Predictive Analysis of Rapid Spread of Heart Disease with Data Mining

*

Radhanath Patra
Dept. of Electronics, G.I.E.T. University, Gunupur
Odisha,India
1radhanath.patra@gmail.com

Bonomali Khuntia
Dept. of CSC, Berhampur University, Berhampur
Odisha,India
bonomalikhuntia@gmail.com

***Abstract*—cardiovascular disease is the leading cause of mortality for both sexes in worldwide. Heart disease is increasing at a rapid rate in both older and younger generation of males and females now days. So in need demand of right strategies, development and implementation of effective health monitoring policies should be emphasized to combat the epidemic of heart related diseases. So early detection and treatment with the use of both conventional and innovative technique must be preferred In this paper we have used the UCI machine learning repository Cleveland heart disease database having 303 instance and 76 attributes. For the proposed method we have used the Information gain concept for selection of best attribute and processes the selected features using weka and python. This paper identifies the gap of research on prediction of heart disease based on python Anaconda navigator, spyder and weka platform on which we have much emphasized. The various techniques, processes which have used to train the model of heart datasets such as feature selection, numpy, pandas library, decision tree classifier, KNN classifier, entropy, gini- index, confusion matrix.The result shows that decision tree classifier is most effective and appropriate for prediction of UCI repository Cleveland heart dataset.**

***Index Terms*—KNN, SVM, DT, UCI, CVD**

I. INTRODUCTION

Some statistics says death of around 630,000 people in the united state (US) each year due to cardio vascular diseases (CVD) and as per analysis it has been found that at least one person dies per minute due to heart problem in US [1]. In recent years the epidemic of cardiovascular disease in Asia specific region is most alarming. In India the death rate due to cardiovascular diseases (CVD) rose by around 34 percentages in last three year in both rural and urban areas [2].The diagnosis of cardio vascular disease in last three year in both rural and urban areas is started by considering patients history and symptoms. Variety of test may be implemented for diagnosis of heart disease. These test re classified into two types such s non invasive test and invasive test. The various processes under the non invasive test re described as

i. Electrocardiogram (ECG or EKG): the test can monitor the hearts electrical activity and help doctors to find any irregularities.
ii. Echocardiogram :Ultrasound test that helps doctor to visualize the picture of heart structure
iii. Stress test: doctor can monitor the hearts activity through the physical exertion test
iv. CT scan: provides highly detailed X-RAY image structure of heart
v. Heart MRI: provides the detailed image of heart and blood vessels

Similarly invasive test includes cardiac catheterization and coronary angiography respectively for clear picturization of delicate arteries and capillary surrounding the heart as well as respond of heart through the sending of electrical pulses with attached electrodes. Heart disease encompasses a wide range of cardiovascular problems. Type of heart disease includes Arrhythmia: due to abnormality rhythm of heart, atherosclerosis: caused for hardening of the arteries, cardiomyopathy: condition causes the heart muscle to harden or grow weak, coronary artery disease (CAD): caused by the buildup plaque in the hearts arteries.

II. ROLE OF DATA MINING TECHNIQUE IN MEDICAL FIELD

Data mining plays an important role in building an intelligent model system based up on variety of crucial algorithm to detect and predict the severity of various life threatening diseases [3]. Through the predictive analysis and early detection mechanism of different data mining technique, the various epidemic diseases can be initially analyzed, diagnosed, controlled and prevented [4]. Through the data analysis processes the large data set can be easily processes through the various algorithms. The useful and required pattern gets trained and processed for predicting the accurate result. So data mining technique is really helpful in medical data such that mortality rate can be seriously declined and controlled and more patients life gets improved [5].

A. Two major and useful application of data mining

Measuring Treatment Effectiveness:Data analysis process is most effective approach in analyzing selected features of dataset and predicting the accurate output. In the data analysis

978-1-5386-8158-9/19/$31.00©2019IEEE

process various applications can be designed which may help in the medical field for finding the disease effectively with no cost.

Filtration of fraud diagnoses case:Data analysis process figures out the symptoms by analyzing the previous recorded data so unnecessary prescription and false di- agnosis case can be surely eradicated [6].

III. DATA ANALYSIS PROCESS

Data Preprocessing:- A few hours of measurements later, we have gathered our training data. Now its time for the next step of machine learning for data preprocessing. Data collection with feature extraction and uploading in proper way is indeed required for our machine learning training. First of all we will put our data together, and then randomize the ordering. It is also a good time to do any pertinent visualization of our data so to help us to see if there are any relevant relationships between different variables that can take advantages of, as well as shows us if there are any data imbalances [7].

