IJIREEICE

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 9, Issue 7, July 2021

DOI 10.17148/IJIREEICE.2021.9703

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

Heart Attack Prediction and Visualization of Contributing Factors Using Machine Learning

Megha Banerjee[1], Reetodeep Hazra[1], Suvranil Saha[1], Megha Bhushan[1], Subhankar Bhattacharjee[2]

Student, Department of Electronics and Communication Engineering, Techno International New Town, Kolkata, India[1]

Assistant Professor, Department of Electronics and Communication Engineering, Techno International New Town, Kolkata, India[2]

Abstract: Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data analysis and machine learning are the most commonly used techniques for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyze huge complex medical data, helping healthcare professionals to predict heart disease. This research paper presents various attributes like age, gender, chest pain, cholesterol etc which are used to predict heart attack, and the model is trained using 4 machine learning algorithms namely- Logistic Regression, Gaussian Naïve Bayes, Decision tree and Random Forest algorithm. It uses the existing dataset from the UCI Heart Disease Data set of heart disease patients. The dataset comprises 303 instances and 76 attributes. This research paper aims to envision the probability of developing heart attacks in patients. The results portray that the highest accuracy score is achieved with Logistic Regression.

Keywords: Heart Attack Prediction, Cardiac Analytics, Machine Learning, UCI Heart Disease Data Set, Logistic Regression.

I.INTRODUCTION

Cardiovascular diseases (CVDs) are one of the top contributors to mortality in the world. According to World Health Organization (WHO), CVDs alone causes nearly 18 million deaths each year. Most of them are because of Acute Myocardial Infarction (Heart Strokes and attacks). Millions of dollars were spent on medical treatment included medications, surgical intervention, and so forth. Instead of spending humongous amount of money in treatment, preventive measures can significantly impact population health in positive way. With the help of Artificial Intelligence (AI), Machine Learning (ML), and Computational Biology, people can benefit from early diagnosis before heart attack and provides them the opportunity to take preventive actions.

Malfunctions thyroid gland can have direct effect on heartbeat, blood pressure and cholesterol level as this hormone influences the force and speed of the heartbeat. Higher cholesterol thickens up the wall of arteries which eventually restricts the blood flow to the heart. When the blood supply to the heart is extremely insufficient or completely cut off, it causes an immediate severe heart attack. Apart from that, higher cholesterol value combining with other major factors can causes severe CVDs.

The focus of this study is to analyze several factor related to heart attacks (like age group, gender, type of chest pain, thyroid, level and type of cholesterol) to design a algorithm that can predict the risk factor for heart attack for a particular person accurately with higher precision than all the previous literatures reported. We deployed several algorithms to analyze their efficiency in order to predict heart attacks and concluded with the best one in terms of accuracy, precision, recall, and F1-score.

The commitment of this article summed up as follows-
a) A clear novel step-by-step description of the full research work.
b)Exploratory data analysis and multiple visualizations to determine relationships between different factors cause heart attacks.
c)The higher risk factors of heart disease in between different sexual classification (male and female) are also analyzed.
d) Performance matrix analyze and optimization techniques of 4 algorithms- Logistic Regression, Gaussian NB, Random Forest, and Decision Tree.

DOI 10.17148/IJIREEICE.2021.9703

The remainder of this paper is summarized as follows: Section II has the details of related works done in this field. Section III describes the methodology of the research followed by the performance evaluation in the Section IV and conclusion in the Section V.

II.RELATED WORKS

Lot of work has been carried out to predict heart disease using **UCI Heart Disease Data Set** [1]. Different levels of accuracy have been attained using various data mining techniques which are explained as follows:

Fahd Saleh Alotaibi, et al., has designed a ML model comparing five different algorithms [2]. Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy. Nagaraj M Lutimath, et al., has performed the heart disease prediction using Naive bayes classification and SVM (Support Vector Machine). The performance measures used in analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that SVM was emerged as superior algorithm in terms of accuracy over Naive Bayes [3].

Hongmei Yan et al. (2006) used Multilayer perception (MLP) with 40 input variables for Input layer and the output layer with 5 nodes. Improved backpropogation algorithm was used to train the system and assessment methods like cross validation, holdout and bootstrapping were applied to assess the system. The hidden layer was obtained by cascading learning process. The experimental results achieved were of 90% accuracy [4]. Chaitrali S. et al. (2012) used 13 attributes like sex, blood pressure, and cholesterol for prediction of heart disease. Two more attributes called smoking and obesity was added. The classification methods used were Decision tree (DT), neural network (NN) and Naïve bayes (NB). Accuracy obtained for these techniques were 100%, 99.62% and 90.74% respectively. Confusion matrix was obtained for 3 classification methods for 13 attribute data sets and 15attribute data set. The accuracy with 15 attribute was 100% for neural network [5].

