Early Prediction of Heart Attack using Machine Learning Algorithms

ASIF RAHMAN SNIGDHA
Computer Science and Engineering
American International University Bangladesh
Dhaka, Bangladesh
asifrahmansnigdho@gmail.com

SYEDA NISHAT TASNIM
Computer Science and Engineering
American International University Bangladesh
Dhaka, Bangladesh
syedanishatt@gmail.com

KAMRAN RAFSAN MIAH
Computer Science and Engineering
American International University Bangladesh
Dhaka, Bangladesh
krafsan01@gmail.com

***Abstract*—Data mining strategy is the foremost important method for analyzing information from totally different areas.The point of this proposal is to discover out whether the data in existing of any other investigate of wellbeing disease, but we need to form a genuine visualization of heart disease. The most point of this ponder to plan a few designs of how numerous individuals of our nation kicked the bucket fair since of heart illness. The dataset for this proposition utilized from Researchers which is an open information entry of heart failure clinical record data set. To improve this, consider the k-means Clustering have been too utilized. This ponder uncovers the relationship between traits of heart failure clinical record data set this design of clinical information records. We believe that by this think about the wellbeing segment will be profited, and they can analyze the heart disappointment quiet effortlessly additionally, they will be cautioning some time recently the heart disease happens.**

***Index Terms*—Data mining,feature selection, formatting, style, styling, insert**

I. Introduction

Heart is the most important part of the body as it pumps blood to the entire body, helps them to function and thus life depends on its efficiency. Heart disease occurs due to the disease of heart blood vessel system within it which depends on number of factors. These factors need to be analysed to detect whether a person is suffering or about to suffer from a heart disease and this mostly starts from heart attack.

Heart disease occurs when the heart is incapable to pump adequately to preserve blood stream to meet the body tissue's needs for digestion system. Common causes of heart disappointment incorporate coronary supply route illness, counting a past myocardial localized necrosis (heart assault), tall blood weight, atrial fibrillation, valvular heart disease, overabundance liquor utilize, contamination, and cardiomyopathy of an obscure cause. In individuals with persistent steady gentle heart disappointment, treatment commonly comprises of way of life adjustments such as halting smoking, physical work out, and dietary changes, as well as drugs. Overall, around 2 percent of adults have heart failure. The hazard of passing is approximately 35 percent the primary year after conclusion, whereas by the moment year the hazard of passing is less than 10 percent for those who stay lively.

Early location of heart infections can avoid the death rate, individuals are not mindful approximately the discovery of heart disease prior due to need of information. But now there are different data mining technique to reduce heart disease, most often different research must classify to make research of hearth disease become very easy. In a word heart failure of a human is not only a bad impact but also a costly disease. Now in our research we want to make an efficient way to find the death and surviving people of heart failure. Cure of this issue is data mining strategy. Information mining is the method of analyzing large set of information and summarizing into valuable data.

II. Literature Review

Data is basically a record of information, in this case it is a record of past medical cases. If the data of past medical cases could be used to find the pattern and the initial symptoms of the disease, it would be easier to identify and cure before the case becomes serious.

My Chau Tu , Dongil Shin and DongKyoo Shin, [1]. In this paper three algorithms were performed, decision tree C4.5 algorithm, bagging with decision tree C4.5 and bagging with Naïve Bayes. Out of the three, bagging with Naïve Bayes gave the best performance .Since the bagging algorithm used here is straightforward, more improvement is required.

K.Sudhakar, Dr. M. Manimekalai [2] Only the very few specific variables were used here which effects the heart. This paper shows that they applied three different data mining classification techniques i.e. Neural Networks, Decision Trees, and Naive Bayes to predict the heart disease. Three of the methods were compared to find the best method of prediction.

Identify applicable funding agency here. If none, delete this.

Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M [3] In this research paper Genetic algorithm, K-means algorithm, MAFIA algorithm and Decision tree classification where used.The purpose of this research was to apply different data mining technique and find the most effective attribute out of the fourteen attribute used.

Mai Shouman, Tim Turner, Rob Stocker [4] Decision tree is one of the successful method of data mining that can be used for better analysis. This research paper contains different discretization techniques, multiple classifiers voting technique and different Decision Trees type in the diagnosis of heart disease patients. Later, reduced error pruning was used to reduce the error from the result found using decision tree. The decision tree accuracy is not being effected by the supervised discretization methods neither with or without voting. Though the accuracy of different types of Decision Tree increased by applying voting.

