Classification of Health Care Data using Decision Tree

Madhu H. K[1]
Assistant Professor, Department of MCA
Bangalore Institute of Technology, Bengaluru, India.

Dr. D. Ramesh[2]
Professor & Head, Department of MCA
Sri Siddhartha Institute of Technology, Tumakuru

Abstract: In the present situation, voluminous health care data is getting accumulated at various places in various formats convenient to the ERP designed. Analyzing these data for prediction, assessment, classification and identification is a challenge faced by medical professionals, doctors and researchers. This requires intelligent system to auto analyze data and recommend appropriate action. In this research work, an attempt has been made to implement a 2-class classifier using random forest decision tree algorithm which is applied to UCI health care heart disease data set, which supports the present situation with faster and efficient classification of data for further action by hospitals and healthcare professionals at the initial stages.

Keywords: Random forest decision tree, Classification of health care data analytics.

I. INTRODUCTION

The health care data collected over various ERP models used in hospitals are dumped in data centers and are also structured by researchers for testing, proposed model have become very common. Assessing these data through machine learning / Data science algorithms has become the need of the hour. Many procedures done by doctors / diagnosis by pathologists are automated. Initial diagnosis / identification / classification of data has been successfully conducted through various smart devices available in market today.

Data classification [3,5,6,8,9,10,11] is a process of classifying the samples into predefined classes based on the supervised knowledge of the class. Complexity of any classification depends on the size of the attributes/features, samples, number of classes defined.

Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. The results obtained from Decision Trees are easier to read and interpret. Feature to access detailed patients‟ profiles is only available in Decision Trees.

In this research work an attempt has been made to implement random forest decision tree, on the data set available in UCI repository heart disease data set.

Random forest decision tree is a classification algorithm in a supervised approach. A detailed step has been discussed in implementation section from the available data set, 80% is used for training and 20% is used for testing the proposed model. The outcome of this model is the confusion matrix and ROC curve. Accuracy are mentioned in result analysis section.

II. LITERATURE SURVEY

Many works have been reported in literature, the use of various machine learning technique on health care data for identification, recognition and classification.

The following survey demonstrates how machine learning is able to provide a medical diagnosis, for detecting diseases like heart diseases, [1,7,12,15,17,18] dengue, [4,14] Parkinson,[2] dermatological disease [19], kidney disease [13] and cancer [16].

There have been several works reported in the literature which uses decision trees for medical data classification.

Showman et al. [1] presented enhanced decision tree to identify heart disease patients. The accuracy achieved is 76% when enhanced decision tree was implemented and with multiple classifier voting technique the accuracy is 74.1% for gain decision tree.

The aim of Ramani et al. [2] was to verify the efficacy of applying different classifiers to the Parkinson's Dataset. The Random Tree Algorithm correctly classifies the dataset for Parkinson's disease and provides 100 percent of it. The Linear Discriminant Analysis, C4.5, CS-MC4 and K-NN produce results of accuracy above 90%. The error rate for K-NN is just 0.02566.

An assessment analysis on various single decision tree classifiers was discussed by Hasan et al. [3]. For Wisconsin's breast cancer results, the LMT classifier provided the highest precision, i.e., 74.23 percent, followed by NB Tree and Random Forest (71.13 percent) and Random Tree (70.10 percent). FT provided the highest precision for the Pima Indian diabetes dataset, which is 80.84 percent, followed by Hoeffding tree (80.08 percent), LMT (79.31 percent), and NB tree (78.16 percent).

Tanner et al. [4] proposed a decision tree algorithm for the prediction of serious dengue disease, which was estimated for 1200 patients with data collected within the first 72 hours of disease, and categorized the data into four separate groups, such as probable dengue, probable dengue, probable dengue and probable dengue. With a sensitivity and specificity of 71.2 percent and 90.1 percent respectively, the total error rate calculated after k-fold cross validation was 15.7 percent.

