International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Heart Disease Prediction using Machine Learning Techniques

Shaik Razia, J. Chinna Babu, K. Hemanth Baradwaj, K. S. S. R. Abhinay, Anusha M

Abstract. ***Nowadays, heart disease has become a major disease among the people irrespective of the age. We are seeing this even in children dying due to the heart disease. If we can predict this even before they die, there may be huge chances of surviving. Everybody has various qualities of beat rate (pulse rate) and circulatory strain (blood pressure). We are living in a period of data. Due to the rise in the technology, the amount of data that is generated is increasing daily. Some terabytes of data are being produced and stored. For example, the huge amount of data about the patients is produced in the hospitals such as chest pain, heart rate, blood pressure, pulse rate etc. If we can get this data and apply some machine learning techniques, we can reduce the probability of people dying. In this paper we have done survey using different classification and grouping strategies, for example, KNN, Decision tree classifier, Gaussian Naïve Bayes, Support vector machine, Linear regression, Logistic regression, Random forest classifier, Random forest regression, linear descriptive analysis. We have taken the 14 attributes that are present in the dataset as an input and applying on the dataset which is taken from the UCI repository to develop and accurate model of predicting the heart disease contains colossal (huge) therapeutic (medical) information. In the proposed research, the exhibition of the conclusion model is acquired by using utilizing classification strategies. In this paper proposed an accuracy model to predict whether a person has coronary disease or not. This is implemented by comparing the accuracies of different machine-learning strategies such as KNN, Decision tree classifier, Gaussian Naïve Bayes, SVM, Logistic regression, Random forest classifier, Linear regression, Random forest regression, linear descriptive analysis***

Keywords: Linear Regression, Random Forest, Decision Tree.

Manuscript published on November 30, 2019.
* Correspondence Author
Shaik Razia *, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India. Email: razia28sk@gmail.com
J.Chinna Babu, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India. Email: bchinna032@gmail.com
K. Hemanth Baradwaj, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.Email: khbaradwaj7@gmail.com
K. S. S. R. Abhinay Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, IndiaEmail: abhinaykothagundi@gmail.com
Anusha M Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.Email: anushaaa9@gmail.com

I. INTRODUCTION

Today people are leading a life that involves a lot of pressure and stress from a very young age. In addition, the level of individuals who are dry and dependent on cigarettes increases substantially. This causes cardio disease, cancer and soon to be diseases. The challenge behind these diseases is their prediction. There are different estimates of heart rate and pulse of every person. However, clinically proven, the pulse rate should be 60 to 100 beats for each moment and the pulse must be in the range 120/80 to 140/90. Coronary disease is one of the main causes of death. The amount of individuals affected by coronary disease makes up in both men and women regardless of age. In any case, various components like gender, diabetes, BMI add to the disease. In this paper, we have tried to predict and analyze heart disease by considering age, gender, blood pressure, cholesterol and soon parameters. Since various elements are associated with cardio disease, the expected test of this disease is. Some of the major symptoms of a heart attack are:

- Chest tightness.
- Shortness of breath.
- Nausea, Indigestion, Heartburn, or abdominal pain.
- Sweating and Fatigue.
- Pressure in the upper back Pain that extends to the arm.

Types of heart disease: The heart signifies "cardio". In addition, all heart diseases concern a class of heart diseases. There are different types of coronary disease:

- Coronary heart infection.
- Angina pectoris
- Decreased heartbeat.
- Cardiomyopathy
- Congenital heart infection.

Dataset Description Table

SLNO	Name of the Attribute	Description
1	Age	Age in Year
2	sex	Value 1 and 0 for male and female
3	Cp	Chest pain type Value 0: typical angina Value 1: atypical angina Value 2: non- anginal pain Value 3: asymptomatic
4	Trestbps	Resting blood pressure
5	Chol	Cholesterol of serum in mg/dl
6	Fbs	Fasting blood sugar Value1: True Value 0: False
7	Resteeg	Resting electrocardiographic results
8	Thalach	Maximum heart rate achieved
9	exang	Exercise induced angina (1 = Yes, 0 = No)
10	Old peak	ST depression induced by exercise relative to rest
11	Slope	The slope of peak exercise of ST segment. Value 0: up sloping Value 1: flat Value 2: down sloping
12	Ca	Number of major Vessels (0-3) colored by fluoroscopy
13	Thal	0 = normal, 1: fixed defect, 2: reversible defect
14	Target	Diagonal of heart disease Value 0: no Risk (< 50% diameter narrowing) Value 1: Risk (> 50% diameter narrowing).

