2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) | 978-1-7281-8337-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362842

Heart Disease Prediction using Machine Learning Techniques

Vijeta Sharma
Computer Science, DST-Centre for Interdisciplinary Mathematical Science, Institute of Science,
Banaras Hindu University
Varanasi, India
vijeta.it@gmail.com

Shrinkhala Yadav
Computer Science, DST-Centre for Interdisciplinary Mathematical Science, Institute of Science,
Banaras Hindu University
Varanasi, India
Shrinkhalayadavbhu2018@gmail.com

Manjari Gupta
Computer Science, DST-Centre for Interdisciplinary Mathematical Science, Institute of Science,
Banaras Hindu University
Varanasi, India
manjari@bhu.ac.in

***Abstract*— As per the recent study by WHO, heart related diseases are increasing. 17.9 million people die every-year due to this. With growing population, it gets further difficult to diagnose and start treatment at early stage. But due to the recent advancement in technology, Machine Learning techniques have accelerated the health sector by multiple researches. Thus, the objective of this paper is to build a ML model for heart disease prediction based on the related parameters. We have used a benchmark dataset of UCI Heart disease prediction for this research work, which consist of 14 different parameters related to Heart Disease. Machine Learning algorithms such as Random Forest, Support Vector Machine (SVM), Naive Bayes and Decision tree have been used for the development of model. In our research we have also tried to find the correlations between the different attributes available in the dataset with the help of standard Machine Learning methods and then using them efficiently in the prediction of chances of Heart disease. Result shows that compared to other ML techniques, Random Forest gives more accuracy in less time for the prediction. This model can be helpful to the medical practitioners at their clinic as decision support system.**

Keywords*— *Heart Disease, Machine Learning, Artificial Neural Network, Support Vector Machine, Classification

I. INTRODUCTION

Healthcare is one of the primary focus for humanity. According to WHO guidelines, good health is the fundamental right for individuals. It is considered that appropriate health care services should be available for regular checkup of one's health. Almost 31% of all deaths are due to heart related disease in all over the world. Early detection [1] and treatment of several heart diseases is very complex, especially in developing countries, because of the lack of diagnostic centers and qualified doctors and other resources that affect the accurate prognosis of heart disease. With this concern, in recent times computer technology and machine learning techniques are being used to make medical aid software as a support system for early diagnosis of heart disease. Identification of any heart related illness at primary stage can reduce the death risk. Various ML techniques are used in medical data to understand the pattern of data and making prediction from them. Healthcare data are generally massive in volumes and complex in structure. ML algorithms are capable to handle the big data and mine them to find the meaningful information. Machine Learning algorithms learn from past data and do prediction on real time data. This sort of ML framework for coronary illness expectation can encourage cardiologists in taking quicker actions so more patients can get medicines within a shorter timeframe, thus saving large number of lives.

Machine Learning is a branch of AI research [2] and has become a very popular aspect of data science. The Machine Learning algorithms are designed to perform a large number of tasks such as prediction, classification, decision making etc. To learn the ML algorithms, training data is required. After the learning phase, a model is produced which is considered as an output of ML algorithm. This model is then tested and

ISBN: 978-1-7281-8337-4/20/$31.00 ©2020 IEEE

validated on a set of unseen real time test dataset. The final accuracy of the model is then compared with the actual value, which justify the overall correctness of predicted result.

Lots of efforts has already been done to predict the heart disease using the ML algorithms by authors [3-5], but this is an additional effort to do the experiment on benchmarking UCI heart disease prediction dataset while comparing the four popular ML technique to check the most accurate ML technique.

The paper is structured as follows: section 2 contains the details of ML techniques used in this research work. Section 3 shows the methodology, section 4 summaries with result of this work and section 5 list out the conclusion.

