

Efficient System for Heart Disease Prediction by applying Logistic Regression

¹S.Adithya Varun, ²G.Mounika, ³Dr. P.K. Sahoo, ⁴K. Eswaran

^{1,2,3,4}Dept. of CSE, SNIST, Hyderabad, Telangana, India

Abstract

In today's modern lifestyle people are effecting by different health issues, one among them is heart disease which may be incipient from a very early age. Cardiovascular disease remains as the number one cause of death globally. The main objective of this paper is to identify the presence or absence of heart disease for an individual. In the healthcare industry, it is very difficult to discover whether an individual is affected by heart disease or not by a physician. It requires a careful understanding of patient's data, and the identification of those parameters which cause the disease all of this is considered as a difficult task. Additional tools are required for making the clinical decision of heart disease. For the implementation of this work we referred to the Kaggle dataset1, which comprises 14 features (attributes) with class label, are identified as a cause for heart disease. In this paper logistic regression algorithm is applied for the Heart Disease prediction in order to improve the system's efficiency when compared with Naïve Bayes (NB) and Random Forest (RF).

Keywords

Logistic Regression, Cost Function, Regularization, Gradient Descent, Artificial Intelligence (AI).

I. Introduction.

Heart Disease is also called as a cardiovascular disease which means there is a problem with the functioning of heart and blood vessels. The heart is considered as the main center for the cardiovascular system. The heart pumps blood to all parts of the body. The blood carries oxygen, which is required for all the body cells. The cardiovascular disease occurs when the heart and blood vessels are not working properly. There are some other problems along with the cardiovascular disease. Arteriosclerosis which generally means hardening of arteries, the arteries become thicker and no longer are flexible. Atherosclerosis which means narrowing of arteries, so less blood flow through those buildups is called plaque. Heart attacks generally occurs when the blood clots or the blockage of blood flow to and from the heart.

'Global Hearts', a new program launched on 22nd September, 2016 by the World Health Organization (WHO) to overcome the deaths of cardiovascular disease and also heart attacks [2]. According to the world health organization, men and women are equally affected by heart disease. Based on the statistics WHO estimated 17.9 million people are died due to cardiovascular heart disease in 2016 [3] which represents 31% of the global deaths, 85% of these deaths are due to heart attacks and stroke. Here the data is more important for predicting heart disease. Now –a-days the hospitals are maintaining records of patients to manage their management and patient's information for future use. This system generates a huge amount of data the major focus is the knowledge identification which is to discover the hidden pattern that was not formerly known. Once those patterns are found that are used to support for making effective clinical decisions. For the clinical decision, they are using the data mining, which is an old approach. There is a need to expertise the existing which can be replaced by artificial intelligence.

A. Key idea

Machine learning is the application of artificial intelligence, which provides the system's ability to learn automatically and improve the systems. The learning process begins from observation such as direct experience, examples to look for patterns in data in order to make a better decision in future based on the examples. Machine learning can also be used in the medical field. Diagnosis of disease is based on doctor intuition and experience of the doctor, there may be a chance of inappropriate diagnosis of disease. Here, artificial intelligence or machine learning plays a vital role in making a clinical decision. We are providing additional knowledge to the doctor or physician by using a tool in which the doctor can analyze the patient's data visually. The world health organization estimates 400 million people have no access to the most basic health care [4]. AI has the potential to change this scenario. Many industries like medical, manufacturing aerospace and chemical that are already utilizing the benefits of data mining. Experts usually feel that there is need for new technology for perfect decision making. This new technology enables to predict the pattern using machine learning. The main cause that increases the possibility of heart attacks are concentrated use of alcohol, high blood pressure, cholesterol in the blood, lack of physical exercise, a large amount of certain fats, high sugar level and unhealthy diet.

II. Literature Survey

There is a number of prediction systems proposed for different diseases and implemented using different techniques. Previous works on heart disease with different authors studied and implemented different methods and analyzed the results. For the implementation of the work, they have considered the data set from the UCI data repository which can also be collected from the kaggle. The authors performed the classification and prediction technique on the data set.

