

Predicting Heart Disease Using Standalone Application

¹Pinjarla Poornamohan, ²Suresh Kumar Meenige

^{1,2}Dept. of CSE, KIET, Kakinada, AP, India

Abstract

Heart related diseases or Cardiovascular Diseases (CVDs) are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. So, there is a need of reliable, accurate and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases. This paper presents a survey of various models based on such algorithms and techniques and analyze their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), NaïveBayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers.

Keywords:

Machine Learning

I. Introduction

According to the World Health Organization, every year 12 million deaths occur worldwide due to heart disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years[1]. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high turn reduces the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future heart disease by analysing data of patients which classifies whether they have heart disease or not using machine learning algorithm[2]. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease [3].

III. Objective of Research

- Classification prediction model to just about all of the foremost extremely cited distance measures within the connected on heart condition datasets.
- Nearest Neighbor classifiers are the most effective method and take into account alternative classifiers, as well as neural

networks and J.48 algorithm.

- To improve analysis study area by increasing our search area to incorporate deletion.
- Neural network approach is captures the optimization values and uses these values to represent the statistic measurement.
- PSO optimization selection model to detect the key modification points in anytime series, and uses these points to represent the whole statistic.
- NN classification approach is to interrupt a variable statistic instance into multiple univariate-time series data then every time series is processed individually into disjoint segments and also the aggregate distance is generated.

III. Performances Metrics Analysis

It describes an evaluation metrics for heart disease prediction model. The table contains Mean Absolute error, Root Relative square Error, Root Relative Square Error and Accuracy values of SVM, KNN, RF, J.48 and MLP classification algorithm.

Conclusion

The proposed technique is producing an enhanced concept over the heart disease prediction within novel data mining techniques; SVM, RF, NB, MLP and j48 the weighted association classifier.

IV. Approach Methodology

A. Classification Algorithms

Classification is a supervised learning procedure that is used for predicting the outcome from existing data. This paper proposes an approach for the diagnosis of heart disease using Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes 0 for failure and 1 for success[3].

Naïve Bayes

Naïve Bayes classifier is a supervised algorithm. It is a simple classification technique using Bayes theorem. It assumes independence among attributes. Bayes theorem is a mathematical concept that is used to obtain the probability. The predictors are neither related to each other nor have correlation to one another[7]. All the attributes independently contribute to the probability to maximize it. Many complex real-world situations use Naive Bayes classifiers $P(X/Y) =$

$$P(Y/X) \cdot P(X)/P(Y),$$

$P(X/Y)$ is the posterior probability, $P(X)$ is the class prior probability, $P(Y)$ is the predictor prior probability, $P(Y/X)$ is the likelihood, probability of predictor.

$P(X/Y)$ is the posterior probability, $P(X)$ is the class prior probability, $P(Y)$ is the predictor prior probability, $P(Y/X)$ is the likelihood, probability of predictor.

Support Vector Machine

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized[8]. The goal of SVM is to divide the datasets into classes to find a maximum marginal.

K-Nearest Neighbour

The K-Nearest Neighbour algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbour. It is a type of instance based learning. The calculation of distance of an attribute from its neighbours is measured using Euclidean distance. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them[6]. K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy[13].

Random Forest

Random Forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast[9]. In the random forest classifier, the more the number of trees higher is the accuracy. It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets [16].

XGBoost

XGBoost is an optimized distributed gradient model designed to be highly efficient, flexible and portable. It is a decision tree based ensemble Machine Learning algorithm that uses gradient boosting framework. It provides an optimized gradient boosting algorithm through parallel processing, tree pruning, handling missing values and regularization to avoid overfitting or bias[7].

Gradient Boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees.

Gradient boosting trees can be more accurate than random forests. Because we train them to correct each other's errors, they're capable of capturing complex patterns in the data.

Algorithm Used: Gradient Boosting classifier

Boosting is an ensemble method that combines several weak learners into a strong learner sequentially. In boosting methods, we train the predictors sequentially, each trying to correct its predecessor. Gradient Boosting.

