Dashboard in R

Prediction of Heart Disease Using

Cleveland Dataset from UCL Machine Learning

Wajeeha Mushtaq

Department of Computer Science & Engineering UET, Lahore,

Email

2018CS26@student.uet.edu.pk

**Abstract**— A large amount of data is accumulated by the health-care industry. This data contains effective patterns that enable efficient decision-making. These patterns often go unexplored. Various machine learning methods can be incorporated in such a situation. This work uses various machine learning methods such as Decision Tree ID3, Decision Tree C4.5, and Random Forest to realize heart failure chances in the medical database. Results show that the Decision Tree method predicts the diagnosis of heart disease most effectively from the patient's data. The dataset used to carry on this research work is taken from the popular UCI repository and is known as the Cleveland Dataset. It is implemented on the R platform.

1. Introduction

The data collection of different diseases is very important. Medical and health areas are among the most important sections in industrial societies. Blood pressure, cholesterol, pulse rate, stress, food habits are the factors that contribute to Heart disease. The main functioning organ of the human body is the heart. If the human heart isn't functioning it will affect the entire human body. Some risk factors of cardiopathy organs are family background, stress level, cholesterol level, Age, food diet. Inhaling of tobacco. Blood vessel unit overstretched indicates the danger level of the body pressure. The pressure at the middle muscle of the heart exaggerates the level of lipids over time in the blood that causes heart disease. Lipids settle in the arteries and block the flow of blood to the cardiac organ.

Inhaling tobacco is the root cause of cardiac arrest that ends up in death. As the pressure increases in thickening the blood. The Naïve mathematician technique is employed to predict the center illness through likelihood. The Neural Network provides the reduced error of the prediction of cardiopathy. By monitoring the activity of the patients continuously the death rate can be reduced. The main cause of heart stroke is due to blockage in arteries. It has many other names such as cardiovascular disease and arterial hypertension. Approximately, there are almost 26 million people around the world affecting heart disease. The worrying point is, this ratio is expected to increase rapidly in the coming years if precautions are not taken efficiently. Apart from making lifestyle healthy and diet control, the right time diagnosing and comprehensive analysis are other essential factors, which can ultimately save lives. Therefore, this paper has taken a small step towards saving the lives of HF patients and describes a way to improve the performance of diagnosing the patients on the bases of their medical history.

Therefore, this dashboard attempts to improve the performance of the classifiers by doing experiments using multiple machine-learning models to make better use of the dataset collected from different medical databases. The dashboard will help the doctors to diagnose the disease in the patients by analysis of dataset using different analysis techniques like the random forest, decision tree id3, decision tree c4.5. It is widely used in healthcare fields due to its descriptive and predictive power. It can predict health insurance fraud, healthcare cost, disease prognosis, disease diagnosis.

1. METHODOLOGY
2. Dataset

Cleveland dataset, which is collected from the UCI repository. The system uses 303 datasets for the evaluation. The system contains the following subtasks.

1. Data Overview

The complete description of each attribute and the number of values for each attribute is shown in Table 1 below.

