Prediction of Heart Disease using Machine Learning Techniques

[1]S. Durga Devi,
[1] Assistant Professor , Department of CSE, CBIT, Hyderabad, India

Abstract : **It might have happened so many times that you or someone yours need doctors help immediately, but they are not available due to some reason. The Health Prediction system is an end user support and online consultation project. Here I propose a system that allows users to get instant guidance on their health issues through an intelligent health care system online. The system is fed with various symptoms and the disease/illness associated with those symptoms. The system allows user to share their symptoms and issues. It then processes users symptoms to check for various illness that could be associated with it. Here the prediction model of heart disease is implemented using two machine learning techniques one is Linear Regression and another one is Neural Network. Linear Regression is used to determine critical factors that influence on heart diseases and Neural network model is used to predict of heart diseases. If the system is not able to provide suitable results, it informs the user about the type of disease or disorder it feels user's symptoms are associated with. It also consists of doctor address, contacts along with Feedback and administrator dashboard for system operations..**

Index Terms **- Linear Regression, Machine Learning, Neural Network.**

I. INTRODUCTION

Now a days, health disease is increasing day by day due to life style, hereditary. Especially, heart disease has become more common these days, i.e. life of people is at risk. Heart disease is the biggest cause of death now a days. Blood pressure, cholesterol and pulse rate are the major reason for the heart disease. Some non-modifiable factors are also there such as smoking, drinking also reason for heart disease. The heart is an operating system of our human body. If the function of heart is not done properly means, it will affect other human body part also. Some risk factors of heart disease are Family history, High blood pressure, Cholesterol, Age, Poor diet, Smoking. When blood vessels are overstretched, the risk level of the blood vessels is increased. This leads to the blood pressure. Blood pressure is typically measured in terms of systolic and diastolic pressure. Systolic indicates the pressure in the arteries when the heart muscle contracts and diastolic indicates the pressure in the arteries when the heart muscle is in resting state. The level of lipids or fats increased in the blood are causes the heart disease. The lipids are in the arteries hence the arteries become narrow and blood flow is also become slow. Age is the non-modifiable risk factor which is also a reason for heart disease. Each individual has different values for Blood pressure, cholesterol and pulse rate. But according to medically proven results the normal values of Blood pressure is 120/90 and pulse rate is 72.

Here i used machine learning techniques to guess the most accurate illness that could be associated with patient's symptoms. If the system is not able to provide suitable results, it informs the user about the type of disease or disorder it feels user's symptoms are associated with. If user's symptoms do not exactly match any disease in our database, is shows the diseases user could probably have judging by his/her symptoms. It also consists of doctor address, contacts along with Feedback and administrator dashboard for system operations. In doctor module when doctor login to the system doctor can view his patient details and the report of that patient. Doctor can view his personal details.

Admin can add new disease details by specifying the type and symptoms of the disease into the database. Based on the name of the disease and symptom the machine learning algorithm works. Admin can view various disease and symptoms stored in the database. This system will provide proper guidance when the user specifies the symptoms of his illness.

2. Basics and Backgrounds

2.1 Multiple Linear Regressions

Multiple Linear Regression is used to determine critical factors of heart diseases. Simple regression is composed of one dependent variable and one independent variable .Multiple regression analysis is composed of one dependent variable and more than independent variable. It is formulated as follows [1]:

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ...+\beta nXn + \varepsilon$

where:

- y represents the dependent variable
- x1,x2, …,xn are the independent variables
- βi is the regression coefficient
- ε is the random error component.
- β0 is the y intercept

Y represents the dependent variable (degree of influence on heart disease process) and x1, x2… xn are the independent variables (factors that influencing on heart disease).

2.2 Neural network overview

Neural networks are one of the learning algorithms used within machine learning. It is represented of the human brain's information processing mechanism. NN is applied on more applications such as pattern recognition, diagnosis of diseases and data classification, through a learning process. NN has many types of networks such as feed-forward network and back-propagation network. NN is composed of input layers (factors that influence on heart disease), hidden layers and output layers (decision, (YES or NODisease)), as follows:

• **Input Layer** – The input units represent the raw data that is fed into the network.

• **Hidden Layer** – The hidden unit is specified by the input units and the weights on the connections between the input and the hidden units.

