IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 6, June 2020

Prediction of Heart Disease Using Machine Learning Algorithms and Ensemble Learning

J. Phani Prasad[1], T.Venkatesham[2]

Assistant Professor, Information Technology, AGBS, Hyderabad, India[1,2]

Abstract: Heart Disease or Cardio Vascular Disease (CVD) is the key factor that leads to majority of deaths across the world from the past, therefore we require accurate and appropriate treatment as well as diagnosis system, and already lot of machine learning techniques is applied on large data sets in medicine field to analyse the data. Many researchers also have been using various machine learning algorithms to help doctors and medical practitioners to Diagnose the Heart diseases. This paper gives the survey of various classification algorithms like Naive Bayes, Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF) and Logistic Regression (LR) and the execution of the heart data set is depicted using Weka tool.

Keywords: Decision Trees, Heart Disease, Logistic Regression, Random Forest, Support Vector Machines.

I. INTRODUCTION

In our Human Body one of the major organ which is responsible for functioning of blood and to all body parts is heart, one of the most common disease in India is heart attack, if heart stops functioning then the total blood circulation system in our body stops which will lead to serious health condition it may even cause death.

The following are the types of heart disease that are commonly occurring in the world in which Cardio Vascular Disease is a class of diseases that involve the heart or blood vessels. CVD includes Coronary Artery Diseases (CAD) such as angina and myocardial infarction also known as heart attack. Coronary Heart disease is the other type of heart disease where the usual cause is the build-up of plaque. This causes coronary arteries to narrow, limiting blood flow to the heart.

The following are the symptoms of heart attack

1. Chest pain: The chest pain happens because of the blockage in the vessel of coronary part of the body
2. Pain in Arm: The pain usually starts in the chest part and slowly it creeps to the arm
3. Tiredness: The person feels like sweating throughout and looks very tired
4. Lack of Oxygen: level of the oxygen drops causes dizziness and out of balance
5. Bradycardia: in this patient will have a slower heart rate of over 50-60 bpm
6. Hypertension: In this the patient heart rate varies from 100-200 bpm and over some times

The other causes of heart attack are smoking and eating habits and obesity etc, in one of the survey it was estimated that nearly 17 million people in the world on an average are dying because of heart disease, and in India the estimate is about 40 million people who are affected by this disease, and about four lakh open heart surgeries are done per year. The data set used was the e Cleveland heart dataset from the UCI Machine Learning Repository as it is widely used by the Pattern design community. The dataset consists of 303 individual clinical reports in which 164 did not have any disease. In this dataset there is a total of 97 female patients in which 25 people are the affirmative case, also there are 206 male patients in which 114 are diagnosed with the disease. The paper is organised as Section II describes the algorithms and technologies used generally in analysis and prediction of data, Section III describes the Weka tool and Results and Section IV conclusion.

About the Data set: Cleveland Heart Dataset (UCI machine learning)

There are 12 attributes we have taken for the analysis

- Age
- Sex
- Chest pain type(4 values)
- Resting blood Pressure
- Fasting blood sugar > 120 mg/dl
- Maximum heart rate achieved
- Serum cholesterol in mg/dl

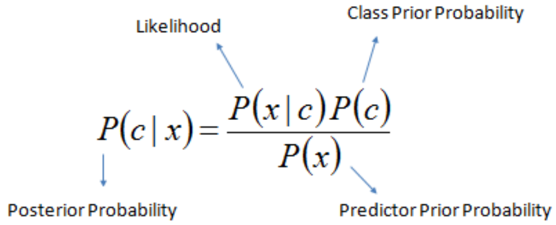
DOI 10.17148/IJARCCE.2020.9618

- Resting electrocardiographic results (values 0,1,2)
- old peak = ST depression induced by exercise relative to rest
- The slope of the peak exercise ST segment
- Number of major vessels (0-3) colored by fluoroscopy
- Thal :3 = normal ; 6= fixed defect ; 7= reversible defect

II. ALGORITHMS AND TECHNOLOGIES USED

A. Naive Bayes

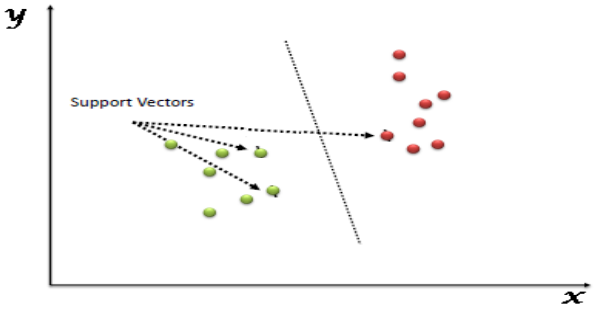
This algorithm is one the simple and better classification technique to classify the objects correctly, this is based on the Bayes theorem this technique uses independent attributes or variables for computation of the data and conclusions it also uses the conditional probability feature. Even though if there is any dependency all the attributes will contribute independently.


Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

B. Support Vector Machine

A support vector machine is a popular supervised machine learning algorithm which is used as a classifier and predictor; it uses a hyper plane which is used to differentiate classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall.



C. Decision Trees:

Decision tree is a supervised learning algorithm, it is used for classification tasks in which the categorical variables are mainly used it divides the main population in to two or more similar sets based on more predictor values in this method the entropy is computed for each and every attribute and later the data set is split based on minimum entropy or maximum information gain.

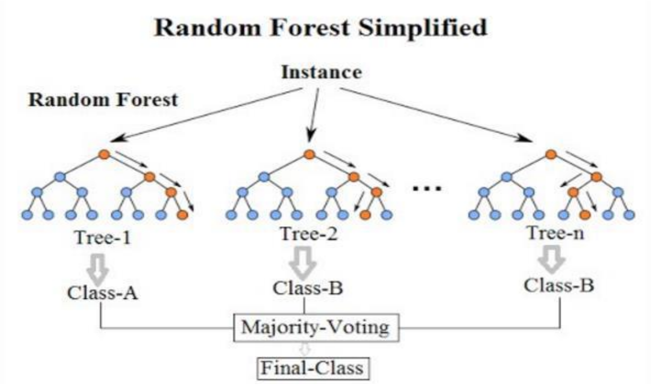
The formulas for Entropy and Information gain are given as follows:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_{v)}$$

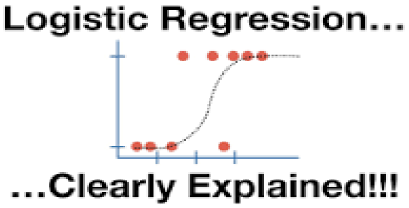
D. Random Forest:

This technique is a classification as well as regression oriented technique, as the name suggest it is a combination or collection of several decision trees before finalizing an output criteria, This technique main goal is it believes more number of trees will cover right decision. For classification it uses a voting system to decide the output and for regression it takes mean of all the outputs of each of decision trees it works well with large and high dimensional data. It is also classified under ensemble Learning.



E. Logistic Regression:

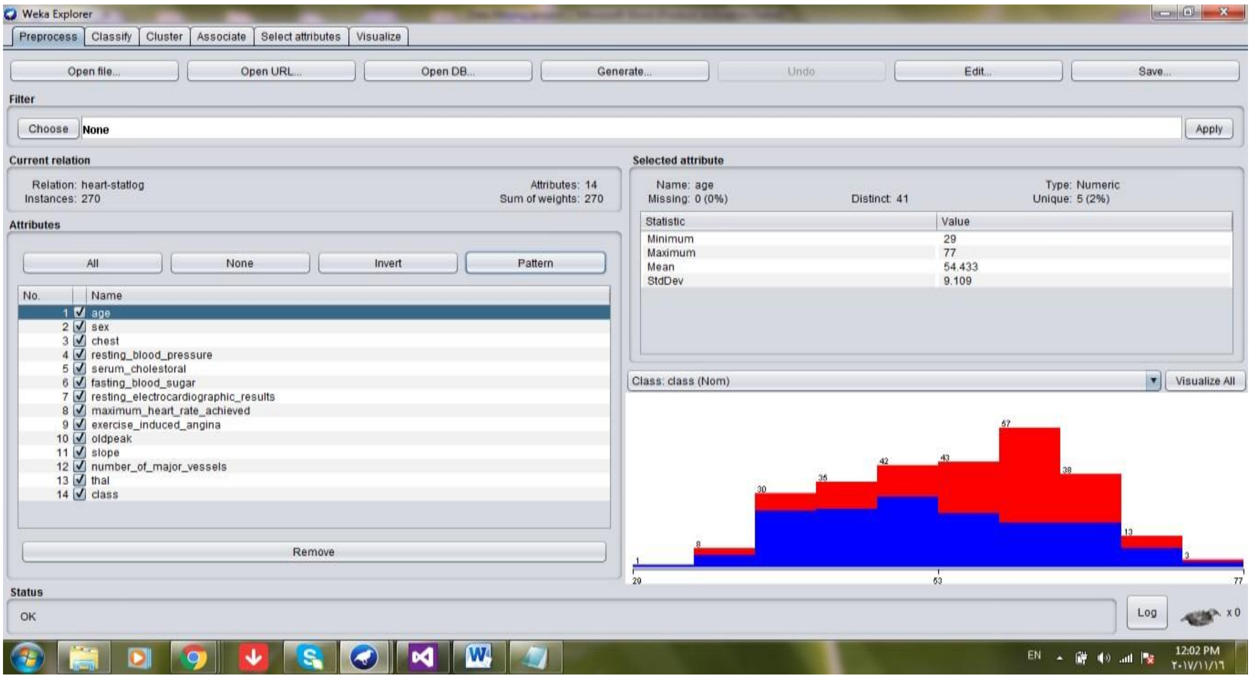
Logistic regression is a binary regression model which is used for categorical data and even in the case of continuous data in statistics logistic regression is used to model the probability of certain class or event existing like well or sick, pass or fail etc. we use a sigmoid function to model the data and infer the results in the form of a curve.