Purusothama et al. applied various classification algorithms for disease prediction model in diagnosing the Heart Disease. The two types of models i.e. primary model is single model and secondary model were used and compared to train the data. Data analysis was done using these two models. For both the single model and combined model, authors have considered the classification techniques only. The results attained by comparing the algorithms such as association rule, KNN, ANN, Naive Bayes, hybrid approach with the accuracy of 76%, 58%, 86%, 69% and 96% respectively. The author recommended that hybrid data mining algorithms performs well and promising accuracy results were attained in heart disease diagnosis [6]. Avinash Golande et al., studies various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree; KNN and K-Means algorithms that can be used for classification and their accuracy were compared [7]. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

Detrano et al. [8] proposed a logistic regression classifier-based decision support system for heart disease classification and obtained a classification accuracy of 77%. Gudadhe et al. [9] used multilayer perceptron (MLP) and support vector machine algorithms for heart disease classification. They proposed classification system and obtained accuracy of 80.41%. Kahramanli and Allahverdi [10] designed a heart disease classification system used a hybrid technique in which a neural network integrates a fuzzy neural network and artificial neural network. And the proposed classification system achieved a classification accuracy of 87.4%. Palaniappan and Awang [11] designed an expert medical diagnosing heart disease system and applied machine learning techniques such as Naive Bayes, decision tree, and ANN in the system. The Naive Bayes predictive model obtained performance accuracy 85.12%. The second best predictive model was ANN which obtained an accuracy of 88.12%, and decision tree classifier achieved 80.4% with correct prediction.

Reetodeep et al. [12] proposed a machine learning model to detect respiratory diseases like bronchiectasis, pneumonia, bronchiolitis, chronic obstructive pulmonary disease, upper respiratory tract infection, and healthy from the recorded lung sounds at early stages by using Mel-frequency cepstral co-efficients. The model reported an accuracy of 92.39%. Megha et al. [13] also proposed a model to study the use of convolutional neural network to classify heart sounds into normal and abnormal categories. The paper also reports 5 designated categories of heart sounds such as artifacts, murmur, extra heart sound, extra systole and normal. The research achieved an accuracy of 85%.

III.METHODOLOGY

A. Dataset Description

The dataset for this research project is taken from **UCI Heart Disease Data Set** [11]. The contributor of the data set includes Hungarian Institute of Cardiology, University Hospital Zurich, University Hospital Basel, V.A. Medical Center Long Beach and Cleveland Clinic Foundation. The dataset contain 302 actual patient entries with 13 different factors such as age, sex, number of major vessels, previously noted chest pain type and specific location, cholesterol, blood sugar, previous electrocardiographic results, smoking history, etc. The data was presented in text format with no null values. For training and testing purpose, we divided the dataset in 80:20 ratios for training and testing respectively.

B. Data Visualization

(i) Age vs Density Graph:

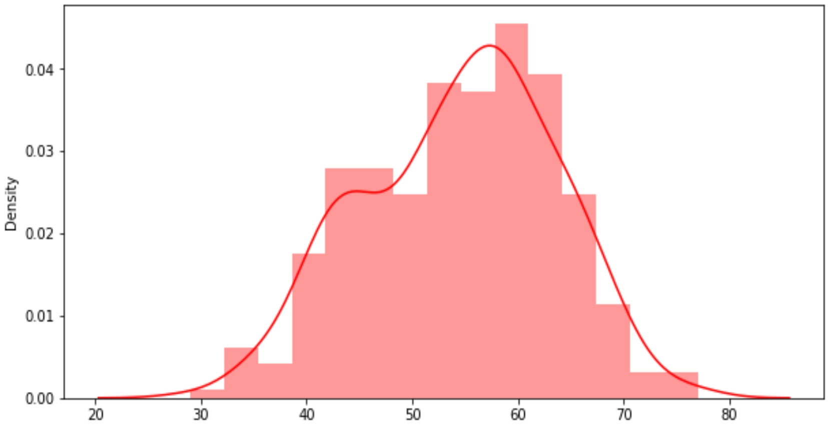

Fig. 1: Age vs. Density Graph

The graph in Fig. 1 represents age vs. density. By using visualization techniques, we found out that the minimum value for age is 29 and the maximum value for age is 77. The mean value for age is also calculated which resulted 54.42.