II.LITERATURE SURVEY

Benjamin EJ et.al [1] say that there are seven key factors for heart disease such as smoking, physical inactivity, nutrition, obesity, cholesterol, diabetes, and high blood pressure. They also discussed the statistics of heart disease including stroke and cardiovascular disease.

Abhay Kishore et.al [2] on their experimentation showed that recurrent neural network gives good accuracy when compared to other algorithms like CNN, Naïve Bayes, and SVM. Subsequently, neural systems perform well in cardio illness expectations. They additionally accomplished a framework that could anticipate quiet cardio failures and illuminate the client as most punctual conceivable.

M. Nikhil Kumar et.al [3] used various algorithms– Decision tree, random forest, Na-ïve Bayes, KNN, Support vector machine, logistic model tree algorithm, and Naive Bayes calculation gave great outcomes when contrasted with different calculations. They made use of the UCI repository of the heart disease dataset. Also, the J48 algorithm took less time to build and gave good results.

Amandeep Kaur et.al [4] compared various algorithms such as artificial neural network, K–nearest neighbor, Naïve Bayes, Support vector machine on heart disease prediction.

Sahaya Arthy et.al [5] analyze the existing works on heart disease prediction, which uses data mining. The data mining systems are ordinarily utilized in coronary illness expectations. They additionally talk about the databases utilized, for example, the cardio illness dataset from UCI store, instruments utilized, for example, Weka, Quick Excavator, Information dissolve, Apache Mahout, Clatter, Bottom, R and soon. They conclude that the use of single algorithm results in better accuracy in prediction. But the use of hybridization of two or more algorithms can enhance and improve the heart disease prediction with good accuracy

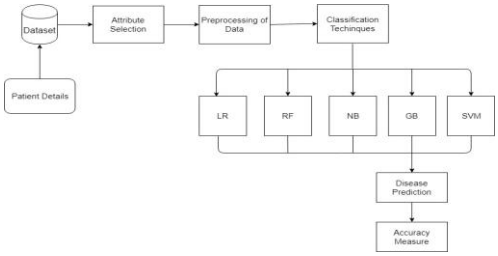
III. METHODOLOGY

In this paper, we used various machine learning algorithm methods for predicting heart disease of patients using medical dataset. applied the different machine learning techniques and found out the algorithm that gives the best accuracy and our target value contains 0 and 1 which tells us that if 0 means the patient is safe and 1 indicates the patient is at risk.

We have used 10 different machine-learning algorithms and predicted heart disease using the best algorithm with more accuracy. The different algorithms we have used in our project are as follows:

1. KNN (K nearest neighbor classifier)
2. Decision tree
3. SVM (Support Vector Machine)
4. Random Forest Classification
5. Linear regression
6. Random Forest Regression
7. Logistic regression
8. Linear descriptive analysis
9. Gaussian Naïve Bayes

The following is the flowchart that shows the proposed methodology:



We imported a few libraries for the venture:

Numpy: To work with arrays

Pandas: To work with the CSV record and information framework

Matplotlib: To draw diagrams using Pyplot, mark the parameters using rcParams and match them to cm.rainbow

Warnings: Ignoring all warnings appearing in the magazine due to past / future devaluation of an element

train_test_split: Splitting the dataset into training and testing information

Standard Scaler: To scale every single highlight, with the goal of best adjusting the machine-learning model to the dataset

a. K-Neighbors Classifier

This classifier scans for the classes of K nearest neighbors of a given data point and relies on the larger part class, it distributes a class at this data point. Regardless, the number of neighbors can be shifted. I separated them from 01 to 30 neighbors and judged the test scores for each situation. Until then, I draw a line drawing the number of neighbors and the test scores obtained for each situation.

b. Support Vector Classifier

This classifier aims to create hyperplanes that can separate classes anyway by adjusting the partition between the data focus and the hyperplane that might be common. There are parts of the subject for which the hyperplane is selected. I tried four classes to be linear, shift, RBF and sigmoid respectively.