II. Related Works

Lots of research work have been done for assessment of the classification accuracies of different machine learning algorithms by using the Cleveland heart disease database which is uninhibitedly accessible at an online data mining repository of the UCI. Authors of [6] achieved 77% prediction accuracy by applying logistic regression algorithm on this dataset. In this study, authors [7] did an enhanced work by doing comparison of global evolutionary computation approaches and thus they observed higher prediction accuracy. Authors Bayu Adhi Tama, et.al [8] in their work suggested a research related to the identification of diabetes malady with utilization of ML procedures. This disease was viewed as incredibly a thrust area of ML. Roughly 285 million individuals around the globe were experiencing diabetes as per a study directed by International Diabetes Federation (IDF). As a matter of fact, detection of type 2 diabetes at beginning phase isn't a simple undertaking, yet research done by the authors, in which data mining was used on the grounds that it gives the best results, helped in the disclosure of information from accessible data. In their research, they utilized SVMs for the mining of related information of various patients from the previous records. The on-time acknowledgment of type 2 diabetes gave assistance in the taking of legitimate treatment and avoid the risk of expanding.

Yu-Xuan Wang, et.al. have explored different applications that demonstrated the significance of the ML methods in various areas [9]. They proposed a new technique for the designing of a working framework. The approach used the distinct machine learning procedures. After getting the proper result from the data miner, the whole information assembled from the structure was inspected. In light of the various tests, it was seen that proposed approach gave proficient results. Zhiqiang Ge, et.al, (2017) proposed a work on analytics and data mining applications, which was done prior. These procedures were used in business area for various purpose of perspectives. Here they have explored 8 unsupervised and 10 supervised learning algorithms [10]. In their research, they showed an application work for the semi-supervised type learning algorithms. In industry method, it was seen that roughly 90%-95% applications utilized both the unsupervised and supervised machine learning procedures. Consequently, it was portrayed that the Machine Learning methods play an indispensable part in the planning of different novel applications for domains like medical services and industry.

III. Machine Learning Techniques

We have chosen four popular ML techniques to develop the heart disease prediction model. Details of these techniques are as follows:

A. Support Vector Machine

Support Vector Machine [11] is a classification technique of Machine learning to, which is used to analyze data and discover patters in classification and regression analysis. SVM is typically mull over when data is characterized as two class problem. In this strategy, data is characterized by finding the best hyper plane that isolates all data points of one class to the other class. The higher separation or edge between the two classes is, the better is the model, considered. The data points lying on limit of the margin are called as support vectors. The actual basis of SVM is mathematical methods used to design complex real-world problems. We have chosen SVM for this experiment because our dataset - Cleveland Heart Disease Dataset CHDD has multi class to predict based on various parameters. In SVM, the mapping of training data is to be done with a function called kernel (Kernels of SVM), these are - linear kernel, quadratic kernel, polynomial kernel, Radial Basis Function kernel, Multilayer Perceptron kernel, etc. Apart from the kernel's functionalities in SVM, few more methods are available such as quadratic programming, sequential minimal optimization, and least squares.

While building up the model with SVM, most challenging thing is kernel selection and method selection to evade the issue of overfitting and underfitting. Since our dataset is having enormous number of parameters and instances too. So, we had choice of selecting the RBF or linear kernel. Thus, final model developed by SVM requires tested and validated against actual data.

B. Decision Tree

Decision Tree algorithm [12] in Machine Learning is used to develop the Classification models. This classification model is based on the tree-like structure. This comes under the category of supervised learning, where the target result is already known. Both the categorical and numerical data can be applied on Decision tree algorithm. Decision tree consists of root node, branches and leaf nodes. Data is evaluated on the basis of traversing path from the root to a leaf node. For our dataset - CHDD, a total of 283 tuples were assessed down the decision tree. They potentially came to a positive or negative assessment for the heart disease prediction. These were compared to the actual parameters to check for the false positives/false negatives which show the accuracy, specificity, and sensitivity of the model.

C. Naive Bayes

This supervised machine-learning algorithm is based on the Bayes' Theorem [13], which consider that features are statistically independent to each other's. The Naïve Bayes Classifier [14] is used with high dimensionality of inputs data. Naive Bayes method is highly useful in computer vision application. In particularly, it has proven itself to be a classifier with good results.

