Heart Disease Prediction System Using Linear Regression, Smoreg, And Rep Trees Algorithms

R.Nanthini, P.Pandi Selvi

Abstract: ***The broad area of data mining covers most of the fields of research. Its role in medical diagnosis is very motivative to the researchers. It is very easy for the medical practitioners to analyze and treat the disease of the patients at an early stage. The proposed work deals with predicting heart disease of the patients at an early stage. The method was organized in three stages, Data collection, Data preprocessing and Data classification. The dataset for the work was collected from UCI repository. The collected sample was first preprocessed to clean unwanted information from the dataset. Classification operation is then performed on the preprocessed data. Classification is carried out with three different techniques, Linear regression model, SMOreg and REP trees. The results of the three methods were compared based on Root mean squared error and the Absolute error and are tabulated.***

Index Terms: Data Mining, Predictive Data Mining,Linear Regression model, SMOreg, REP trees.

I. INTRODUCTION

There is a major advancement in the area of research in medicine and computer science. On the Health care side, it provides various services which are used to improve the health and the long life of the people. One of the most common health issues among the people is the heart disease. The symptoms of the disease vary for each and every individual. Therefore, it is necessary analyze the symptoms in a proper way, and predict the disease at an earlier stage [13].Apart from the analysis of the medical specialist, there exists a need for further improvement. There occurs the necessity of a computer based system, which helps in predicting the disease at an earlier stage. One of the major research fields, which carry out this task, is the field of predictive data mining. It has the capability of combining various techniques to extract the needed information and its relations from the database [15]. The database regarding the heart disease patients can be collected from various hospitals that maintain a record of all the patients. Different techniques are available in the area of data mining to perform the task such as, Naive bayes, neural network, decision trees, classification, etc. With these techniques, it was possible to identify the disease at an early stage with better accuracy and was able to reduce the risk of heart attack[11].

Revised Manuscript Received on August 05, 2019

R.Nanthini, Department of Computer sceince, Alagappa University/ Dr.Umayal Ramannathan College for Women/ Karaikudi, India.

Dr.P.Pandiselvi, Department of Computer Science, Alagappa University/ Dr. Umayal Ramanathan College for Women/ Karaikudi , India.

II. LITERATURE SURVEY

In 2017, Ch.Sai Chaitanya, et al [2], introduced a new system data mining methods. They used three different techniques to the prediction process. K-means clustering algorithm, Maximal Frequent Item set algorithm and the C4.5 algorithm. They were able to obtain better predictive results with their approach.

In 2018, R.Nanthini et al[10]., aimed to analyze the prediction system in the field of healthcare industry for heart disease. It was clear from the survey, the usage of various data mining techniques and their predictive results, that helps in analyzing the disease at an early stage.

In 2017, Ramin Assari, Parham Azimi and Mohammad Reza Taghva [12]., made a detailed analysis on how to predict the heart disease at an early stage, identify the risk and treat the patients accordingly. They too proposed a new model based on the obtained rules.

In 2018, Shakuntala Jatav, et al [15], introduced an algorithm for the prediction of Kidney, Diabetes and Liver disease. They used two major techniques SVM and the Random Forest, the results obtained by the above methods were also compared.

In 2018, Uma N Dulhare [17]., introduced a model with Naive Bayes classifier and the Particle Swarm Optimization. The results proved that the approach with PSO provided better results in predicting the disease at an early stage.

III. PROPOSED METHOD

Various data mining techniques exist for predicting the severity of the heart disease in patients at an early stage. Even then there exist cases, where the predictive results have to be improved. In order to carry out the task the authors proposed a new technique. The technique compares the results of three different classification techniques, linear regression model, SMOreg, REP trees.

The steps involved in the process are as follows,

(i) Data Collection.
(ii) Data Preprocessing.
(iii) Data Classification.

(i)*Data Collection:* The sample dataset of heart disease patients, for the work was collected from UCI repository. The data that was collected

contain some noise. Hence it undergoes the preprocessing stage to eliminate noise.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	old peak	slope	ca	thal
54	1	4	120	0	0	0	155	0	0	2	7	2
54	1	4	130	0	1	0	110	1	3	2	7	3
54	1	4	180	0	0	0	150	0	1.5	2	7	1
55	1	2	140	0	1	1	150	0	-2	1	7	0
55	1	4	115	0	0	0	155	0	-1	1	7	1
55	1	4	120	0	0	1	92	0	-3	1	7	4
55	1	4	140	0	0	0	83	0	0	2	7	2
56	1	3	120	0	0	0	97	0	0	2	7	0
56	1	3	125	0	1	0	98	0	-2	2	7	2
56	1	3	155	0	0	1	99	0	0	2	3	2
56	1	4	115	0	0	1	82	0	-1	1	7	1
56	1	4	120	0	0	1	100	1	-1	3	7	2
56	1	4	120	0	0	1	148	0	0	2	7	2
56	1	4	125	0	1	0	103	1	1	2	7	3
56	1	4	140	0	1	0	121	1	1.8	1	7	1
57	1	3	105	0	1	0	148	0	-3	2	7	1
57	1	4	110	0	1	1	131	1	1.4	1	7	3
57	1	4	140	0	0	0	120	1	2	2	6	2
57	1	4	140	0	0	0	100	1	0	1	6	3

Figure 1. Sample Dataset

(ii)Data Preprocessing:

The first step in any data mining technique is the preprocessing stage. The data that were collected may not be in an order. It may contain any unnecessary details that are not required. Hence they have to be eliminated in the initial stage itself. In the heart disease dataset the unwanted attribute values were omitted and only those attributes that play a key role in prediction were selected. Among various attributes only 14 attributes were considered.