Dimensionality Reduction: - The next workflow is to choose a model. Researchers and data scientists have created many intelligent models over the years. They eliminated the repeated and collinear features with set of feature extraction rules. It can be feature selection and feature extraction [8].

Feature Selection: - The process of finding set of unique feature sole responsible for data processing which comes under various strategies like filter strategy, wrapper strat- egy and the embedded strategy.

Feature Extraction:- The transformation of high dimen- sional data in to lower dimension.. The data transforma- tions can be linear, but many non-linear dimensionality reduction techniques also exist.

Advantages of Dimensionality Reduction: - Time and space gets improved and performance of machine learn- ing algorithm gives better result .most importantly low dimensions data are easy to analyze and having better computation process [9].

IV. DATAMINING APPLICATION

Applications:- This techniques is sometimes used in neuroscience is maximally informative dimensions which finds a lower dimensional representation of a dataset such that valuable information as of the original data is preserved.

Model Learning: - The process of training a machine learning model involves providing an ML algorithm with training data to learn form. The algorithm finds pattern in the training data that map the input data attributes to the target and its output and machine learning model that captures these patterns. To train a datasets, we need to specify the following conditions:- Input training data source. Data attributes with reference to the target to be predicted. Required data transformation instructions. Learning parameters to control the learning algorithm. Now approaching to the training parameters, we have to provide specific parameters which will be able to provide better output in training process of a predictive model [10]. So in need parameters are:- Maximum model size. Maximum number of analysis over training data. More number of process. Regularization type Regular- ization amount.

Model Testing: Once we have defined our problem and prepared our data, we need to apply machine learning algorithm to the datasets in order to solve the problem. We can spend more time choosing, training and tuning algorithm. We have to reduce the time complexity to obtain the more accurate result to reach our goal.

Performance Measure- The best model and appropriate selection parameter always results best performance of set of analyzed data and it provides the best output [11].

Test and Train Datasets-Most general approach is to split the entire dataset in to train and test data. Useful algorithm will be implemented on the training datasets and will be tested against the test set. The selection process may be from a random split of data or may involve complicated sampling methods. So the outcome of training model is applied with test data for predicting the result [12].

Cross Validation: -The most sophisticated approach that we could use to fold the entire dataset in to set of equal sized groups and our build model is used to train all folds of data and performance measures are averaged to yield the result of a specific problem [13].

Testing Algorithm:- With the given dataset the training and testing process is generally applied but after the training process for a sample of information test data should be considered to measure the performance of testing algorithm [14].

V. CLEVELAND HEART DATABASE DESCRIPTION

As per recent medical report factors like smoking, cholesterol and diabetes if gets controlled in a country, patients suffering from heart disease can be surely declined nearly up to 15 percentage[5].Robert detrano collected this clevand data base that consists of 303 instances of 76 attributes.In UCI machine learning repository Cleveland heart disease is the most used dataset by the data mining researchers and out of 76 attributes Researchers have used only 13 attributes for prediction and analysis of heart diseases. The attributes are as described as the

1) age in years
2) sex(1=male.0=female)
3) chest pain type resting blood pressure,
4) serum
5) cholesterol
6) fasting blood sugar
7) resting electrocardiographic result
8) maximum heart rate achieved
9) exercise included angina(exhang)
10) old peak
11) slope
12) number of major vessel
13) thal
14) num(target)

are predicted attribute for diagnosis of heart diseases..

VI. PROPOSED METHOD

In this paper our main objective is data analysis with feature selection as preprocessing technique [1]. After the feature selection process, certain attribute based on the preprocessing tech- nique can be removed for better analysis and prediction. So for feature selection one of the recognized and viable methods we have used here is Entropy and Information gain concept [5].

A. Design flow

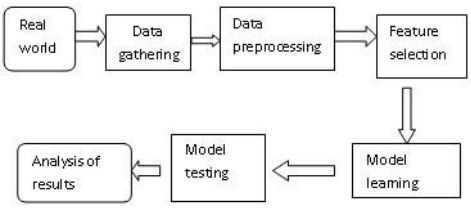

Fig. 1. Flow Process

B. Entropy, Information gain

It is measure of randomness in the information being processed. Higher the entropy more is the information content.For feature selection entropy plays a important role because the attributes having zero information gain can be deleted thus the useful and necessary attributes are considered for data analysis [11].