(ii) Gender Representation:

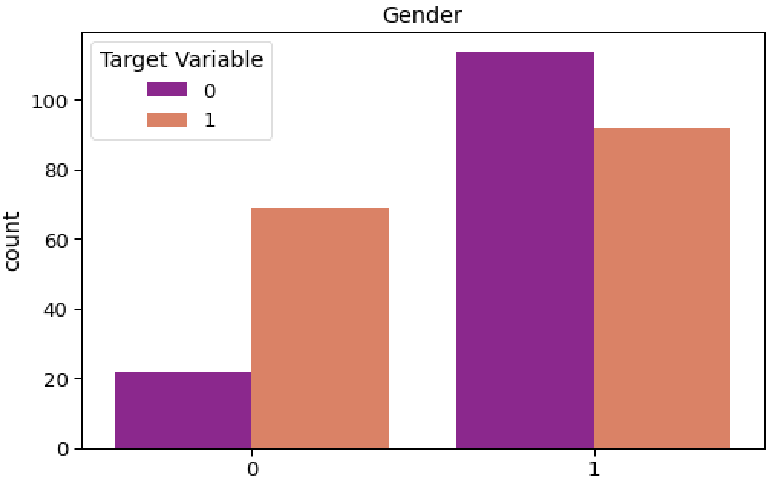

Fig. 2: Count vs. Sex

In Fig. 2, the number '0' represents Female and '1' represents Male. The gender proportion is imbalanced with male accounts for ~69% and female ~31%.

(iii) Chest Pain:

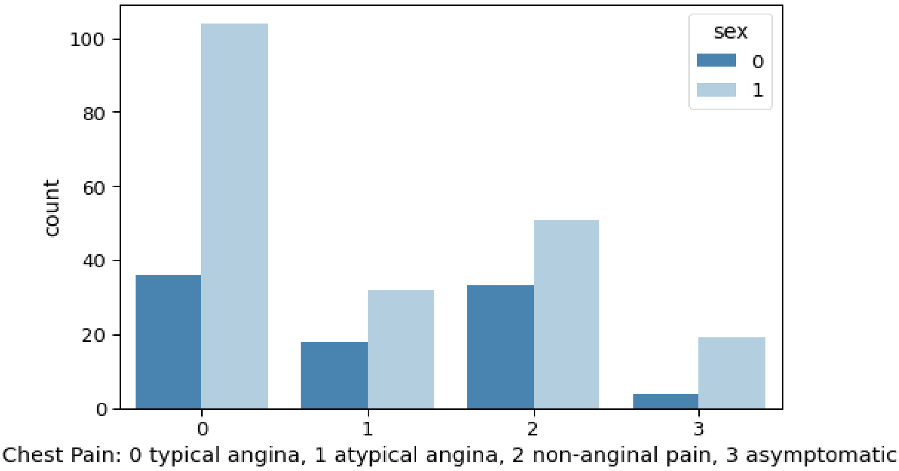

Fig. 3: Chest Pain Type

In Fig. 3, value 0 represents typical angina, value 1 represents atypical angina, value 2 represents non-anginal pain and value 3 represents asymptomatic type. As shown in the graph above, males have higher count than females across all types of chest pain because of the gender proportion in the dataset. But the probability of a female having a chance of heart attack is 75.82% whereas the probability of a male having a chance of heart attack is 44.93%

(iv) Cholesterol vs Age:

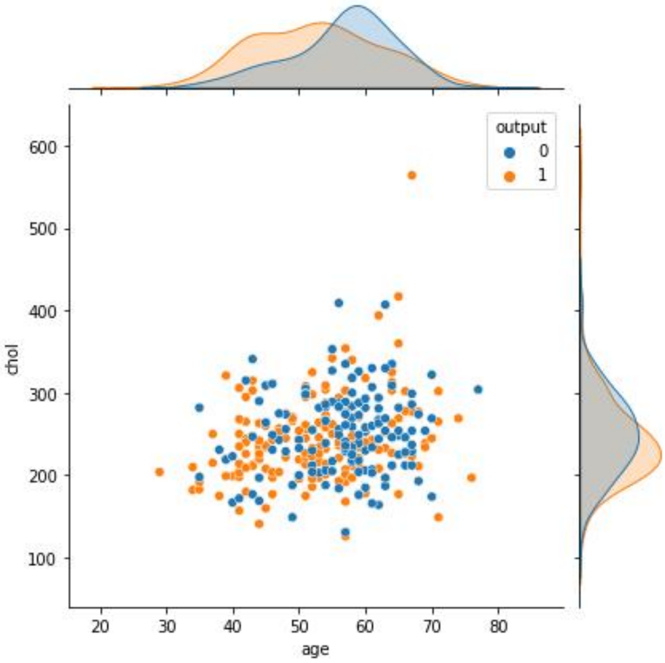

Fig. 4: Cholesterol vs. Age Graph

In the dataset the value of cholesterol is present in the unit of mg/dl fetched via a BMI sensor. In Fig. 4, output 0 represents a person with no cholesterol and output 1 represents a person with cholesterol. We can see that persons in between the age group of 50 to 60 have higher chances of having cholesterol whereas persons in between the age group of 60 to 70 have higher chances of having no cholesterol.

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 9, Issue 7, July 2021

DOI 10.17148/IJIREEICE.2021.9703

(v) Thalachh vs Trtbps:

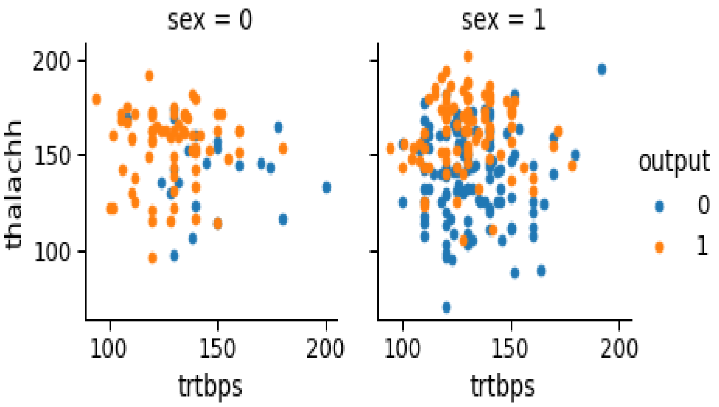

Fig. 5: Thalachh vs. Trtbps Graph

Here Thalachh represents maximum heart rate achieved and Trtbps represents resting blood pressure (in mm Hg). We used the scatter plot technique to find the graph and found that the maximum value of Trtbps is 178 whereas the minimum value of Thalachh is 103.

(vi) Thalachh vs exng:

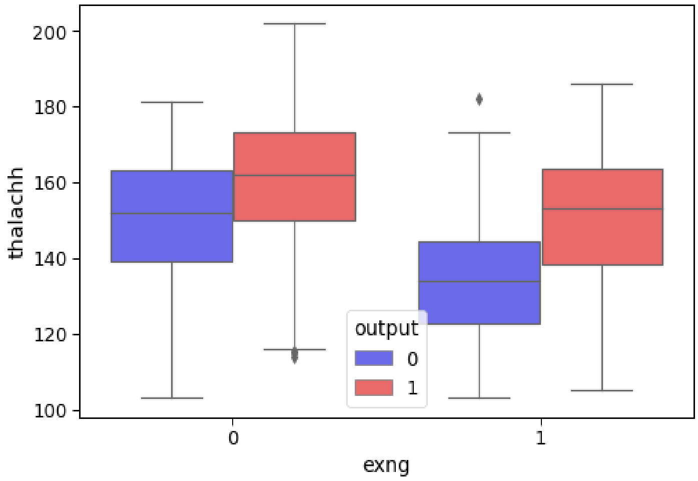

Fig. 6: Thalachh vs. Exng Graph

Here Exng represents exercise induced angina. The output of 0 represents no and the output of 1 represents yes.

