Algorithm To Predict Heart Disease Using Machine Learning

Sayed Aftab Ahamed[1], Sharath Kumar S R[2], Girish Mantha[3]

Dept. of Information Science & Engineering

J N N College of Engineering Shimoga-India

Abstract

In this new generation of social and economic changes the industries are giving a huge preference for the new technologies of industrial revolution for industrialization. In this period, day by day the technologies are growing very fast. Nowadays, the information's are used as technology this is known as nothing but knowledge. Here are the combinations of algorithms of data analytics converts the stored data into an knowledge. In this system various machine learning algorithms are used and the data which predicts the patient whether patient is having heart disease or not. The main theme of our project is the heart disease prediction using machine learning algorithm. Models based on supervised machine learning algorithms such as Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression. Keywords: Support Vector Machine, K-Nearest Neighbor, Decision Tree, Logistic Regression.

Keywords: Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression

I. Introduction

In human body, heart is one of the most important and muscular organ. The function of the heart is to pump blood around the body to keep us alive. The normal resting heart rate for adults range from 60 to 100 beats a minute and one lakh times per day. According to our lifestyles, work condition, stressful life and bad food habit leads to increase the rate of heart related diseases. Nowadays, Coronary heart disease is leading for death related issues around worldwide. As per survey 3.8 million in men and 3.4 million in women are used to die every year due to coronary heart diseases. Nowadays, we are using theold patient records to know the chances of heart related diseasesin new patients.This is the way of machine learning techniques to know the possibilities of heart diseases in new patients to save the millions of life in the early stages.

The paper is divided into seven sections. Brief introduction is discussed in Section I.We are gone through the paper and carrying the literature survey is presented as Section II. The dataset of patients is presented in Section III. The proposed system is presented in Section IV. The implementation and algorithm of the proposed system is discussed in Section V. Section VI provides the snapshots of the results obtained with the proposed system. In Section VII the paper is concluded.

II. Literature Review

The proposed [1] predicting of heart disease of a person is required the importance of vertical system integrating a sensor made of machine learning algorithm. In order to increase the accuracy of the model mainly the project is based on human heart rate. The system uses the sensor data and applying in ML algorithm then they predict the chances of heart diseases. System proposes [2] The collected data which is changed into knowledge data by using the various algorithms. The logistic algorithm is used in this paper and patients classified through the information given by this data and can be known whether they are having heart disease or not. This paper also includes Navie Bayes algorithm for getting accuracy result. And also finally judge the result through comparing models and confusion matrix.

The system [3] has two purposes which are rising diagnosing accuracy and reducing classification delay. The WPCA gives with the successful cacophonous criteria that has been given into the generic algorithms. The system successfully recognizes the disease and its subtypes such as normal and mild to maintain with highest accuracy. The system proposes [4] the new chance for predict heart attack provides by machine learning and deep learning algorithms. Deep learning works on the principal of higher level hierarchy to lower features.

III. Dataset of Patients

The Framingham heart study is dedicated to identifying common factors or characteristics that contribute to cardiovascular diseases. Core research in the dataset focuses on cardiovascular and cerebrovascular diseases. This dataset contains the information of patients about heart disease.

Total number of patient records in the data set is 4240 and columns are 16. The attributes of the dataset are Gender, Age, Education, Current Smoker, Cigarettes per day(Cigs per day), BP Meds, Prevalent Stroke, Prevalent Hyp, Diabetes, Total Cholesterol(tot Chol), systolic blood pressure(sysBP), diastolic blood pressure(diaBP), body mass index(BMI), heart rate of the patient, glucose level, Ten year coronary heart diseases(CHD).

IV. Proposed System

The proposed system can be used for heart disease prediction. Initially, we gather the data of the patient's medical history and store it in table or excel format. And then we check whether there is any missing value in the data accordingly and choose the appropriate algorithm which provides a highest accuracy accordingly. After we split the data into training set and testing set. Further, convert the columns to a suitable format and normalize all the numerical data to fall in a similar range. Then feed the preprocessed training data to algorithm and train the models. This will give the final model. Apply the final model to the test data and check the accuracy of the algorithm and select the algorithm of highest accuracy.

