Heart disease prediction using data mining

R Fadnavis¹, K Dhore², D Gupta³, J Waghmare⁴ and D Kosankar⁵

¹Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India
²,³,⁴,⁵Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India

E-mail: rafatycce@gmail.com

Abstract. There is huge amount of information accessible within the healthcare systems. But there do not exist enough analysis tools to mine uncovered, unusual but useful patterns in data. Data mining has been used successfully in various fields to discover hidden patterns and trends, alerting about the hidden anomalies in the data or simply helping in the decision making process. This paper how classification techniques in data mining can be applied for heart disease prediction. To predict and alert about any future coronary ailment in the patients techniques like Naïve Bayes, and Decision tree are applied and efficiency of these algorithms is compared. The dataset taken is Cleveland dataset with 14 attributes.

Keywords: Data mining, Naïve Bayes, Decision tree, health care systems, heart disease

1. Introduction

Heart diseases are one of the most prominent causes of death all around the world. It brings down the profitability of an individual and furthermore builds the costs on medical care. India has suffered huge financial loss due to heart related or cardiovascular diseases in last decade. Thus, feasible and accurate prediction of heart related diseases is very important.

In recent years, awareness to make use of the advancement of data mining technologies in healthcare systems is growing. Data mining is a process of discovering patterns and trends in large datasets using statistical methods, machine learning, and database systems. Knowledge extracted by data mining process is frequently utilized for different applications, for example healthcare services industry. This paper proposes automated technique for diagnosing heart infections dependent on earlier information. like symptoms, manifestations and body conditions of the patient. Different parameters that indicates the risk of heart illness, are smoking tendency, cholesterol percentage, any prior family history of heart diseases, lack of physical exercise, heftiness, and hypertension. In recent years, healthcare industry generates great deal of knowledge about patients, their disease, treatments and so on. But the major issue faced by Healthcare industry is providing good quality of service. Quality of service refers to diagnosis of illness accurately and giving powerful medicines to patients. Poor diagnosis can prompt awful results which are unsatisfactory.

1.1. Overview of the Disease

The heart is critical organ of human body. It siphons blood to all aspects of human life systems. On the off chance that it neglects to work accurately, at that point the cerebrum and different organs will quit working and inside couple of moments the individual may pass on. Life is totally hooked in to
efficient working of the guts. Heart disease refers to disease of blood vessels systems within heart. Various parameters that show the risk of heart diseases are:

1. Smoking tendency
2. Poor diet
3. Family history of heart illness
4. High blood pressure
5. High cholesterol
6. Obesity
7. Lack of physical exercise
8. Hypertension

Diagnosis is a complex task which requires expertise, High skills. Mostly diagnosis in general is based on patients current test results and doctors experience. Some of common heart disease includes cardiac arrest, Stroke, heart failure due to congestion, disease related to Coronary artery and congenital heart disease. The symptoms to predict a heart disease depends upon the types of heart disease mentioned above. Each type will have its own symptoms. For example, Chest pain is one of the symptoms for coronary artery, but all people would not have the same symptom. Some may have other symptom like indigestion or acidity. The doctor confirms the heart disease with the diagnosed report of the patient and various other parameters. Most common types of heart diseases, are listed in Table I:

| Heart Disease Types              | Description                                                                 |
|----------------------------------|-----------------------------------------------------------------------------|
| Coronary heart disease           | Block in the coronary blood vessels, which reduces blood and oxygen supply to heart. |
| Angina pectoris                  | Chest pain will occur due to the insufficient supply of blood to heart.     |
| Congestive heart failure         | Heart is not able to pump enough blood.                                     |
| Cardiomyopathy                   | Weakening or a change in the heart muscle.                                  |
| Congenital heart disease         | Defect in the structure of the heart or great vessel present at time of birth |
| Arrhythmias                      | Problem with the rate or rhythm of the heartbeat                           |
| Myocarditis                      | Inflammation of heart muscle, can affect heart muscle or heart’s electrical system |