• **Output Layer** – The attitude of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

2.3 Heart Diseases

Heart Disease is one of the most serious diseases of the world. There are many factors that lead to Heart Disease, shown below in Table 1.

Table1
Preliminary factors that influence on heart diseases

No	Factors	Description
1	Age	Years
2	Gender	Male/ female
3	Family history	Present/ not Present
4	High Blood pressure	Yes/No
5	Cholesterol	Mg/Dl
6	Diabetes	Yes/No
7	Obesity	Yes/No
8	Smoking	Yes/No
9	Class	Yes/No Heart disease

3. Machine learning algorithm for predicting heart disease

3.1 Diagnose heart disease using Multiple Linear Regression technique

MLR is used to identify critical factors that are influencing on heart diseases. MLR provides squared error (RMSE), the relative absolute error (RAE), the relative squared error (RSE) and the coefficient of determination (CD). MAE is a quantity used to measure how close predictions are to the eventual outcomes. RMSE can be compared between models. whose errors are measured

in the same units. RAE can be compared between models whose errors are measured in the different units. RSE can be compared between models whose errors are measured in the different units.

Algorithm 1. The Multiple Linear Regressions to Determine Critical Factors that Influence on Heart Disease.

1. Input: α←(dependent variable degree of influence on heart disease process),
2 .σ←(independent variables factors that influencing on heart disease)
3.Output: β←(critical factors list of heart disease)
4. LR=Linear Regression
5. CD=Coefficient of Determination
6. FW=Feature Weights
7.Create the LR model based on the number of α and σ
8.Evaluate the LR model.
9. Check the value of CD. CD is calculated as follows:

$$\mathbf{R^2=\frac{SSR}{SST}\ =1-\frac{SSE}{SST}}$$

10. R^2=CD
11. If (CD < 0.5)4
12. Change the explanatory factors;
13.Go to step 11;
14.Else
15. Approve the model
16. End If
17. Check FW for each variable to determine β
18. If (FW < 0.05)
19. Approve β
20. Else
21. Refuse the other factors
22. End If
23. Return β

where:

• *fi*, is the prediction and ϖ the true value.

• ϖ, is the mean of yi.

• Sum of Squares Regression, SSR=Σ*i* (*fi*−ϖ)2

• Sum of Squares Total, SST=Σ*i* (*wi*−ϖ)2

• Sum of Squares Error, SSE=Σ*i* (*wi*−*fi*)

3.2 NN algorithm for predicting Heart disease

MLR gave five factors out of nine factors those are effected on heart disease. The NN uses these five factors as input in a network. The NN gives the output (class, “HD”, “NOHD”). Accuracy level of NN is calculated by using True Positive(TP), False Positive(FP), True Negative (TN), False Negative (FN) from Heart disease data set.

Steps of the NN algorithm is as follows:

Algorithm 2: Neural Network algorithm to predict the Heart Disease

1. Input: α←(Input layer thirteen critical factors that influence on CKD),
2. σ←(hidden layers three hidden layers),
3. W←(Random weights in the neural network model)

4. Output: β←(Decision of the prediction (class,(CKD, NOCKD)))
5. NN=Neural Network
6. V=Accuracy of the NN model
7.TP=True Positive
8. TN=True Negative
9. FP=False Positive
10. FN=False Negative
11. Divide the chronic kidney data into training, test, and validate
12. Build the NN model initially with α, σ, W and β
13 .Train the NN model by using two classes NN
14. Check the value of V. V is calculated as follows:

$$\mathbf{V}= \frac{(TP+TN)}{(TP+TN+FP+FN}$$

15.If (V > 0.9)
16 .Approve NN model
17 .Go to step 19
18 .Else
19. Change W between α and σ
20. Change the number of σ
21. Go to step 8
22. End If
23. Determine V through test and validate the network
24. Return β

---

4. Results

This system is build and implemented using Java and MYSQL. The result shows whether the patient has heart disease or not. The following are the critical factors influencing on the heart disease are shown in Table2 are found in MLR algorithm.

