© 2017 IJSRST | Volume 3 | Issue 7 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Predicting the onset of Cardiovascular Diseases using Machine Learning Techniques

Qamar Rayees Khan[1], Parvez Abdulla[2]
[1]Department of Computer Sciences, Baba Ghulam Sahah Badshah University, Rajouri (J&K), India
[2]Department of Management Studies, Baba Ghulam Shah Badshah University, Rajouri (J&K), India

ABSTRACT

Cardiovascular diseases are one of the major diseases that consume many lives each year. The early prediction of the disease can help save a lot of lives, thus arising the need for better intelligent systems. Data mining is one of the popular fields in Computer sciences that mines an enormous amount of data to generate knowledge. Researchers are using these data mining and machine learning techniques to analyse the data related to health and predict the probability of the onset of any disease. In this research article, we have proposed a model trained using supervised machine learning algorithms to predict the onset of the cardiovascular disease. The various algorithms that were used to train the machine learning model are Sequential Minimal Optimization, Decision tress, Naïve Bayes and Random Forest. These algorithms were implemented using WEKA tool. The best performing algorithm was SMO that attained an overall accuracy of 83.4 % in predicting this deadly disease.

Keywords : Heart Disease, Weka, Machine Learning, Information Gain, SMO

I. INTRODUCTION

Heart is quite possibly the most crucial organ of human body; thus its care is very necessary for the body to perform well. According to the world Health Orgnaisation, Non communicable diseases contribute more in total deaths allover the world with 41 million people losing their lives which is about about 71% of the total worldwide deaths [1] Heart disease is considered as one of the primary causes of death world wide. Cardio vascular disease contribute over 18 million of death worldwide. [2] The various factors that contibute to the heart disease are smoking, abuse of liquor, caffeine, stress, and actual idleness alongside other physiological components like stoutness, hypertension, high blood cholesterol etc. The prior heart conditions are also inclining factors for coronary illness. The proficient, precise and early clinical analysis of coronary illness assumes a vital part in taking preventive measures to decrease the graph of death. Further, lack of facilities and specialists in rural areas are the major causes for the increase in the death rate.

Data mining refers to the mining of important information from massive datasets releated to any field like education, business, medical and various other fields of interest. Machine learning is one of the branches of Artificial Intelligence that can help in learning the complex interactions of various relationshiops in the data and help in building the model that is effective in predicting the diseases and reducing the errors that occur using the traditional approach.

As, the factors related to the cardiovasular diseases are many and independent, there arise a need for the efficient and robust model that can draw the relationships between various contributing parameters and select the most appropriate parameters for prediction of the disease. Various classification techniques like SMO, Decision trees, Random Forest and Naïve Bayes are used in training the machine learning model for predicting the heart diseases based on different parameters. This research paper is divided into four sections. Section I introduces the research problem. Section II reviews the relevant research in this domain. Section III describes the methodology adopted for training the machine learning model. Section IV discusses about the Results. Section V concludes the research work with the future direction.

II. RELATED WORK

Coronary illness stays the central reason for death on the planet. As Heart is the core organ that is responsible for pumping the blood. Clinical analysis should powered by Artificial intelligence techniques for the better prediction.

Monika Gandhi et.al.[3] analysed and predicted the dataset using three machine learning algorithms viz. Naïve Bayes, Decision Tree and Neural Network Alghorithms. They used feature Engineering to scale down different features and analysed its impact on machine learning model.

AH Chen et al. [4]introduced a coronary illness forecast framework that can help specialists in anticipating coronary illness status dependent on the clinical information of patients. Thirteen significant clinical highlights like age, sex, chest torment type were chosen. ANN was used to predict the heart disease based upon above parameters. The Artificial Neural Network contains input layer, hidden layer and output layer. There were thirteen neurons, six neurons and two neurons respectively present in the output layer. The accuracy achieved by the model was 80%.

Manpreet et.al.[5] used structural equation modeling abd Fuzzy cognitive map techniques for their proposed heart disease prediction model. The dataset used in the experimentation contained twenty attributes . The data set was divided into 80:20 ratio for training and testing the model. The model achieved 74% accuracy in predicting the disease.

Prajakta et.al.[6]worked to built an intelligent prototype based on big data techniques that can predict the attacks in timely manner. The System extracted the hidden patterns and relationships in the dast using the backend historical sisease database supplied to it. Researchers used the Hadoop for distributed processing and storage of the big data. The 13 attributes were used for experimentation. The various machine learing techniqurs that were usedfor trauining the model are Neural Networks, Naïve Bayes and Decision Trees.

