International Journal of Advanced Science and Technology
Vol. 29, No. 9s, (2020), pp. 2702-2709

Exploratory Analysis to Predict Heart Disease Occurrence through Machine Learning Approaches

[1*]Dr. P.K.A. Chitra, [2]Dr. P. Udaykumar,
[1]*Associate Professor, Department of Computer Science and Engineering, St. Martin's Engineering College, Secunderabad*
[2]*Professor and Head, Department of Computer Science and Engineering, St. Martin's Engineering College, Secunderabad,*
[1]*pkachitra@yahoo.com* [2]*uday.uday08@gmail.com*

Abstract

Heart disease is one of the leading diseases in the world that affects human life. Prediction of heart disease at early stage with symptoms is crucial to save life. In real fie, diagnosis of heart disease through traditional medical history hasn't been reliable. Classifying healthy people and people with heart disease can be done with non-invasive-based methods such as machine learning. In the proposed study, machine Learning methods like Decision Tree, Naïve Baye's, k – Nearest Neighbour, Support Vector Machine (RBF and Linear Kernel), Logistic Regression and Artificial Neural Network are used for the analysis of prediction of Heart Disease patients from normal persons using heart disease dataset from UCI repository. Popular machine learning algorithms are implemented and tested with k – fold cross-validation, and the performances of those methods are evaluated with classifiers performance evaluation metrics such as classification accuracy, specificity, sensitivity. Proposed system can easily identify and classify people with heart disease from healthy people. Naïve Baye's Shows good performance in overall in terms of Accuracy, Sensitivity and Specificity. Support Vector Machine with RBF kernel and Logistic Regression also show good performance in terms of three performance measure next to Naïve Baye's method. Decision Tree method gives average performance. The primary motive of this work is to predict heart diseases with high accuracy rate.

Keywords: *Heart Disease Prediction, SVM, Decision Tree, Naïve Bayes, Artificial Neural Network, Logistic Regression, k – Nearest Neighbour*.

1. Introduction

Cardiovascular diseases are rapidly increasing all over the world from the past few years and even if it has found as the most important source of death, it has been announced as the most manageable and avoidable disease [1]. When heart does not pump the blood around the body efficiently, the heart disease occurs. High blood pressure is also one of the main causes of heart disease; a survey says that, in 2011 to 2014, the commonness of hypertension in the world was about 35%, which is also a cause of heart disease. Likewise, there are many more reasons for heart disease such as obesity, not taking in proper nutrition, increased cholesterol and lack of physical activity. For prevention, awareness of heart diseases is to be made among people; around 47% of people die before admission to hospital and it shows that they don't act on early warning signs. Nowadays, lifespan of a human being is reduced because of heart diseases. In 2013, World Health Organization (WHO) developed prevention of non-communicable diseases (NCDs) in which 25% of relative reduction is from cardiovascular diseases and it is being ensured that at least 50% of patients with cardiovascular diseases have access to relevant drugs and medical counselling by 2025 [2]. In 2016, 17.9 million people died just because of cardiovascular diseases, which is 31% of deaths around the world. A major challenge in heart disease is its detection at early stage and is difficult to predict that a person will get heart disease or not [3]. A survey of World Health Organization (WHO) says that medical professionals are able to predict just 67% of heart disease at early stage that leads to a vast scope of research in this field [4]. In India, access to good doctors and hospitals in rural areas is very low, in 2016, WHO report says that, just 58% of the doctors have medical degree in urban areas and 19% in rural areas.

