INTELLIGENT HEART DISEASE PREDICTION SYSTEM USING KNN

Project report submitted by

AKKAASH P

KISHORE B

MOHAMED AFSAL S

BOOMATHI S

RAMYA P

Under the Guidance of

Mr. Praveen Kumar, assistant professor

MAM College of Engineering and Technology

Department of Computer Science Engineering

MAM College Of Engineering And Technology- 621105

Content

1. INTRODUCTION

2. RELATED WORK

3. PROPOSED WORK

4. RESULT AND DISCUSSION

5. CONCLUSION

Estimation of the Probability of Heart Disease for Patients Using KNN Model of Machine Learning

Akkaash, Mohamed afsal, Kishore, Boomathi, Ramya
Dept. of Computer Science and Engineering
MAM collage of engineering and technology, India

Abstract*— This paper presents a complex polynomial problem on Estimation of the probability of heart disease for patients using KNN model of Machine Learning, which is linked with factors related to human life and depicts one of the major cause of birth and death rates of humans. Though one can access numerous statistical reports and data resources related to heart disease and its risk factors its a complicated task to be performed. In this paper, the data classification is based on supervised machine learning algorithm which results in accuracy, time taken to build the algorithm.

Keywords*— Heart Disease, KNN, Machine Learning.

I. INTRODUCTION

Nowadays, the problem of heart disease is increasing day by day and people are getting more concern about their health. Heart disease depicts the range of conditions that affect the heart. Range of risk factors or symptoms that are classified under the heart disease are age, sex, smoking, high blood pressure, diabetes, obesity, stress and also due to some inappropriate lifestyle are leading towards this disease [1]. The case study is under concern to diagnosis tool that can extract the relevant information from the recorded signals for classification according to the method used by and is equipped with the graphical user interface for ease of use [2].

The term “heart disease” is often used with the word “cardiovascular disease” which depicts the condition related to blocked blood vessels that can cause heart attack [3]. Other factors that affect the heart’s muscles are also considered as heart disease.

The symptoms of heart disease depends on what kind of depends on the type of heart disease. They are different for both men and women. In particular, men are more likely to be affected by chest pain whereas, for women, there are other symptoms along with those for men like nausea and extreme fatigue. In brief, symptoms are chest pain, chest pressure, shortness of breath, numbness, weakness or coldness in your legs or arms if the blood vessels are blocked in those parts, pain in neck, jaw, throat, upper abdomen or back [4]. These diseases can sometimes be found early with regular evaluations in consideration with proper discussion with doctors. Heart may beat too fast or slow or with some irregularity is known as a heart arrhythmia which is an abnormal heartbeat.Symptoms for this includes fluttering in chest, racing heartbeat (tachycardia), slow heartbeat (bradycardia), chest discomfort, shortness of breath, lightheadedness, dizziness, syncope (fainting).

The defects in heart from birth are serious congenital heart defects, usually diagnosed soon after birth. Symptoms for this are cyanosis infection (affects inner membrane which separates chambers an(pale Gray or blue skin colour), swelling in legs, abdomen or areas around the eyes. Less serious congenital heart defects did not become evident until later in childhood or adulthood. Symptoms for this include getting short of breath or tiring during exercise or activity, swelling in hands, ankles or feet. Cardiomyopathy is a heart disease caused by weak heart muscle. Include symptoms are swelling of limbs, fatigue, irregular heartbeats, dizziness, fainting. Endocarditis, and valves of the heart) includes symptoms such as fever, weakness, difficulty in breathing, swelling in legs and abdomen, irregular heartbeats, cough, skin rashes or spots. Symptoms for valvular heart disease which damaged the four valves of the heart are fatigue, uneasiness in breathing, irregular heartbeat, swollen limbs, chest pain, fainting [6]. Heart disease is easier to treat when detected early, so one must take care of proper discussion with doctors for the concern of health. Many forms of heart disease can be prevented or treated with healthy lifestyle choices. The most commonly used, popular and easy to understand classifier is k-nearest neighbour. KNN algorithm categorized the data using similarities among them such as the distance function by identifying the nearest neighbours. Using the Euclidean distance formula:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

