# Comparative Analysis of Heart Disease Classification Algorithms Using Big Data Analytical Tool

Sinkon Nayak(✉), Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray

KIIT Deemed University, Bhubaneswar, India
sinkonnayak07@gmail.com, mkgourisaria2010@gmail.com,
{manjushafcs, siddharthfcs}@kiit.ac.in

**Abstract.** Immense volume of data has been generated from unlike sources like health care, social media, business applications, manufacturing industries and many more. HealthCare plays a pivotal role in Big Data. Spotting and safeguarding of the diseases at a primitive stage are very much crucial. Heart disease specifically implies the condition of the heart that contracts or obstructs blood vessels which result in heart attack, chest pain or stroke. This paper emphasizes on the diagnosis of heart diseases at a primitive stage so that it will lead to a successful cure of the diseases. In this paper, diverse data mining classification method like Decision tree classification, Naive Bayes classification, Support Vector Machine classification, and k-NN classification are used for identification and precaution of the diseases at an early stage so that it can be curable and preventable.

**Keywords:** Big data · Healthcare · Heart disease · Decision tree classification · Svm classification · Naive bayes classification · K-NN classification

## 1 Introduction

HealthCare is the maintenance and betterment of the health by the help of diagnosis, prevention, and treatment of any kind of diseases. But a major challenge faced is to provide better care and clinical services at an affordable cost. By the help of various predictive analysis, the cost will get diminish and can get better clinical care. Cardiovascular disease refers to the trouble occur with heart. It specifically implies the condition of the heart that contracts or obstructs blood vessels which result in heart attack, chest pain or stroke. So various data mining classification techniques such as Decision tree classifier, Naive Bayes classifier, Support Vector Machine classifier, and k-NN classifier are used to spot and prevent the diseases at an primitive stage.

This paper is organized into section as follows. Section 2 summarizes heart disease. Section 3 provide a brief description of literature survey of heart related disease. The work flow steps are discussed in Sect. 4. Section 5 is all about the concise discussion of the classification techniques such as Naive Bayes, Decision tree, SVM, k-NN. Dataset collection attributes elucidation, comparison study is discussed in Sect. 6.

S. Smys et al. (Eds.): ICCNCT 2019, LNDECT 44, pp. 582–588, 2020.
https://doi.org/10.1007/978-3-030-37051-0_65

Section 7 is all of the result analysis. Section 8 is the conclusion, summarizes a brief overview of the content.

## 2 Heart Disease

Any abnormality in heart results to heart disease. Heart disease affects the structure and function of the heart. There are various types of abnormality observed in the heart such as narrowing arteries, heart attack, aberrant rhythms of the heart, crushing of heart, disease related to a heart valve and heart muscle etc. The abnormal function of the heart is because of various factors such as blood sugar level, cholesterol level, blood pressure, etc. From the various study, the death rate of Cardio Vascular Diseases is 272 people per 100 000 population in India and globally it is 235 per 100 000 population. 610,000 number of people deceased because of heart-related problems in the United States every year.

## 3 Literature Survey

Aditya Sundar et al. describes classification techniques for prediction and evaluates the performance of Naive Bayes classification technique and WAC (Weighted Association Classifier) by using different performance measure [1]. Sellappan Palaniappan et al. describe various data mining classification algorithm Naive Bayes, Decision tree and Neural network to predict heart disease [2]. Dangare et al. in their paper describe early anticipation of heart related illness by the help of Neural network, Decision tree and Naive Bayes and determine their accuracy [3]. J Thomas et al. in their paper describes the classification techniques k-NN classification, Naive Bayes Classification, Decision tree classifier and Neural network method to predict the danger level of a diligent to have a heart-related illness or not [4]. Swathy Wilson et al. in their paper conclude that decision tree with k means clustering yield improved accuracy as compared to others [5]. A Nishara Banu et al. did the study of Association Rule Mining, Classification, and Clustering for spotting heart-related disease. They showed that the designed spotting structure is able to spot the heart attack efficiently [6]. Shabana Asmi et al. add some attributes for spotting the heart-related unwellness which results in high accuracy by the help of association rules [7]. Beant Kaur et al. used various Genetic and data mining algorithm for the spotting of heart-related illness. Their result shows that Genetic Algorithm gives an accuracy of 73.46% [8]. Sashikant Ghumbre et al. in their study used SVM classifier and Radial basis function network for heart disease diagnosis and got the result that SVM is best for identification [9].