Data mining is the process of extracting knowledge from huge amount of data and is an essential step in discovering knowledge from databases. There are numbers of databases like data marts, data warehouses all over the world and data Mining is needed to extract hidden information from these large amount of database. Data mining is also called as Knowledge Discovery Database (KDD) that has four main techniques: Classification, Clustering, Regression, and Association rule. The fields like the medical field, business field, and educational field have a vast amount of data, thus these fields data can be mined through those techniques more useful information. Data mining is mainly needed in healthcare field to extract useful information from a large amount of data. In this work, a heart disease data set is used. The main aim of this work is to predict the possibilities of occurrence of heart disease patients and is performed through data mining classification techniques. The classification technique is used for classifying the entire dataset into two categories namely yes and No. Classification technique is applied to the dataset through the machine learning classification algorithm namely Decision tree classification and Naïve Bayes Classification models. These models are used to enhance the accuracy level of the classification technique. This model performs both the classification and prediction methods. These models are performed using python Programming Language.

The contribution of the proposed work is to design a machine-learning-based intelligent decision support system for diagnosis of heart disease at early stage. Various predictive models such as logistic regression, k-Nearest Neighbour, Artificial Neural Network, Support Vector Machine, Decision Tree, Naive Bayes and Random Forest have been used for classification of people with heart disease and healthy people. The proposed work has been trained and tested on Cleveland heart disease dataset, 2016, UCI data-mining repository and is available online. All classifiers performances have been checked on full features in terms of classification accuracy and the study suggests which prediction model is feasible with for designing high-level medical intelligent system for heart disease that accurately classifies heart disease and healthy people. The remaining parts of the paper are structured as follows: in Section 2, the background information regarding heart disease dataset briefly reviews the theoretical and mathematical background of classification algorithms in machine learning and discusses cross-validation method and performance evaluation metrics. Section 3, discusses the dataset description and the methods used in the prediction modelling. In Section 4, experimental results are discussed in detail. Finally, Section 5 discusses the conclusion of the paper.

2. Existing System

Heart Disease is a silent killer that leads to the death of the person without obvious symptoms. Medical diagnosis plays a vital role and yet complicated task that needs to be executed efficiently and accurately. Data mining techniques finds patterns and consistency in datasets and with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes or classes. Existing system [4] works on sets of both Deep learning and data mining [5]. Analysis of risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having Heart Disease. Statistical analysis has identified risk factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hypertension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar, etc. but by using all the existing systems the accuracy is very less [4].

3. Materials and Methods

3.1 Data set

The dataset was downloaded from UCI Repository. There are 303 records in the dataset and contains 14 continuous attributes and the goal is to predict the presence of heart disease in the patient. Here are the 14 attributes from the dataset along with their descriptions: *age*: The person's age in years (Continuous), *sex*: The person's sex (1 : male, 0 : female), *cp*: The chest pain experienced (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic), *trestbps*: The person's resting blood pressure (Continuous), *chol*: The person's cholesterol measurement in mg/dl (Continuous), *fbs*: The person's fasting blood sugar (> 120 mg/dl, 1 : true; 0 : false), *restecg*: Resting electrocardiographic measurement (0 : normal, 1: having ST-T wave abnormality, 2 : showing probable or definite left ventricular hypertrophy by Estes' criteria), *thalach*: The person's maximum heart rate achieved (Continuous), *exang*: Exercise induced angina (1 : yes; 0 : no), *oldpeak*: ST depression induced by

exercise relative to rest (Continuous), *slope*: the slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: down sloping), *ca*: The number of major vessels (0–3), *thal*: A blood disorder called thalassemia (3 : normal; 6 : fixed defect; 7 : reversable defect), *target*: Heart disease (0 : no, 1 : yes). Features *cp*, *thalach*, *slope* are highly correlated with target ($p > 0.5$).

