RESEARCHERID
THOMSON REUTERS

[Lavanya, 5(8): August 2018] ISSN 2348 – 8034
DOI- 10.5281/zenodo.1400571 Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

HEART DISEASE PREDICTION USING RANDOM FOREST ALGORITHM

Thota Lavanya[*1], Nimmala Satyanarayana[2] & Manasa.K[3]
[*1]Assistant Professor, Department of CSE
[2&3]Associate Professor, Dept. of CSE, CVR College of Engineering

ABSTRACT

Present days one of the major application areas of machine learning algorithms is medical diagnosis of diseases and treatment. Machine learning algorithms also used to find correlations and associations between different diseases. Nowadays many people are dying because of sudden heart attack .Prediction and diagnosing of heart disease becomes a challenging factor faced by doctors and hospitals both in India and abroad. In order to reduce number of deaths because of heart diseases, we have to predict whether person is at the risk of heart disease or not in advance. Data mining techniques and machine learning algorithms play a very important role in this area. Many researchers are carrying out their research in this area to develop software that can help doctors to take decision regarding both prediction and diagnosing of heart disease. In this paper we focused on how data mining techniques can be used to predict heart disease in advance such that patient is well treated. We used different algorithms for comparative analysis but random forest algorithm has shown highest accuracy in prediction. We used Random forest machine learning algorithms supported by WEKA to predict heart disease in advance. Dataset contains 303 samples and 14 input features as well as 1output feature. The dataset is available in UCI Machine Learning Repository; we used 65% data for training and 35% data for testing. The algorithm has shown 0.763 precision and 0.935 recall in predicting negative class tupples.

Keywords: *Classification, Heart disease machine learning, C4.5, J48 algorithm, Random Forest algorithm.*

I. INTRODUCTION

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc [1]. The large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. As per world health organization (WHO) latest statistics the highest mortality rate of people, both in India and as well as in abroad is due to heart disease. So it is vital time to check this death toll by correctly identifying the disease in initial stage. It is really a headache for all doctors both in India and abroad. Now a day's doctors are adopting many scientific technologies and methodology for both identification and diagnosing not only common disease, but also many fatal diseases. The successful treatment is always attributed by right and accurate diagnosis. Doctors may sometimes fail to take accurate decisions while diagnosing the heart disease of a patient, therefore heart disease prediction systems which use machine learning algorithms assist in such cases to get accurate results [2]. In this article, we especially focused on important attributes like, high blood pressure, abnormal blood lipids, use of tobacco, obesity, physical inactivity, diabetes, age, gender, family generation, etc to predict whether person is suffering with heart disease or not. Many heart strokes are happening because of accumulation of cholesterol in blood vessels or blood clot in blood vessels in arteries which supply blood to the heart muscles [3]. Internal and external view of heart is as shown in figure1, figure 2 given below.