1.2. Data source
The publicly available Cleveland[21] and Statlog[22] heart disease databases from UCI Repository are used. It consists of 303 records and 270 records respectively. The data set contains 14 different types of attributes categorised into 2 types as:

- Input attributes(Nonclass attributes)- 13 input attributes
- Predictable attribute (class attribute) as num which signifies whether patient has heart disease or not

2. Literature Survey
Different Data mining approaches has been applied in literature for prediction of heart disease. Following section provides the survey of these methods:
In one of the paper, the author provides an in-depth survey of all aspects of heart malfunctioning management. It discusses about various machine learning methods for the assessment of heart failure. It predicts about the presence of untimely events such as death or rehospitalizations.

The paper titled “Analysis of Data Mining Techniques for Heart Disease Prediction”, discusses about different classification techniques applied to find out a more accurate technique for the heart disease prediction. Authors have applied classification methods such as BayesNet, SMA, Kstar, multilayer perceptron and decision tree based J48 algorithm and they found that the performance of BayesNet and SMA classifiers to be optimum.

The paper titled “Predicting the likelihood of heart failure with a multi level risk assessment using decision tree”, discusses about risk factors of developing heart failure at different levels. They have identified some important risk factors and used a decision tree classifier C4.5 for predicting risk at different levels. Accuracy they have received was found to be nearly 87 percent.

Different data mining as well as machine learning methods has been suggested by different authors for predicting various diseases like diabetes, tumor, cancer, heart problem, etc. one of the paper, discusses how authors have used different classification methods like support vector machine, Bagging approach, decision tree based classifiers like Random forest and predicted about existence or non existence of heart disease.

Some researchers have applied ensemble algorithms such as bagging, boosting stacking and majority voting. Different methods provided different accuracies. A comparative result indicated that majority voting outperformed with maximum accuracy. To improve the performance of ensemble methods, feature selection techniques also were taken into account.

In the paper titled “Efficient Heart Disease Prediction System”, A rule based classification model was used. The rules were divided into different categories and the priority was associated with each of these types. 10 fold cross validation was used. Accuracy in testing and training phase were computed and compared.

3. Algorithms and methodologies used

Following section discusses algorithms used for predicting heart disease:

3.1 Naïve Bayes

Naïve Bayes classifier is called a statistical classifier since it performs probabilistic classification. Predicts class membership probabilities. Naïve bayes is useful for very large sets and easy to build. Bayes theorem finds a posteriori probability from likelihood and prior probability. Classification is to derive the maximum posteriori, i.e., the maximal P(C|X). This can be derived from Bayes’ theorem:

\[
P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}
\]

Informally, this can be viewed as

posteriori = likelihood x prior / evidence, so

P(C|x) is the posterior probability of class
P(C) is the prior probability of class.
P(X|C_i) is the likelihood
P(X) is the prior probability
Since P(X) is constant for all classes, only

\[
P(C_i|X) = P(X|C_i)P(C_i)
\]
3.2. Decision Tree

Decision trees are supervised learning method used for classification problems. It predicts the categorical class label. It classifies data based on training set and values in a classifying attribute and uses it in classifying new data. Tree is constructed in top down recursive and divide and conquer manner. In tree representation leaf nodes represents class labels and non leaf nodes represents features and branches represents conjunction of features that lead to classification. Some of the decision tree algorithms are ID3, C4.5, CART, CHAID.

4. Implementation results and discussion

In this work, two supervised learning algorithm Naïve Bayes and Decision tree has been used to predict existence of heart illness based on person's medical data. Since the decision consist of two classes only, so it is binary classification problem. The Cleveland heart disease data set that has been used here consist of 14 attributes. Details of the attributes are as follows:

Attributes\(^{[21]}\)

Input attributes for the system will be as follows:

1. Age (in Year) – continuous variable
2. sex {0:Female, 1:Male} - categorical variable
3. cp (Chest Pain Type) categorical variable - It has values 1 to 4 each value indicating different type of chest pain: value 1: typical angina value2: atypical angina value3: non angina pain value 4: asymptomatic
4. restbp (Resting Blood Pressure) (mmHg) - continuous variable
5. chol (Serum Cholesterol) - continuous variable
6. fbs (Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) - categorical variable
7. restecg - resting electrocardiographic results
   - Value0: normal
   - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05mV)
   - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalch (Maximum Heart Rate)
9. exang (Exercise Induced Angina){0:False, 1:True} - categorical variable
10. slope (Slope of Peak Exercise){1:Up, 2:Flat, 3:Down} - discrete variable
11. ca (Number of Major Vessels){0,1,2,3} - categorical variable
12. thal (3=Normal, 6=Fixed defect, 7=Reversible defect)
13. oldpeak (ST depression) - continuous variable
14. num (Diagnosis of Heart Disease){0,1} - categorical variable

In above attribute list, first 13 are nonclass attributes and last attribute num is a class attribute. Num attribute is a binary and can have value 0 indicating no heart disease and 1 indicates existence of heart disease.

Before applying classification model on dataset, it needs to be preprocessed. So cleaning and filtering are applied on these records in order to remove the irrelevant data from the dataset. The test-train data split was done using sklearn package with test size set as 0.2 (i.e., 20% of the total dataset). The model is then fitted to the newly formed dataset to predict accuracy, thereafter the algorithms namely Decision Tree and Naïve Bayes are applied. The Naïve Bayes algorithm is applied using Gaussian Distribution which uses some of the most naïve statistical functions such as mean and standard deviation to make predictions. Decision tree classification model was fitted and accuracy was found to be 81.97% shown if figure1.
5. MODEL FITTING

```python
In [44]: from sklearn.metrics import accuracy_score

Decision Tree Classifier

In [45]: from sklearn.tree import DecisionTreeClassifier
   ...: max_accuracy = 0
   ...: for k in range(100):
   ...:     dt = DecisionTreeClassifier(random_state=k)
   ...:     dt.fit(X_train, Y_train)
   ...:     y_pred_dt = dt.predict(X_test)
   ...:     current_accuracy = round(accuracy_score(Y_pred_yt, y_pred_dt),2)
   ...:     if current_accuracy > max_accuracy:
   ...:         max_accuracy = current_accuracy
   ...:         best_k = k
   ...: print(f"The accuracy score achieved using Decision Tree is: \"{max_accuracy}\" %")

The accuracy score achieved using Decision Tree is: 85.87 %
```

**Figure 1.** Decision tree classifier results.

With Naïve Bayes classifier, accuracy was found to be 85.25% as shown in figure 2

```python
In [46]: from sklearn.naive_bayes import GaussianNB

Naïve Bayes Classifier

In [47]: nb = GaussianNB()
   ...: nb.fit(X_train, Y_train)
   ...: Y_pred_nb = nb.predict(X_test)

In [48]: Y_pred_nb.shape

Out[48]: (61,)

In [49]: score_nb = round(accuracy_score(Y_pred_nb, y_pred_nb),2)
   ...: print(f"The accuracy score achieved using Naïve Bayes is: \"{score_nb}\" %")

The accuracy score achieved using Naïve Bayes is: 85.25 %
```

**Figure 2.** Naïve bayes classifier results.

So from above prediction it can be seen that accuracy of Naïve Bayes is more than that of Decision Tree.

5. Conclusion
Various classification techniques have been used in literature for prediction of heart disease. A significant accuracy has been achieved for such models. This paper discusses implementation of two classification methods; Naïve Bayes and decision tree applied on Cleverand and Statlog datasets with results showing higher accuracy using Naïve Bayes algorithm. Each classification algorithm has its own advantages and disadvantages, so a combination of such models may be used to achieve higher
accuracy and to achieve scalability, such models may be applied on large data sets and can be validated.