TABLE I  
DATA OVERVIEW AND ATTRIBUTE DESCRIPTIONS

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | Attribute Description | Discrete Values |
| 1. | **Age -** The first attribute is defining the age of the person. [Minimum Age: 29, Maximum Age: 77] | Multiple values between 29 and 77 |
| 2. | **Sex -** Attribute number two describes the gender of a person. [―0‖ means Female and ―1‖ means Male] | 0, 1 |
| 3. | **CP -** The third attribute is defining the level of chest pain (CP) a patient suffering from when reached to the hospital. There are four kinds of distinct values defined for this attribute, where each value is describing a level of chest pain | 0, 1, 2, 3 |
| 4. | **RestBP -**The next attribute describes the blood pressure (BP) figure for the patient while admitted to the hospital. [Minimum BP: 94, Maximum BP: 200] | Multiple values between 94 and 200 |
| 5. | **Chol-** This column is showing the cholesterol level recorded while admitting the patient in the hospital. [Minimum Chol: 126, Maximum Chol: 564] | Multiple values between 126 and 564 |
| 6. | **FBS -** The next attribute is describing the fasting blood sugar level in the patient. It has binary classified values. The values are depending on, if the patient has more than 120mg/dl sugar = 1, if not = 0. | 0, 1 |
| 7. | **RestECG -** This parameter is showing the result of ECG from 0 to 2. Where each value is showing the severity of the pain. | 0, 1, 2 |
| 8. | **Heart Beat-**The maximum value of heartbeat counted at the time of admission [Minimum: 71, Maximum: 202] | Multiple values between 71 and 202 |
| 9. | **Exang-**This parameter was used to understand about, does exercise induce angina or not. If yes, the value will be ―1‖, and ―0‖ for not. | 0, 1 |
| 10. | **Slope-** The condition of the patient during peak exercise. This value defined into three segments [Upsloping, Flat, Down sloping] | 1,2,3 |
| 11. | **CA-**This attribute is showing the status of fluoroscopy. It is showing that how many vessels are colored. | 0,1,2,3 |
| 12. | **Thal-** This parameter is another kind of test required for the patient having chest pain or breathing difficulty. Four kinds of values showing the result of Thallium test | 0,1,2,3 |
| 13. | **Target –** This is the last column in the dataset. This Target column is also known as the Class column or Label column. This column describes the number of categories, (classes) defined in the data file. As per the dataset is taken in this experiment. There are two different types of classes (0,1), where ―0‖ means there are no chances of Heart Failure, whereas ―1‖ imply that there are strong chances of heart failure in a patient. The value ―0‖ and ―1‖ is based on the other 13 parameters described in this dataset above | 0,1 |
| 14. | **Old Peak-** The next attribute is defining the patient’s depression status. It is assigned as different real number values fall between 0 and 6.2. | Multiple real number values between 0 and 6.2 |

1. ALGORITHM

The algorithms we used for heart failure disease are random forest, decision tree ID3, decision tree C4.5

1. Decision Tree: Decision Tree It’s a tree-like classification model, which built a structure consisting of branches and nodes on the bases of evidence collected for each attribute during the model learning phase. The decision tree's branches and nodes connect according to the number of entities described in the dataset. The forwarding process uses the number of values dedicated to each attribute. Furthermore, following the rules described on each branch and node, it decided for each transaction. Finally, according to the decision node, the class label will be assigned to the record. This procedure is iterative and repeats until each transaction got a class category. Therefore, this algorithm converts the attributes into branches and nodes and selects one of the attributes as a decision node, which also known as a class label. The class label in rapid miner can select while importing the dataset.
2. Random Forest: Random forest is the next model selected and implemented in this research. This model is from a classification family; therefore, it is also known as the supervised learning algorithm. During the learning phase, this model first generates multiple random trees called a forest [35]. For example, a dataset contains ―x‖ number of attributes, it first selects some feature randomly known as ―y‖. Using all features; (i.e. ―y‖), it produces nodes using the best rift method. Furthermore, the algorithm will work for creating a complete forest by repeating the previous steps. Then during the prediction process, the algorithm tries to combine the trees using estimated outcome and voting procedure [36]. The purpose of merging the random trees through voting in a forest is to opt-out of the highest forecasted tree, which can enhance the prediction accuracy for future data.
3. IMPLEMENTATION:

This research used 3 machine learning models, using a predictive approach to forecast the chances of heart failure in a patient admitted in the hospitals. Therefore, as described above the dataset taken from the UCL machinery dataset, which has been collected from multiple locations. The dataset has a list of 14 attributes, which collectively used for diagnosing heart disease in a patient. For experiment execution, the RSTUDIO tool is used which is software for r language.

1. RESULTS

The random forest gives the chances of heart disease by analyzing data in terms of confusion matrix as shown in figures:

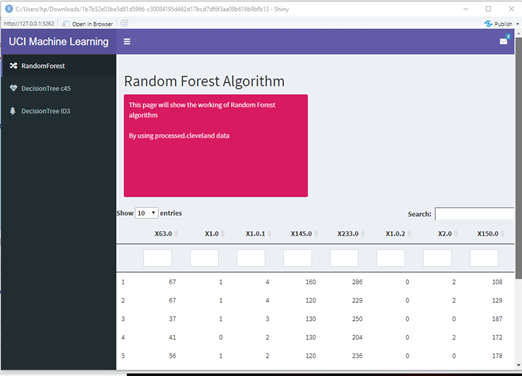
****

Fig. 1 Confusion Matrix of Random Forest

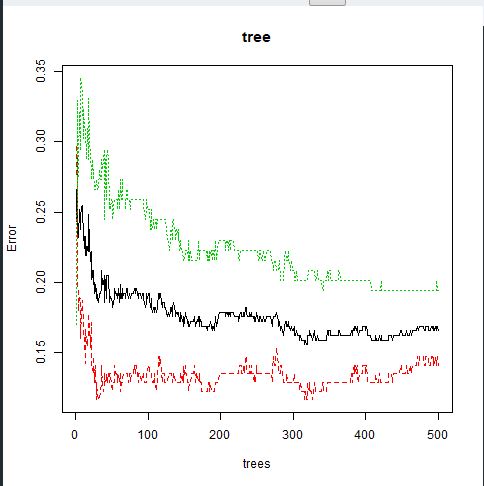


Fig. 2 Graph of Random Forest

The ID3 algorithm is used by training on a data set to produce a decision tree that is stored in memory. At runtime, this decision tree is used to classify new test cases by traversing the decision tree

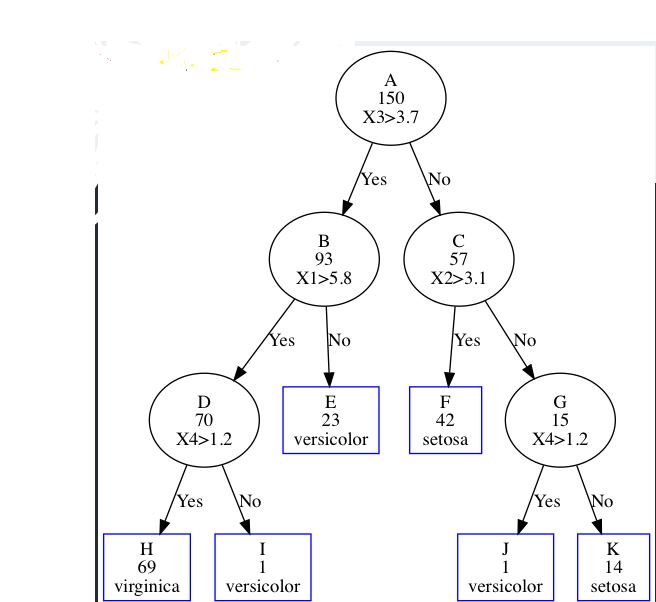


Fig. 3 Graph of Decision Tree ID3

Decision tree C4.5 give outcome in terms of a tree for getting heart failure

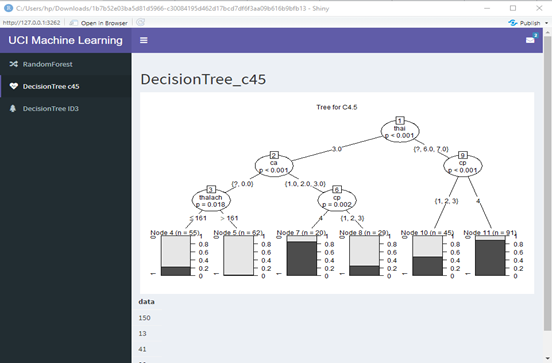
****

Fig. 4 Decision tree C4.5

1. MODEL PERFORMANCES AND COMPARISONS:

The model performance in the form of a confusion matrix is displayed in Table 2. A confusion matrix is a table used for describing the performance of a classifier that executed on given test data where the ―True values are considered known data values.

TABLE 2.

MODEL COMPARISONS THROUGH CONFUSION MATRIX

|  |  |  |  |
| --- | --- | --- | --- |
| **DECISION TREE** | **True (0)** | **True (1)** | **Class Precision** |
| Prediction (1) | 458 | 36 | 92.2% |
| Prediction (0) | 32 | 486 | 93.1% |
| Class recall | 99.28% | 93.10% |  |
| **RANDOM FOREST** | **True (1)** | **True (0)** | **Class Prediction** |
| Prediction (1) | 436 | 55 | 88.80% |
| Prediction (0) | 55 | 467 | 89.46% |
| Class recall | 88.8% | 89.46% |  |

These tables show that the Decision Tree model best predicts the Cleveland Dataset. It yields the best accuracy and in the minimum period.

1. CONCLUSIONS

In this project, we explore three machine learning methods with fourteen properties for predicting heart disease according to the Cleveland dataset. The absolute quality of the results is calculated from the accuracy and the time taken by the machine learning models. From the above discussion, it is concluded that the Decision Tree performs the best out of Random Forest in terms of accuracy as well as time taken for diagnosing disease. In the future, the results indicated that the system can be useful and helpful for the doctors and heart surgeons for timely diagnoses the chances of a heart attack in a patient.

1. REFERENCE

[1]https://archive.ics.uci.edu/ml/datasets/Heart+Disease