David W. Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar [5] In this case Big data is shown as an option that can be used rather than assumption using decision tree as it might not always give a positive result.Big data is basically obtaining new information by analysing large amount of data.It is suggested in many health care purposes but the uptake is relatively limited. Analysis is also required to predict which patients are at risk of adverse events of several types which might cause the present carrying disease. By evaluating genetic and genomic information, laboratory data, information on vital signs, and other data it is possible to detect the disease at the early stage.

Chaitrali S. Dangare , Sulabha S. Apte [6] The Healthcare industry have vast amount of information but not all the data are not mined for which it is difficult to find out unknown patterns and thus effects decision making. In one of the cases to predict the likelihood of getting heart disease only 13 attributes are used with obesity and smoking which is one of the rule. Decision Trees, Naive Bayes, and Neural Networks are studied on Heart disease database out of which Neural Networks comes up with the highest accuracy but as it is also said that these processes are still expandable. Getting introduced to a different perspective and gathering the knowledge from the data is the core reason of data mining. Classification method was used for the extraction of multipara metric features by assessing HRV (Heart Rate Variability) from ECG, data pre-processing and heart disease pattern and only data of 670 peoples were used. The amount of people who are used for the data are not sufficient and are not specified as age matters.

K. SRINIVAS, B. KAVITHA RANI [7] In this case only three techniques are used Naïve Bayes, Artificial neural network, and J48 decision tree algorithms, whereas there are other techniques that could have been used to get the maximum amount of accuracy and on the other side decision tree and Naive Bayes method are mostly assumptions. There are some data mining tools like WEKA, Rapid Miner, TANAGRA, MATLAB that has been used in this research paper.

H. Benjamin Fredrick David and S. Antony Belcy [8] In this case fourteen attributes are used including type, description and range of each of the attribute. It seems by the number of the factors that are considered is not sufficient. Random Forest, Decision trees and Naive Bayes are used in this paper. Random Forest provides better results compared to Decision tree and Naive Bayes. Data mining is done to explore the hidden pattern of the data set and analysing the future state so that proper treatment can take place. There is a scope to increase the accuracy of the Decision Tree and Bayesian Classification for the better result.

V. Sabarinathan, V. Sugumaran [9] This paper is entirely based on heart disease and to classify the features, mainly decision tree with J48 algorithm has been used. 85 percent accuracy has been achieved by the research that correctly predicts the cause of heart disease. The paper is mainly focused on three primary attributes, Thalassemia, Chest pain type and number of major vessels color. The resting electrocardiographic results and maximum heart rate achieved has been suggested to neglect while diagnosis as it does not improve the performance of the results.

Atul Kumar Pandey, Prabhat Pandey,K.L. Jaiswal and Ashish Kumar Sen [10] In this paper heart disease prediction model has been developed by using fourteen attribute containing fasting blood sugar, resting ecg, max heart rate, exercise induced angina, old peak, slope, number of vessels colored that can assist medical professionals. Decision tree has been used followed by unpruned, pruned and pruned with reduced error pruning approach. The accuracy of Pruned J48 Decision Tree with Reduced Error Pruning Approach gives better result compared to the simple Pruned and Unpruned approach. The result shows that out of all the attributes considered, fasting blood sugar is the most important attribute which gives better classification against the other attributes but it effects the accuracy.

Beant Kaur, Williamjeet Singh [11] In this paper, the author has mainly reviewed the usage of data mining technique on the prediction of heart disease. The complexity to analyze the huge amounts of data generated for prediction of heart disease is difficult if it is analyzed using traditional methods whereas Data mining provides the methodology and technology to transform these data into useful information. The paper that has been reviewed showed that neural networks given the accuracy of 100 percent in prediction of heart disease and on the other hand, Decision Tree has also performed well with 99.62 percent accuracy by using 15 attributes. Thus different technologies shows different accuracy depend on the number of attributes taken and tool used for implementation.

Gandhi, M., Singh, S. N. [12] This paper is mainly about knowledge abstraction from the data mining technique used in the most research paper. Naive Bayes, Neural network, Decision tree algorithm are the data mining methods that has been analyzed on medical data sets using algorithms.The basics and the algorithm of the three data mining method has been explained in the simple form to be easily understandable.