(vii) Fbs vs count:

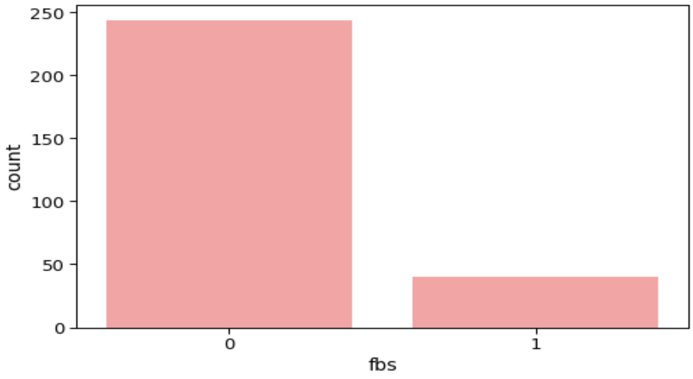

Fig. 7: Count vs. Fbs Graph

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526
Vol. 9, Issue 7, July 2021
DOI 10.17148/IJIREEICE.2021.9703

Here Fbs represents fasting blood sugar > 120 mg/dl. From this graph, 85.91% people have no blood sugar whereas 14.08% people have fasting blood sugar.

(viii) Cp vs. Age:

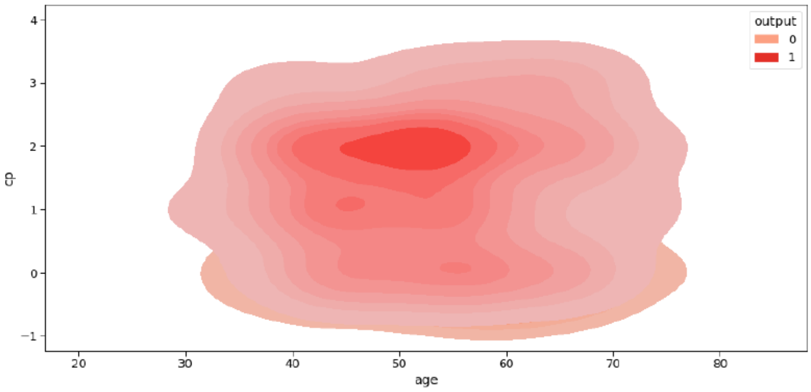

Fig. 8: Cp vs. Age Graph

Here Cp represents Chest Pain. From the visualization we can say that persons of age 50 with non-anginal chest type has the highest risk of having a heart attack.

(ix) Heatmap:

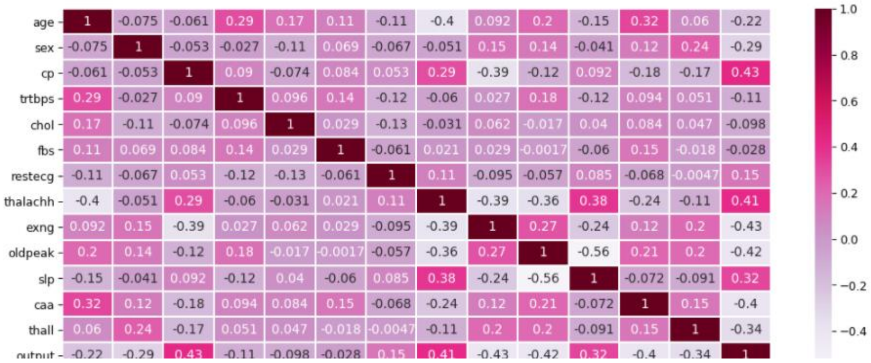

Fig. 9: Heatmap

From the heatmap, we can say that the data correlation shows us that cp, thalachh, and slp have the highest correlation with output which is our target variable.

C. Proposed Algorithms

In this project, 4 machine learning algorithms are used namely Decision Tree, Random Forest, Gaussian Naive Bayes and Logistic Regression. A Decision Tree (DT) represents a tree like structure where each number considered being a branch with an outcome. DT is a fundamental component of Random Forest, which are among the most powerful ML algorithms available today. DT uses a layered splitting process, where at each layer the information data is split into two or more groups so that elements of the same group are as homogenous as possible to each other. A decision tree is a classification as well as a regression technique. It works great when it comes to taking decisions on data by creating branches from a root, which are essentially the conditions present in the data, and providing an output known as a leaf.

The Random forest is basically a supervised learning algorithm which can be used for regression and classification tasks both. It is one of the most used algorithms because of its simplicity and stability. While building subsets of data for trees, the word "random" comes into the picture. A subset of data is made by randomly selecting x number of features (columns) and y number of examples (rows) from the original dataset of n features and m examples.
Random forests are more stable and reliable than just a decision tree.