At this point when I had a score for each, I used the rainbow method to choose different colors for each bar and drew a visual graph of the points obtained by each.

c. Decision Tree Classifier

This classifier decides on the subject of a decision tree, according to which it tells the class with respect to each datum point. Here, we can change the offending number considered when constructing the model. I run the features from 01 to 50 (hard and fast features in the dataset after the move parts are included). At the point when we have a score, we have the option to plot a line chart and see the effect of the amount of features in the model score.

d. Random Forest Classifier

This takes along the possibility of classifier decision trees. This creates a forest of trees where each tree is limited by outright assurance of facilities from outright facilities. Here, we can change the amount of trees that will be used to predict the class. I get test scores over 10, 100, 200, 500 and 1000 trees. Next, I look at these scores on a resolved proposal as to who gave the best result. You can see that I have not set explicitly with respect to x [100, 200, 300, 400, and 500. This would show a coherent plot from 100 to 500, which would be difficult to open. Subsequently, to solve this issue, I initially used [10, 20, 30, 40, 50] with respect to x. Until then, I renamed those using xticks.

e. Logistic Regression

Logistic regression is a kind of regression analysis in statistics utilized for expectation of the result of a straight out ward variable from a lot of indicator or free factors. In logistic regression, the reliant variable is continually parallel. Logistic regression is primarily utilized for expectation and computing the likelihood of accomplishment.

f. Naïve Bayes

In probability, the Bayes theorem (called Bayes law after Thomas Bess) deals with the prohibitive and periphery probabilities of two unpredictable occasions. It is routinely used to record the probabilities given after observations [2]. For example, a patient may be considered to have certain signs (side effects). Bayes' hypothesis can be used to apply the possibility that a proposed conclusion is true, given that assumption.

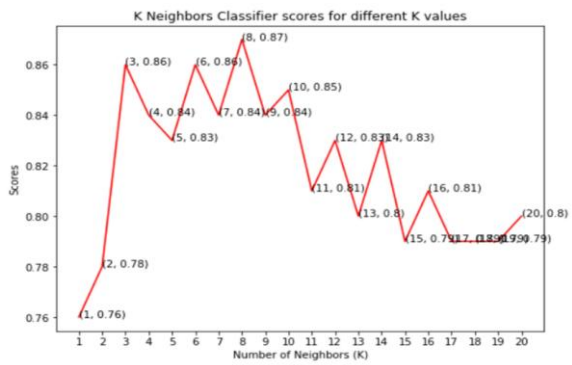
$$P(A/B) = P(B/A)\ P(A)\ /\ P(B)$$

g. Linear Regression

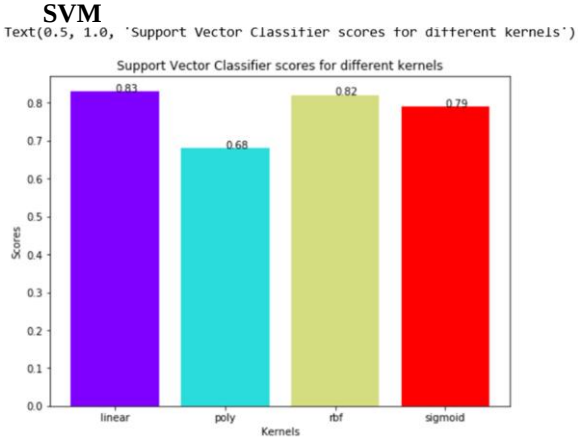
Linear Regression is one kind of AI calculation, and it depends on directed (supervised) learning. It plays out a regression task. Regression model Regression model targets expectation values based on independent variables. It is mostly used for discovering the relationship between variables (ward) and forecasting (independent). Linear Regression is to anticipate the at least two variables with the relationship between subordinate variables (Y) and the independent variables (X).