(iii)Data Classification: Classification process was carried out with three different techniques, Linear Regression Model, SMOreg and the REP Trees. The results obtained from the three techniques were compared, based on their level of prediction.

IV. RESULTS AND DISCUSSION

Linear Regression Model

Regression is one of the simplest methods to use. Linear regression is a standard statistical method that computes the coefficients or "weights" of a linear expression, and the predicted value is the sum of each attribute value multiplied by its weight [18].

SMO Reg

Support Vector Machines were mainly used for binary classification problems [3]. The alteration of SVM for regression is termed the Support Vector Regression. They have the capability of normalizing the input data before they were used within the application.

REP Tree

It is used as an efficient algorithm for Classification. The classification algorithm is used in situations where there is limited amount of data and to estimate a statistical quantity. It best fit with models that have low bias and high variance. The most commonly used algorithm and the default is the REP tree.

The efficiency of the three different classification techniques can be analyzed from the experimental results. The classification graph based on age for the given dataset is as shown below.

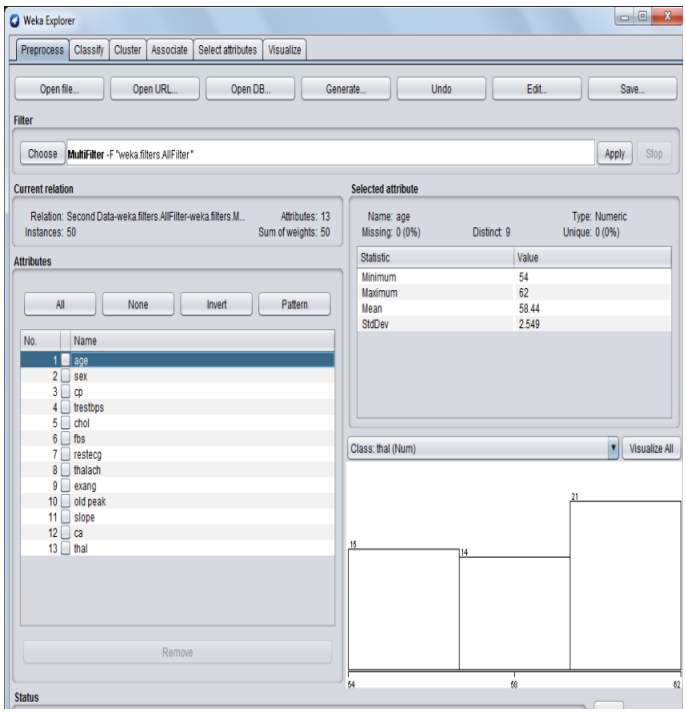
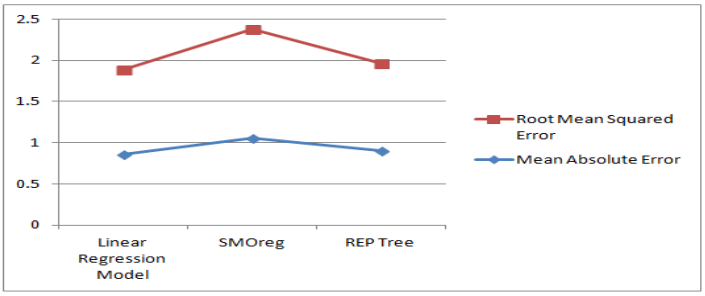


Figure 2: Graph showing classification based on age.

The results obtained from the three different models were tabulated as follows

Table 1: Comparitive Results Of Three Different Classification Techniques.

Model	Mean Absolute Error	Root Mean Squared Error
Linear Regression Model	0.8604	1.0253
SMOreg	1.0584	1.3174
REP Tree	0.9	1.0608



V. CONCLUSION

Predictive data mining covers the broad area of research. The various steps involved in the proposed method were, Data collection, Data preprocessing and Data classification. After preprocessing, Classification of the given dataset was carried out with three different techniques, linear regression model, SMOreg and REP trees. regression The results obtained from the three methods were compared based on Root mean squared error and the Absolute error. From the results, it is well evident that Linear model and REP trees provided best results in predicting the data. As a future work, the authors proposed to improve the method by including more number of datasets and to use a hybrid technique for classification in prediction.

