IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PROGNOSTIC ANANLYSIS ON HEART DISEASE

Ms Sukshitha R[1] and Ms Poornima[2]

Department of Statistics, Mangalore University, Karnataka

ABSTRACT: In human principal part of the body is heart. It is the most important organs in our body. Irregularity in our heart can effect destress in our body. Malfunctioning of our heart is termed as Cardiac disease or heart disease. The major heart diseases are cardiomyopathy (heart Muscle disease, congenital heart disease, Coronary Artery disease (CAD), heart Arrhythmias, heart failure, heart valve disease, Pericardial disease. Heart attack & heart failure are the most common heart diseases. Finding the causes of heart disease is the most important factor to our study. In this paper we have to study the prevention or minimizing the chances of getting heart disease and also we have to study the relationship between gender and heart disease. Heart disease is affected by several factors such as smocking, Cholesterol, diabetes, blood pressure, drug misuse, stress. Excessive use of alcohol and many more. Handling of imbalance data process is great challenge for the researchers to identify the significant factors. Detection of significant variable using classification model will reduce burden of doctors. The main focus of the study is to identify the major factors associated with significant variable & identity the better classification model. The prediction performance of different classification models is compared based on accuracy measure.

Key word: classification, chisquare test, logistic regression

1. INTRODUTION:

In human principal part of the body is heart. Heart is made up of two upper and lower chambers. Upper chambers are right and left atria and lower chambers are right and left ventricles. The atria collects incoming blood and ventricles pushes the blood out of the heart. The malfunctioning of our heart is termed as heart disease or cardiac disease. In India 52% of all cardiovascular deaths are occurred to the people who are aged below 70. For every 33 seconds in our country one person dies according to survey conducted by Times of India. Heart diseases can be likely to occur genetically. Blood pressure, heart disease and other related conditions with family history have more chance of risk with heart. Familial Hypertrophic Cardiomyopathy is a most common heart disease which can occur by inheritance for the person of any age. Symptoms of heart disease are chest pain or discomfort in shoulder, arm, back, neck or jawline due to excess exercise or emotional stress. Sometimes people may feel like heartburn, shortness of breath and sometimes there will be no symptoms. Usually men have a greater chance than women to attack of heart disease. Men in younger age suffer greater risk of heart attack. Due to smoking, excessive consumption of alcohol, drug abuse there is a greater chance of getting heart disease. Elder people are having more chance of getting heart disease. According to an estimation four out of five individuals dies due to coronary heart disease. To overcome this problem we should take steps to prevent heart disease.

2. LITERATURE REVIEW:

There are several works have done related to disease prediction systems by using different machine learning algorithms and data mining technique.

K. Polaraju et al, had proposed a Prediction on Heart Disease using Multiple Regression Model which proves that MLR model is appropriate for predicting heart disease. The work is performed using training data set consists of several instances with some attributes which has mentioned.. Based on the results, it is clear that the accuracy of classification Regression algorithm is best compared to other algorithms.

Marjia et al. developed heart disease prediction. Based on different factor like SMO and Bayes Net achieve better performance than KStar. The accuracy performances achieved by those algorithms are not that much satisfactory. So the accuracy performance has been improved to give better decision to diagnosis disease.

3. METHODS:

CLASSIFICATION MODEL:

3.1 Logistic regression:

Logistic regression analysis which gives association between a categorical dependent variable and a set of independent variables, and estimates the probability of occurrence of an event by fitting data to logistic curve. Binary logistic regression is used when the dependent variable is dichotomous and the independent variables is categorical. The numerical values of 0 and 1 are assigned to the outcomes of the binary variables which are dummy variables. Here the value indicate the presence or absence of categorical effect. If a categorical variable have more than two categories, it can be represented by dummy variables, having one variable for each category.