Table2
Critical factors influencing on heart disease

No	Factors	FW
1	Age	-0.0002
2	High Blood pressure	0.01
3	Cholesterol	0.004
4	Diabetes	-0.001
5	Obesity	0.03

4.1 Checking Accuracy

Different types of studies have been done to focus on prediction of heart disease. Various Machine Learning techniques are used for diagnosis and achieved different accuracy level for different models. The goal is to have high accuracy, as well as high precision and recall metrics. These can be easily converted to true-positive (TP) and false-positive (FP) metrics.

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

Table 3
Comparison of accuracy for different models

Model	Precision	Recall	Accuracy
Support Vector Machine	0.78	0.67	74%
Decision Tree	0.88	0.50	85%

MLR+NN	0.95	0.53	95%

4.2 Best Accuracy Model

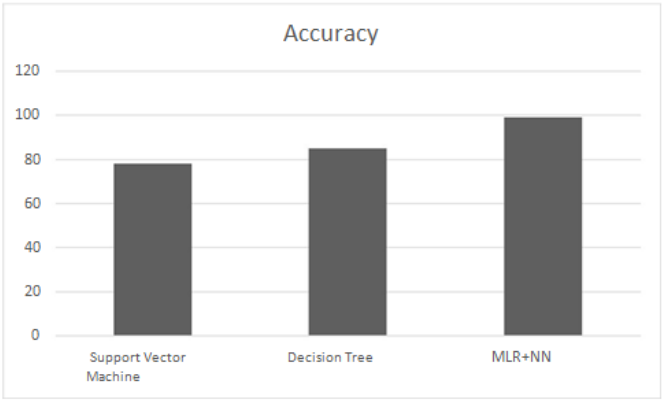

Fig-1 Comparing Accuracy for models

5. Conclusion and Future Work

This project introduces an approach that employs Machine Learning Techniques to identify the disease associated with the given symptoms. After the user provides the symptoms in the front end, in the background MLR and NN algorithms works to predict the associated disease. It is a convenient way of predicting and focuses on issues relating to their feasibility, utility, efficiency and scalability.

User can search for doctor's help at any point of time. User can talk about their Disease and get instant diagnosis. Doctors can get more clients online. It is very useful in case of emergency. It can be enhanced as a computerized system alone does not ensure accuracy, and the warehouse data is only as good as the data entry that created it. The system is not fully automated, it needs data from user for full diagnosis.

References

[1] N.R. Darwish, A.A. Mohamed, A.S. Abdelghany, A hybrid machine learning model for selecting suitable requirements elicitation techniques, IJCSIS 14 (6) (2016)
[2] S. Sharma, Cervical cancer stage prediction using decision tree approach of machine learning, IJARCCE 5 (4) (2016) 345–348.
[3] C.B. Kumar, M.V. Kumar, T. Gayathri, S.R. Kumar, Data analysis and prediction of the hepatitis using Support Vector Machine (SVM), IJCSIT 5 (2) (2014) 2235–2237.
[4] T. Prerana, N. Shivaprakash, N. Swetha, Prediction of heart disease using machine learning algorithms – Naïve Bayes, Introduction to PAC Algorithm, comparison of algorithms and HDPS, IJSE 3 (2015) 90–99.
[5] P. Tintu, R. Paulin, detect breast cancer using fuzzy c means techniques in Wisconsin Prognostic Breast Cancer (WPBC) Data Sets, IJCAT 2 (5) (2013)614–617.
[6] A.M. Hamad, Lung cancer diagnosis by using fuzzy logic, ijcsmc 5(3) (2016) 32-41.
[7] C. Arjun, S. Anto, Diagnosis of Diabetes Using Support Vector Machine and Ensemble Learning Approach, IJEAS 2 (11) (2015) 68–72.
[8] K. Parikh, N. Hawanna, P. Haleema, R. Jayasubalakshm, Virtual machine allocation policy in cloud computing using CloudSim in Java, IJGDC 8 (1) (2015) 145–158.
[9] N.R. Darwish, A.A. Mohamed, B.S.M. Zohdy, Applying swarm optimization techniques to calculate execution time for software modules, IJARAI 5 (3) (2016)12–17.
[10] A. Salekin, J. Stankovic, Detection of chronic kidney disease and selecting important predictive attributes, ICHI 8 (2016) 1–9.
[11] K.R. Padmanaban, G. Parthiban, Applying machine learning techniques for predicting the risk of chronic kidney disease, ijst 9 (29) (2016) 1–5.