Sairabhi et.al.[7] used two machinelearning algorithms viz. k means and Naïve Bayes to predict the heart disease. The researchers used the dataset from Cleveland heart disease Database. The clustering was performed using the two centroids. The model achieved an accuracy of 93% in predicting the heart disease.

Indira et.al. [8] used probabilistic Neural network based approach to predict the heart disease. Clustering was performed using k-means clustering algorithm. The proposed system Of probabilistic Neural networkls was compared to existing algorithms such as Decision trees and naïve Bayes. The ROCCh method was used to evaluate the performance of the system. The model achieved an accuracy of 94.6 % of accuracy.

S.Pravabathi et.al.[9]presented the review of the research being carried out for prediction of heart diseases. The various algorihms that were reviewed for training the machione learning model are Decision trees, SVM, K-means, K-nearest neighour asaand ANN's. The researchers concluded that c4.5 classifier was better at disgonosis and predicting the diseases that neural networks and SVM's. It was also found that Naïve Bayes and Decsiion trees achieved an accuracy of 95% in predicting and diagonosis of cardiovascular diseases.

III. Proposed Methodology

The methodology to classify and predict heart disease consists of four majorsteps is shown in Figure 1. The first step consists of obtaing the relevant data. The data is collected from Cleveland heart disease database. The dataset consists of 304 records with 14 different asttributes.The screen shot of the dataset used and various parameters are shown below in Figure 2 and Figure 3. respectively. Information gain algorithm is used to select the most contributing parameters in step 2. The third step consists of building the machine learning model using various machine learning techniques. The last step is about evaluating the machine learning model by using various evualation metrics.

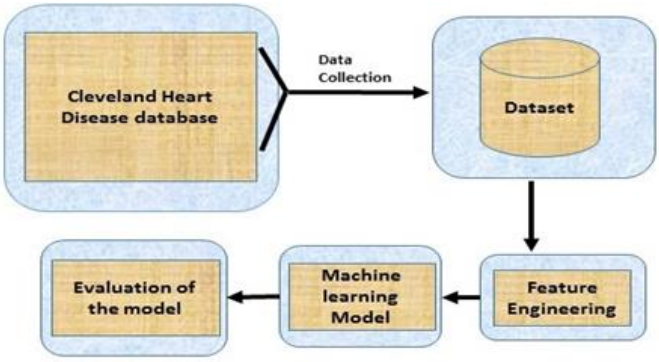


Figure 1. Proposed Methodology for predicting Heart Diseases.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1
25	61	1	2	150	243	1	1	137	1	1	1	0	2	1
26	40	1	3	140	199	0	1	178	1	1.4	2	0	3	1
27	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
28	59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
29	51	1	2	110	175	0	1	123	0	0.6	2	0	2	1
30	65	0	2	140	417	1	0	157	0	0.8	2	1	2	1
31	53	1	2	130	197	1	0	152	0	1.2	0	0	2	1
32	41	0	1	105	198	0	1	168	0	0	2	1	2	1
33	65	1	0	120	177	0	1	140	0	0.4	2	0	3	1
34	44	1	1	130	219	0	0	188	0	0	2	0	2	1
35	54	1	2	125	273	0	0	152	0	0.5	0	1	2	1
36	51	1	3	125	213	0	0	125	1	1.4	2	1	2	1
37	46	0	2	142	177	0	0	160	1	1.4	0	0	2	1
38	54	0	2	135	304	1	1	170	0	0	2	0	2	1
39	54	1	2	150	232	0	0	165	0	1.6	2	0	3	1
40	65	0	2	155	269	0	1	148	0	0.8	2	0	2	1
41	65	0	2	160	360	0	0	151	0	0.8	2	0	2	1
42	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
43	48	1	1	130	245	0	0	180	0	0.2	1	0	2	1
44	45	1	0	104	208	0	0	148	1	3	1	0	2	1
45	53	0	0	130	264	0	0	143	0	0.4	1	0	2	1

heart(1)

Figure 2. Screenshot of the Dataset used

S.No	Feature name	Description
1	**Age**	Age in years
2	**Sex**	Gender of the individual
3	**cp**	Chest pain type. (1 = typical angina 2 = atypical angina 3 = non — anginal pain 4 = asymptotic)
4	**Trestbps**	Resting Blood Pressure
5	**Chol**	Serum Cholestrol
6	**Fbs**	Fasting Blood Sugar
7	**Restecg**	Resting ECG (0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hyperthrophy)
8	**Thalach**	Max heart rate achieved
9	**Exang**	Exercise induced angina (1 = yes 0 = no
10	**Oldpeak**	ST depression induced by exercise relative to rest
11	**Slope**	Peak exercise ST segment (1 = upsloping 2 = flat 3 = downsloping)
12	**Ca**	Number of major vessels (0–3)
13	**Thal**	displays the thalassemia (3 = normal 6 = fixed defect 7 = reversible defect)
14	**Num**	Class Attribute (0= healthy heart, 1=Having heart disease)

Figure 3. Various parameters and the description of the dataset

IV. Experimental Results

The dataset of 304 attributes are divided into training and testing datasets in the ratio of 70:30. The model is validated using 10 cross validation. Four machine learning algorithms as Sequential Minimal optimization (SMO), Decision tress, Naïve Bayes, Random Forest and Decision tree (J48) are used to train the model. The detailed results generated by applying these algorithms are shown in Table 1.