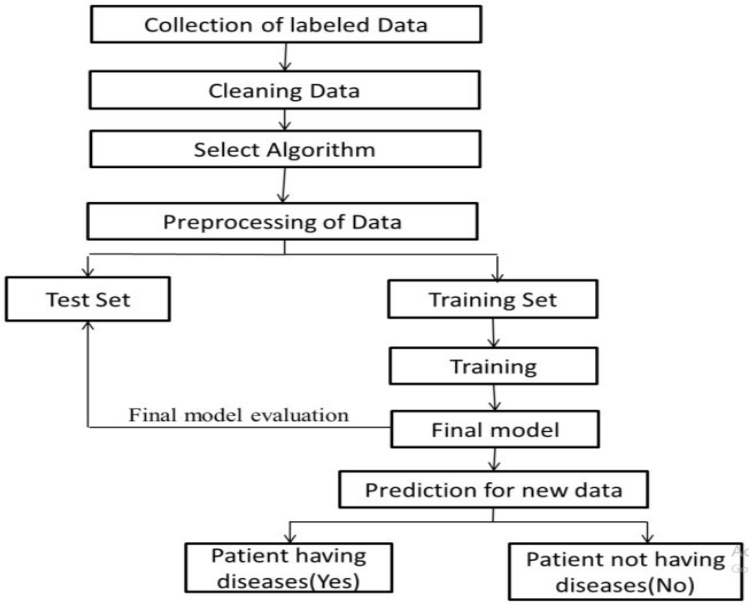

Fig.1 Proposed system

Once all the models are trained, compare the accuracies of all the model and select and save the model with highest accuracy in pickle format. This file is later used for prediction in real time. Create web page to enter the patient details which accepts all the patient details like (BP, Cholesterol etc). Web page is created using Flask library from python. Once user enters all the values of patient in web page and click on Predict button, the values are processed and sent to machine learning file (stored in. pickle format). Pickle file consumes the value and predicts the chances of Heart disease.

If Machine learning model prediction will be displayed back in the web page accordingly. If ML model predicts it as 1 (Heart Disease), Then the model calculated the % of chances and display in Web page as "You have 90% of Heart Disease. Please contact doctor Immediately.!". If ML model predicts it as 0 (No Heart Disease), then we display the message in webpage as "You are Safe! Stay Healthy!".Used algorithms in the project are Decision Tree, Support vector machine(SVM), k-nearest neighbors (KNN), Logistic regression algorithm. Fig.1 shows Proposed system.

V. Algorithm and Implementation

1) Decision Tree: Decision tree is one of the most popular algorithm. Decision tree algorithm falls under the category of supervised learning algorithms. Decision tree is used to build classification and regression models in the form of a tree structure like leaves are final outcomes and nodes are where the data is split.The accuracy of Decision tree is 80.11%.

Steps of Decision Tree Algorithm:

1. Calculate entropy for dataset.

2. For each attribute

    1. Calculate entropy for all its categorical values.

2. Calculate information gain for the attribute.

3. Find the attribute with maximum information gain.

4. Repeat it until we get the desired tree.

5. After the tree built on the training data, used it for prediction on test data.

6. Calculate confusion matrix and accuracy of the model.

2) SVM Algorithm: SVM are one of the SL models with associated learning algo,. thatdoes analyze the data used for classification and regression analysis. The accuracy of SVM is 88.44%.

Steps of SVM Algorithm:

Step 1: Find the points closest to the both classes. This points are support vectors.

Step 2: Find the proximity between diving planes and support vectors. The distance is known as margins.

3) Logistic Regression Algorithm: It is a machine learning algorithm used for the classification analysis, a predictive analysis algorithm based on the concept of probability. The accuracy of logistic regression algorithm is 69.02%.

Steps of logistic regression algorithm:

Given each training instance:

1. Using the current values of coefficients calculate the prediction.

2. Based on the error in prediction calculate the new coefficient values.

3. Repeat the process till the model is accurate for the threshold number of iterations.

4. Continue to update the model for training instances and correcting errors until the model is accurate enough or cannot be made any more accurate.

*4) K-nearest Neighbors:*It is one of the simplest algorithmsused for regression and classification problem in machine learning. Based on similarity measures (e.g. distance function) the KNN algorithm use data and classify new data points. The accuracy of KNN algorithm is 85.14%.

Steps of K-nearest neighbor's algorithm:

Step 1: For implementation of an algorithm, we require data set. In the first step of KNN, training data and test data is to be loaded.

Step 2: In this step, the value of K need to be chosen i.e. the nearest data points. K can be any integer value.