DOI 10.17148/IJIREEICE.2021.9703

Naive Bayes is a classification technique based on the Bayes theorem. It is a simple but powerful algorithm for predictive modeling under supervised learning algorithms. The technique behind Naive Bayes is easy to understand. Naive Bayes has higher accuracy and speed when we have large data points. Gaussian Naive Bayes is a variant which supports continuous values and has an assumption that each class is normally distributed.

Logistic regression is a machine learning technique used for binary classification problems. It is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables; the independent variables can each be a binary variable a continuous variable.

IV.PERFORMANCE EVALUATION

A. Metrics

A confusion matrix is an idea of predict results on a classification problem where output can be of two or more classes. The quantity of true and false predictions are added up with tally esteems and separated by each class. The metrics used for the comparison purposes are accuracy, precision, and recall.

TABLE 1: **CLASSIFICATION REPORT**

		Actual Data	
Predicted Data	**Decision Tree**	20	6
		5	26
	Random Forest	20	6
		4	27
	Logistic Regression	19	7
		1	30
	Gaussian Naïve Bayes	20	6
		3	28

Accuracy is a proportion of correctly anticipated perception to the absolute number of perceptions. The accuracy of the proposed scheme defined as,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

To evaluate the model performance, precision and recall are also used where TP refers to true positive, TN refers to true negative, FP refers to false positive, and FN refers to false negative.

Recall quantifies the number of positive class predictions made out of all actual positive class values in the confusion matrix. It is also called sensitivity or True positive rate (TPR) and defined as,

$$Recall = \frac{TP}{TP + FN}$$

Precision quantifies the number of positive class predictions that actually belongs to the positive class and defined as,

$$Precision = \frac{TP}{TP + FP}$$

F1 score is a harmonic mean of precision and recall. It is a simple form that balances both precision and recall in one number and defined as,

$$Recall = 2 \times \frac{Precision + Recall}{Precision \times Recall}$$

All the parameters mentioned in the equation (1), (2), (3) and (4) are mentioned in the Table II.

B. Results and Discussion

We divided the dataset in the ratio of 80:20; 80% for the training set and 20% for the test set because the training set is large enough to yield statistically meaningful results and also it is a representative of the whole data set and we have used

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

DOI 10.17148/IJIREEICE.2021.9703

4 machine learning algorithms namely Logistic Regression, Gaussian Naïve Bayes, Random Forest and Decision Tree classifier to see the comparison and figure out the highest accuracy provided by these 4 algorithms.

TABLE II: RESULT CLASSIFICATION OF THE ALGORITHMS USED

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.88	0.87	0.86	0.86
Gaussian Naïve Bayes	0.84	0.84	0.84	0.84
Random Forest	0.82	0.83	0.82	0.82
Decision Tree	0.81	0.81	0.81	0.81

So based on the division of training and testing set, different algorithms are proposed. For statistical analysis, the confusion matrix is developed. Based on the confusion matrix, the percentage of accuracy, recall, precision, and F1 score are achieved. The values are given in Table I above.

Table III describes the related works and the accuracies of the related works. The accuracy of the model presented here consisting of 4 machine learning algorithms- Logistic Regression, Gaussian Naïve Bayes, Random Forest and Decision Tree classifier are placed at the top of the table. Logistic Regression reported the highest accuracy of 88%.

TABLE III: RELATED WORKS AND THEIR ACCURACY

Sl. No.	Model	Accuracy	Reference
1.	Logistic Regression	88%	Proposed Model
2.	Gaussian Naïve Bayes	84%	Proposed Model
3.	Random Forest	82%	Proposed Model
4.	Decision Tree	81%	Proposed Model
5.	K-Nearest Neighbor	58%	[6]
6.	Artificial Neural Network	86%	[6]
7.	Logistic Regression	77%	[8]
8.	Multilayer Perceptron	80.41%	[9]
9.	Decision Tree	80.40%	[11]
10.	Naïve Bayes	85.12%	[11]
11.	Convolutional Neural Network	83%	[13]

Algorithms such as K-Nearest Neighbor and Artificial Neural Networks are used in [6] for building a disease prediction model in diagnosing the Heart Diseases. The author also recommended that hybrid data mining algorithms performs well and promising accuracy results were attained in heart disease diagnosis. But the accuracies achieved by these two models are very less in comparison with other works. It is observed from [8] that using Logistic Regression, they achieved an overall accuracy of 77%. They built a classifier-based decision support system for heart disease classification. We also applied Logistic Regression algorithm where our accuracy came 88% which outperformed their accuracy. Multilayer Perceptron algorithm is used for heart disease classification by Gudadhe et al. [9]. But it reported an accuracy of 80.41% which is much lesser than our proposed models.