Ensemble Methods:

Ensemble Methods are the entities in which multiple models such as classifiers can be combined to solve particular classification problems. They improve the performance and will boost the accuracy of data. Examples are Random Forest, Ada Boost.

III. WEKA AND RESULTS



Filter: None

Current relation

Relation: heart-statlog
Instances: 270
Attributes: 14
Sum of weights: 270

Attributes

No.		Name
1	☑	age
2	☑	sex
3	☑	chest
4	☑	resting_blood_pressure
5	☑	serum_cholestoral
6	☑	fasting_blood_sugar
7	☑	resting_electrocardiographic_results
8	☑	maximum_heart_rate_achieved
9	☑	exercise_induced_angina
10	☑	oldpeak
11	☑	slope
12	☑	number_of_major_vessels
13	☑	thal
14	☑	class

Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 41
Type: Numeric
Unique: 5 (2%)

Statistic	Value
Minimum	29
Maximum	77
Mean	54.433
StdDev	9.109

Class: class (Nom)

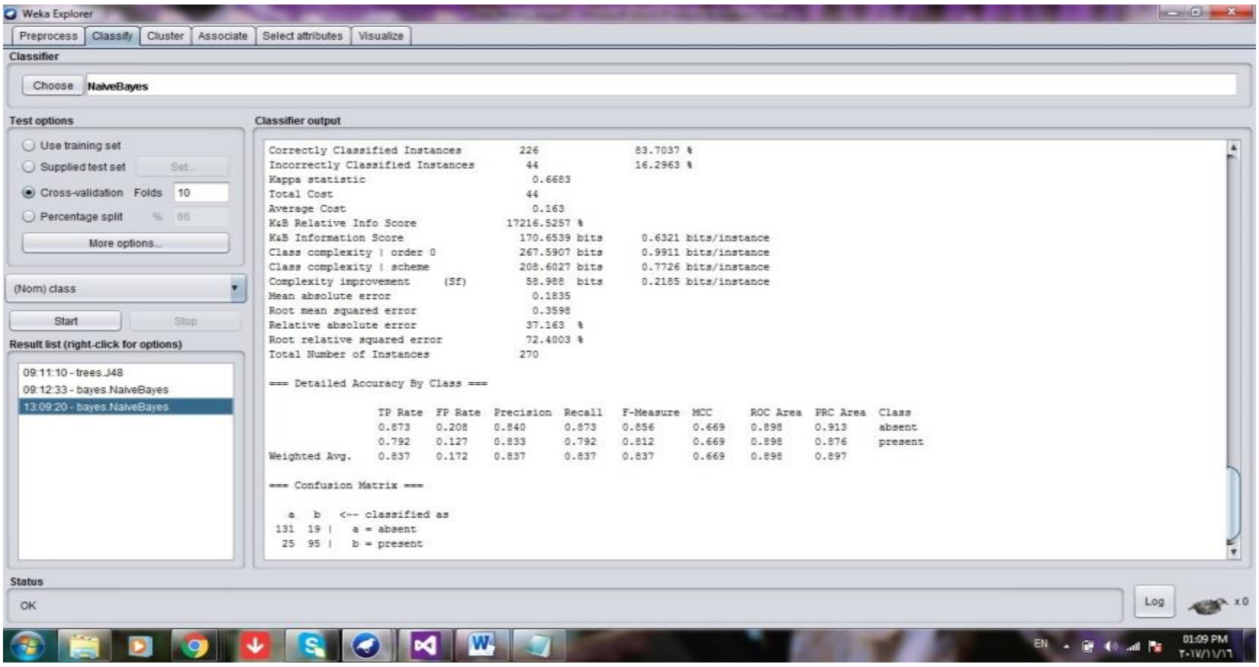
Status

OK

Fig1: Pre-processing of Heart Disease Data using Weka

Weka open source software provides tools for data pre-processing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. Weka (Waikato environment for knowledge analysis) is a mainstream suite of machine learning programming written in Java, created at the University of Waikato, New Zealand.

Using Weka we can pre-process the data, analyse the data and run the data with the help of various Machine Learning algorithms like J48, Naive Bayes classification, and we can use Association analysis even using Aprori and FP growth etc. Clustering and many more



Classifier: NaiveBayes

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

09:11:10 - trees.J48
09:12:33 - bayes.NaiveBayes
13:09:20 - bayes.NaiveBayes

Status: OK

```
Correctly Classified Instances         226               83.7037 %
Incorrectly Classified Instances        44               16.2963 %
Kappa statistic                          0.6683
Total Cost                              44
Average Cost                             0.163
K&B Relative Info Score              17216.5257 %
K&B Information Score                  170.6539 bits      0.6321 bits/instance
Class complexity | order 0             267.5907 bits      0.9911 bits/instance
Class complexity | scheme              208.6027 bits      0.7726 bits/instance
Complexity improvement     (Sf)         58.988  bits      0.2185 bits/instance
Mean absolute error                      0.1835
Root mean squared error                  0.3598
Relative absolute error                 37.163  %