Gradient Boosting is the grouping of Gradient descent and Boosting. In gradient boosting, each new model minimizes the loss function from its predecessor using the Gradient Descent Method. This procedure continues until a more optimal estimate of the target variable has been achieved.

Unlike other ensemble techniques, the idea in gradient boosting is that they build a series of trees where every other tree tries to

correct the mistakes of its predecessor tree.

Components of Gradient Boosting

- Loss function
- Weak Learners
- Additive Component

STEPSTOGRADIENTBOOSTINGCLASSIFICATION



Gradient Boosting Model

STEP 1: Fit a simple linear regression or a decision tree on data [= , =]

STEP 2 : Calculate error residuals by subtracting predicted target value from actual target value. [= -]

STEP 3 : Fit a new model on the error residuals as the target variables keeping the input variables same. []

STEP 4 : Add the predicted residuals to previous predictions [= +]

STEP 5 : Fit the next model on the remaining residuals. [= -]

Repeat steps 2 to 5 until the model starts overfitting or there is no change in the residuals sum Example

Step 1: Make initial guess using log of the odds of target variable.

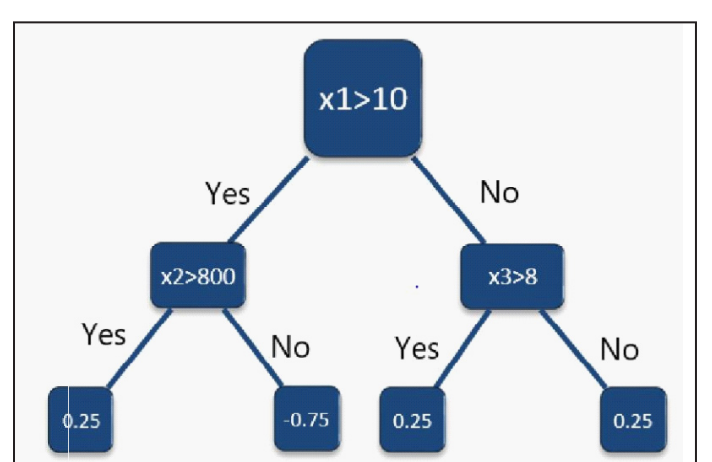
$$odds = \log\left(\frac{P(Y = 1)}{P(Y = 0)}\right) = \log\left(\frac{3}{1}\right) = \log(3)$$

To do classification, we apply softmax transformation.

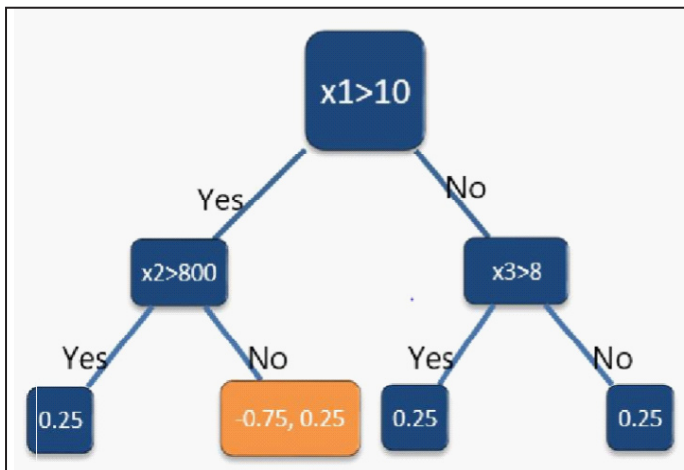
$$P(Y = 1) = \frac{e^{odds}}{1 + e^{odds}} = \frac{3}{1 + 3} = 0.75$$

Step 2: Calculate error residuals or pseudo residuals by subtracting prediction from the observed values

Step 3: Compute Classification tree.