Table 1. Accuracy achieved by various classifiers

Classifiers	**Accuracy**	**Precision**	**Recall**	**F-Measure**
SMO	**83.4 %**	**84.1 %**	**83.5 %**	**83.3 %**
Naïve Bayes	82.8 %	83.3 %	82.8 %	82.6 %
Random Forest	81.8 %	81.9 %	81.8 %	81.7 %
Decision Tree (J48)	78.5 %	78.5 %	78.5 %	78.5 %

Although many researchers use same algorithms to predict the heart diseases, but from our experiments it was revealed that applying the same type of algorithms on same data set can reveal different accuracies

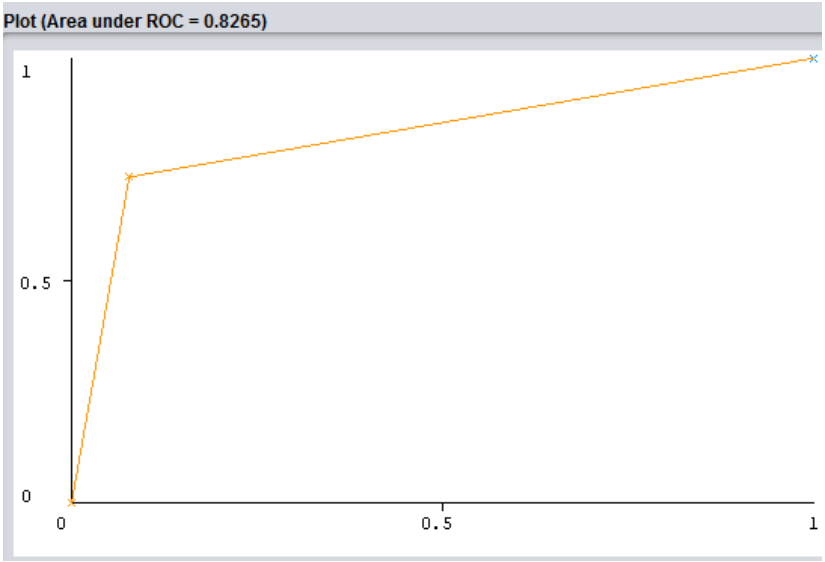
depending on the attributes we take. After applying the information gain algorithm, only the most contributing attributes / parameters were selected that resulted in dimensionality reduction and increase in precision, recall and overall accuracy of the model.

```
Attribute Evaluator (supervised, Class (nominal): 14 target):
        Information Gain Ranking Filter

Ranked attributes:
 0.2127  13 thal
 0.2013   3 cp
 0.1617  12 ca
 0.1585  10 oldpeak
 0.1422   9 exang
 0.1297   8 thalach
 0.1157  11 slope
 0.0602   1 ï»¿age
 0.0591   2 sex
 0         5 chol
 0         6 fbs
 0         4 trestbps
 0         7 restecg

Selected attributes: 13,3,12,10,9,8,11,1,2,5,6,4,7 : 13
```

Figure 4. Most contributing parameters selected using Information gain Algorithm

The screenshot of the best performing algorithm SMO are shown below in Fig. 5 and Fig. 6.

Figure 5. Detailed Result of SMO in predicting the patients having Heart ailment.