The observations inferred from this project are:

a) When comparing sex with the output, we can see that, the gender 0 is having very high chance of getting heart attack, while the gender 1 is having around 50% chance of getting the heart attack.
b) Comparing exercise Induced Angina with output, we can see an anomaly that, of the people who do not get pain due to physical activity are having more chance of getting the heart attack.
c) The number of major vessels with output is also having an anomaly, when the number of vessels is very less that is when 0, the chances of getting heart attack are around 75% and also when the number of major vessels is 4, there is 80% chance for getting heart attack The intermediate number of vessels are having less percentage of people getting heart attack.

DOI 10.17148/IJIREEICE.2021.9703

d) We can see from chest pain type that, the atypical angina is having more than 80% of people getting heart attack Followed by people with non-anginal and asymptomatic, where both of them are having more than 70% chance of getting heart attack.
e) The fasting blood sugar does not seem to have an effect on output, as people with blood sugar and also people without blood sugar is also having almost equal chances of getting heart attack.
f) When comparing Resting Electrocardiogram with output, we can see that, people with normal electrocardiogram is also having around 46% chances of getting heart attack. Also, when there is abnormality in the Resting Electrocardiogram, there is around 63% chance of getting heart attack.
g) While comparing slow peak with output, we can see that, the slow peak value 2 is having very high chance of getting heart attack; the other two slope values too have some significant chance of getting heart attack
h) From thallium stress test result we can see that, the value 2 is having 78% chance of getting heart attack and also value 0 is having 50% chance, the other two values are having less chance.

The primary motivation behind performing data cleaning and data visualization is to improve the prediction performance and guarantee quicker forecast. The accuracy detailed in the proposed work is superior to the cutting edge facilities. The related works where different machine learning models have been reported for predicting heart diseases and preventing heart attacks, our proposed work outperforms their work by a large amount.

V.CONCLUSION

In this work, UCI Heart Disease dataset has been used to analyze cardio logical data. The data included multiple factors such as age, sex, blood pressure, previous cardiac history, etc. We performed multiple data visualization to analyze what factor influences or contribute the most in the cardiac arrest. Lastly we compared four different highly reputed algorithms in the field of health analytics- Decision Tree, Random Forest, Gaussian Naïve Bayes, and Linear Regression. The Linear Regression algorithm showed the highest level of accuracy and precision (88% and 87% respectively). This is the highest accuracy reported than all the previous work done using the same dataset with different techniques.

In future, we would work to deploy this algorithm into a physical biomedical device and train, so that it can be used to determine the chances of heart attack of real patients.

ACKNOWLEDGMENT

The authors acknowledge Dr. Subhankar Bhattacharjee and Techno International New Town for providing the facilities and resources without which this research would not have been possible.

REFERENCES

[1].David W. Aha (1988) UCI Machine Learning repository [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
[2].Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.
[3].Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.
[4].Hongmei Yan, Yingtao Jiang, Jun Zheng, Chenglin Peng and Qinghui Li, " A Multilayer perceptron-based medical decision support system for heart disease diagnosis", ELSEVIER 2006.
[5].Chaitrali S. Dangare Sulabha S Apte, "Improve study of Heart Disease prediction system using Data Mining Classification techniques", International journal of computer application, 2012.
[6].Purusothaman, G., &Krishnakumari, P. "A survey of data mining techniques on risk prediction: Heart disease", Indian Journal of Science and Technology, 8(12), 1, 2015.
[7].Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.
[8].R. Detrano, A. Janosi, and W. Steinbrunn, "International application of a new probability algorithm for the diagnosis of coronary artery disease," American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989.
[9]. M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in Proceedings of International Conference on Computer and Communication Technology (ICCCT), pp. 741–745, Allahabad, India, September 2010.
[10].H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," Expert Systems with Applications, vol. 35, no. 1-2, pp. 82–89, 2008.
[11].S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in Proceedings of IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2008), pp. 108–115, Doha, Qatar, March-April 2008.
[12].R. Hazra and S. Majhi, "Detecting Respiratory Diseases from Recorded Lung Sounds by 2D CNN," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-6.
M. Banerjee and S. Majhi, "Multi-class Heart Sounds Classification Using 2D-Convolutional Neural Network," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-6.