6. References

[1] Tripoliti E, Papadopoulos T, Georgia Karanasiou S, Fotiadis D. 2017, Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques Computational and Structural Biotechnology Journal, Volume 15, Pages 26-47

[2] Sultana M, Haider A and ShorifiUddin M. 2016 “Analysis of Data Mining Techniques for Heart Disease Prediction”, 978-1-5090-2906-8/16/$31.00 IEEE

[3] Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. ESC 2015 guide- lines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J2016; (ehw128). http://dx.doi.org/10.1093/eurheartj/ehw128

[4] Aljaaf AJ, Al-Jumaily D, Hussain AJ, Dawson T, Fergus P, Al-Jumaily M. 2015 Predicting the likelihood of heart failure with a multi level risk assessment using decision tree. Third international conference on technological advances in electrical, Beirut, Lebanon;

[5] Austin PC, Tu JV, Ho JE, Levy D, Lee DS. 2013 Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J ClinEpidemiology;66:398–407. http://dx.doi.org/10.1016/j.jclinepi.2012.11.008.

[6] Isler Y. 2016 Discrimination of systolic and diastolic dysfunctions using multi-layer perceptron in heart rate variability analysis. Comput Biol Med [accepte dpaper]. http://dx.doi.org/10.1016/j.compbiomed.2016.06.029

[7] Beulah Christalin Latha, CarolinJeeva, 2019 Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques Informatics in Medicine Unlocked, Volume 16, 100203

[8] Masetic Z, Subasi A. 2016 Congestive heart failure detection using random forest classifier. Comput Methods Programs Biomed;130:54-64.http://dx.doi.org/10.1016/j.cmpb.2016.03.020.

[9] Sharma P, Saxena K, Sharma R 2016, Efficient Heart Disease Prediction System Procedia Computer Science, Volume 85, Pages 962-969

[10] Sharma P, Saxena K, Sharma R May 2015. Efficient Heart Disease Prediction System using Decision Tree” in IEEE International Conference on Computing Communication and Automation (ICCCA-2015).

[11] Sharma P, Saxena K, Sharma R 2015, Heart Disease Prediction System Evaluation Using C4.5 Rules and Partial Tree” in Springer, Computational Intelligence in Data Mining, pp. 285-294, DOI 101007/978-81-322-2731-1_26.

[12] Chhikara, S & Sharma, P 2014 Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases, I JRASET, PP 396–402.

[13] Selappan P, Awang R 2008, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, 978-1-4244-1968-5/08/$25.00 ©IEEE, 2008

[14] Soman K.P., ShyamDiwakar M. and Madhavdas P., “Efficient classification and analysis of Ischemic Heart Disease using Proximal Support Vector Machines based Decision Trees”, 0-7803-7651-x/03/$17. ©2003 IEEE, 2003

[15] Rajesh, T Maneesha, Shaik Hafeez, Hari Krishna, (2018)" Prediction of Heart Disease Using Machine Learning Algorithms", International Journal of Engineering & Technology, 7 (2.32) 363-366

[16] Chandra D, 2014 “Diagnosis of Heart Disease Using Data Mining”, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 1678-1680

[17] Danagere C, Apte S, June 2012 “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 – 888)

[18] Yang G, Ren Y, Pan Q, Ning G, Gong S, Cai G, et al. :2010 A heart failure diagnosis model based on support vector machine. 2010 3rd international conference on biomedical engineering and informatics (BMEI), vol.3.p.11058.http://dx.doi.org/10.1109/BMEI.2010.563961

[19] Vanisree K, JyothiSingaraju, April 2011 Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. Int J Comput Appl;19(6). (0975 8887)

[20] Yanwei X, Wang J, Yonghong G Zhihong Z. 2007 Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease. Convergence Information Technology.p. 868–72.

[21] Cleveland heart disease dataset : https://archive.ics.uci.edu/ml/datasets/heart+dataset
[22] Statlog dataset:https://archive.ics.uci.edu/ml/datasets/statlog+(heart)

[23] Data Mining : Concepts and Techniques ©2011 Jiawei Han, Micheline Kamber, and Jian Pei University of IllinoisatUrbana-Champaign&Simon Fraser University