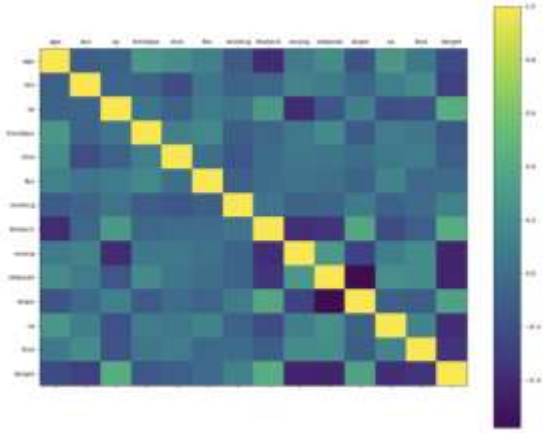

Figure 1 – Correlation among covariates and target

3.2 Machine Learning Methods

Machine learning classification algorithms are used to classify the heart patients and healthy people. Some popular classification algorithms and theoretical background are presented briefly in this work.

3.2.1 Support Vector Machines

SVM is a machine learning method, sued for classification problems [5]. It uses a maximum margin strategy that transform to a complex quadratic programming problem. In a binary classification problem, the instances are separated with a hyper plane

$$w^T x+b=0,$$

where w and d are dimensional coefficient vectors, which are normal to the hyper plane of the surface, b is offset value from the origin, and x is data set values. The SVM gets results of w and b. w can be solved by introducing Lagrangian multipliers in the linear case. 0e data points on borders are called support vectors. 0solution of w can be written as

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i,$$

where n is the number of support vectors, y_i are target labels, x_i are features. W and b are calculated and the linear discriminant function is written as

$$g(x) = sgn\left(\sum_{i=1}^{n} \alpha_i y_i x_i^T x + b\right)$$

For nonlinear distribution, kernel and decision function [ ] are written as

$$g(x) = sgn\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right)$$

3.2.2 Naïve Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets [6]. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|x) = P(x_1|c)\ X\ P(x_2|c)\ X\ \ldots.X\ P(x_n|c)\ X\ P(c)$$

Where

$P(c|x)$ – Posterior Probability of target given predictors
$P(c)$ – Prior Probability of target
$P(x|c)$ – Likelihood of predictors given target
$P(x)$ – Prior Probability of Predictors

3.2.3 Logistic Regression

A logistic regression is a classification algorithm [7]. Logistic regression for binary classification problem predicts the value of target variable y when y ∈ [0, 1], 0 is negative class and 1 is positive class. Logistic regression for multi classification predicts the value of target y when y ∈ [0, 1, 2, 3]. In order to classify two classes 0 and 1, a hypothesis h(θ) = θ^T X will be designed and threshold classifier output is h θ(x) at 0.5. If the value of hypothesis hθ(x) ≥ 0.5, it will predict y = 1 which mean that the person has heart disease. If value of h(x) < 0.5, then predict y = 0, that the person is healthy. Hence, the prediction of logistic regression under the condition 0 ≤ h(x) ≤ 1 is done. Logistic regression using sigmoid function can be written as follows:

$$h\theta(x) = g\left(\theta^T X\right),$$

Where $g(z) = \frac{1}{(1+x^{-z})}$ and

$h\theta(x) = \frac{1}{(1+x^{-z})}$, cost function can be written as $h(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost\ (h\theta(x^{(i)}), y^{(i)})$

3.2.4 Decision Tree

Decision tree is a simpler classifier which is easy to implement and which does not require any domain knowledge. The main advantage is decision tree method can be applied to huge data which is to interpret. Decision tree is a tree like structure which consists of arcs, nodes, and branches [10,11,12]. The arcs connect from one node to another. The branch has attributes, an internal node has a test on which the attribute is used for, and leaf node consists of the classes which are predicted from decision tree.to make an appropriate decision the traversing starts from root node to leaf node. Iterative Dichotomiser 3 is also known as ID3 algorithm used in the field of data mining, one of the most important algorithms of decision tree algorithms. The main aspect of ID3 algorithm is to select the appropriate attribute to test at each node in a tree which is an iterative process. It uses top down approach which traverses from top node to leaf node at the bottom which completes the decision tree. This is the basic algorithm used for classification in data mining. The algorithm for building decision trees called C4.5 [ ], which employs a top-down, greedy search through the space of possible branches with no backtracking. C4.5 uses Entropy and Information Gain to construct a decision tree. A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. To build a decision tree, two types of entropy using frequency tables are need to be calculated and are as follows:

$$E(S) = \sum_{i=1}^{n} -p_i log_2 p_i$$

3.2.5 Artificial Neural Network

The Multilayer Perceptron Neural Network (MLPNN) consists of one input layer, one or more hidden layers and one output layer[9]. In MLPNN, the input nodes pass values to the first hidden layer, and then nodes of first hidden layer pass values to the second and so on till producing outputs. Input layer is set with 13 nodes, hidden layer with 16 nodes and output layer with 2 neuron to predict yes or no.

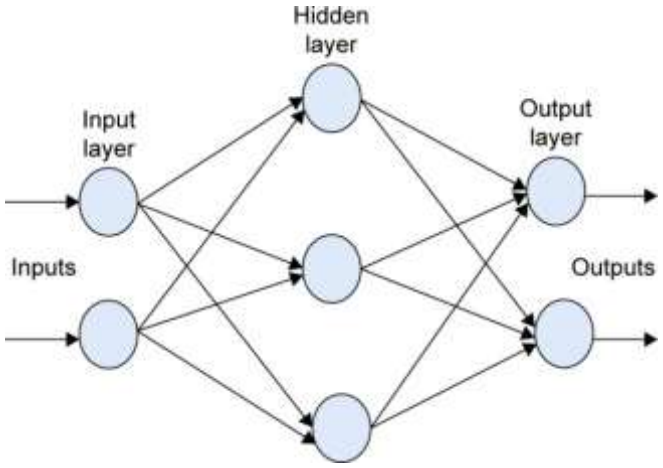

Figure 2 – Multilayer perceptron neural network.

3.2.6 k – Nearest Neighbour

K nearest neighbor (KNN) is a simple classification method which stores all cases and classifies new cases based on similarity measure. KNN algorithms have been used since 1970 in many applications like statistical estimation and pattern recognition etc. KNN is a non-parametric classification method which is broadly classified into two types as structure-less NN techniques and other as structure based NN techniques. In structure less NN techniques whole data is classified into training and test sample data. Euclidean distance is calculated from training point to sample point, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are based on structures of data like orthogonal structure tree (OST), ball tree, k-d tree, axis tree, nearest future line and central line [8]. Nearest neighbor classification is used mainly when all the attributes are continuous. There are various ways of measuring the similarity between two instances with n attribute values. Every measure has the following three requirements. Let dist (A, B) be the distance between two points A,B, then, (i) dist(A,B) $\geq$ 0 and dist (A, B) =0, if and only if A = B, (ii) dist (A, B) = dist (B, A) and (iii) dist (A, C) $\leq$ dist (A,B) + dist (B, C). The shortest distance between any two data points is a straight line.

3.3 Performance Evaluation Metrics

In order to check the performance of the classifiers, various performance evaluation metrics were used in this work. Confusion matrix is used; every observation in the testing set is predicted in exactly one box. It is a 2 × 2 matrix as this is a binary class problem. Additionally, it gives two types of correct prediction of the classifier and two types of classifier of incorrect prediction. Table 1 shows the confusion matrix.

Table 1: Confusion Matrix

	Predicted Heart Disease Patient (1)	Predicted Healthy Person (0)
Actual Heart Disease Patient (1)	TP	FN
Actual Healthy Person (0)	FP	TN

TP: predicted output as *True Positive (TP);* Heart Disease patients are correctly classified and patients have heart disease.
TN: predicted output as *True Negative (TN),* Healthy Person is correctly classified and the Person is healthy.
FP: predicted output as *False Positive (FP)*; Healthy Person is incorrectly classified that they do have Heart Disease (Type 1 error).

FN: predicted output as *False Negative (FN);* Heart Disease patient is incorrectly classified that the subject does not have heart disease as the person is healthy (Type 2 error).