$$Entropy(p1, p2, p3.., pn) = -p1log(p1) - p2log(p2).....$$

where p1,p2. ..pn are the attributes.
Information gain:=(Entropy of distribution before splitting)-(Entropy of distribution after it)
Entropy:

$$H(X) = -\sum p(X)\log p(X) \quad (1)$$

Information Gain:

$$I(X, Y) = H(X) - H\ (X|Y) \quad (2)$$

As per the medical report it has been observed and analyzed that apart from 14 attributes other attributes are also responsible for the heart disease which has not been considered for analysis and in further considering that concept and analogy into mind we have done feature analysis and selection using information gain concept and the attribute having infor- mation gain zero is deleted and 51 attributes including target class is selected for the data analysis. It has been found through the algorithm of information gain and same has been cross verified by the weka-classifier-attributor selector. The weka processor and python tool has been used for classification using different data mining technique [16]. As per the literature survey decision tree holds good for the classifier in medical disease and accuracy finding and it has been found that using python the decision tree classifiers results with best accuracy as compared to other data mining technique [1] .Decision tree is a graphical representation of a sequential decision process that forms the tree like structure using classification and regression model. It starts at the root node and breaks down a dataset into smaller subsets while resulting development of decision tree. It creates tree with decision nodes and leaf nodes. Decision tree algorithm maximizes the information gain. An attribute with highest information is tested first.

VII. RESULT AND SECTION

WEKA TOOL: Weka is one of the mostly used data mining tool for classification technique. In Weka tool different data mining algorithm technique can be applied to view the result and Selected classifier based up on the performance and accuracy is considered. As per analysis and performance view: J48, KNN, RBF, NAIVE BAYES is considered. The result shows that the j48 classifier has a better accuracy in Weka tutorial as compared to other classification technique. In considering the above result analysis feature selection process is applied in weka tutorial and thus the Select attribute option of Weka tutorial infogain evaluator with ranker method is selected to process the 303 instant of 76 attribute for filtering the attribute having zero information. After filtering 51 attributes are selected with num: attribute as the target value. That 51 attributes contains the valid and necessary information for data processing.

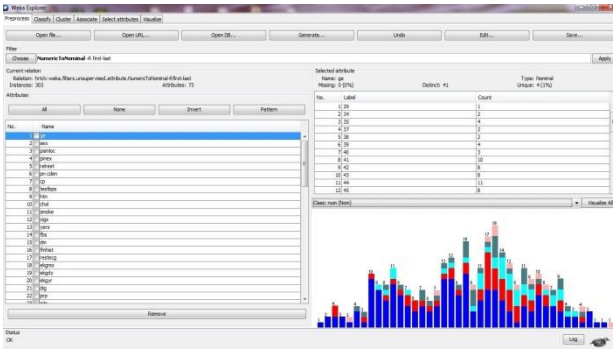

Fig. 2. Preprocessing of Raw Data

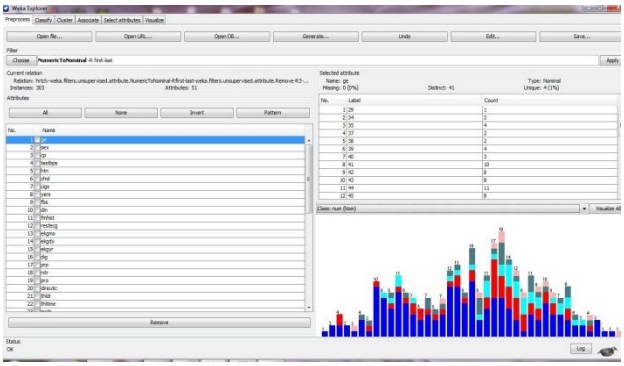

Fig. 3. Selected Features

VIII. PYTHON TOOL

Python is one of the simple and powerful programming languages for data analysis and visualization. Different machine

learning algorithm is used in python using scikit-learn for data analysis.Here we have used the decision tree classifier,KNN and SVM for data analysis.With decision tree classifier,to achieve the best performance, parameters for simulation in Python tool is mentioned as:criterion = entropy, random state = 9,max depth=5, min samples leaf=6 with and the value of test size= .15.similarly value of test size= .25 is set for the KNN and SVM classifier.

A. Comparision between weka tool and python tool

WEKATOOL	Accuracy	TP Rate	FP Rate	Precision	Recall
J48	87.12%	87.1%	1.7%	87.12%	87.12%
KNN	62.4%	62.4%	25.6%	56%	62%
RBF	63.69%	63.7%	17.4%	63%	64%
NAIVE BAYES	65.01%	.65%	12.8%	65%	65%

PYTHONTOOL	Accuracy	Precision	Recall	f1-score
Decision Tree	93.4%	95%	93%	92%
KNN	76.3%	72%	76%	72%
SVM	63.15%	40%	63%	49%

B. output

The result shows that the j48 classification technique decision tree has a better accuracy in Weka tutorial as compared to other classification technique. In considering the above result analysis feature selection process is applied in weka tutorial and thus the Select attribute option of weka tutorial info gain evaluator with ranker method is selected to process the 303 instant of 76 attribute to filter the attribute having zero information. The same result is also cross verified through the MATLAB. The selected 51 attributes have significant contribution towards the prediction of heart disease. These 51 attributes is processed in python for Classification and result found is improving. The python tool used for classification of Cleveland heart database .The model parameter value and processing is done for the 50 attributes of CSV file. The result is quite satisfactory in decision tree other than other algorithm in data mining in python.