Sellappan Palaniappan, Rafiah Awang are the authors they have used three data mining classification modeling techniques. These techniques dig out the hidden information from the heart disease database [5]. For accessing the model they have used DMX query language. The model is trained on train data and tested with test data to evaluate the results. These authors used Lift chart and classification matrix method for the evaluation of the effectiveness of the model. The system extracts the hidden knowledge from the historical heart disease database. The authors implemented the proposed system based on the.net framework which is a web-based prediction system that can be used easily and is reliable, expandable and scalable. The most effective model for heart disease prediction is Naïve Bayes followed by the Decision Tree. Naïve Bayes showed a better accuracy than the Decision Tree. [6].

Mai Showman, Tim Turner, Rob Stocker are the authors who applied the K-Means method that is combined with the Decision Tree for the Heart Disease prediction system. For the implementation of this work, they have applied Initial centroid selection techniques in order to boost the model accuracy [7]. For diagnosing the heart disease they have used the Decision Tree classifier. In order to generate the initial centroids based on the number of actual samples in the data set the Random Row,

Random Attribute, Inliers, outlier and Range methods are used [8]. The author collected the data from the UCI data repository. Finally, the author compares the performance with previously applied decision tree implemented formerly on the same data with decision-tree and K-Means technique that is used to get better accuracy. The proposed method showed an accuracy of about 83.9% [9].

Kanika Pahwa and Ravinder Kumar applied the hybrid technique for selecting the features on the heart disease dataset for prediction. Firstly the author approached the feature selection technique using the SVM-RFE along with the gain ratio for getting the subset of feature and removing the irrelevant and redundant feature. Identifying the features is important for prediction. They have applied Naïve Bayes and Random Forest on the subset of features for classifying the data set into present or absence of heart disease. The results have shown a better accuracy for both when applied with selected features. The accuracy achieved by Naïve Bayes is 84.15%, while the other Random Forest achieved 84.16%, for which it got almost similar accuracy when applied with selected features for both the methods [10].

III. Data set

Attribute description

Table 1: Dataset description

Attribute	Description
Age	how many years old Range[25 -110]
Gender	having value 1 for 'Male' and value 0 for 'Female'
Chest pain	value 1 for 'typical angina' value 2 for 'atypical angina' value 3 ' non-angina' value 4 'asymptomatic'
Resting Blood Pressure	blood pressure mm Hg range [60 - 200]
Cholesterol	cholesterol measured in mm/dl range[120 -600]
FBS (Fasting Blood Sugar)	value 1 for '>120 mg/dl' value 0 for '<=120 mg/dl'
Resting ECG (Electro Cardio Graph)	value 0 for 'normal' value 1 for 'abnormal' value 2 for 'showing probable'
Maximum Heart rate	heart count per minute range[60-200]
Exercise Induced Angina	value 0 for NO value 1 for YES
Old peak	ST curve relative to Rest range[0-6]
Slope	The slope of peak exercise ST segment value 1 for 'upsloping' value 2 for 'flat' value 3 for 'down sloping'
Coronary Artery	count of major vessels colored by floursospy range[0-3]
Thalassemia	Bone marrow expand relative congestive heart failure value 3 for 'normal' Value 6 'fixed defect' value 7 'reversible defect'
Class Label	0 for 'absence of heart disease' 1 for 'presence of heart disease'

IV. Proposed System

Heart Disease Prediction System uses Logistic Regression methodology for building the training model. Logistic Regression was first originally developed by David Cox [11]. Logistic Regression is a statistical analysis technique that is used for predicting the data value based on the prior observation of the data set. The logistic regression model predicts the dependent data variable by analyzing the relationship between one or more existing independent variables. Logistic Regression is one of the important tools for prediction, which can also be used for classifying and predicting the data based on the historical data. The implemented model is a binary Logistic model that has dependent variables with only two possible outcomes i.e., one is a positive value and another one is the negative value which is having 0 or 1 as class label. The logistic regression is implemented on the coronary artery disease [12]. Python code for the implementation of logistic regression algorithm done by Anthony made few changes according to need and requirements [13]. The proposed model is implemented as the following: It mainly consists of two major phases: regularized cost function and regularized gradient descent. Cost Function is used for calculating the maximum likelihood estimation. Gradient descent is an iterative process for getting coefficients from training data. The process is repeated until we get the optimal parameters of train data. The model is trained with the optimal coefficient. Whenever a test data has been passed to the model based on the parameters is able to identify whether the person is having heart disease or not, it tests the data using the sigmoid function. The cost function is the method that is used for reducing the errors of the predicted label and the actual label. Gradient descent function is the method that is used for calculating the coefficient until we obtain a minimum value of the class label.