This is an example of classification tree with just two leaves. However, Gradient Boosting often has more than 5 leaves and many leaves can have multiple values. Therefore, Gradient Boosting uses transformation for Classification. Consider the following tree:



Therefore, the value of second leaf is given by the following transformation.

$$= \frac{\sum \text{Residual}}{\sum [\text{Previous Probability} * (1 - \text{Previous Probability])]}$$

$$= \frac{-0.75 + 0.25}{0.75(1 - 0.75) + 0.75(1 - 0.75)} = -1.33$$

Step 4: Make the prediction.

$$y_{\text{prediction}} = \text{odds} + \text{learning_rate} * \text{residual}$$

$$y_{\text{prediction}} = \log(3) + 0.1 * (-1.33) = 0.965$$

Learning rate defines the contribution of the new tree. Now new log odds prediction can be converted to probability using softmax function.

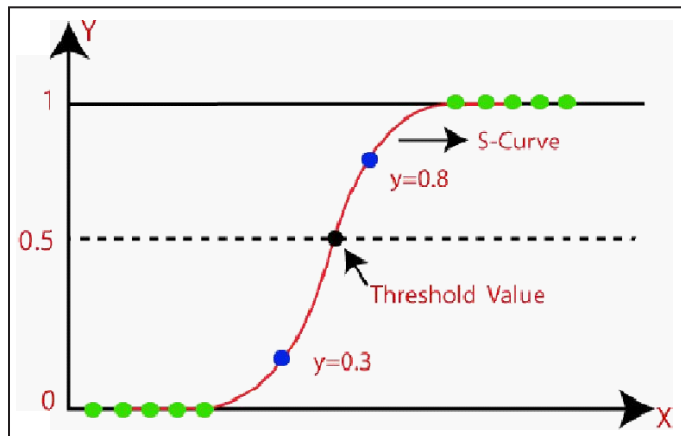
$$P(Y = 1) = \frac{e^{0.965}}{1 + e^{0.965}} = 0.724$$

As you can see, the probability has diminished from previous log odds ratio.

Step 5: Repeat steps until the model starts over-fitting or there is no change in the residuals sum.

Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an “S” shaped logistic function, which predicts two maximum values (0 or 1).



Logistic Function (Sigmoid Function):

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1.

The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the “S” form. The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1 - y):

$$\frac{y}{1-y}; 0 \text{ for } y=0, \text{ and infinity for } y=1$$

Type of Logistic Regression used in this model Binomial In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

VI. Result and Analysis

The aim of this research is to analyse the performance of various classification algorithms and in doing so find the most accurate algorithm for predicting whether a patient would develop and heart disease or not. This research was done using techniques of Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, XGBoost on the UCI dataset[4]. Dataset was split into training and test data and models were trained and the accuracy was noted using Python

VII. Conclusion

In this paper, we propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to

several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

References

- [1] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. *Clin Epidemiol.* 2011;3:67.
- [2] Weng SF, Reys J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE.* 2017;12(4):e0174944.
- [3] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. *Int J Eng Technol.* 2018;7(2.8):6847
- [4] Avinash Golande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, *International Journal of Recent Technology and Engineering*, Vol 8, pp.944-950,2019.
- [5] T.Nagamani, S.Logeswari, B.Gomathy, Heart Disease Prediction using Data Mining with Mapreduce Algorithm, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 22783075, Volume-8 Issue3, January 2019.
- [6] Internet source [Online]. Available (Accessed on May 1 2020): <http://acadpubl.eu/ap>
- [7] H. Jindal, S. Agrawal, R. Khera, R. Jain and P. Nagrath, "Heart disease prediction using machine learning algorithms", *ICCRDA 2020, IOP Conf. Series: Materials Science and Engineering*, 1022 (2021) 012072, DOI:10.1088/1757-899X/1022/1/012072.
- [8] P. Motarwar, A. Duraphe, G. Suganya, M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning", *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, IEEE, 2020
- [9] V. Sharma, S. Yadav, M. Gupta, "Heart Disease Prediction using Machine Learning
- [10] Techniques", *2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, IEEE, 18-19 Dec. 2020.