Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

Comparative Analysis of Heart Disease Prediction using Machine Learning Classification Techniques

Naveen Reddy Navuluri
Department of Electronics and communication
Maulana Azad National Institute of Technology
Bhopal, India

***Abstract*—BHeart is the main component of the human body and without it the body can't function. It provides the flow of blood to different organs and body parts. It purifies the blood by removing the carbon dioxide(co2). It is also known as cardiovascular disease, it creates many risk factors for a human, including death. Researchers apply several data mining and machine learning techniques to analyse huge complex medical data, helping healthcare professionals to predict heart disease.[1] The research paper presents various attributes available from the dataset which has 300 instances and 14 features/attributes which will be used to perform the given problem. The main idea for the paper is to do a comparative research on machine learning classification techniques and to show which is the best performing algorithm to predict the heart disease at a much earlier phase to avoid the repercussions that would be faced by the patients later.**

Keywords—Comparative analysis, machine learning,Heart disease prediction, random forest, Classification

I. INTRODUCTION

Diseases are the biggest threat to human life and cardiovascular are the most dangerous ones. Thee disease are increasing on day-to-day basis. The treatment of heart problems has recently been stated in a study that has received huge attention in the medical system worldwide. On median, 17.7 million deaths result from heart disease, which counts for about 31% throughout the world in 2016, according to World Health Organization (WHO).The number of cardiac cases, which is the focus of this study, shows that 82 percent of cases are from low and middle-income countries, 17 million people are under 70 years old and susceptible to non-infectious diseases, 6.7 million people are affected by stroke, and 7.4 million people have heart disease (WHO, 2016)[3]. The machine learning techniques that are used by the authors are Random Forest classifier, decision tree classifier , KNN , Support vector machine and Naive Bayes classifier.

Determining the probability of having the cardiac disease is tough, and we need a proper dataset for achieving a proper accuracy for success of determination. Hence, a classification comparison is performed to achieve so.

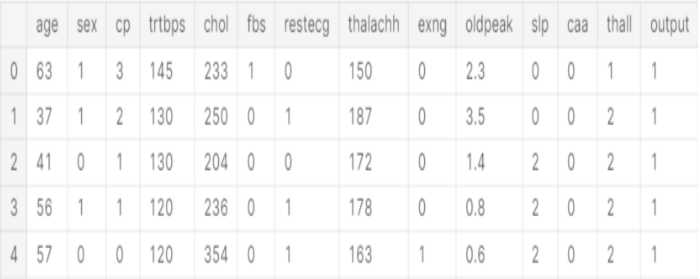
II. LITERATURE OVERVIEW

There are many research works done on heart diseases diagnosis, The results revealed that LR had a higher accuracy of 85.68 percent than XGBoost, which had a lower accuracy of 84.46 percent. Bhatet et cetera After that, he devised a model for diagnosing heart illness that combines a multilayer perceptron network (MLP) with a backpropagation method. The suggested model has a reduced error and an enhanced accuracy of 80.99 percent, according to the results. To forecast heart illness, Abushariah et al. used ANN and an adaptive neuro-fuzzy inference system (ANFIS). ANFIS has the lowest accuracy of 75.93 percent, while ANN has the highest accuracy of 87.04 percent. Hasanet and his colleagues[4]. [5]used seven machine learning algorithms: LR, ANN, KNN, NB, SVM, DT, and RF with three feature selections: minimal-redundancy-maximal-relevance (mRMR), Relief, and Shrinkage and Selection Operator (LASSO) to predict heart disease. LR with Relief achieved the highest accuracy of 89% compared to other techniques

III. METHODOLOGY

A. DATA PREPARATION

The dataset is collected from a internet know dataset known as the heart attack analysis and prediction dataset. Because the data is difficult to come by, the only way to run the model and make a forecast was to use data from a reliable source. The dataset contains various attributes such as Age,gender, fasting blood sugar, serum cholestrol , maximum heart rate achieved, rating blood pressure , exercise indexed angina, oldpeak , number of vessels etc. The length of the dataset is 300 with 14 different attributes mentioned above.