A. Sankari Karthiga , M. Safish Mary and M. Yogasini [13] In this research paper, Data mining algorithms such as decision tree and Naïve Bayes has been used for predicting heart attacks. The result in this research paper shows 99 percent accuracy on the favour of decision tree. The paper mainly showed the comparison between two data mining method used to get as accurate result as possible. A specific dataset from UCI repository has been used in this research paper for the accuracy, thus it might not give the same amount of accuracy in different data set .

K. Thenmozhi, P.Deepika [14] The research in this paper is done by using different decision tree eg-ID3, C4.5,C5.0, J48. Additionally, Gini Index is done to measure the impurity of data, to reduce the bias resulting Gain Ratio is used, to reduce error after decision tree extraction reduced error pruning is done.None of the other data mining technique has been used in this paper rather than decision tree for the simplicity and accuracy purpose.

The papers that has been studied for this research shows different datamaing process and sequence of workings that has been performed to get possible accurate result to support the intention of initiating the research. The research that has been done in this paper is different from the ones that has been studied. Here the data has been divided into two sets, categorical attributes and numerical attributes, python was used for categorical data and Fisher's Correlation for numerical to find the specific attributes which effects the heart most. Random Forest Tree, function SMO, meta.AdaBoosts M1 and Nave Bayes were applied on the specific independent attribute and the output dependent variable for the result.

METHODOLOGY

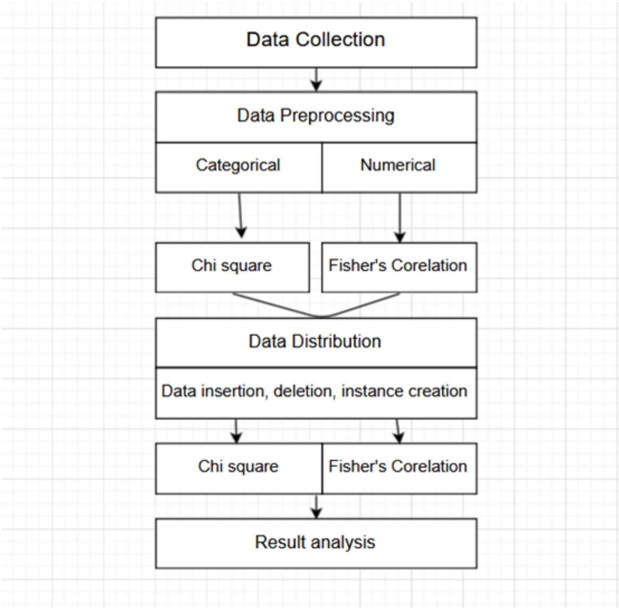
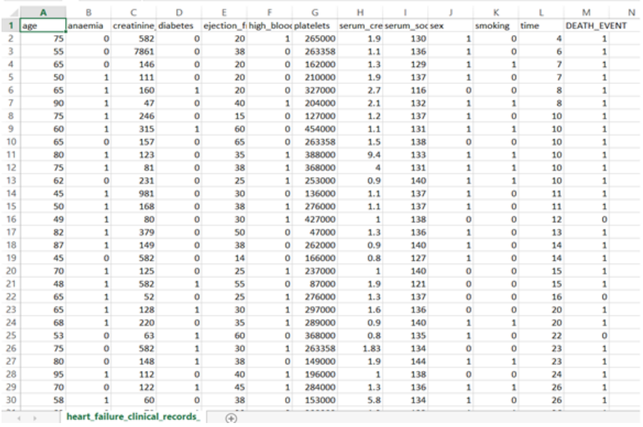