Step 3: Following steps are to be performed on each point of test data –

1. Using either Euclidean, Manhattan or Hamming distance the distance between test data and each row of training data to be calculated. Euclidean distance is the most commonly used.
2. Based on distance value they will be sorted in ascending order.
3. Top K rows are chosen in this step from the sorted array.
4. The class for the test point will be assigned based on most frequent class of these rows.

Step 4: End

VI. Result

The proposed system focuses on prediction of heart diseases. Fig.2 Graph shows accuracy of the algorithm such as Decision Tree, Logistic Regression, SVM and KNN Algorithm. The SVM algorithm of the proposed system gives the accuracy of 88.44%. The best accuracy of the proposed system is Support Vector Machine Algorithm.

Fig.3 describes the comparison graph of SVM algorithm between parameters. Compare to other algorithm SVM gives best accuracy rate. The model with highest accuracy is saved in .pickle format. The values are processed and sent to machine learning file. Pickle file consumes the value and predicts the chances of Heart disease. In the SVM algorithm the parameters are taken as default but in our paper to improve the model accuracy, these are the parameters need to be tuned.

Three major parameters include Kernel (kernel="rbf"): The function of the kernel is to undertake low dimensional input space and transform that into a higher-dimensional space. It is useful innonlinear separation problem.C(C=10): C is the penalty parameter; it represents the misclassification or error term. The bearable error rate is informed to SVM optimization by misclassification or error term. The trade-off between decision boundary and misclassification term can be controlled. Gamma(gamma=1.0): defines how it is going to influences the calculation of plausible line of separation. Tuning of the hyper-parameters of an estimator is donethis parameter are not directly learnt within the estimators. In scikit-learn, are passed as arguments to the constructor of the estimator class.

Fig.4 describes the Heart Disease Prediction Page, it contains the details of the heart patients such as gender, age, current smoker etc. Fig.5 describes the Result page not or having heart disease, after giving all the details of the patient it gives the patient is not having heart disease with "You are Safe! Stay Healthy!" in figure and it gives the patient having heart disease with accuracy and displays "You have 95.74% chances of Heart Attack. Please contact Doctor Immediately!" in Fig.6

Fig.2 shows Accuracy of the various algorithm

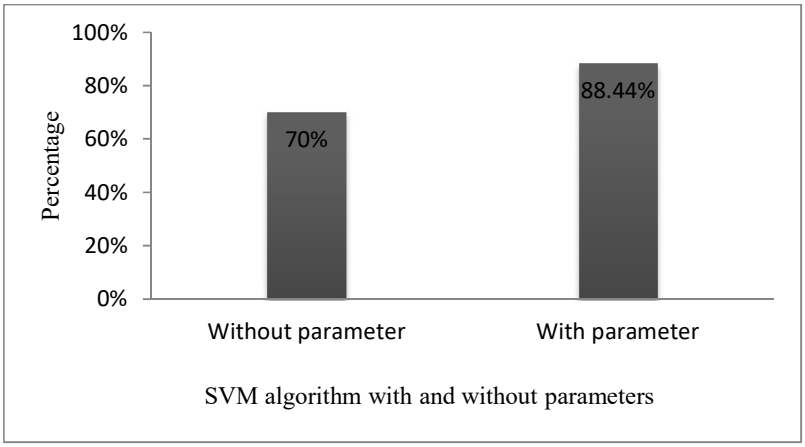

Fig.3 Comparison graph of SVM algorithm between parameters.