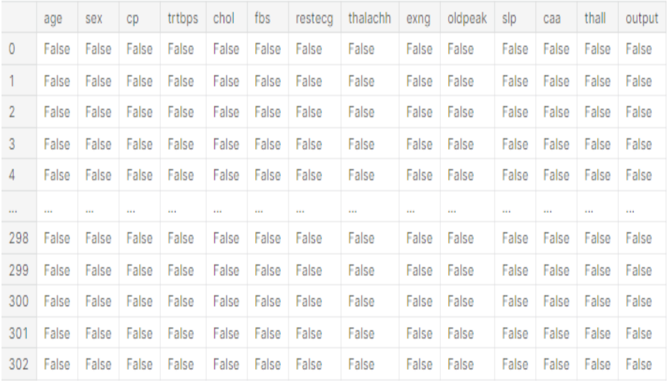


	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 1: Dataset Overview

B. DATA PREPROCESSING

Real world things contains errors in it so does our data so for this the preprocessing is a good step to improve it. The speed of the method is determined on whether or not the data has been preprocessed. Better the preprocessing done better will be the result of the model which will one use. Firstly the author checks for all the null values and then remove the id column which won’t hinder the results.



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	False	False	False	False	False	False	False	False	False	False	False	False	False	False
299	False	False	False	False	False	False	False	False	False	False	False	False	False	False
300	False	False	False	False	False	False	False	False	False	False	False	False	False	False
301	False	False	False	False	False	False	False	False	False	False	False	False	False	False
302	False	False	False	False	False	False	False	False	False	False	False	False	False	False

Figure 2: Checking Null Values in the dataset

C. FEATURE SELECTION

Features are important component for getting proper results from the algorithm used. Visualisation helps us to see the different features and how they would make an impact on the results. Figure 3 shows which gender is more affected by the disease. Again figure 4 shows the visualisation of which age is more impacted by the disease which give a great factor to the model for the deciding weather the patient will be prone to the disease or not. Finally a coorelation matrix would help the authors to get an idea of the relations between all the features/attributes which can be seen in figure 5.

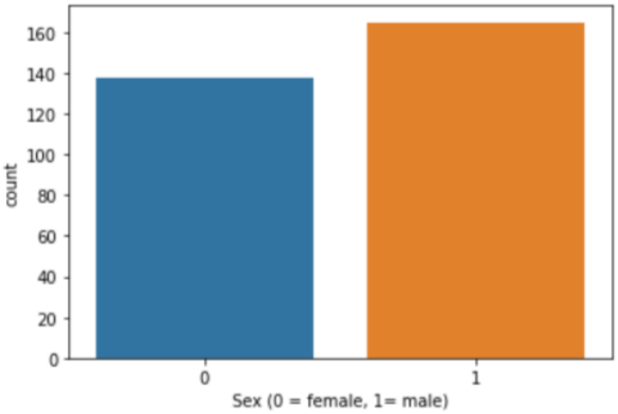


Figure 3:

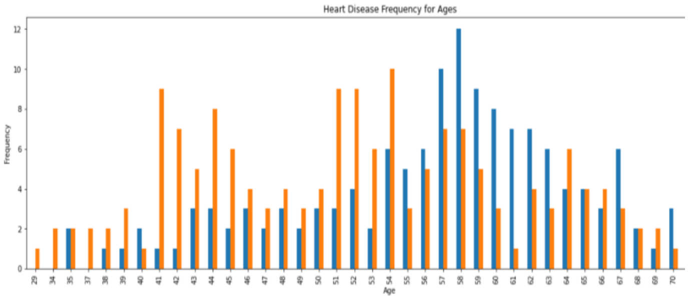


Figure 4:

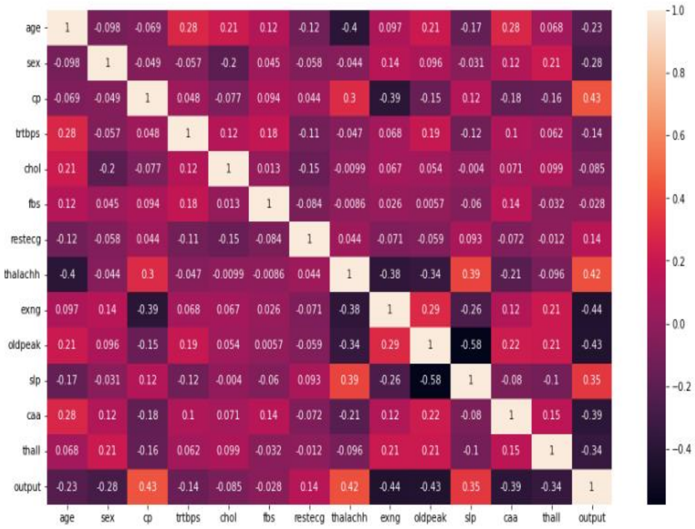


Figure 3: Correlation map between the features

D. MODEL ARCHITECTURE

1) RANDOM FOREST CLASSIFIER

It's a method that falls under the ensemble model umbrella. Combining classification and regression techniques, it may be utilised to create a good prediction model. Decision trees are employed as the basis estimators in this study. Decision trees are a poor predictor on their own, but they improve when paired with other decision trees. In classification tasks, decision trees vote on how to categorise a single instance of input data, and in regression tasks, they output the class that is the mode of the classes or the mean of forecasts. We can avoid parameter tweaking and decrease overfitting this way

2) SUPPORT VECTOR MACHINE

Support Vector Machines employ a linear model to implement nonlinear class boundaries. Support vectors (lines or hyperplanes) are used to distinguish the target classes. To deal with a nonlinear situation, the model applies several transformations to the input using a mapping function before training a linear SVM model to classify the data in a higher-dimensional feature space.