IV. RESULTS

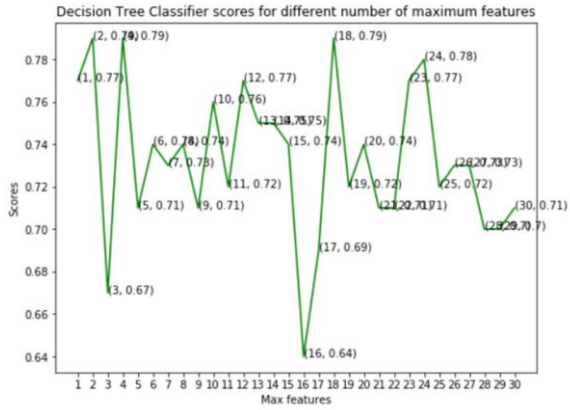
a. KNN



b. SVM

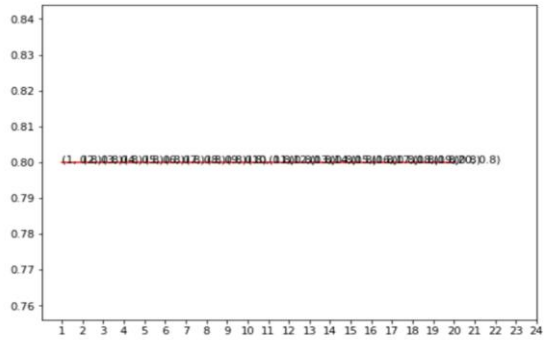


c. Decision tree

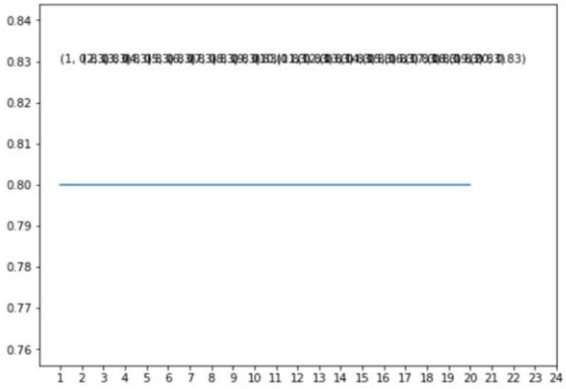


d. Gaussian Naïve Bayes

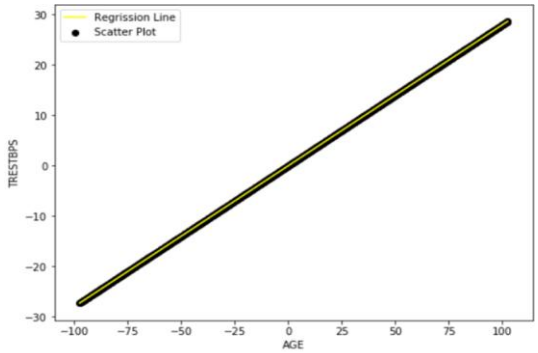
Accuracy of GNB classifier on training set: 0.86
Accuracy of GNB classifier on test set: 0.80