## 4 Work Flow Design

Figure 1 describes the workflow and methodology for the prediction of heart disease. We have taken the dataset from UCI/Kaggle in CSV format then preprocess the data has been done which includes data transformation, data cleaning, and data integration.

After preprocessing data mining classification algorithms such as Decision Tree, SVM, Naive Bayes, k-NN are applied for the prediction and comparison of the classification techniques based on their performance.

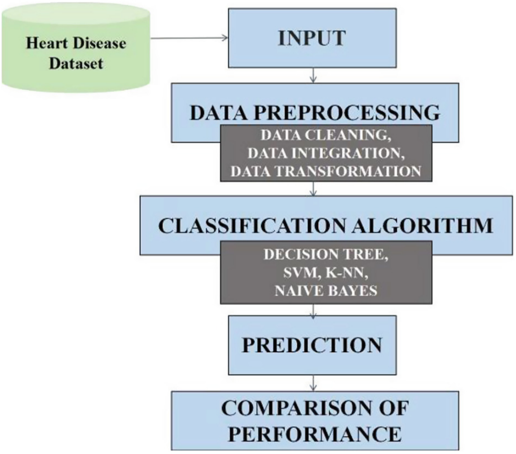

Heart Disease Dataset
INPUT
DATA PREPROCESSING
DATA CLEANING, DATA INTEGRATION, DATA TRANSFORMATION
CLASSIFICATION ALGORITHM
DECISION TREE, SVM, K-NN, NAIVE BAYES
PREDICTION
COMPARISON OF PERFORMANCE

**Fig. 1.** Work flow for prediction of heart disease

## 5 Classification Algorithm

Classification belongs to a supervised learning method, which predicts a class for each object and assigns them to a target class [12]. The main goal is to prognosticate the target class for each data in a data set accurately.

### 5.1 Decision Tree

A decision tree is basically a tree-like structure in which branch nodes denotes attribute, terminal nodes denote class labels and branches denotes the outcome. Testing criteria are applied on the source node and branch nodes and on the basis of testing criteria result the data will follow the branch till it reaches the leaf node or class label (Table 1).

**Table 1.** Pros and cons of decision tree classification techniques

| Pros | Cons |
|---|---|
| Fast, simple to understand and interpret | Computationally expensive to train |
| Robust and need less computation for classification | Overfitting and prediction of continuous variable is not suitable |

### 5.2 Naive Bayes Classifier

Bayesian classification is based on Bayes theorem. Thus it is a classifier based on probabilities it fails in case of continuous attribute because of frequency count is not possible (Table 2).

**Table 2.** Pros and cons of Naive Bayes classification techniques

| Pros | Cons |
|---|---|
| Simple and Easy to implement and robust | Class conditional independence thus less accuracy |
| Fast to train and classify as it need a single scan and space efficient also | Dependencies among attributes can not be taken into consideration |

### 5.3 Svm

A Support Vector Machine (SVM) is based on decision planes on decision margins. A decision plane can be defined as which split up between a bunch of objects, belongs to, unlike class. SVM is used mutually for classification as well as regression analysis [9, 10] (Table 3).

**Table 3.** Pros and cons of SVM classification techniques

| Pros | Cons |
|---|---|
| Training of dataset is easy | Need good kernel function |
| Scale well for high dimensional data | Sensitive to noisy data |