II. RELATED WORK

Several significant research papers have made use of artificial intelligence, neural networks and machine learning algorithms like logistic regression, KNN, random forest classifier etc. These are explained below. In [7] has developed a Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modelling technique, namely Naive Bayes using features such as age, sex, blood sugar, blood pressure etc. To predict the likelihood of the person getting a heart disease or

not. In [8], machine learning algorithms for improving sensitivity and specificity in predicting Ischaemic heart disease has been proposed. The algorithm takes in to account the misclassification cost. They achieve the sensitivity of 89% and the specificity of 55%
(Bayesian classifier) and specificity of 63% sensitivity of 89% (neural network). In [9] provides an overview of the development of intelligent data analysis from machine learning algorithm perspective in medicine. It emphasizes on Naive Bayesian classifier, neural networks and decision tree classifiers in historical overview. It provides a comparison between some state-of-art systems representative from each branch of machine learning when applied to several medical diagnostic tasks.
The future trends are illustrated by two case studies first recently developed methods second describes an approach to using machine learning to verify some unexplained phenomena. In [10] they used three popular data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) extracted from a decision tree or rule-based classifier to develop the prediction models using a large dataset.
They also used 10-fold cross-validation methods to measure the unbiased estimate. Accuracy scores were 83.49%(CART), 72.93%(ID3) and 82.50%(DT). In [11] used five machine learning algorithms, namely Support Vector Machines (SVM), AdaBoost using trees as the weak learner, Logistic Regression, a naive Bayes event classifier, and a variation of a Likelihood Ratio Test adapted to the specific problem. Their results show that with under 30% false alarm rate, the detection rate could be as high as 82%. After reviewing the papers we tried to design a machine learning algorithm using KNN classifiers to improve accuracy in predicting heart disease probability and reducing the errors.

III. Proposed Work

In this paper, we propose a method for estimation of the probability of heart disease using a k-nearest neighbour model of machine learning to improve the accuracy of heart disease data set. We used the statistical data set for heart disease from UC Irvine Machine Learning Repository dated from 1988 and comprises of four databases: Cleveland (303 instances), Hungary (294), Switzerland (123), and Long Beach VA (200). 76 attributes were provided by each database with the inclusion of predicted attributes. Later on after getting corrupted the Cleveland dataset consists of only 14 attributes per instance. Total instances for these 14 attributes, for all four databases are 299 out of which 297 from Cleveland alone. For simplicity referring the dataset to Cleveland dataset.
Description of 14 attributes is given below : The last row ("num") is the attribute to be predicted Attributes of the heart disease patients in the Cleveland data set; the last row (heart disease status) is the response variable; the Type column indicates whether an attribute is binary (bin), integer (int), categorical (cat), or continuous (con) to gain some insight into the power of these attributes to discriminate between disease (num>0) and no disease (num=0), let's look at their distributions over the corresponding subsamples (139 patients with disease, 160 without).
We ranked them accordingly by finding their importance using extra tree classifier inside this feature. Highly ranked datasets among them are used under this case study and less important factors were excluded which results from the increment in the accuracy.
The results give an idea on the inclusion of highly ranked features in machine learning model, but under consideration that they do not take association into account.
List of variables available inside feature are:

1. age: continuous
2. sex: categorical, 2 values {0: female, 1: male}
3. cp (chest pain type): categorical, 4 values {1: typical angina, 2: atypical angina, 3: non-angina, 4: asymptomatic angina}
4. restbp (resting blood pressure on admission to hospital): continuous (mmhg)
5. chol (serum cholesterol level): continuous (mg/dl) 6. fbs (fasting blood sugar): categorical, 2 values {0: <= 120 mg/dl, 1: > 120 mg/dl}
7. restecg (resting electrocardiography): categorical, 3 values {0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy}
8. thalach (maximum heart rate achieved): continuous 9. exang (exercise-induced angina): categorical, 2 values {0: no, 1: yes}
10. oldpeak (ST depression induced by exercise relative to rest): continuous
11. slope (slope of peak exercise ST segment): categorical, 3 values {1: upsloping, 2: flat, 3: downsloping}
12. ca (number of major vessels coloured by fluoroscopy): discrete (0,1,2,3)
13. thal: categorical, 3 values {3: normal, 6: fixed defect, 7: reversible defect}
14. num (diagnosis of heart disease): categorical, 5 values {0: less than 50% narrowing in any major vessel, 1-4: more than 50% narrowing in 1-4 vessels}