D. Random Forest Classification

Random Forest [15] is a troupe of unpruned classification-based trees. It gives amazing performance with concern to number of real-life problems, as it is non effective to noise in the dataset and risk of overfitting is also very less. In comparison to many other tree-based algorithms, it works faster than others and generally improves accuracy for testing and validation data. Random forests are the aggregation of the predictions of individual decision tree algorithm. There are various choices to tune the performance of random forest when constructing a random tree.

IV. METHODOLOGY

Below steps shows the method through which chest disease prediction model has developed.

A. Data Collection

The Cleveland Heart Disease Dataset accessible online on the UCI Repository has been used in our research work [16]. The 14 attributes considered are as follow:

TABLE I. ATTRIBUTES

S. No.	Attribute	Desc.	Mean Value
1	age	in years	54.434
2	Sex	Male, Female	0.696
3	cp	Angina, abnang, notang, asympt	0.942
4	trstbps	Resting Blood Pressure in mm hg	131.612
5	chol	Serum Cholesterol in mg/dl	246
6	fbs	fasting blood sugar- 1 if >120 mg/dl, 0 if <120 mg/dl	0.149
7	restecg	Electrocardiographic Results	0.53
8	thalach	Maximum Heart Rate observed	149.114
9	exang	exercise with angina has occurred	0.337
10	oldpeak	ST depression induced through exercise	1.072
11	slope	slope of the ST segment	1.385
12	thal	Number of major vessels ranging from 0 - 3 color by fluoroscopy	0.754
13	ca	Heart status	2.34
14	Target	Output Class	

So, there are 1025 instances in total, which have been used in this research work. Table 1 shows the mean value of each attribute.

The dataset used have missing values too, which is handled through data preprocessing using some statistical techniques. Class Value 1, is interpreted as "tested positive for the disease", and Class value 0, means "tested negative for the disease". The dataset has been divided into certain percentage, such as 80% data has been considered as training data and rest 20% as testing data. Table II shows a sample training data and Table III is having the sample testing data.

B. Data Preprocessing

Various attributes of the original dataset contain missing values which can lead to imprecise result and may reduce the accuracy of model. To overcome this problem, replace missing value way is an optimal solution by using a method "mean of column". This method replaces 0 with either taking the average of neighborhood values or mean values of it [17]. Then accordingly it manipulates the 0 value with the newly the calculated value. After that the values of the dataset were changed from Numeric to Nominal in order to make the dataset compatible with the ML Techniques used.

TABLE II. SAMPLE TRAINING DATA

52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1

TABLE III. SAMPLE TESTING DATA

52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	156	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1

C. Building Model

The model is built in Weka Data Mining Tool. The Waikato Environment for Knowledge Analysis (WEKA) [18] is an open source machine learning software which developed by Waikato University, New Zealand. The software easily processes various standard data mining tasks such as data preprocessing, clustering, classification, regression, visualization and feature selection. It gives an easy environment to load data in the form of files, URLs or databases. Attribute Relation File Format (ARFF) [19], CSV, C4.5's and Lib SVMs file formats supported by the software. It analyzes and visualizes the confusion matrix, true positive, precision, recall and false negative etc. in a convenient way. It is an open source as well as platform independent, portable

and GUI-based software and packed with a large collection of advanced machine learning techniques such as deep learning. image processing algorithms etc. Four accuracy measures have been considered for comparison of the four models, they are as follows:

- Precision or Positive Predictive Value

It is the average probability of relevant retrieval. Precision = Number of true positives/Number of true positives + False positives.

- Recall

It is the average probability of complete retrieval. Recall= True positives/True positives + False negative.

- Accuracy

The accuracy of a classifier is given as the percentage of total correct predictions divided by the total number of instances. Accuracy = [Number of True Positives + True Negatives]/[Total Instances]

- ROC Area

ROC depicts the performance trade-off between the true positive rate (TPR) and false positive rate (FPR) of a classification model.