### 5.4 K-NN

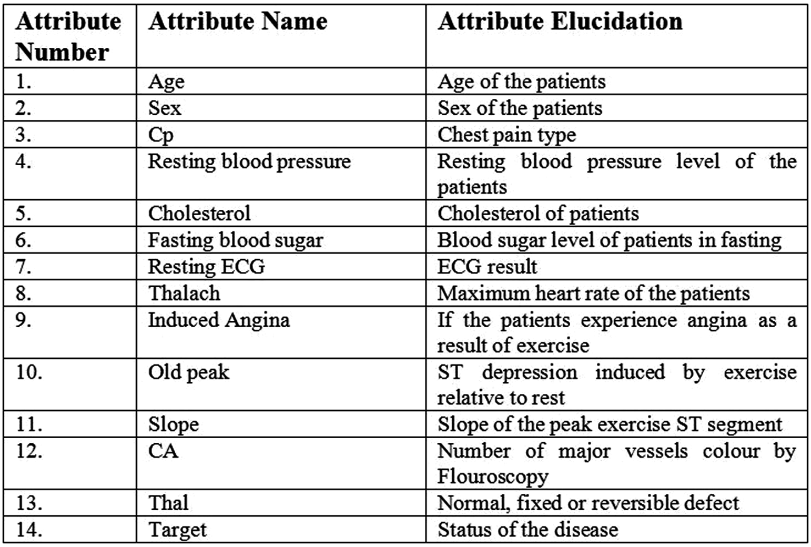
k-NN classifier is the most instance-based method for classifying data. In k-NN the target function may be either discrete valued or real value. k-NN stores all available records and classifies them on the basis of similarity measures (Table 4).

**Table 4.** Pros and cons of k-NN classification techniques

| Pros | Cons |
|---|---|
| Can be applied to data of any distribution and modelling is not expensive | Computationally expensive because it need huge number of sample for accuracy calculation |
| Very simple and intuitive | Depends on K value and affected by irrelevant attributes |
| For large sample size it work good | Affected by irrelevant attributes |

## 6 Data Set Elucidation

Heart disease dataset is collected from kaggle/UCI machine learning repository in which there are 14 attributes and 303 patients record [11] (Fig. 2).



| Attribute Number | Attribute Name | Attribute Elucidation |
|---|---|---|
| 1. | Age | Age of the patients |
| 2. | Sex | Sex of the patients |
| 3. | Cp | Chest pain type |
| 4. | Resting blood pressure | Resting blood pressure level of the patients |
| 5. | Cholesterol | Cholesterol of patients |
| 6. | Fasting blood sugar | Blood sugar level of patients in fasting |
| 7. | Resting ECG | ECG result |
| 8. | Thalach | Maximum heart rate of the patients |
| 9. | Induced Angina | If the patients experience angina as a result of exercise |
| 10. | Old peak | ST depression induced by exercise relative to rest |
| 11. | Slope | Slope of the peak exercise ST segment |
| 12. | CA | Number of major vessels colour by Flouroscopy |
| 13. | Thal | Normal, fixed or reversible defect |
| 14. | Target | Status of the disease |

**Fig. 2.** Detail description of dataset [11]

## 7 Comparison Table of Different Classification Techniques

Table 5 gives the comparison of data mining classification algorithms based on various performance measure.

**Table 5.** Comparison of different classifier with respect to Accuracy, Sensitivity, Specificity, PPV, NPV and AUC

| Classifier | Accuracy (%) | Sensitivity (%) | Specificity (%) | PPV (%) | NPV (%) | AUC |
|---|---|---|---|---|---|---|
| Decision tree | 84.91 | 36.95 | 61.81 | 44.73 | 53.96 | 0.8275 |
| SVM | 88.68 | 39.86 | 58.20 | 44.36 | 53.64 | 0.8848 |
| Naive Bayes | 96.23 | 39.13 | 57.57 | 43.54 | 53.07 | 0.9899 |
| k-NN (Roc) | 58.49 | 50 | 48.49 | 44.80 | 54.7 | 0.6283 |
| k-NN (Acc) | 62.26 | 50 | 49.09 | 45.09 | 54 | 0.6217 |

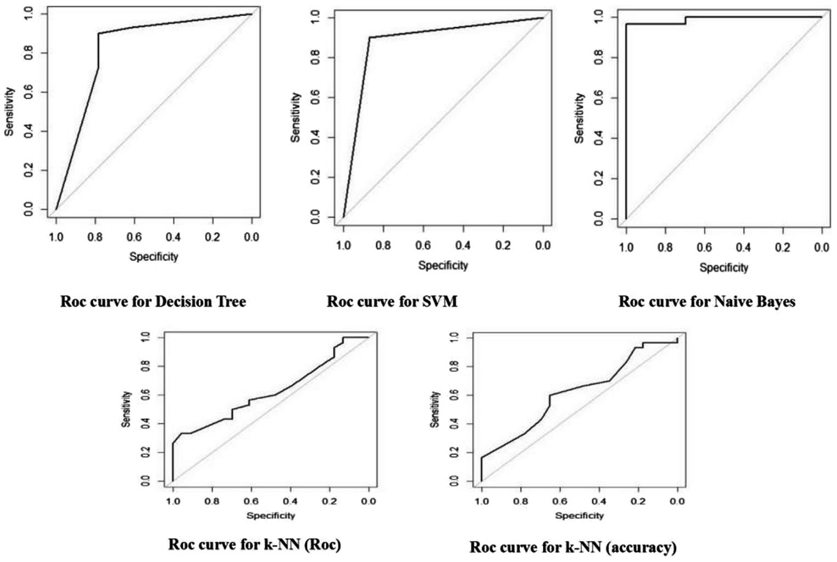
Confusion Matrix is exploited to compute Accuracy, Sensitivity, Specificity, Area under curve and ROC curve. Confusion Matrix for classification of heart disease is shown in Table 6.