Fig. 1. workflow

a) Data Sources: The data set was collected from https://archive.ics.uci.edu/ml/index.php The dataset format was .csv file. Around 300 patient's data was stored in that file.In this dataset there are 13 attributes such as age, anaemia,creatinine phosphokinase, diabetes,ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, death event.These attributes are two types Independent attributes dependant attributes.The only dependent attribute is 'Death Event'. These dependant and independent attributes have categorical and numerical values. anaemia,diabetes,high blood pressure,sex,smoking,death event are categorical attributes. In these attributes '1' is represented as 'Yes' and '0' is represented as 'No'. Rest of the other attributes are Numerical.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	anaemia	creatinine_	diabetes	ejection_fi	high_blooc	platelets	serum_cre	serum_soc	sex	smoking	time	DEATH_EVENT	
2	75	0	582	0	20	1	265000	1.9	130	1	0	4	1	
3	55	0	7861	0	38	0	263358	1.1	136	1	0	6	1	
4	65	0	146	0	20	0	162000	1.3	129	1	1	7	1	
5	50	1	111	0	20	0	210000	1.9	137	1	0	7	1	
6	65	1	160	1	20	0	327000	2.7	116	0	0	8	1	
7	90	1	47	0	40	1	204000	2.1	132	1	1	8	1	
8	75	1	246	0	15	0	127000	1.2	137	1	0	10	1	
9	60	1	315	1	60	0	454000	1.1	131	1	1	10	1	
10	65	0	157	0	65	0	263358	1.5	138	0	0	10	1	
11	80	1	123	0	35	1	388000	9.4	133	1	1	10	1	
12	75	1	81	0	38	1	368000	4	131	1	1	10	1	
13	62	0	231	0	25	1	253000	0.9	140	1	1	10	1	
14	45	1	981	0	30	0	136000	1.1	137	1	0	11	1	
15	50	1	168	0	38	1	276000	1.1	137	1	0	11	1	
16	49	1	80	0	30	1	427000	1	138	0	0	12	0	
17	82	1	379	0	50	0	47000	1.3	136	1	0	13	1	
18	87	1	149	0	38	0	262000	0.9	140	1	0	14	1	
19	45	0	582	0	14	0	166000	0.8	127	1	0	14	1	
20	70	1	125	0	25	1	237000	1	140	0	0	15	1	
21	48	1	582	1	55	0	87000	1.9	121	0	0	15	1	
22	65	1	52	0	25	1	276000	1.3	137	0	0	16	0	
23	65	1	128	1	30	1	297000	1.6	136	0	0	20	1	
24	68	1	220	0	35	1	289000	0.9	140	1	1	20	1	
25	53	0	63	1	60	0	368000	0.8	135	1	0	22	0	
26	75	0	582	1	30	1	263358	1.83	134	0	0	23	1	
27	80	0	148	1	38	0	149000	1.9	144	1	1	23	1	
28	95	1	112	0	40	1	196000	1	138	0	0	24	1	
29	70	0	122	1	45	1	284000	1.3	136	1	1	26	1	
30	58	1	60	0	38	0	153000	5.8	134	1	0	26	1	

heart_failure_clinical_records_

Fig. 2. Patient's Dataset

b) Data Pre-Processing: First we applied chi square using python for the categorical data and SPSS fisher's co-relation for the numerical data to find out the significant attributes which are responsible for heart attack. None of the test gave us any significant attribute. For chi square significant values are normally less than 0.05 and for SPSS fisher's co-relation significant values are above 9. Output discriminant test were also done but none of the fisher's linear value was greater than 9.

None of the attributes were found significant for heart attack after applying chi-square and fisher's correlation. From fisher's co relation test Creatinine phosphokinase is the least significant attribute that causes heart attack. We applied data distribution for creatinine phosphokinase and found most of the values ranges between 0 to 3000. So we deleted entire row that has 2000 up creatinine phosphokinase values. In the modified dataset we again applied chi-square for categorical attributes and fishers co-relation for numerical attributes. None of the categorical attributes gave us any significant result. From spss fisher's co-relation serum creatinine serum sodium gave us the best result. Serum creatinine value is 7.952 serum sodium value is 7.453 .

A new dataset was created containing 'Serum creatinine', 'serum sodium' and death 'event attributes'. Random Forest

Tree, Function SMO, meta. AdaBoostsM1,Nave Bayes were applied to the new dataset to find accurate result.