3.2 DECISION TREE:

A decision tree nothing but flowchart-like tree structure, where the topmost node in tree is the root node, test on an attribute denoted by internal node, an outcome of the test represented by every branch, and each terminal node holds a significant variable. Given a tuple which associated unknown significant variable, the attribute values of the tuple are tested against the choice of tree. A path is traced from the root to a terminal node, that holds the class predicted values for that tuple. This is appropriate for exploratory knowledge discovery since construction of decision tree doesn't need any domain knowledge or parameter setting. Decision trees can handle multidimensional data. The average of dependent variable values in a tuple is taken as the predicted value for all those tuples. CART algorithm applies the Gini index as the attribute selection measure.

3.3 K-NN CLASSIFICATION:

KNN compare the given test tuple with training tuple which is similar to it, so we can say it is based on learning by analogy. Once given an unknown tuple, a K- nearest neighbor classifier search the pattern area for the k-training tuple that area unit nearest to the unknown tuple. Closeness is outlined in terms of distance matric. For K-nearest neighbor classification, the unknown tuple is allotted to the foremost category among its K-nearest neighbors. K-nearest neighbor classifiers can also be used for prediction, that is, to return a real valued prediction for a given unknown tuple. A good value for K, number of nearest neighbors, can be found experimentally or k may be taken as

$k = \sqrt{\text{number of training tuple}}$

3.4 NAÏVE BAYES CLASSIFIER:

Bayesian classifiers are statistical classifiers. This is the algorithm to predict the probability of given tuple that belongs to a particular class. Bayesian classification is based on Bayes' theorem. A simple Bayesian classifier is known as the 'naïve Bayesian classifier' to be comparable in performance with decision trees and selected neural network classifiers. When naïve Bayesian classifier applied to a large

database, it gives high accuracy and speed. According to this algorithm, effect of an predictor on a given class is purely independent of the values of the other attributes. It is known as class conditional independence. To predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of X is Ci if and only if it is the class that maximizes P(X|Ci)P(Ci).

3.5 RANDOM FOREST:

Random forests is a classification method which uses ensemble learning method. Here to correct overfitting problem of decision tree, Random forest is one of the best alternative. It adds additional randomness to the model. Random forest search for the best feature among a random subset. So that in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. Another important quality of the random forest algorithm is that it is very easy to find out the importance of each feature on the prediction. While seeing at the feature importance you can decide which features to possibly drop because they don't contribute enough to the prediction process. The hyper parameter in random forest are used to increase the predictive power of the model as well as the speed of model.

3.6 SUPPORT VECTOR MACHINE:

SVM is a supervised machine applicable for classification and regression problems. This algorithm creates a hyperplane which separates the data into two classes. SVM is an algorithm that takes the data as an input and outputs, a line that separates those classes if possible. According to the SVM algorithm, the data points which are closest to the line from both the classes are called support vectors. The aim is to maximize the margin,that is distance between the line and the support vectors. The maximize margin is nothing but optimal hyperplane. Thus, SVM tries to create a choice boundary in such the way that the separation between the 2 categories is as wide as possible.

3.7 CHI-SQUARE TEST FOR INDEPENDENCE:

The Chi-square test of independence is a statistical hypothesis used to determine whether two categorical variable are likely to be related or not. The chi-square test of independence are based on the observed frequencies. Which is the numbers of observations in each combined group. The test is used to compare the observed frequencies to the expected frequencies.

The hypothesis is given below:

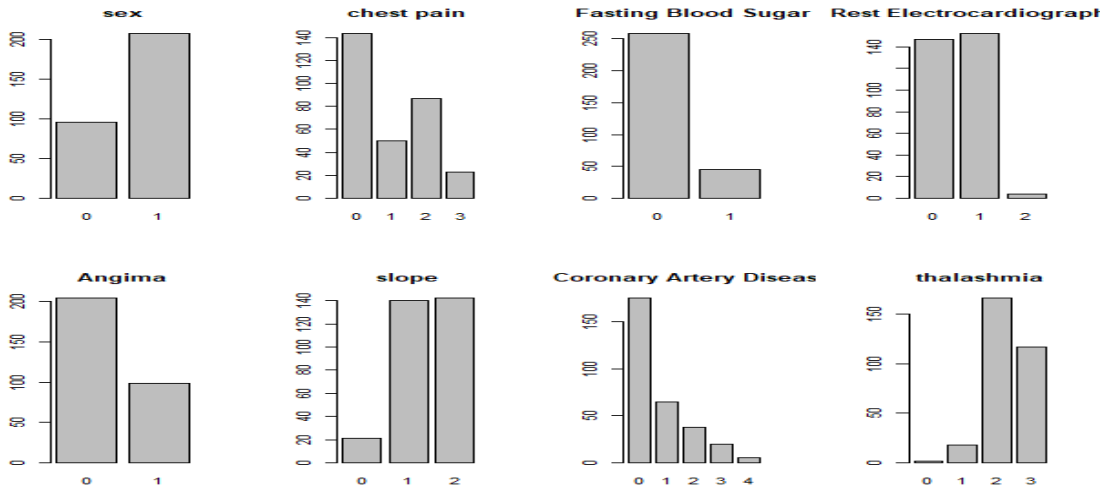
H0: The two categorical variables are independent.

H1: The two categorical variables are independent.

4. ANALYSIS AND DISCUSSION:

In human principal part of the body is heart. It is the most important organs in our body. Irregularity in our heart can effect destress in our body. Malfunctioning of our heart is termed as Cardiac disease or heart disease. The data is collected from the website www.kaggle.com. The data includes 303 records and 14 attributes.

Mode of categorical variables included in the study:



Logistic regression

According to logistic regression model, risk factors for heart disease are Sex, Chest pain, typical angina, atypical angina, non typical angina, the person's resting blood pressure, the person's maximum heart rate, exercise induced angina, ST depression induced by exercise relative to rest, slope, Coronary artery disease, thalassemia, reversible effect.

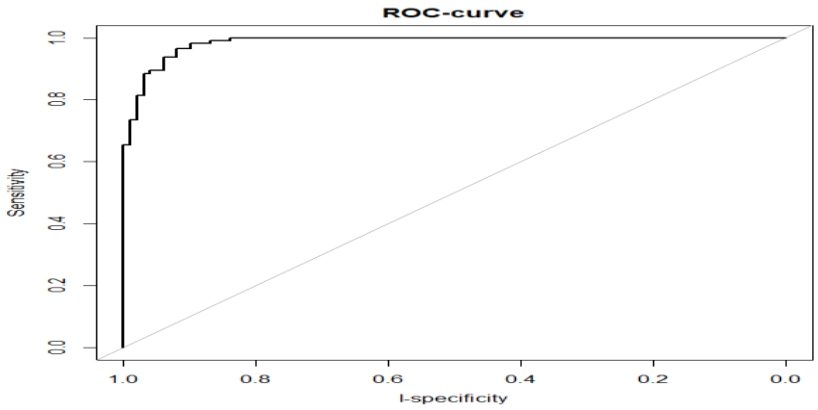


Fig – 1

The area under the curve is 0.9859. Which means that the test is 98% good in a given clinical situation. The test separates the data into two groups (survive or not) with 98% accuracy.

DECISION TREE:

According to the above decision tree thalassemia, exercise induced angina, sex, age, Coronary artery disease, The person's maximum heart rate achieved are important variables which are the reasons for heart disease.