3) NAIVE BAYES CLASSIFIER

It is another classification technique in which there is collection of algorithms based on the bayes theorem. It is a classifier so it is used to discriminate different objects based on certain features. The main task of bayes is classification task with the help of the bayes theorem.

4) DECISION TREE CLASSIFIER

A decision tree is a tool for making decisions that uses a tree-like model of options and their possible consequences, such as chance event outcomes, resource costs, and utility. It's one way to demonstrate an algorithm that's entirely made up of conditional control statements.

5) KNN

We would first choose the number of clusters k, and then assume the cluster's centroid. Any random item, or the first k objects in a series, might be used as the starting centroid.

So the procedure is broken down into three steps: first, we find the centroid's coordinates, then we calculate how distant each item is from the centroid, and finally we group the objects depending on the minimum distance. We can get a centroid by following the technique.

IV. EXPRIMENTAL RESULTS

The process of picking a dataset and preprocessing it to make a good one in order to increase the accuracy of the models chosen. The models chosen have excellent accuracies, as seen in the table below. Starting with Random Forest classifier which gives ~89 percent accuracy followed by naive bayes which gives ~94 percent, Decision tree gives ~86 percent, support vector machine gives ~92 percent and finally k nearest neighbhor gives ~89 percent. Hence, after observing we get to know that naive bayes performs the best.

V. CONCLUSION

Hence, the authors finally observe that naive bayes gets the best accuracy compared to support vector machine, decision tree, knn, random forest classifier after performing all the modes.The accuracy would be used by the other researchers to while choosing the best model out amongst the other models and would definitely help in treating the breast cancer most of the time.

VI. FUTURE SCOPE

Furthermore, we can use the following dataset to test out different regression models and neural networks and see how does it perform.Secondly, we can try this algorithm on a different dataset to know how does it perform and what problems it faces during the testing of the model.The research intended by the authors would help in the development of better and more productive and trustable prediction method of illness, which will in turn not help the medical community but also to many other communities and people in the world.

VII. REFRENCES

[1] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y

[2] M. Sanz, A. Marco del Castillo, S. Jepsen et al., “Periodontitis and cardiovascular diseases: consensus report,” Journal of Clinical Periodontology, vol. 47, no. 3, pp. 268–288, 2020.

[3] World Health Organization. http://www.who.int/cardiovascular diseases/en. 2019.

[4] T. T. Hasan, M. H. Jasim, and I. A. Hashim, “Heart disease diagnosis system based on multi-layer perceptron neural networkand support vector machine,” International Journal of Current Engineering and Technology, vol. 77, pp. 2277–4106, 2017.

[5] Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, Eman M. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method", Complexity, vol. 2021, Article ID 6663455, 10 pages, 2021.

[6] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.

[7] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 704-706

[8] M. T. Islam, S. R. Rafa and M. G. Kibria, "Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means," *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1-6

[9] S. Ambekar and R. Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697423..

[10] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333

[11] S. Modi and M. H. Bohara, ”Facial Emotion Recognition using Convolution Neural Network,” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1339-1344, doi: 10.1109/ICICCS51141.2021.9432156

[12] Purushottam, K. Saxena and R. Sharma, "Efficient heart disease prediction system using decision tree," International Conference on Computing, Communication & Automation, 2015, pp. 72-77, doi: 10.1109/CCAA.2015.7148346.

[13] A. Chauhan, A. Jain, P. Sharma and V. Deep, "Heart Disease Prediction using Evolutionary Rule Learning," 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT), 2018, pp. 1-4, doi: 10.1109/CIACT.2018.8480271.

[14] A. Lakshmanarao, A. Srisaila and T. S. R. Kiran, "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 994-998, doi: 10.1109/ICICV50876.2021.9388482.

[15] S. Bhoyar, N. Wagholikar, K. Bakshi and S. Chaudhari, "Real-time Heart Disease Prediction System using Multilayer Perceptron," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-4, doi: 10.1109/INCET51464.2021.9456389.