Lavanya et al. [5] suggested evaluating the performance of decision tree induction classifiers in terms of accuracy and time complexity on different medical data sets.

In their research, Srinivas et al. [6] briefly explored the possible use of classification-based data mining techniques such as rule-based, decision tree, Naïve Bayes and tomassive health data volume from the Artificial Neural Network.

Hota et al. [7] analyzed different techniques for machine learning to identify heart data downloaded from the UCI repository database. 84.82 percent accuracy, 87.1 percent sensitivity, 83.24 percent precision are the comparative performance of different DTs using FST.

Vijayarani et al. [8] concentrated primarily on finding the best algorithm for classification based on the accuracy of classification and performance factors for execution time. It is found from the experimental results that the SVM's efficiency is higher than the Naive Bayes classifier algorithm.

An effective intelligent medical decision support system based on data mining techniques was developed by Zriqat et al. [9]. With an accuracy rate of 99.0%, followed by Random Forest, the decision tree outperforms other classifiers.

Sharma et al. [10] proposed research focused on applying various data mining classification techniques to the public health dataset for evaluating the health care system using different machine learning methods such as WEKA and RapidMiner.

Krishnaiah et al. [11] briefly explored the possible use of classification-based data mining techniques for large health data volumes, such as rule-based, decision tree, Naïve Bayes and Artificial Neural Network.

Dangare et al. [12] analyzed heart disease prediction systems using a greater number of input attributes. The accuracy of Neural Networks, Decision Trees, and Naive Bayes is 100%, 99.62%, and 90.74%, respectively, as per our outcomes.

Vijayarani et al. [13] predicted kidney diseases by using the Support Vector Machine (SVM) and Artificial Neural Network in their research work (ANN).

Fathima et al. [14] examined the use of the technique of machine learning-SVM to classify Dengue, one of the Arboviral diseases. In this study, they briefly investigated that Decision tree and SVM are most productive for heart disease from the above techniques.

Vijayarani et al. [16] predicted liver diseases using classification algorithms in their research work. Naïve Bayes and support vector machines are the algorithms used in this work (SVM).

Apte[17] has studied heart disease prediction systems using a higher number of input attributes. Neural Networks, Decision Trees, and Naive Bayes are 100 percent, 99.62 percent, and 90.74 percent respectively, as per outcome accuracy.

Rao, et al. [18] developed a Heart Disease Prediction System (DSHDPS) Decision Support System using data mining modeling methodology, namely Naïve Bayes. Using medical profiles such as age, sex, blood pressure and blood sugar, it can predict the risk of heart disease in patients.

Manjusha et al. [19] suggested a method with the help of the Naïve Bayesian theorem that allows data patterns to be collected.

From survey it is evident that many intelligent models are implemented various researches to automate the process of identification / recognition and classification. The major limitation is on the accuracy achieved and more of the methodology proposed are application dependent. In this research the approach is to generate the methodology for any health care data.

III. PROPOSED METHODOLOGY

Analysis of health care data is based on the nature of the data and range of the attributes which specifies the classes. Most of the health care data analytical model are supervised models and the proposed machine models have to learn the nature / characteristics of the data to identify the classes.

Decision tree learning models is also a supervised model which helps in learning discrete valued target function. The function is represented in a decision tree. The advantage is interpretation and visualization of data in the classes in simple and easy to understand. The limitation is that decision tree requires preprocessing of data that is data preparation to handle numerical and categorical data and non-linear parameters. This is one of the most accurate method used by many researchers as it allows researchers to decide which features address deriving the decision process and how each feature is linked based on the priority and outcome.

This predictive model is designed to split the node to two sub nodes and each sub node created increases the homogeneity among the samples. This process is recursively / interactively computed until the set targeted homogeneous nature of the class is identified.

The accuracy of the model is tested through precision, FI-score and ROC-curve.

In order to describe the classifiers' performance in the digital diagnoses problem, four basic characteristics (numbers) are considered on which derivative measurement metrics are defined.