3.3.1 Classification Accuracy

The overall performance of the classification system as follows:

$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\ X\ 100\%$$

3.3.2 Sensitivity

It is a ratio of recently classified heart patients to the total number of heart patients. The sensitivity of classifier for detecting positive instances is known as "true positive rate." In other words, sensitivity (true positive fraction) confirms that if a diagnostic test is positive and the subject has the disease. It can be written as follows:

$$Sensitivity = \frac{TP}{TP + FN}\ X\ 100\%$$

3.3.3 Specificity

A diagnostic test is negative and the person is healthy and is mathematically written as follows:

$$Specificity = \frac{TN}{TN + FP}\ X\ 100\%$$

3.4 Validation Method of Classifiers

K-fold Cross-Validation (CV) method and three performance evaluation metrics are used in this research work. In k-fold cross-validation, the data set is divided into k equal size of parts, in which k – 1 group are used to train the classifiers and remaining part is used for checking outperformance in each step. The process of validation is repeated k times and the classifier performance is computed based on k results. For CV, different values of k are selected. In this experiment, k = 10 is used, because its performance is good. In the 10-fold CV process, 90% data were used for training and 10% data were used for testing purpose. The process was repeated 10 times for each fold of process, and all instances in the training and test groups were randomly divided over the whole dataset prior to selection training and testing new sets for the new cycle.

4. Results and Discussion

Machine Learning methods, Decision Tree, Naïve Baye's, k – Nearest Neighbor, Support Vector Machine (RBF and Linear Kernel), Logistic Regression and Artificial Neural Network are used for the analysis of prediction of Heart Disease patients from normal persons. The Heart Disease dataset from UCI repository is used for the analysis here. 10 – Fold cross validation is used. Performances of all prediction methods over Heart Disease dataset is presented in Table.2.

Table 2: Performance Comparison of all Methods

Classifier Name	Classification Performance Evaluation Metrics		
	Accuracy (%)	Sensitivity (%)	Specificity (%)
Decision Tree	78	74	76
Naïve Bayes	89	82	84
K – Nearest Neighbour (k=9)	76	74	77
SVM (Linear Kernel)	86	85	82
SVM (RBF Kernel)	76	74	78
Logistic Regression	82	80	82
Artificial Neural Network	76	74	72

(13, 16, 2)			

Naïve Baye's gives 89% accuracy, 82% sensitivity and 84% specificity. SVM with RBF kernel shows 86% accuracy, 85% sensitivity and 82% specificity. Logistic regression is good at handling non-linear dependencies; it shows 82% accuracy 80% sensitivity and 82% specificity. k-NN, Artificial Neural Network, Support Vector Machine with Linear Kernel shows low 76% accuracy, 74% sensitivity. Artificial Neural Network shows low specificity (72%). Naïve Baye's Shows good performance in overall in terms of Accuracy, Sensitivity and Specificity. Support Vector Machine with RBF kernel and Logistic Regression also show good performance in terms of three performance measure next to Naïve Baye's method. Decision Tree method gives average performance as 78% accuracy, 74% sensitivity and 76% specificity. k – Nearest Neighbor method is tried with 6 k values. Cluster size starts from 1, then 3, 5, 7,9,13 were tried and are shown in figure 3. Among all cluster sizes k – NN shows good performance for k = 9.