e. Linear Descriptive Classifier

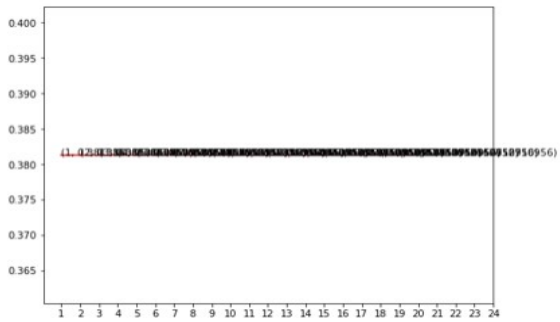


f. Linear Regression

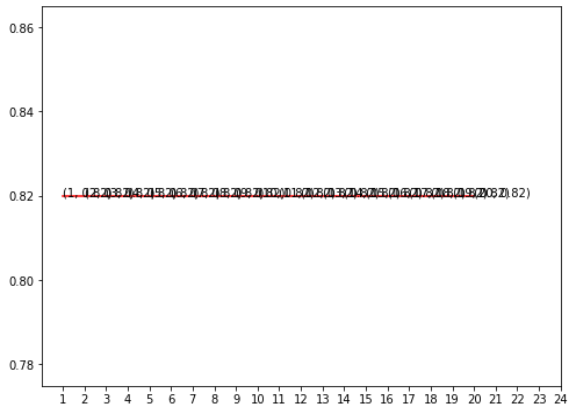


g. Random Forest Regression

```
Accuracy of random Forest Regressor on training set: 0.86
Accuracy of random Forest Regressor on test set: 0.43
```

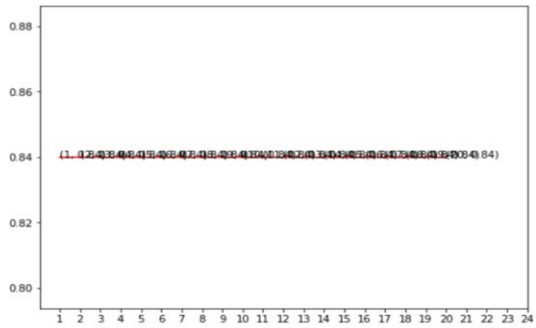


h. Random Forest Classification

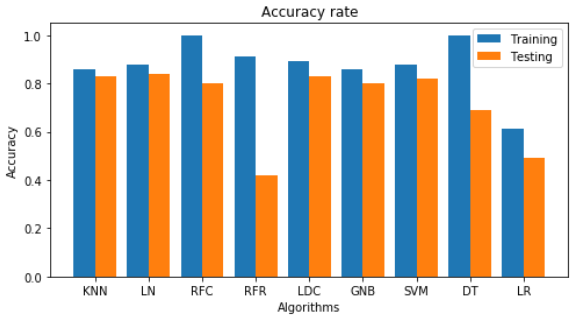


i. Logistic regression

```
Accuracy of Logistic regression on training set: 0.88
Accuracy of logistic regression on test set: 0.84
```



j. Accuracy Rate



V. CONCLUSION

This paper discusses the various machine learning techniques for predicting heart disease such as KNN, SVM, Decision tree classifier, Random forest, Linear regression, Logistic regression, Naïve Bayes, Linear descriptive analysis, Gaussian naïve Bayes. Out of all the algorithms, we found the best algorithm Decision tree that gives accurate results and used in the future for predicting heart disease at the earliest level and can take preventive measures. The Research work contributes the correlative application and analysis of various machine-learning algorithms in the python software, which is used to early expectation of the heart disease. Future work includes distinctive ensemble methods of these algorithms that can progress to better execution with more parameter setting using diverse machine-learning algorithms.

REFERENCES

1. Benjamin EJ et.al Heart Disease and Stroke Statistics 2018 At-a-Glance (2018)
2. Abhay Kishore, Ajay Kumar, Karan Singh, Maninder Punia, Yogita Hambir Department of Computer Engineering, Army Institute of Technology, Pune, Maharashtra Professor, Department of Computer Engineering, Army Institute of Technology, Pune, Maharashtra Heart Attack Prediction Using Deep Learning (2018)
3. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak Department of CSE, VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India Prediction Heart Diseases using Data mining and machine learning algorithms and tools. (2018)
4. Amandeep Kaur, Jyoti Arora Dept of CSC Desh Bhagat University, Punjab, India heart disease prediction using data mining techniques: a survey (2018)
5. A. Sahaya Arthy, G. Murugeshwari A survey on heart disease prediction using data mining techniques (April 2018)
6. SHAIK RAZIA, M.R.Narasingarao, "Development and Analysis of Support Vector Machine Techniques for Early Prediction of Breast Cancer and Thyroid" JARDCS (Journal of Advanced Research in Dynamical and Control Systems), ISSN: 1943-023X, Vol.9.Sp.Issue:6 page no: 869-878, 2017.
7. SHAIK RAZIA, M.R.NARASINGARAO "A Neuro computing frame work for thyroid disease diagnosis using machine learning techniques" JATIT (Journal of Theoretical and Applied Information Technology), ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195, Vol.95. No.9. Pages 1996-2005, 15th May 2017.
8. Cmak Zeelan basha, Azmira Krishna, Pradeep Raj Savarapu,"Automatic Detection of Lung Infection", International Journal of Recent Technology and Engineering , Vol.8, Issue-3, pp.200-203,Sep. 2019
9. KH Kishore, B Venkataramani, R Sanjay, VS Rajan "High-gain inductorless wideband Balun-LNA using asymmetric CCC &BIST using RMS detectors" AEU-International Journal of Electronics and Communications 105, 135-144.