TPR = [Number of True Positives]/[Number of True Positives + False Negatives]

FPR = [Number of False Positives]/[Number of False Positives + True Negatives]

10-fold Cross Validation [20] is used to divide the data into two sections which are training and testing datasets.

D. Accuracy measurement of Model

Precision, Recall, ROC (Receiver Operating Characteristic) and % accuracy performance parameters of all the models have been considered for the analysis and comparison. Table IV shows the performance accuracy of models. Classifier is usually considered “GOOD” when ROC value less than 0.80, “FAIR” when ROC value is 0.77. The ROC value very close to 1 is considered as best model with high accuracy.

V. RESULT

At the end of our experiment, result shows that SVM and Random Forest performed very well in comparison to Gaussian Naive Bayes and Decision tree. Model developed with SVM gives 98% accuracy which is 8% greater than the Naïve Bayes and approximately 13% greater than Decision tree. In the same way, model build with Random Forest gives best prediction result with 99% accuracy, which is itself more accurate than our second best SVM model for heart disease prediction. Unfortunately, we did not find decision tree suitable for our data.

TABLE IV. PERFORMANCE MEASURE OF MODELS

Model	Precision	Recall	ROC Area
SVM	0.995	0.995	0.995
Random Forest	0.997	0.997	1.00
Decision Tree	0.851	0.848	0.889
Naïve Bayes	0.904	0.904	0.966

VI. CONCLUSION

Through this research we have attempted to analyze the various machine learning techniques and anticipate if someone in particular, given different individual attributes and indications, will get coronary illness or not. The primary thought process of our report was to looking at the exactness and analyzing the reasons behind the variation of different algorithms. We have used Cleveland dataset for heart diseases which contains 1025 instances and used percent split to divide the data into two sections which are training and testing datasets. We have considered 14 attributes and implemented four different algorithms to examine the accuracy. By the end of the implementation part, we have discovered that Random Forest is giving the maximum accuracy level in our dataset which is 99 percent and Decision Tree is playing out the least with an accuracy level of 85 percent. Probably for other instances and different datasets other algorithm may work in better manner however for our situation, we have discovered this outcome. Also, on the off chance that we increment the number of training data, maybe we can find more accurate result but it will take more time to process and the system will be slower than now as it will be more perplexing and will be handling more data. In this way, considering these potential things we took this choice, which is better for us to work with.

REFERENCES

[1] https://www.who.int/hrh/links/en/

[2] https://en.wikipedia.org/wiki/Machine_learning

[3] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.

[4] S. Dhar, K. Roy, T. Dey, P. Datta and A. Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Diseases," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777531.

[5] C. Raju, E. Philipsy, S. Chacko, L. Padma Suresh and S. Deepa Rajan, "A Survey on Predicting Heart Disease using Data Mining Techniques," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2018, pp. 253-255, doi: 10.1109/ICEDSS.2018.8544333.

[6] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid,S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” The American journal of cardiology, vol. 64, no. 5, pp. 304-310, 1989.

[7] B. Edmonds, “Using localised ’gossip’ to structure distributed learning,” 2005.

[8] Fsdfsdf BayuAdhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, “Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine”, Vol. 11, issue 3, pp. 12-23, 2008.

[9] Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, “Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization”, Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.

[10] ZhiqiangGe, Zhihuan Song, Steven X. Ding, Biao Huang, “Data Mining and Analytics in the Process Industry: The Role of Machine Learning”, 2017 IEEE Translations and contentmining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.

[11] https://en.wikipedia.org/wiki/Support_vector_machine

[12] https://en.wikipedia.org/wiki/Decision_tree_learning

[13] https://en.wikipedia.org/wiki/Bayes27_theorem

[14] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[15] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[16] https://archive.ics.uci.edu/ml/datasets/heart+disease

[17] https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/slides/Class5-DataMiningWithWeka-2013.pdf

[18] https://wekatutorial.com/

[19] https://www.cs.waikato.ac.nz/ml/weka/arff.html

[20] https://weka.8497.n7.nabble.com/10-fold-cross-validation-in-WEKA-td34105.html