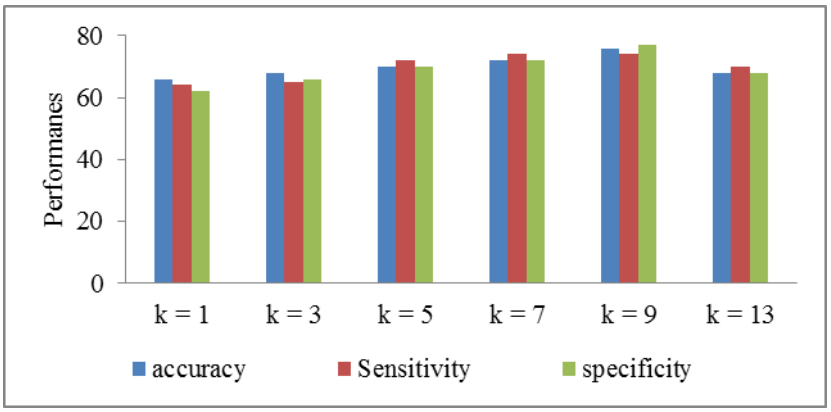

Figure 3 - Performance of K-NN on different values of k.

5. Conclusion and Future work

The main objective of this study is to predict weather a patient is affected with Cardiac disease or not. It is difficult to determine manually the odds of getting heart disease based on the risk factors provided as attributes in this work. This work aimed only to the application of classification methods for heart disease perdition. Machine Learning methods, Decision Tree, Naïve Baye's, k – Nearest Neighbor, Support Vector Machine (RBF and Linear Kernel), Logistic Regression and Artificial Neural Network are used for the analysis of prediction of Heart Disease patients from normal persons. The Heart Disease dataset from UCI repository is used for the analysis here. 10 – Fold cross validation is used. Naïve Baye's Shows good performance in overall in terms of Accuracy, Sensitivity and Specificity. Support Vector Machine with RBF kernel and Logistic Regression also show good performance in terms of three performance measure next to Naïve Baye's method. Decision Tree method gives average performance. The primary motive of this work is to predict heart diseases with high accuracy rate. Still, the methods implemented here shows average accuracy. If these systems can be implemented with data pre-processing methods, still there is a hope to improve accuracy.

References

1. Golande, A., & Kumar, P. (2019). Heart Disease Prediction Using Effective Machine Learning Techniques. International Journal of Recent Technology and Engineering (IJRTE), 8(1).
2. Prabhakaran, D., Jeemon, P., Sharma, M., Roth, G. A., Johnson, C., Harikrishnan, S & Dhaliwal, R. S. (2018). The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990–2016. The Lancet Global Health, 6(12), e1339-e1351.
3. Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99-104.

4. Bharti, S., & Singh, S. N. (2015, May). Analytical study of heart disease prediction comparing with different algorithms. In International Conference on Computing, Communication & Automation (pp. 78-82). IEEE.
5. H.-L. Chen, B. Yang, J. Liu, and D. Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," Expert Systems with Applications, vol. 38, no. 7, pp. 9014–9022, 2011.
6. Sana Bharti, 2015. Analytical study of heart disease prediction compared with different algorithms; International conference on computing, communication, and automation (ICCA2015).
7. Sarath Babu, 2017.Heart disease diagnosis using data mining technique, international conference on electronics, communication and aerospace technology (ICECA2017).
8. Monika Gandhi, 2015. Prediction in heart disease using techniques of data mining, International conference on futuristic trend in computational analysis and knowledge management (ABLAZE- 2015)
9. Gudadhe M, Wankhade K, Dongre S. Decision support system for heart disease based on support vector machine and Artificial Neural Network. Computer and Communication Technology (ICCCT), 2010 International Conference on; 2010. pp. 741–745.
10. P. W. Wagacha, "Induction of decision trees," Foundations of Learning and Adaptive Systems, vol. 12, pp. 1–14, 2003.
11. Yasoda, K., Ponmagal, R.S., Bhuvaneshwari, K.S. K Venkatachalam, " Automatic detection and classification of EEG artifacts using fuzzy kernel SVM and wavelet ICA (WICA)" Soft Computing Journal (2020).
12. VenkatachalamK,KarthikeyanNK,"Effective Feature Set Selection and Centroid Classifier Algorithm for Web Services Discovery",Indonesian Journal of Electrical Engineering and Computer Sceince,Vol 5,2017.