ISSN 2348 – 8034
Impact Factor- 5.070

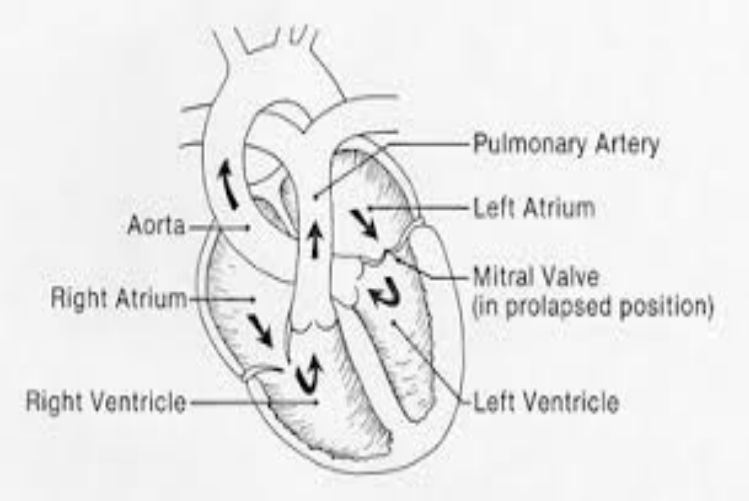

Figure 1. Internal view of the heart

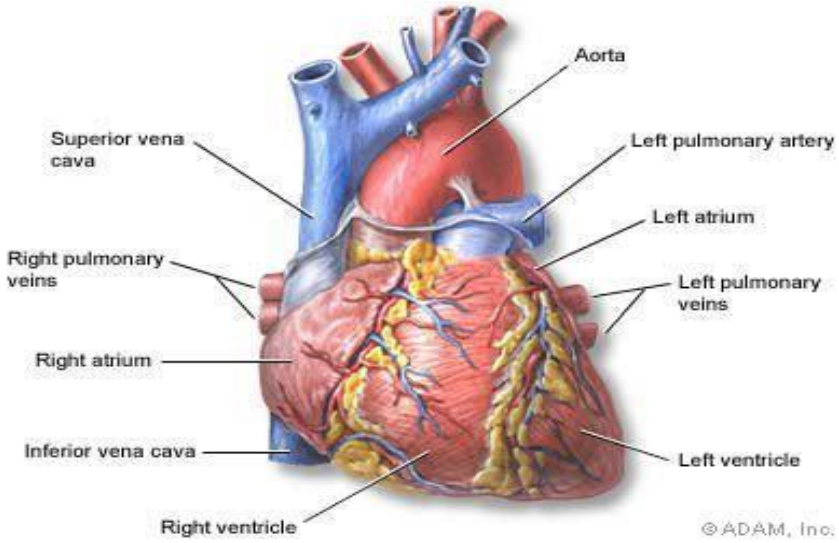

Figure 1.External view of the heart

II. MACHINE LEARNING ALGORITHMS

In this paper we our focus is how we can train the Machine to learn from the medical data so it can predict and treat the disease. Learning can be defined in general as a process of gaining knowledge through experience. We humans start the process of learning new things from the day we are born. This learning process continues throughout our life where we try to gather more knowledge from our surroundings and through our experience [4]. Machine Learning (ML) is a sub-field of AI whose concern is the development, understanding and evaluation of algorithms and techniques to allow a computer to learn [4]. ML intertwines with other disciplines such as statistics, human psychology and brain modeling. Human psychology and neural models obtained from brain modeling help in understanding the workings of the human brain, and especially its learning process, which can be used in the formulation of ML algorithms. Since many ML algorithms use analysis of data for building models, statistics plays a major role in this field [5]. ML algorithms need a dataset, which is collection of records are instances where each instance consist of attributes. The input attributes are the information given to the learning algorithm and the output

attribute contains the feedback of the activity on that information. The value of the output attribute is assumed to depend on the values of the input attributes.

Machine learning algorithms are broadly classified as Supervised and unsupervised learning algorithms. In supervised learning instances and its predefined classes are there. The model predicts the class membership of an instance. In unsupervised learning only instance are there based on the similarities between the instances, they are segmented as groups. In this paper we used supervised learning techniques to predict the class label of test instances. Supervised learning algorithms also called as Classification Algorithms [6].

2.1 Classification algorithms

Algorithms that classify a given instance into a set of discrete categories are called classification algorithms [4]. These algorithms work on a training set to come up with a model or a set of rules that classify a given input into discrete output values. Most classification algorithms can take inputs in any form, discrete or continuous. Although some of the classification algorithms require all of the inputs also to be discrete. The output is always in the form of a discrete value. Decision trees and Baye's nets are examples of classification algorithms. In this paper we used Random forest classifier for prediction, in the random forest approach; a large number of decision trees are created. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model. Random forest algorithm also works on the principle of decision tree; rest of the section explains how decision tree algorithm works.

2.1.1 The decision tree algorithm

J48 is a decision tree learner based on C4.5. The C4.5 is an update of the ID3 algorithm. A decision tree classifies a given instance by passing it through the tree starting at the top and moving down until a leaf node is reached [5]. The value at that leaf node gives the predicted output for the instance. At each node an attribute is tested and the branches from the node correspond to the values that attribute can take. When the instance reaches a node, the branch taken depends on the value it has for the attribute being tested at the node. The ID3 algorithm builds a decision tree based on the set of training instances given to it. It takes a greedy top-down approach for the construction of the tree, starting with the creation of the root node. At each node the attribute that best classifies all the training instances that have reached that node is selected as the test attribute. At a node only those attributes are considered which were not used for classification at other nodes above it in the tree. To select the best attribute at a node, the information gain for each attribute is calculated and the attribute with the highest information gain is selected. Information gain for an attribute is defined as the reduction in entropy caused by splitting the instances based on values taken by the attribute.

