Heart Disease Prediction Using Machine Learning

Pushpa R.N[1], Sreedevi S[2], Thaseen Bhashith[3]
[1,2,3]*Assistant Professor, Department of Computer Science and Engineering, JNNCE, Shivamogga,Karnataka,India*

Abstract- The large amount of data is generated in medical organizations (hospitals, medical centers), but this data is not properly used. There is a wealth of hidden information present in the datasets. This unused data can be converted into useful data by using different data mining techniques. This paper presents a classifier approach for detection of heart disease and shows how Naive Bayes can be used for classification purpose. This system will categorize medical data into five categories namely no, low, average, high and very high. If unknown sample comes then the system will predict the class label of that sample. Hence two basic functions namely classification (training) and prediction (testing) will be performed
Keywords – World Health Organization, K-Nearest Neighbors

I. INTRODUCTION

The heart is one of the main organs of the human body. It pumps blood through the blood vessels of the circulatory system. The circulatory system is extremely important because it transports blood, oxygen and other materials to the different organs of the body. Heart plays the most crucial role in circulatory system. If the heart does not function properly then it will lead to serious health conditions including death. Common risk factors of heart disease include high blood pressure, abnormal blood lipids, and use of tobacco, obesity, physical inactivity, diabetes, age, gender and family generation. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, where heart disease is the major cause of deaths. For example, in 2008, 17.3 million people died due to Heart Disease. WHO estimated by 2030, almost 23.6 million people will die due to Heart disease.
The rest of the paper is organized as follows. Proposed heart disease prediction algorithms are explained in section II. Experimental results are presented in section III. Concluding remarks are given in section IV.

II. PROPOSED ALGORITHM

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer only simple queries but they cannot answer complex queries. Diagnosing of heart disease is one of the important issues. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

2.1. Naïve Bayes Algorithm -

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and allows capturing uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.
Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$. Look at the equation below:
$P(c|x)=P(x|c)P(c)/P(x)$
Above,
$P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes.)
$P(c)$ is the prior probability of class.
$P(x|c)$ is the likelihood which is the probability of predictor given class.
$P(x)$ is the prior probability of predictor.
Steps in algorithm are as follows:
Each data sample is represented by n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2… An.
Suppose that there are m classes, C1, C2……Cm, given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned if and only if:
$P(Ci|X)>P(Cj|X)$ for all $1< = j< = m$ and $j!= i$

Thus P(Ci|X) is maximized. The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,
As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = Si /s , where Si is the number of training samples of class Ci, and s is the total number of training samples on X. That is, the Naive probability assigns an unknown sample X to the class Ci.

2.2. KNN Algorithm -
K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use a data and classify new data points based on a similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors.
Steps in KNN algorithm:
Load the data.
Initialize K to the chosen number of neighbors.
For each example in the data,
Calculate the distance between the query example and the current example from the data.
Add the distance between query example and the current example from the data.
Sort the ordered collection of distances and indices from the smallest to largest (in ascending order) by the distances.
Pick the first K entries from the stored collection.
Get the labels of the selected K entries.
If regression, return the mean of the K labels.
If classification, return the mode of the K labels.

2.3. Logistic regression algorithm -
Logistic Regression is one of the most used Machine Learning algorithms for binary classification. It is a simple Algorithm that one can use as a performance baseline, it is easy to implement and it will do well enough in many tasks. The building block concepts of Logistic Regression can also be helpful in deep learning while building neural networks.
There are three types of logistic regression algorithm: Binary, Multi and ordinal.
Pseudocode for Multiclass logistic regression algorithm:
Divide the problem into n+ 1 binary problem.
For each class,
Predict the probability the observations are in that single class.
Prediction=<math>max (probability of the classes)

2.4. Decision Tree Algorithm Pseudocode -
Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too.
Decision Tree Algorithm Pseudocode:
Place the best attribute of the dataset at the root of the tree.
Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
Repeat step 1 and step 2 on each subset until one find leaf nodes in all the branches of the tree.

III. EXPERIMENT AND RESULT

Figure 1 demonstrates the graph which gives the information about the percentage of people having the heart disease and not having heart disease in the given dataset. 303 instances are there in the dataset that has been provided in this paper. It has been estimated that 54.1% of the instances are labeled to not to have heart disease and 45.9 % of the instances are labeled to have the heart disease.