Result

In this result part we used three classifiers to find the best result of this significant data, to find the most significant data we use chi square and fishers co algorithm, but in fishers co algorithm we get the most significant data which is serum creatinine and serum sodium., now we must classify the dataset.

c) Random Forest Tree: When we classified data using the Weka explorer using random forest tree, only the MCC the ROC area remained unchanged in every outcome. Even though we used the same data set, other results (TP rate, FP rate, Precision etc.) varied from one to another.

d) Function SMO: When we used functions SMO, the results of MCC and ROC area remained same in every aspect similar to when we used the random forest tree. But in functions SMO the results are smaller in number compared to the results when we used random forest tree. We have to keep in mind that we used the same data set in both situations.

e) meta. AdaBoostsM1: When we used meta.AdaboostM1 the results were higher than Functions SMO and random forest tree comparing the numbers. The results of MCC and ROC area were same every time even when we used meta.AdaBoostsM1. The results are constant in every method we have used so far. The results vary in different methods but there is one constant result for every method used.

f) Nave Bayes: When we used Nave Bayes the results were average between Functions SMO and random forest tree comparing the numbers. The results of MCC and ROC area were same every time even when we used Nave Bayes. The results are constant in every method we have used so far. The results vary in different methods but there is one constant result for every method used. Among all of those classifier we can say meta.AdaboostM1 is the best classifier among them.

	Random Forest Tree	SMO Function	MetaAdaBoostM1	Nave Bayes
TP Rate (Recall)	0.506	0.022	0.382	0.236
FP Rate	0.203	0.005	0.094	0.063
Precision	0.536	0.667	0.654	0.636
F1 Score	0.520	0.043	0.482	0.344
Accuracy	70.4626%	68.6833%	74.0214%	71.5302%
ROC Area	0.723	0.509	0.704	0.684

References

[1] Author, My Chau Tu , Dongil Shin and DongKyoo Shin (2009) A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing.China . doi:10.1109/dasc.2009.40

[2] Author, K.Sudhakar, Dr. M. Manimekalai (2014) Study of Heart Disease Prediction using Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering.4(1),1157-1160

[3] Author, Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M (2017). Heart disease diagnosis using data mining technique. International Conference on Electronics, Communication and Aerospace Technology, Manjeri, India. doi:10.1109/iceca.2017.8203643

[4] Author, Mai Shouman, Tim Turner, Rob Stocker (2011)Using Decision Tree for Diagnosing Heart Disease Patients .9-th Australasian Data Mining Conference, Australia.

[5] Author, David W. Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar (2014). Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients, HEALTH AFFAIRS 33, NO. 7, 1123–1131 https://doi.org/10.1377/hlthaff.2014.0041

[6] Author, Chaitrali S. Dangare , Sulabha S. Apte (June 2012) .Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques.International Journal of Computer Applications (0975 – 888) Volume 47– No.10

[7] Author, K. SRINIVAS, B. KAVITHA RANI (2014). Data Mining Tools and Techniques Analysis on Heart Disease Prediction. International journal For Innovative Engineering and Management Research 03(01), 16–23. http://www.ijiemr.org/downloads.php?vol=Volume-03issue=ISSUE-01

[8] Author, H. Benjamin Fredrick David and S. Antony Belcy (October 2018). HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES. ICTACT JOURNAL ON SOFT COMPUTING, 09(01),1817-1823 DOI: 10.21917/ijsc.2018.0253

[9] Author, V. Sabarinathan, V. Sugumaran (2014) Diagnosis of Heart Disease Using Decision Tree. International Journal of Research in Computer Applications Information Technology. 2(6) 74-79.

[10] Author, Atul Kumar Pandey, Prabhat Pandey,K.L. Jaiswal and Ashish Kumar Sen(2013) A Heart Disease Prediction Model using Decision Tree. Journal of Computer Engineering (IOSR-JCE),12(6), 83-86.

[11] Author, Beant Kaur, Williamjeet Singh(2014) “Review on Heart Disease Prediction System using Data Mining Techniques” International Journal on Recent and Innovation Trends in Computing and Communication .2(10) 3003-3008.

[12] Author, Gandhi, M., Singh, S. N. (2015,February). Predictions in heart disease using techniques of data mining. 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE). doi:10.1109/ablaze.2015.7154917

[13] Author, A. Sankari Karthiga , M. Safish Mary and M. Yogasini(2017) Early Prediction of Heart Disease Using Decision Tree Algorithm.International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST).3(3)

[14] Author, K. Thenmozhi, P.Deepika(2014) . Heart Disease Prediction Using Classification with Different Decision Tree Techniques. International Journal of Engineering Research and General Science . 2(6),6-11.