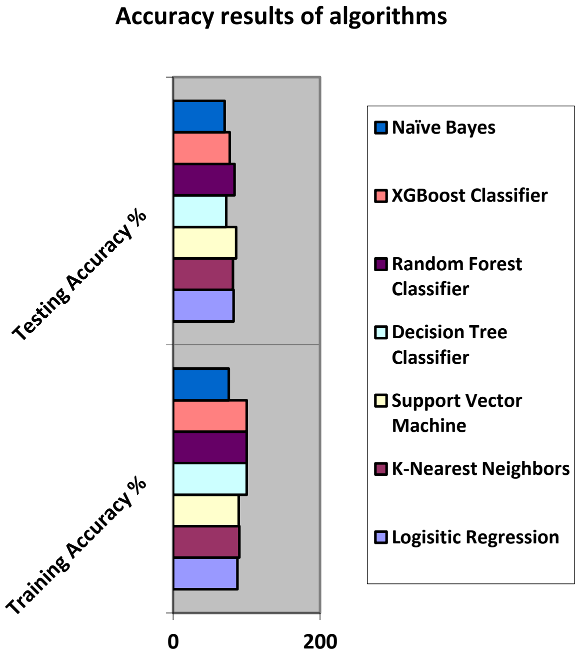


Fig. 7.Accuracy of algorithms

VIII. CONCLUSION AND FUTURE SCOPE

Heart illnesses is a considerable killer in tamilnadu and all through the globe, software of promising generation like gadget studying to preliminary prediction of coronary heart illnesses can have a profound effect on society. The primary diagnosis of cardiovascular sickness can useful resource in making choices on lifestyle changes in high-risk suffers and successively reduce the complications, which is probably a terrific milestone inside the subject of drugs. The variety of people going through coronary heart illnesses on a boost annually. The activates for its early analysis and treatment. The employment of suitable generation assist for the duration of this regard can have an impact on be exceedingly useful to the clinical fraternity and sufferers. The predicted attributes ensuring in cardiopathy in sufferers are to be had inside the dataset which includes 76 functions and 14 crucial functions which can be beneficial to evaluate the device are decided on amongst them,. If all of the functions taken into the consideration then the performance of the device the author receives is a smaller amount to increase the performance characteristic choice is finished for the duration of this functions ought to be decided on for evaluating the version which components greater accuracy. The correlation of a few capabilities in the dataset is sort of identical then they may be removed. If all of the attributes gift inside the dataset are taken beneath attention then the performance decreases considerably. All the seven device studying techniques accuracies are in comparison supported which one prediction version is generated. Hence the purpose is to apply diverse assessments metrics like confusion matrix, Accuracy, Precision, Recall, F1-Measurewhich all these had measured the disorder efficiently. Comparing the entire seven algorithm The SUPPORT VECTOR MACHINE algorithm offers the quality of 85.71% of accuracy.

REFERENCES

1. Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.
2. Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018. [CrossRef]
3. Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.
4. C.L. Tsien, H.S.F. Fraser, W.J. Long and R.L. Kennedy "Using classification trees and logistic regression methods to diagnose myocardial infarction" in Proc. 9th World Congr., Inf., vol. 52, pp. 483-497, 2001.
5. Aakash Chauhan"Heart Disease Prediction using Evolutionary Rule Learning" in Conference: 2018 4th International Conference on "Computational Intelligence & Communication Technology (CICT) [CrossRef]

AUTHORS PROFILE

Prof J. Gowri, MCA., M.Phil., Ph.D., with 16 years of teaching experience and Published 10 journals publications.

R. Kamini currently pursing 5th year in M.sc(Software systems) at Sri Krishna Arts And Science College, Affiliated to Bharathiyar University, Coimbatore.

G .Vaishnavi currently pursing 5th year in M.sc(Software systems) at Sri Krishna Arts And Science College, Affiliated to Bharathiyar University, Coimbatore.

S. Thasvin currently pursing 5th year in M.sc(Software systems) at Sri Krishna Arts And Science College, Affiliated to Bharathiyar University, Coimbatore.

C. Vaishna currently pursing 5th year in M.sc(Software systems) at Sri Krishna Arts And Science College, Affiliated to Bharathiyar University, Coimbatore.

Retrieval Number: 100.1/ijitee.H91480711822
DOI: 10.35940/ijitee.H9148.0711822