The four attributes are:

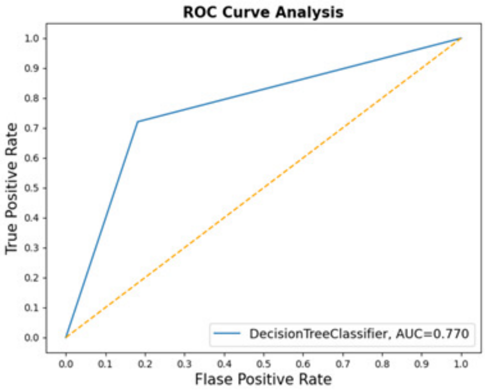
1. True Positive (TP)– True class correctly classification patients with disease,
2. True Negative (TN)– True classes correctly classification as healthy patients,
3. False Positive (FP) – False classes healthy patients identified under patients with disease.
4. False Negative (FN)– False class patients with disease classified as healthy.

Based on these characteristics the metrics are,

- Accuracy-the ratio of patients appropriately identified to the total number of patients (Accuracy = (TP+TN)/ Total number of samples or (TP+FP+FN+TN))
- Precision: the percentage of patients with the disease classified correctly to the total patients classified as having the disease. How many patients listed as having disease actually have the disease is the intuition behind precision (Precision = TP/(TP+FP)).
- Recall-ratio of appropriately identified patients with the disease to patients with the disease. The idea behind recall is how many patients with disease are categorized as having the disease. (Recall = TP/TP+FN).

- F1-score: good F1-score and that would be indicative of a good Precision and a good Recall value as well. (F1-score = (2 * ((TP / (TP+FP) * TP / (TP+FN))) / ((TP / (TP+FP) + (TP / (TP+FN))).

- Receive Operating Characteristic (ROC)-



It is a area under the curve with target 1 to represent perfect classifier and 0.5 to represent worthless classifier.

This model is set to work on all health care data set available. The proposed model is independent of the attribute features which is the uniqueness of the model.

IV.DATASET

The UCI heart disease dataset, consisting of 303 person data with 13 features and 1 mark, is considered in order to perform the experiment. Characteristics are,

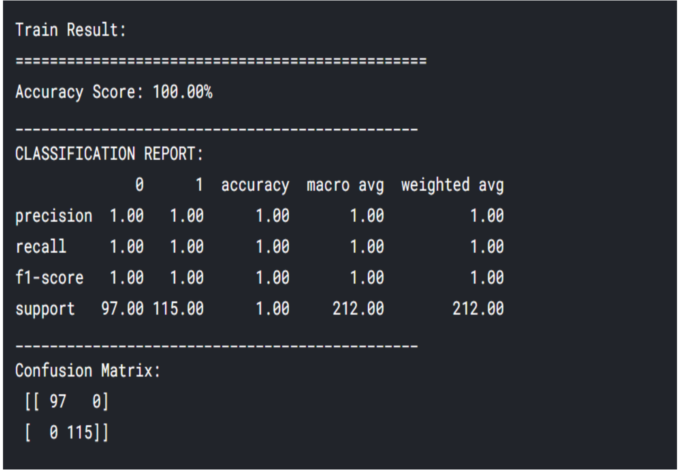
1. Age: Indicates the individual's age.
2. Sex: uses the following format to show the individual's gender: 1 = male 0 = female
3. Class of chest pain (cp): indicates the type of chest pain suffered by the person using the following format: 1 = normal angina 2 = atypical angina 3 = non-anginal pain 4 = asymptotic angina 2 = atypical angina 3 = non-anginal pain 4 = asymptotic
4. Resting Blood Pressure (trestbps): shows an individual's resting blood pressure value in mm Hgg (unit)
5. Cholesterol serum (chol): shows the cholesterol serum:
6. Fasting Blood Sugar (fbs): compares an individual's fasting blood sugar value to 120mg/dl. If blood sugar > 120 mg/dl is fasting, then: 1 (true) else: 0 (false)
7. Resting ECG (restecg): displaying resting electrocardiographic findings 0 = normal 1 = ST-T wave abnormality 2 = left ventricular hypertrophy 2 = left ventricular hypertrophyMax heart rate achieved (thalach): displays the max heart rate achieved by an individual.
8. Angina triggered exercise (exang): 1 = yes 0 = no
9. Exercise-induced ST depression compared to rest (oldpeak): shows the value that is integer or float.
10. Segment of peak exercise ST (slope): 1 = upsloping 2 = flat 3 = downsloping 2 = flat 3
11. Number of fluoroscopy-colored major vessels (0-3) (ca): shows the value as an integer or float.
12. Thal: indicates thalassemia: 1 = normal 2 = fixed defect 3 = reversible defect [label] Diagnosis of heart disease (target): 1 = normal 2 = fixed defect [label] Diagnosis of heart disease (target):
13. It shows whether or not the person suffers from heart disease: 0 = absence 1 = presence.