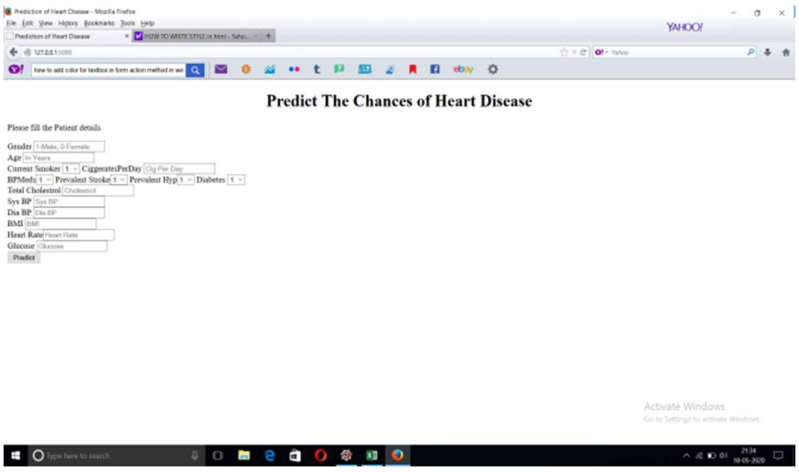

Fig.4 Heart disease prediction page

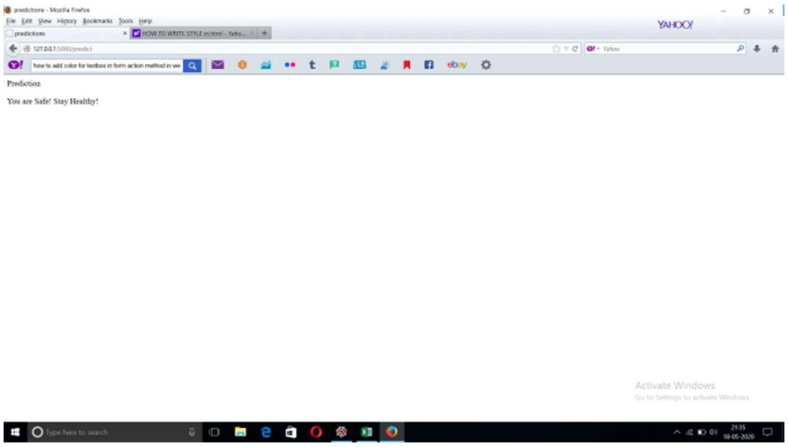

Fig.5 Result page not having heart disease

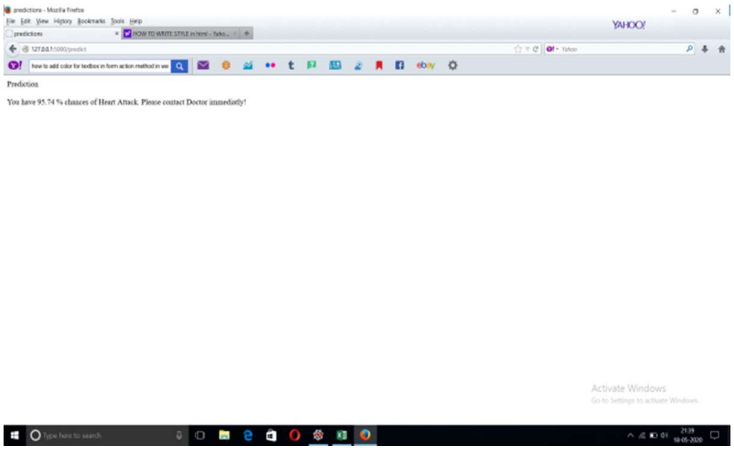

Fig.6 Result page having heart disease

VII. Conclusion

Heart attack is the one of the health problem in human body. Heart disease involve each and every year more people are dying. This paper summarized the methods that is available for prediction of heart disease. To prove the effectiveness of the system we ran experiments with popular algorithms like KNN, Decision tree, SVM, Logistic regression. This experiment is carried out with the accuracy of 88.44% achieved by Support Vector Machine of proposed system. So we can use this Support Vector Machine algorithms to predict whether the patient is having heart related diseases or not.

References

[1] S.Nandhini, Monojit Debnath, Anurag Sharma, Pushkar, "Heart Disease Prediction using Machine Learning", International Journal of Recent Engineering Research and Development (IJRERD), October 2018 , PP. 39-46 .

[2] Reddy Prasad, Pidaparthi Anjali, S.Adil, N.Deepa, "Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning", International Journal of Engineering and Advanced Technology (IJEAT), February 2019.

[3] S. Kavitha, K. R. Baskaran, S. Sathyavathi, "Heart Disease with Risk Prediction using Machine Learning Algorithms", International Journal of Recent Technology and Engineering (IJRTE), November 2018.

[4] Himanshu Sharma, M A Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms", International Journal on Recent and Innovation Trends in Computing and Communication IJRITCC, August 2017.

[5] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE), 2019.

[6] Pahulpreet Singh Kohli, Shriya, "Application of Machine Learning in Disease Prediction", 4th International Conference on Computing Communication and Automation (ICCCA), 2018.

[7] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE), June 2019.

[8] M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar, V. Pavithra, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach", International Journal of Computer Applications, September 2018.