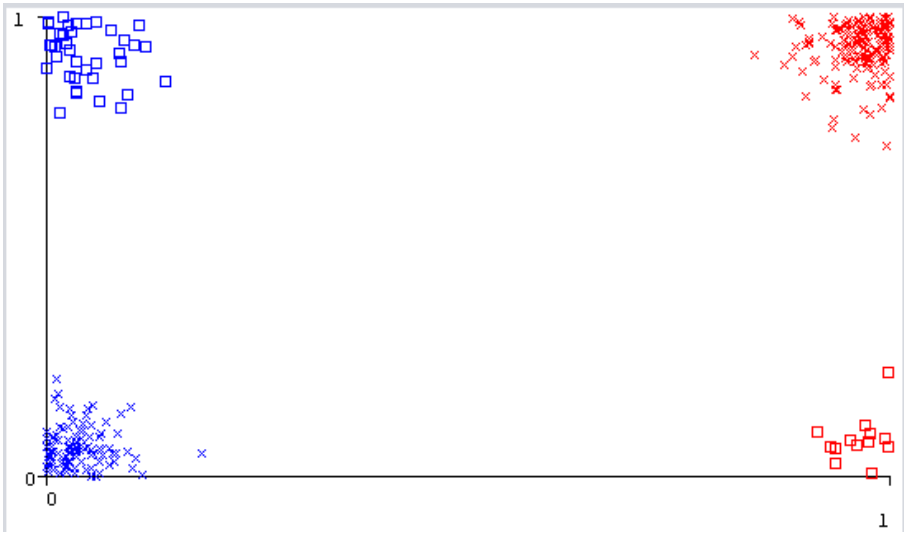


Figure 6. ROC curve depicting True Positive Rate vs False Positive Rate for SMO

```
Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         253               83.4983 %
Incorrectly Classified Instances        50               16.5017 %
Kappa statistic                          0.6625
Mean absolute error                      0.165
Root mean squared error                  0.4062
Relative absolute error                 33.2631 %
Root relative squared error             81.5627 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.732     0.079     0.886      0.732     0.802      0.671     0.827     0.771     0
                 0.921     0.268     0.804      0.921     0.859      0.671     0.827     0.784     1
Weighted Avg.    0.835     0.182     0.841      0.835     0.833      0.671     0.827     0.778

=== Confusion Matrix ===

   a   b   <-- classified as
 101  37 |   a = 0
  13 152 |   b = 1
```

Figure 7. Depiction of Classifier Errors for SMO

V. Conclusion

In this research article, a cardiovascular disease prediction model has been trained using various machine learning algorithms. The model has been trained based upon various contributing parameters like Chest pain type, Resting blood pressure, Maximum heart rate achieved etc.The various algorithms that were used to train the model are Sequential Minimal Optimization, Decision tress, Naïve Bayes and Random Forest. The best accuracy was achieved by SMO. This model after deployement will certainly help the patients having the heart related issue to forestall the untoward happening.By using the machine learning model like this one can help in predicting the patients having heart ailments beforehand and reduce the cost of tests. The better experimental results achieved also reveal that other many medical databases can also be used to predict the various the various diseases using our approach that can inturn help the patients and Doctors.

VI. REFERENCES

[1]. "Cardiovascular diseases." https://www.who.int/health-topics/cardiovascular-diseases

[2]. M. D. Seckeler and T. Hoke, "The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease," Clin. Epidemiol., vol. 3, no. 1, p. 67, Feb. 2011, doi: 10.2147/CLEP.S12977.

[3]. M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," in 2015 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management, ABLAZE 2015, Jul. 2015, pp. 520–525, doi: 10.1109/ABLAZE.2015.7154917.

[4]. A. Chen, S. Huang, P. Hong, C. Cheng, and E. Lin, "HDPS: Heart Disease Prediction System," Comput. Cardiol. (2010)., vol. 38, pp. 557–560, 2011.

[5]. M. Singh, L. M. Martins, P. Joanis, and V. K. Mago, "Building a cardiovascular disease predictive model using structural equation model & fuzzy cognitive map," in 2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016, Nov. 2016, pp. 1377–1382, doi: 10.1109/FUZZ-IEEE.2016.7737850.

[6]. P. Ghadge, V. Girme, K. Kokane, and P. Deshmukh, "Intelligent Heart Attack Prediction System Using Big Data," Int. J. Recent Res. Math. Comput. Sci. Inf. Technol., vol. 2, no. 2, pp. 73–77, 2015, Accessed: Apr. 16, 2021. Online]. Available: www.paperpublications.org.

[7]. S. H. Mujawar, P. R. Devale, and P. G. Student, "Prediction of Heart Disease using Modified K-means and by using Naive Bayes," Int. J. Innov. Res. Comput. Commun. Eng. (An ISO, vol. 3, no. 10, pp. 10265–10273, 2007, doi: 10.15680/IJIRCCE.2015.

[8]. I. S. Fal Dessai, "Intelligent Heart Disease Prediction System Using Probabilistic Neural Network," vol. 2, no. 3, pp. 2319–2526, 2013.

[9]. S. Prabhavathi and D. M. Chitra, "ANALYSIS AND PREDICTION OF VARIOUS HEART DISEASES USING DNFS TECHNIQUES," Int. J. Innov. Sci., vol. 2, no. 1, pp. 2347–9728, 2015, Accessed: Apr. 16, 2021. Online]. Available: www.ijiser.com.