We had 76 features/attributes in the actual dataset, but because of its medically validated significance in evaluating heart disease, we selected only the above 13 features for our analysis.

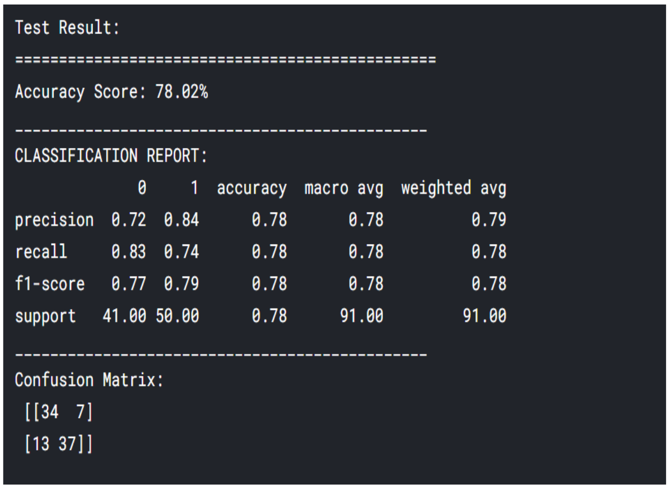
V. RESULTS

The data set is divided into 75 percent training data and 25 percent testing data to determine the efficiency of the decision tree algorithm, and the model is applied. In terms of precision, accuracy, recall, f1-score and support, the model is assessed.

Random forests have shown the highest precision for the given dataset, as seen in the given tables.

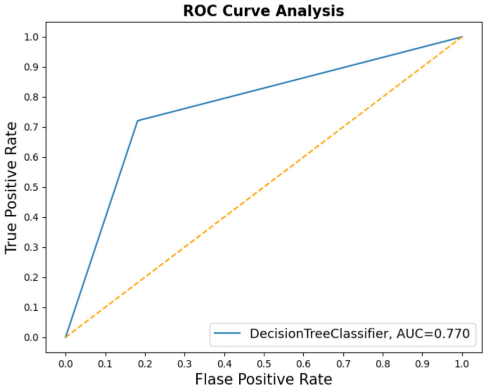


```
Train Result:
================================================
Accuracy Score: 100.00%
_______________________________________________
CLASSIFICATION REPORT:
              0      1  accuracy  macro avg  weighted avg
precision  1.00   1.00      1.00       1.00          1.00
recall     1.00   1.00      1.00       1.00          1.00
f1-score   1.00   1.00      1.00       1.00          1.00
support   97.00 115.00      1.00     212.00        212.00
_______________________________________________
Confusion Matrix:
 [[ 97   0]
 [  0 115]]
```



```
Test Result:
================================================
Accuracy Score: 78.02%
_______________________________________________
CLASSIFICATION REPORT:
              0      1  accuracy  macro avg  weighted avg
precision  0.72   0.84      0.78       0.78          0.79
recall     0.83   0.74      0.78       0.78          0.78
f1-score   0.77   0.79      0.78       0.78          0.78
support   41.00  50.00      0.78      91.00         91.00
_______________________________________________
Confusion Matrix:
 [[34  7]
 [13 37]]
```