A. Cost Function

A minimization function is used that is the cost function. It uses the Log Loss i.e. the logarithmic loss which measures the performance of the model where the prediction input value is the probability between the zero and one. The Log loss is the uncertainty of the prediction which is based on the how much it varies from the actual label. Cost function which helps the learner to correct or change the behavior to minimize the mistakes. The cost function can be estimated by iteratively running the model to compare the estimated predicted value and the known or actual value. The regularized cost function is a method that is used for overcoming the risk of over fitting. Lamda is the parameter which controls the regularization term. The cost function is calculated by the following:

Regularized Cost Function [13]

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2$$

m = number of instances n = number of attributes y = class label
x = train data features
θ = coefficients λ = learning rate

B. Gradient Descent

Gradient descent is an optimization method which is used to find the parameters or the coefficient of the cost function. Gradient descent is a repeated process in order to get the coefficients to minimize the cost function. The Gradient descent is calculated for both the classes to get the pair of a coefficient for both class labels. The goal here is to continue the procedure to try the different value for the coefficient, evaluating their cost and selecting the

new coefficient that is having the slightly lower cost. Considering this coefficient and storing them in the model. Gradient descent is calculates as the following:

Regularized Gradient Descent [14]

$$\theta_j = \theta_j - \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^i) - (y^i)) x_j^i + \frac{\lambda}{m} \theta_j$$

m = number of instances x = train data features
y = class label θ = coefficients λ = learning rate

C. Sigmoid Function

Sigmoid function is the logistic function between. This takes the real input vales and output values between the 0 and 1 for logistic function [12]. This is interpreted as taking log odds and having the output probability. Generally sigmoid function is used to map predictions to probability it is defined as:

Logistic Function [15]

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

x = test data features θ = coefficients

Whenever a test data is passed it calculates the value based on the parameters stored in the model. It calculates the probability of each class label. We return the maximum probability value of the class label x_i .

$$h_{\theta}(x_i) = \frac{1}{1 + e^{-\theta^T x}}$$

The test data contains the thirteen attributes that we need to pass and calculate for both the classes it will return the two values we take the maximum value of two values we will return the class label which is having the maximum probability.