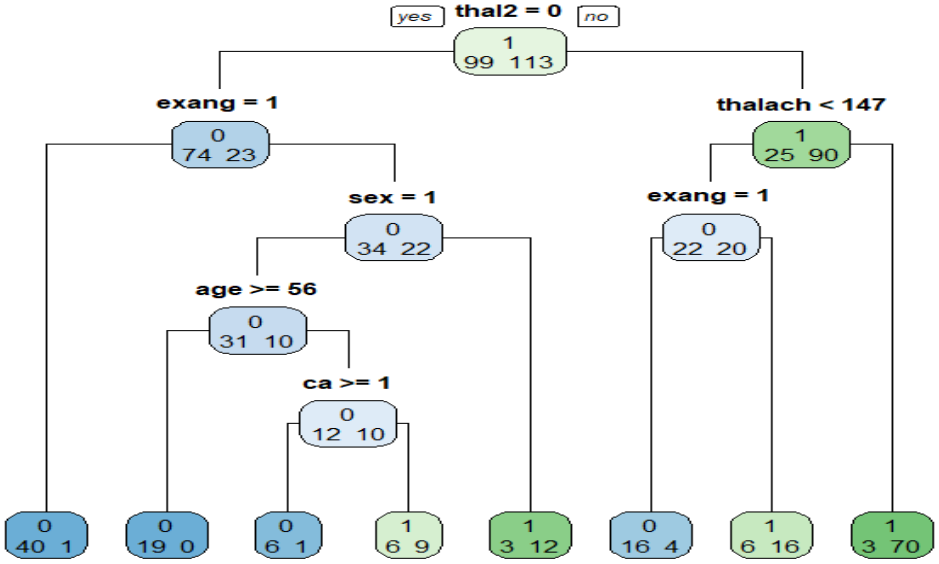


Fig - 2

Comparison of various classification models based on evaluation measures

Method	Accuracy	Sensitivity	Specificity
Logistic Regression	82.41%	81.08%	83.33%
Decision Tree	82.41%	78.84%	69.23%
K-NN Classification	63.73%	73.07%	51.28%
Naïve Baye's Classification	76.92%	82.69%	69.23%
Random Forest	81.31%	82.69%	79.48%
SVM	81.31%	86.53%	74.35%
XG Boost	78.02%	80.76%	74.35%

Table - 1

From the above table, we observe that accuracy, sensitivity and specificity of logistic Regression is highest compared to all other models. i.e, LOGISTIC REGRESSION classifier has a great potential for performance prediction. The generated classification rules can be used to predict whether the patients are getting heart disease or not.

Chi-Square test:

Hypothesis:

H0: there is no association between Gender and the heart disease

V/s

H1: there is an association between Gender and the heart disease

	(Male) 0	(Female) 1
(No Disease) 0	24	72
(Disease) 1	114	93

The chi-square statistic is 32.8362. Since the p value (0.00986) is less than 0.05. Therefore, we reject null hypothesis and conclude that there is an association between Gender and Heart disease.

CONCLUSION:

In human principal part of the body is heart. It is the most important organs in our body. Irregularity in our heart can effect destress in our body. Malfunctioning of our heart is termed as Cardiac disease or heart disease. Since the dataset contains many dependent categorical variables, we have fitted a logistic regression model and selected the significant risk factor responsible for the Disease event using logistic regression. ROC curve is constructed to check the discriminant power of the fitted model and the result shows that fitted model has very high discriminant power. Based on the Accuracy measure we can conclude that LOGISTIC REGRESSION model performs better than all other model under consideration. Based on chi square test we can conclude that there is an association between Gender and the heart disease.

REFERENCE:

- Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Prediction Using Machine Learning and Data Mining July 2017, pp.2137-2159.
- Ashok kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross validation", Springer, 17 September 2016.
- Boshra Brahmi, Mirsacid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining. o Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel, 2015-March 2016, pp. 129-137.
- S.Prabhavathi, D.M.Chitra, "Analysis and Prediction of International Journal of Innovations in Scientific and Engineering Research, vol.2, 1, January 2016, pp.1-7.
- Sairabi H. Mujawar, P.R. Devale, "Prediction of Heart Disease using Modified K-means and by using Naïve Computer and Communication Engineering, vol.3, October 2015, pp. Sharan Monica.L. Sathees Kumar.B, "Analysis of Cardio Vascular Disease Prediction using Data Mining Science".
- Preethi Jayarama Shetty, "Comparing classification model for multilabel imbalanced fetal growth data", vol.10, IJCRT in the year 2022.