The actual number of feature variables (after converting categorical variables to dummy ones) is:
1 (age) + 1 (sex) + 3 (cp) + 1 (restbp) + 1 (chol) + 1 (fbs) + 2 (restecg) +
1 (thalach) + 1 (exang) + 1 (oldpeak) + 2 (slope) + 1 (ca) + 2 (thal) = 18
The response variable (num) is categorical with 5 values, but we don't have enough data to predict all the categories. Therefore we'll replace num with:

14. hd (heart disease): categorical, 2 values {0: no, 1: yes} As our dataset contained some categorical variables which didn't provide enough information so we convert categorical variables with more than two values into dummy variables. Note that variable ca is discrete but not categorical, so we don't convert it.Therefore, we converted cp into cp_1,cp_2 ,cp_3 and restecg into recg_1,recg_2 and slope into slope_1, slope_2 and thal into thal_6 , thal_7.

Features	Importance
age	0.0702
sex	0.0523
restbp	0.0548
chol	0.0609
fbs	0.0168
thalach	0.0897
exang	0.0894
ca	0.0692
cp_1	0.1466
cp_2	0.0182
cp_3	0.0183
recg_1	0.0532
recg_2	0.0283
slope_1	0.0739
slope_3	0.0067
thal_6	0.0164
thal_7	0.1343

Table-1*:* Importance of features using ExtraTree Classifier**.** We selected the following features based on there importance.

sex,restbp, thalach,ca,cp_1,cp_2,cp_3,slope_1,slope_1 and thal_7.We didn't select the age coefficient because it has negative coefficient with ca and thalach which decreases the model accuracy. Standardize our features variable to the same scale. divided the data set into training(79%) and testing(21%) at random state of 29. For our model after testing with different values of K, K=6 provided the best accuracy score and minimum error. We also used 10-fold cross-validation to make sure that there is no overfitting or underfitting in our model. Figure-1 represents the variation of accuracy score with different values of K.Figure-2 represents the variation of accuracy score with 10-fold cross-validation against different values of K.Figure-3 represents a comparison between our model and previous model[12].

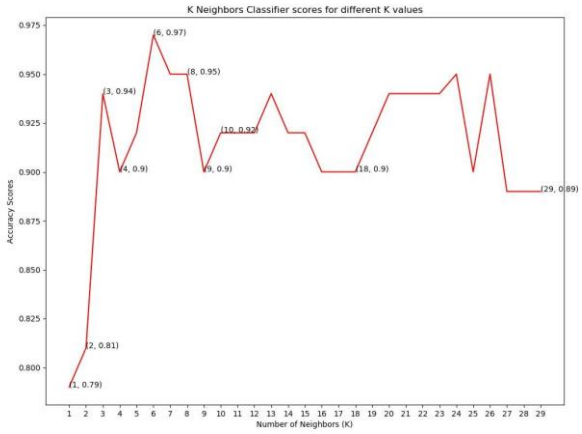

Figure-1: Graph of accuracy score v/s K value

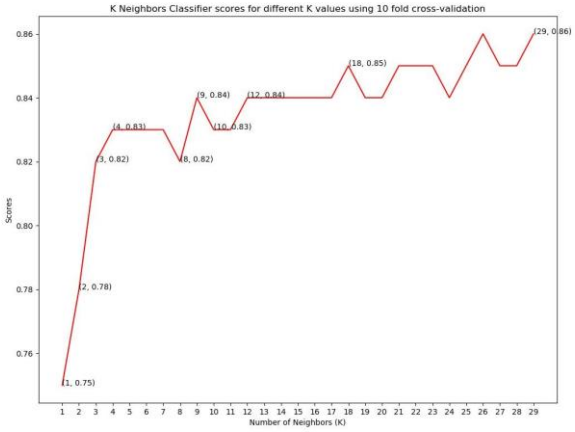

Figure-2: Graph of K Neighbours classifier scores for different K values using 10 fold cross-validation.