2.1.2 Perforamnce measures used for classifier evaluation

The classifier's evaluation is most often based on prediction accuracy (the percentage of correct prediction divided by the total number of predictions). If the error rate evaluation is unsatisfactory, we must return to a previous stage of the supervised Machine learning process. A variety of factors must be observed, perhaps relevant features for the problem are not being considered, may need a larger training set is required, the dimensionality of the problem is too high, the selected algorithm may not suitable or parameter tuning is needed [7].

Table 1. Measures and Formula

Classifier	
Accuracy	$\frac{TP+TN}{(P+N)}$
Classifier Error rate	$\frac{FP+FN}{(P+N)}$
Recall	$\frac{TP}{P}$
Precision	$\frac{TP}{}$

ISSN 2348 – 8034
Impact Factor- 5.070

	(TP +FP)
F-Measure	$\frac{(2\times precision\times recall)}{(precision+recall)}$

Where P is total number of positive records, N is total number of negative records, TP refers to the positive records which are correctly labeled by the classifier, TN is the negative records which are correctly labeled by the classifier, FP is the negative records which are improperly labeled as positive, and FN is the positive records which are incorrectly labeled as negative.

III. EXPERMENTAL SETUP

In this paper we used a decision tree based Random forest classification algorithm which is implemented in WEKA (Waikato Environment for Knowledge Analysis). WEKA is a collection of various ML algorithms, implemented in Java, which can be used for data mining problems. Apart from applying ML algorithms on datasets and analyzing the results generated, WEKA also provides options for pre-processing and visualization of the dataset. It can be extended by the user to implement new algorithms. The details of the attributes considered for experimental analyses are represented in Table 1 .We have taken heart disease Cleveland data set of 303 diagnostic records. Each record is classified to one of the Five class labels, whose values are {0,1,2,3,4}, if predicted class label value is zero means arteries are narrowed below 50%, if predicted value is greater than or equal to 1, arteries narrowed more than 50%, based on the value predicted we can say person is victim of heart disease or not.

Table 2. Attributes Details

Attribute Number	Attribute Name
1	age
2	sex
3	cp
4	trestbps
5	choi
6	fbs
7	restesg
8	thalach
9	exang
10	oldpeak
11	slop
12	ca
13	thal
14	num

Each record consists of 14 attributes as represented in Table 1. Data set is available on online at **https://archive.ics.uci.edu/ml/datasets/Heart+Disease.** Here 65% records are used to train the model remaining 35% records are used to test the model. Table 3 gives the performance details of Random forest algorithm used for classifying BP patient records. Here Table 2 represents confusion matrix of Random Forest classifier.

ISSN 2348 – 8034
Impact Factor- 5.070

Table 3. Class wise Accuracy

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
0	0.935	0.300	0.763	0.935	0.841
1	0.111	0.192	0.091	0.111	0.100
2	0.250	0.113	0.250	0.250	0.250
3	0.077	0.042	0.333	0.077	0.125
4	0.000	0.016	0.000	0.000	0.000

IV. CONCLUSION

In this paper we used random forest classifier to predict whether the patient is suffering with heart disease or not. The algorithm has shown 53.7736 % accuracy in predicting the class label of unknown records. But in predicting the class label of records whose class label value is zero, means people with less than 50% narrowed arteries the algorithm has shown 0.763 precision and 0.930 accuracy, the evaluation criteria proved that, random forest algorithms are more effective and efficient classification techniques for the prediction of heart disease risk among patients. The considerable point about the algorithm used is, it out performs in accuracy while predicting the zero class labels. These machine learning algorithms can be used to predict many disease like heart attack, asthma, diabetes and high blood pressure etc.

REFERENCES

1. *B.L Deekshatulua Priti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm ", M.Akhil jabbar International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.*
2. *S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovasculardiseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.*
3. *American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).*
4. *N. Satyanarayana, CH. Ramalingaswamy, and Y. Ramadevi, 2014. Survey of Classication Techniques in Data Mining, International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 9, November 2014.*
5. *C. Christin, H. C. Hoefsloot, A. K. Smilde, B. Hoekman, F. Suits, R. Bischoff, and P. Horvatovich, "A critical assessment of feature selection methods for biomarker discovery in clinical proteomics, " Molecular & Cellular Proteomics, vol. 12, no. 1, pp. 263-276, 2013.*
6. *High blood pressure prediction based on AAA++ using machine-learning algorithms, Satyanarayana Nimmala, Y. Ramadevi, R. Sahith & Ramalingaswamy Cheruku, Cogent Engineering (2018), 5: 1497114.*
7. *B.L Deekshatulua Priti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm " M.Akhil jabbar International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013*