10. Bhimanpallewar, R., & Narasingarao, M. R. (2018). A prototype model for continuous agriculture field monitoring and assessment. International Journal of Engineering and Technology (UAE), 7(2), 179-182.
11. Shaik Razia, P.Swathi Pryathyusha, N.Vamsi Krishna "A Comparative study of machine learning algorithms on thyroid disease prediction" International Journal of Engineering and Technology(UAE), ISSN No: 2227-524X, Vol No: 7, Issue No: 2.8, Page No: 315-319, March **2018.**
12. P. Gopi Krishna 1 *, K. Srinivasa Ravi"Design and development of Bi-Directional IoT gateway using Zigbee and Wi-Fi technologies with MQTT Protocol" in *International Journal of Engineering & Technology, 7 (2.8) (2018) 125-129*.
13. P. Gopi Krishna 1 *, K. Srinivasa Ravi" Design and development of smart energy meter for effective use of electricity in IoT applications"*inInternational Journal of Engineering & Technology, 7 (2.8) (2018) 115-119.*
14. Avinash Yadlapati, K Hari Kishore "System Level Verification of Advanced Extensible Interface Protocol Using Verilog HDL" Journal of Advanced Research in Dynamical and Control Systems, ISSN No: 1943-023X, Vol No: 10, Special Issue No: 7, Page No: 1359-1365, June 2018.
15. K.Sarath Chandra, K Hari Kishore "Physical Design Implementation of High Performance CMOS Dynamic Latch Comparator" Journal of Advanced Research in Dynamical and Control Systems, ISSN No: 1943-023X, Vol No: 10, Special Issue No: 9, Page No: 323-332, June 2018.
16. . G. Vijaya Padma, K. Hari Kishore and S. Jaya Sindura, "Controlling the Traffic Interactions with High Mobility and Constant Network Connectivity by Vanets", Lecture Notes in Electrical Engineering, ISSN No: 1876-1100, E-ISSN: 1876-1119, pp. 593-601, 2018
17. P Ramakrishna, K Hari Kishore "DESIGN OF AN ULTRA LOW POWER CMOS COMPARATOR FOR DATA CONVERTERS" Journal of Advanced Research in Dynamical and Control Systems, ISSN No: 1943-023X, Vol No: 10, Special Issue No: 7, Page No: 1347-1352, June 2018.
18. A Murali, K Hari Kishore, L Srikanth, A Trinadha Rao and V Suresh, "Implementation of Reconfigurable Circuit with Watch-Points in the Hardware", Lecture Notes in Electrical Engineering, ISSN No: 1876-1100, E-ISSN: 1876-1119, pp. 657-664, 2018.
19. K.Sarath Chandra, K Hari Kishore **"**Electrical Characteristics of Double Gate FINFET under Different Modes of Operation**"** International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Volume-8, Issue No: 6, Page No: 172-175, April 2019**.**
20. Avinash Yadlapati, Hari Kishore Kakarla **"**Low-power design-for-test implementation on phase-locked loop design" Measurement and Control, ISSN: 0020-2940, Volume-52, Issue No: (7-8), Page No: 995-1001, June 2019.