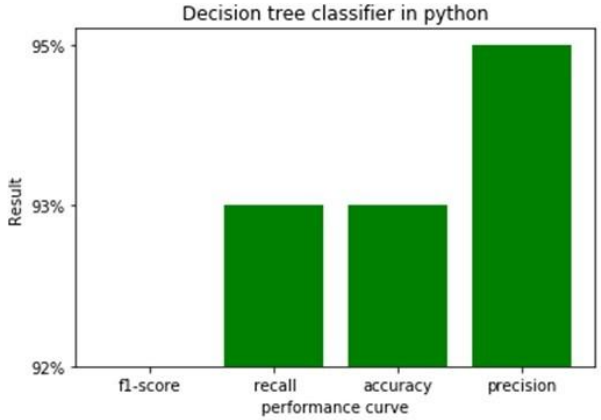

Fig. 4. Decision Tree classifier

IX. CONCLUSION

The result shows that decision tree classification technique holds good for the analysis of medical data classification especially for heart diseases. For better performance and more accuracy deep learning techniques can be applied for the diagnosis of heart disease.

REFERENCES

[1] Randa El-Bialy , Mostafa A. Salamay, Omar H. Karam and M.Essam Khalifa Feature Analysis of Coronary Artery Heart Disease Data Sets", International Conference on Communication, Management and Information Technology (ICCMIT 2015) , Procedia Computer Science 65 (2015) 459-468.

[2] Poornima V, Gladis D "A novel approach for diagnosing heart disease with hybrid classifier". Biomedical Research,vol-29,issue-11,pp.2274-2280,2018.

[3] B. Nithya, Dr. V. Ilango Predictive Analytics in Health Care Using Machine Learning Tools and Techniques ,International Conference on Intelligent Computing and Control Systems ICICCS 2017,pp.492-498.

[4] Salma Banu ,N.K Suma Swamy "Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey", ,International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT) ,2016,pp.256-261.

[5] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi and C. S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 3, pp. 559-566, May 2010.

[6] Chadha,R.,Mayank, S. "Prediction of heart disease using data mining techniques" CSI Transactions on ICT, vol.4,issue(2-4), pp.193198,2016.

[7] Jesmin Nahara, Tasadduq Imama, Kevin S. Ticklea, Yi-Ping Phoebe Chen" Computational intelligence for heart disease diagnosis: A medical knowledge driven approach" Expert Systems with Applications, vol- 40 issue-1,pp.96-104, 2013

[8] Purushottama, Prof. (Dr.)Kanak Saxena , RichaSharma,"Efficient Heart Disease Prediction System" International Conference on Computing, Communication and Automation (ICCCA),pp.72-77,2015.

[9] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale and A. Lamgunde, "Heart disease prediction using data mining techniques," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, 2017, pp. 1-8.

[10] Qurat-ul-ain Mastoi, Teh Ying Wah, Ram Gopal Raj, and Uzair Iqbal "Automated Diagnosis of Coronary Artery Disease: A Review and Workflow"Cardiology Research and Practice, Volume 2018, Article ID 2016282, pp.1-9.

[11] M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden nave bayes classifier," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), Bangalore, 2016, pp. 1-5.

[12] C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputhur, 2017, pp. 1-5.

[13] Babu, S., Vivek, E. M., Famina, K. P., Fida, K., Aswathi, P., Shanid, M., Hena, M. (2017). Heart disease diagnosis using data mining technique. 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2017, pp. 750-753.

[14] N. Gawande and A. Barhatte, "Heart diseases classification using convolutional neural network," 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2017, pp. 17-20.

[15] Purushottam, K. Saxena and R. Sharma, "Efficient heart disease prediction system using decision tree," International Conference on Computing, Communication and Automation, Noida, 2015, pp. 72-77.

[16] A. Khemphila and V. Boonjing, "Heart Disease Classification Using Neural Network and Feature Selection," 2011 21st International Conference on Systems Engineering, Las Vegas, NV, 2011, pp. 406-409.

[17] E. O. Olaniyi, O. K. Oyedotun, A. Helwan and K. Adnan, "Neural network diagnosis of heart disease," 2015 International Conference on Advances in Biomedical Engineering (ICABME), Beirut, 2015, pp. 21-24.