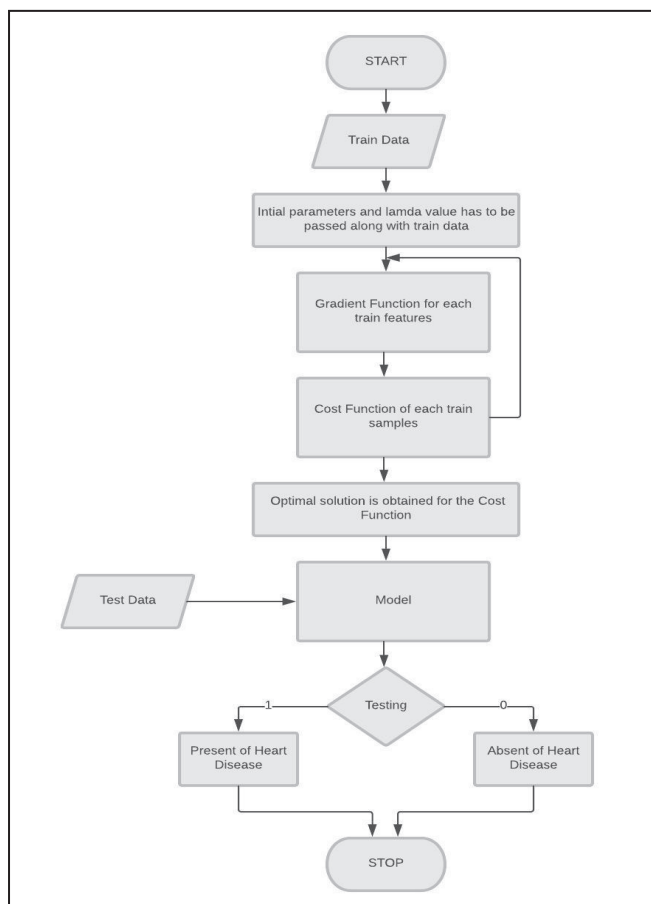


Fig. 1: Data Flow Diagram

V. Results

The accuracy obtained is about 87% after applying the Logistic Regression on the dataset. The dataset is divided into the ratio of 90:10 i.e. 272 instances for training and 31 samples for testing the data. The accuracy obtained for the proposed method is about 87% when compared with the Naïve Bayes which got 83.7% and Random Forest 80%. The applied method is efficient than the Naïve Bayes and Random Forest.

Table 1: Comparison Table With Different Algorithm

S.NO	ALGORITHM	ACCURACY
1	Logistic Regression	87%
2	Naïve Bayes	83.7%
3	Random Forest	80%

A. Confusion Matrix

Confusion matrix generally used for measuring the performance of the algorithm. It shows the way in which the model is confused when it makes predictions. Each row indicates the instances of the actual class label and each column indicates the instances of the predicted class label.

We illustrate the results with confusion matrix which shows the complete details.

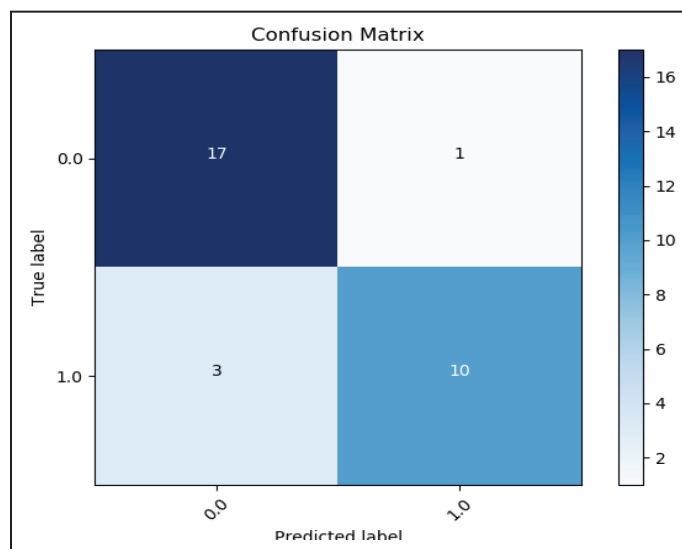


Fig. 2: Confusion Matrix

Table 2: Precision and Recall

Class label	Precision	Recall	F1-Score	Support
0	0.85	0.94	0.89	18
1	0.91	0.77	0.83	13
Avg / Total	0.87	0.87	0.87	31

Recall = TP/(FN+TP)

Precision = TP/(FP+TP)

VI. Conclusion

A proper system has to be developed for predicting the heart disease with an efficient model. In this project, the dataset has 303 instances for training the model which is giving 87% test accuracy. When given with more points for training the model we can improve the test accuracy. Building a perfect model with a good amount of data always gives better results. The proposed

system is giving better when compared with naïve bayes and random forest.