**Table 6.** Confusion matrix for heart disease

| Class label | Classified as present of heart disease | Classified as heart disease not present |
|---|---|---|
| Heart disease present | TP | FN |
| Heart disease not present | FP | TN |

$$\text{Sensitivity} = \mathrm{P}(+|1) = \%\text{ of True Positive}\ :\ \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$$
$$\text{Specificity} = \mathrm{P}(-|0) = \%\text{ of True Negative}\ :\ \mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$$
$$\text{Accuracy} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$$
$$\text{PPV (Positively Predicted Value} = \ \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$$
$$\text{NPV (Negatively Predicted Value)} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FN})$$

Figure 3 represents the Roc curve of different classifier.



Roc curve for Decision Tree

Roc curve for SVM

Roc curve for Naive Bayes

Roc curve for k-NN (Roc)

Roc curve for k-NN (accuracy)

**Fig. 3.** ROC curve of different classifier

# 8 Conclusion

This paper focuses on the early detection and prevention of heart related illness by using several data mining classification method which is implemented by using data analytical tool R. For the prediction of heart disease various classifiers are used and we obtained several performance measurement parameters and observed that the performance is better for prediction in case of Naive Bayes as compared to others. Here we also observed that the performance of classifier varies from each other and also depended upon the platform or analytical tool on which the classification techniques are implemented. In future we would try to implement other techniques in which prediction is more accurate.

## References

1. Sundar, N.A., Latha, P.P., Chandra, M.R.: Performance analysis of classification data mining techniques over heart disease database. Int. J. Eng. Sci. Adv. Technol. **2**(3), 470–478 (2012)
2. Palaniappan, S., Awang, R.: Intelligent heart disease prediction system using data mining techniques. In: 2008 IEEE/ACS İnternational Conference on Computer Systems and Applications. IEEE (2008)
3. Dangare, Chaitrali S., Apte, Sulabha S.: Improved study of heart disease prediction system using data mining classification techniques. Int. J. Comput. Appl. **47**(10), 44–48 (2012)
4. Thomas, J., Theresa Princy, R.: Human heart disease prediction system using data mining techniques. In: 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT). IEEE (2016)
5. Wilson, A., et al.: Data mining techniques for heart disease prediction (2014)
6. Banu, M.N., Gomathy, B.: Disease forecasting system using data mining methods. In: 2014 International Conference on İntelligent Computing Applications. IEEE (2014)
7. Waghulde, Nilakshi P., Patil, Nilima P.: Genetic neural approach for heart disease prediction. Int. J. Adv. Comput. Res. **4**(3), 778 (2014)
8. Kaur, Beant, Singh, Williamjeet: Analysis of heart attack prediction system using genetic algorithm. Int. J. Adv. Technol. Eng. Sci. **3**, 87–94 (2015)
9. Ghumbre, S., Patil, C., Ghatol, A.: Heart disease diagnosis using support vector machine. In: International Conference on Computer Science and İnformation Technology (ICCSIT') Pattaya (2011)
10. Bhatia, S., Prakash, P., Pillai, G.N.: SVM based decision support system for heart disease classification with ınteger-coded genetic algorithm to select critical features. In: Proceedings of the World Congress on Engineering and Computer Science, pp. 34–38, San Francisco, USA (2008)
11. Database. http://archive.ics.uci.edu/ml/datasets/Heart+Disease
12. Han, J., Kamber, M.: *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers (2006)