ROC Curve:



VI. CONCLUSION

The ultimate aim of our work is to identify patients with heart disease more accurately. To get more reliable information, we used the UCI repository dataset. The decision tree classification algorithm was used for the classification of heart disease here. It has been shown from the results that decision trees have a successful outcome in the forecast. Other methods of machine learning can also be used for prediction, e.g., Clustering, Naïve Bayes, Random Forest.

REFERENCES

[1] Mai Showman, Tim Turner, Rob Stocker. “Using Decision Tree for Diagnosing Heart Disease Patients”. Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia. CRPIT Volume 121 - Data Mining and Analytics 2011.

[2] Dr. R. Geetha Ramani G. Sivagami, “Parkinson Disease Classification using Data Mining Algorithms”. International Journal of Computer Applications (0975 – 8887) Volume 32– No.9, October 2011.

[3] Md. Rajib Hasan, Nur Azzah Abu Bakar, Fadzilah Siraj, “SINGLE DECISION TREE CLASSIFIERS’ ACCURACY ON MEDICAL DATA”.

[4] Lukas. Tanner, Mark Schreiber, Jenny G. H. Low. “Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness”.

[5] Lavanya. D, Dr. K. Usha Rani. “Performance Evaluation of Decision Tree Classifiers on Medical Datasets”. Volume 26– No.4, July 2011.

[6] K. Srinivas, B. Kavihta Rani and Dr. A. Govrdhan. K. Srinivas et al. “Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks”. International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.

[7] H.S. Hota, Seema Dewangan. “Classification of Health Care Data Using Machine Learning Technique”. Volume 5 Issue 9|| September 2016 || PP. 17-20.

[8] Dr. S. Vijayarani, Mr. S. Dhayanand. “DATA MINING CLASSIFICATION ALGORITHMS FOR KIDNEY DISEASE PREDICTION”.

[9] Isra’a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh. “A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods”. IJCSIS, Vol. 14.

[10] Tanvi Sharma, Anand Sharma, Prof. Vibhakar Mansotra. “Performance Analysis of Data Mining Classification Techniques on Public Health Care Data”.

[11] V. Krishnaiah et al, “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques”. IJCSIT, Vol. 4 (1), 2013, 39 – 45.

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

[13] Dr. S. Vijayarani AND Mr.S.Dhayanand. “KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS”. IJCBR, Volume 6 Issue 2 March 2015.

[14] A. Shameem Fathima and D. Manimeglai. “Predictive Analysis for the Arbovirus-Dengue using SVM Classification”. International Journal of Engineering and Technology Volume 2 No. 3, March, 2012. ISSN: 2049-3444 © 2012 – IJET Publications UK.

[15] Aqueel Ahmed, Shaikh Abdul Hannan. “Data Mining Techniques to Find Out Heart Diseases: An Overview”. (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012.

[16] Dr. S. Vijayarani and Mr.S. Dhayanand. “Liver Disease Prediction using SVM and Naïve Bayes Algorithms”. IJSETR Volume 4, Issue 4, April 2015.

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

[18] Mrs. G. Subbalakshmi, Mr. M. Chinna Rao. G. Subbalakshmi et al. “Decision Support in Heart Disease Prediction System using Naive Bayes”. Indian Journal of Computer Science and Engineering (IJCSE). ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011.

[19] Manjusha K. K, K. Sankaranarayanan and Seena P. “Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification”. Volume 4, Issue 1, January 2014 ISSN: 2277 128X. International Journal of Advanced Research in Computer Science and Software Engineering.