References

- [1] Kaggle.com. (2018). Heart Disease UCI. [Online] Available: <https://www.kaggle.com/ronitf/heart-disease-uci>.
- [2] "A program is launched to overcome the cardiovascular disease and heart attacks and strokes", World Health Organization.[WHO]. [Online] Available: http://www.who.int/cardiovascular_diseases/global-hearts/Global_hearts_initiative/en/03-Jul-2017.
- [3] [Online] Available: [https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)2016](https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)2016).
- [4] Who.int. (2015). WHO | New report shows that 400 million do not have access to essential health services. [Online] Available: <https://www.who.int/mediacentre/news/releases/2015/uhc-report/en/>.
- [5] Obenshain, Mary K., "Application of data mining techniques to healthcare data", Infection Control & Hospital Epidemiology 25.8 (2004), pp. 690-695.
- [6] S. Palaniappan, R. Awang, "Intelligent heart disease prediction system using data mining techniques", IEEE/ACS International Conference on Computer Systems and Applications, Doha, 2008, pp. 108-115.
- [7] Tajunisha, N., V. Saravanan, "A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets", International Journal of Advanced Science and Technology 27 (2011): pp. 85-94.
- [8] Khan, Dost Muhammad, Nawaz Mohamudally, "A multiagent system (MAS) for the generation of initial centroids for k-means clustering data mining algorithm based on actual sample datapoints." Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on. IEEE, 2010.
- [9] Shouman, Mai, Tim Turner, Rob Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients", Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.
- [10] Pahwa, Kanika, Ravinder Kumar, "Prediction of heart disease using hybrid technique for selecting features", Electrical, Computer and Electronics (UPCON), 2017 4th IEEE Uttar Pradesh Section International Conference on. IEEE, 2017.
- [11] Cox, David R., "The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological) (1958): 215-242.
- [12] Walker, Strother H., David B. Duncan, "Estimation of the probability of an event as a function of several.
- [13] A. segura, "Logistic Regression from scratch | Kaggle", Kaggle.com, 2018. [Online] Available: <https://www.kaggle.com/anthonysegura/logistic-regression-from-scratch>.
- [14] Wiki.fast.ai. (2010). Logistic Regression - Deep Learning Course Wiki. [Online] Available: http://wiki.fast.ai/index.php/Logistic_Regression#Cost_Function.
- [15] Wiki.fast.ai. (2010). Logistic Regression - Deep Learning Course Wiki. [Online] Available: http://wiki.fast.ai/index.php/Logistic_Regression#Gradient_Descent.
- [16] Wiki.fast.ai., (2010), "Logistic Regression - Deep Learning Course Wiki. [online] Available: http://wiki.fast.ai/index.php/Logistic_Regression#Sigmoid_Function.



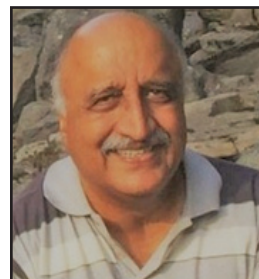
S. Adithya Varun received his Bachelor degree in computer science engineering from Nalla Malla Reddy Engineering College, Hyderabad, India, in 2013, pursuing Masters from Srinidhi institute of science and technology doing research on machine learning.



G. Mounika received her Bachelor degree in computer science and engineering from Mahatma Gandhi Institute of Technology College, Hyderabad, India, in 2015, pursuing Masters from Srinidhi institute of science and technology doing research on machine learning.



Dr. P.K. Sahoo, completed his Ph.D. from Fakir Mohan University, Odisha in Computer Science Engineering. He has 14 years of teaching, research and administrative experience. He has earlier worked as Head of the Department for both CSE and IT in various reputed Engineering Colleges. His Research interest includes Cyber Security, Information Security and Data Mining. He has published around 21 research papers in various reputed journals at national and International levels. He is a certified professional from BalaBit, completed Electronic Contextual Security Intelligence, exam for Intermediate Level (ECSI).



Dr. Kumar joined SNIST from 1999 after leaving BHEL R&D as Additional general Manager. He has Masters and Ph. D, degrees from IIT Kanpur and University of Madras respectively. He now works in the area of Artificial Intelligence involving Neural Networks and Image Processing using pattern recognition methods. He has more than 40 publications in various reputed International Journals and conferences. Many of his publications are referred to by present researchers, even after several decades. He has also won several best teacher awards- the latest being as Best Faculty Award 2013-14, Computer Science from Cognizant for the South India Area.