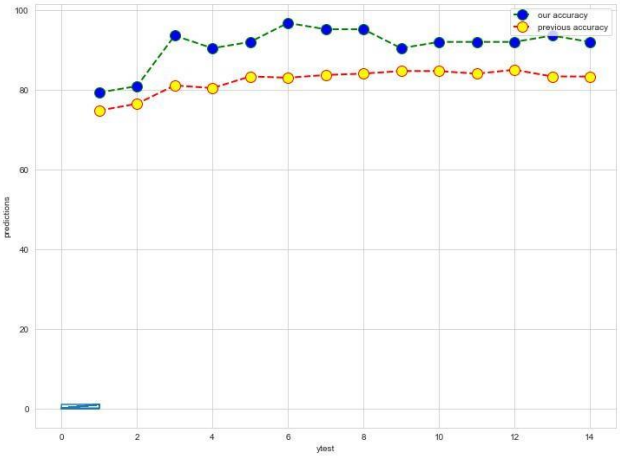

Figure-3: Graph of comparison of previous accuracy and our accuracy

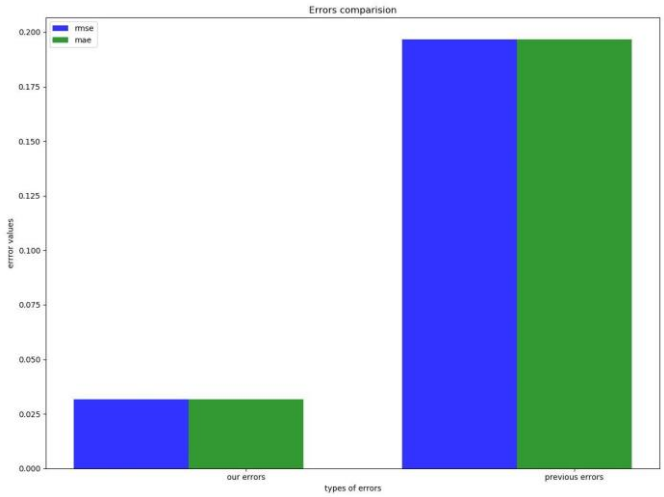

Figure-4: Graph of comparison between the errors between the previous approach and our approach.

IV. Result and Discussion

Our model can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It can also provide decision support to assist doctors to make better clinical decisions or at least provide a "second opinion." currently our model is based on the 9 attributes listed in Table1. This list may need to be expanded to provide a more comprehensive diagnosis system. For some diagnosis, the use of continuous data may be necessary. Figure 3 and figure 4 are the comparison graphs of accuracy scores and error values between previous models and our model .

Another limitation is that it only uses the KNN model of machine learning. Additional data mining techniques can be incorporated to provide a better diagnosis. The size of the dataset used in this research is still quite small. A large dataset would give better results. It is also necessary to test the system extensively with input from doctors, especially cardiologists before it can be deployed in hospitals. Also more updated data set could be used to make the model updated. Because with the period of time many new factors add up to cause heart diseases.

V. Conclusions

In this paper, we develop a heart disease prediction system that can assist medical professionals in evaluating a patient's heart disease based on the clinical data of the patient. First, we converted the categorical data(cp, restecg, slope, thal) into discrete variables to extract more information from these features which increased our model accuracy. In our approach we select 9 important clinical features, i.e. sex, restbp, thalach, ca, cp_1, cp_2, cp_3,slope_1, thal_7. using Extra-tree classifier. Next, we trained our model for these features using the KNN algorithm of machine learning. For training and testing, we used test size = 0.21 and random state = 29 and K=6 which provides the best accuracy score and minimizing the errors, RMSE(Root Mean Square Error) = 0.0317460 and MAE(Mean Absolute Error) = 0.0317460.

We developed a heart disease prediction model which provides the accuracy score of nearly 97%. After this, we used 10 fold cross-validation to avoid any overfitting and underfitting in the model. Which provided the accuracy of 86% for k = 29. The model used in this project will be a good approach in predicting whether the person is having heart disease or not.