Root relative squared error             72.4003 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.873    0.208    0.840      0.873    0.856      0.669    0.898     0.913     absent
                 0.792    0.127    0.833      0.792    0.812      0.669    0.898     0.876     present
Weighted Avg.    0.837    0.172    0.837      0.837    0.837      0.669    0.898     0.897

=== Confusion Matrix ===

   a   b   <-- classified as
 131  19 |   a = absent
  25  95 |   b = present
```

Fig 2: Heart Disease Data Result with Naive Bayes Algorithm Using Weka

IV. CONCLUSION

Based on the above study it is evident that most of the machine learning algorithms are performing well in predicting and diagnosing of cardio vascular or heart diseases, may be some algorithms are poor in their results in terms of performance and accuracy measures, Random Forest and Decision Tree algorithms generally work well on the data related to over fitting, whereas algorithms like support vector machine and Naive Bayes will work for most of the real world problems and on data sets. In future we can extend the usage of these algorithms to high dimensional data.

REFERENCES

[1]. V Ramalingam, Dandapath, Karthik Raja, Heart Disease Prediction Using Machine Learning Techniques: a survey, International Journal of Engineering and Technology, 2018.
[2]. Nandhni, Debnath, Pushkar, Heart disease prediction using machine learning, International Jour of Engineering Research and Development, 2018.
[3]. Tamara saad Mohammad, Heart disease prediction using weka, Baghdad College of economic sciences university, issue 58.
[4]. M A Jabbar, BL Deekshithulu, Heart Disease prediction using Genetic algorithm based trained recurrent fuzzy neural networks ,International conference on theory and application of soft computing with words and perception, 2017.
[5]. M muthuvel, Analysis of Heart disease prediction using various machine learning techniques, international conference on artificial intelligence smart grid, smart applications,2019.
[6]. Abhisheik Rairikar, vedant kulkarni, Heart disease prediction using data mining Techniques, IEEE conference on intelligent computing & control,2017.
[7]. Avinash G, Heart disease prediction using effective machine learning Techniques, International jour of recent technology and engineering, 2019.
[8]. Reddy Prasad, Anjali, Deepa, Heart disease prediction using logistic regression using machine learning,   IJEAT,2019.
[9]. Dinesh Kumar, prediction of cardiovascular disease using machine learning algorithms, IEEE conference 2018.

DOI 10.17148/IJARCCE.2020.9618