REFERENCES

1. M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu," Heart Disease Prediction System using Associative Classification and Genetic Algorithm". International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012
2. Ch.Sai Chaitanya1 | Ch.Nagur Vali2 | V.Sai Vikas3 | K.Sowmya, "Heart Anomaly Forecasting Applying a Series of Data Mining Methods". Proceedings of National Conference on Computing & Information Technology (NCCIT-2017). International Journal for Modern Trends in Science and Technology, Volume: 03, Special Issue No: 02 March 2017. ISSN: 2455-3778. http://www.ijmtst.com
3. Jason Brownie," How to use regression machine learning algorithms in weka Support vector regression". Published on July 22, 2016, in Weka Machine Learning. https://machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/
4. Jason Browinlee, "How to use ensemble Machine Learning Algorithms in Weka". Published on July 27, 2016in Weka Machine Learning. https://machinelearningmastery.com/use-ensemble-machine-learning-algorithms-weka/
5. Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique". IJCSC ISSN:0973-7391. Volume 7, Number4 1 Sep 2015-March 2016 pp.129-137. DOI: 10.090592/IJCSC.2016.018. Available online at www.csjournalss.com
6. S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita, " Prediction of Heart Disease using Data Mining Techniques". Indian Journal of Science and Technology, Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/102078, October 2016. ISSN (Print): 0974-6846, ISSN (Online): 0974-5645
7. A. Malarvizhi, Dr. S. Ravichandran, "Data Mining's Role in Mining Medical Datasets for Disease Assessments – a Case Study". International Journal of Pure and Applied Mathematics. Volume 119 No. 12 2018, 16255-16260, ISSN: 1314-3395 (on-line version). url: http://www.ijpam.eu. Special Issue
8. Mamta Sharma, Farheen Khan, Vishnupriya Ravichandran," Comparing Data Mining Techniques Used For Heart Disease Prediction". International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 06 | June -2017 www.irjet.net p-ISSN: 2395-0072
9. Michael Abernethy,"Introduction and regression, Data mining with WEKA, Part 1. Updated April 28, 2010 – Published April 27, 2010. https://developer.ibm.com/articles/os-weka1/
10.] R.Nanthini, Dr.P.Pandi Selvi, A survey on Data Mining Techniques for Heart Disease Prediction, 2018 IJRAR December 2018, Volume 5, Issue 4 www.ijrar.org (E-ISSN 2348-1269, P- ISSN 2349-5138), IJRAR1944683 International Journal of Research and Analytical Reviews (IJRAR) www.ijrar.org.
11. Neha Chauhan and Nisha Gautam "An Overview of heart disease prediction using data mining techniques". https://docgo.net/health-9.
12. Ramin Assari, Parham Azimi and Mohammad Reza Taghva, "InternatiHeart Disease Diagnosis Using Data Mining Techniques". International Journal of Economics & Management Sciences. DOI: 10.4172/2162-6359.1000415. Int J Econ Manag Sci, an open access journal Volume 6 • Issue 3 • 1000415 ISSN: 2162-6359, 2017.
13. G.Rasitha Banu, J.H.Bousal Jamala," HEART ATTACK PREDICTION USING DATAMINING TECHNIQUE". International Journal of Modern trends in Engineering and research (IJMTER). ISSN (online): 2349-9745, ISSN(PRINT):2393- 8161, 2015. https://www.researchgate.net/publication/312188365_Heart_Attack_prediction_using_Data_Mining_technique
14. Shafquat Perween, Suraiya Parveen, " Analysis of Heart Disease Prediction Techniques". International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 7, Issue 4, April 2018. www.ijsret.org.
15. Shakuntala Jatav and Vivek Sharma, "An Algorithm for Predictive Data Mining Approach In Medical Diagnosis". International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1, February 2018. DOI:10.5121/ijcsit.2018.10102. https://www.researchgate.net/publication/327722009_A_Review_on_Heart_Disease_Prediction_using_Machine_learning_and_Data_Analytics_Approach.
16.] Sujata Joshi and Mydhili K.Nair,"Prediction of Heart Disease Using Classification Based Data Mining Techniques", Springer India 2015, volume 2. https://link.springer.com/chapter/10.1007/978-81-322-2208-8_46.
17. Uma N Dulhare," Prediction system for heart disease using Naive Bayes and particle swarm Optimization". Biomedical Research 2018; 29 (12): 2646-2649ISSN 0970-938X. www.biomedres.info
18. University of Waikato,New Zealand. CC Creative Commons Attribution 4.0 International License. https://www.futurelearn.com/courses/data-mining-with-weka/0/steps/25396.

AUTHORS PROFILE

R.Nanthini is currently pursuing the M.Phil Degree in Computer Science at the Department of Computer Science, Dr.Umayal Ramanathan College For Women, Karaikudi Her research interest includes Heart disease and Data mining.

Dr.P.Pandiselvi is working as the Assistant Professor in the Department of Computer Science, Dr. Umayal Ramanathan College For Women, Karaikudi. She has the sound knowledge in many research fields especially in Data mining Big data, and Analytics. She has published 11 international journals.