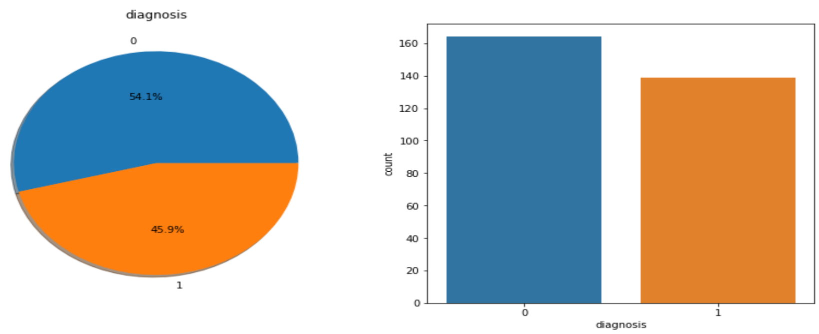

Figure 1. Analysis of the absence and presence of disease in the dataset

Figure 2 demonstrates the sample of the dataset that has been used in the implementation.

```
In [10]: runfile('C:/Users/google/Desktop/fpppp.py', wdir='C:/Users/google/Desktop')
    age  sex   cp  restbp   chol  fbs  restecg  thalach  exang  oldpeak  \
0  63.0  1.0  1.0   145.0  233.0  1.0      2.0    150.0    0.0      2.3
1  67.0  1.0  4.0   160.0  286.0  0.0      2.0    108.0    1.0      1.5
2  67.0  1.0  4.0   120.0  229.0  0.0      2.0    129.0    1.0      2.6
3  37.0  1.0  3.0   130.0  250.0  0.0      0.0    187.0    0.0      3.5
4  41.0  0.0  2.0   130.0  204.0  0.0      2.0    172.0    0.0      1.4

   slope   ca  thal  num
0    3.0  0.0   6.0    0
1    2.0  3.0   3.0    2
2    2.0  2.0   7.0    1
3    3.0  0.0   3.0    0
4    1.0  0.0   3.0    0
```

Figure 2. Sample of the dataset

Figure 3 and Figure 4 demonstrate the graphs depicting the numerical analysis of the given dataset. Each graph represents the behavior of a particular attribute present in the dataset.

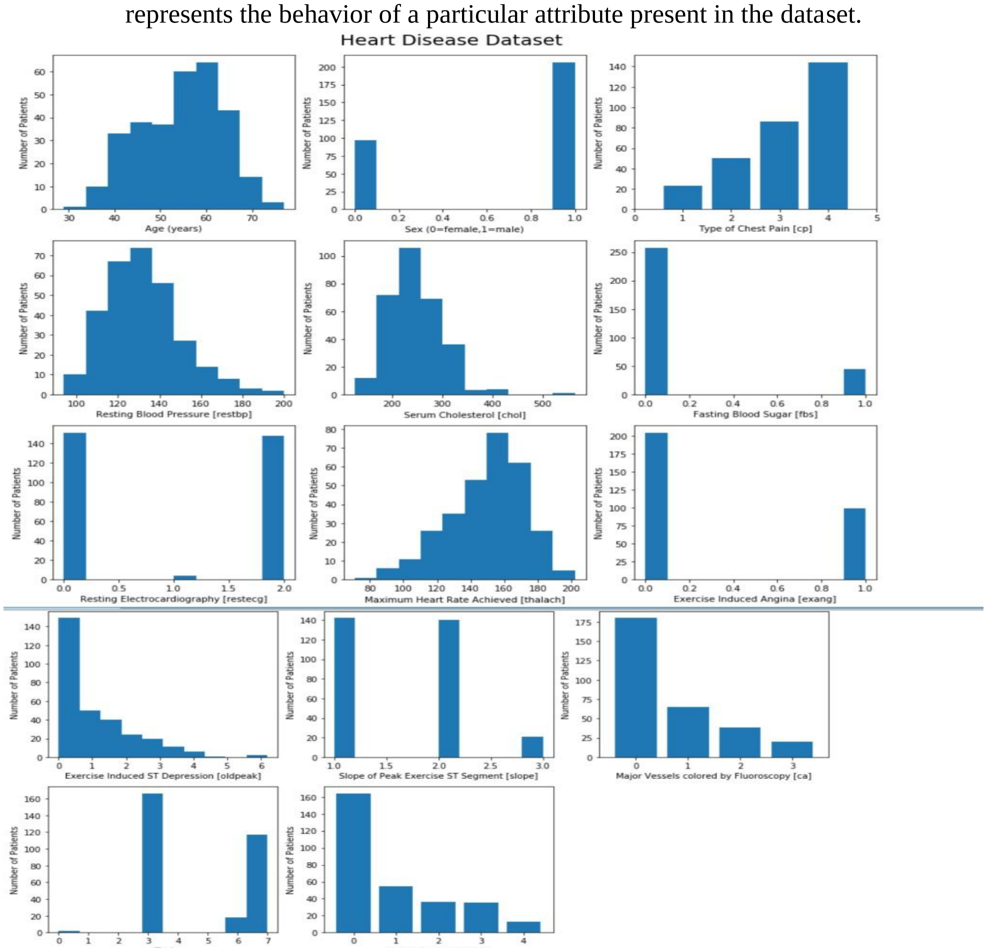

Figure 3. Graphical analysis-1 of data within the dataset

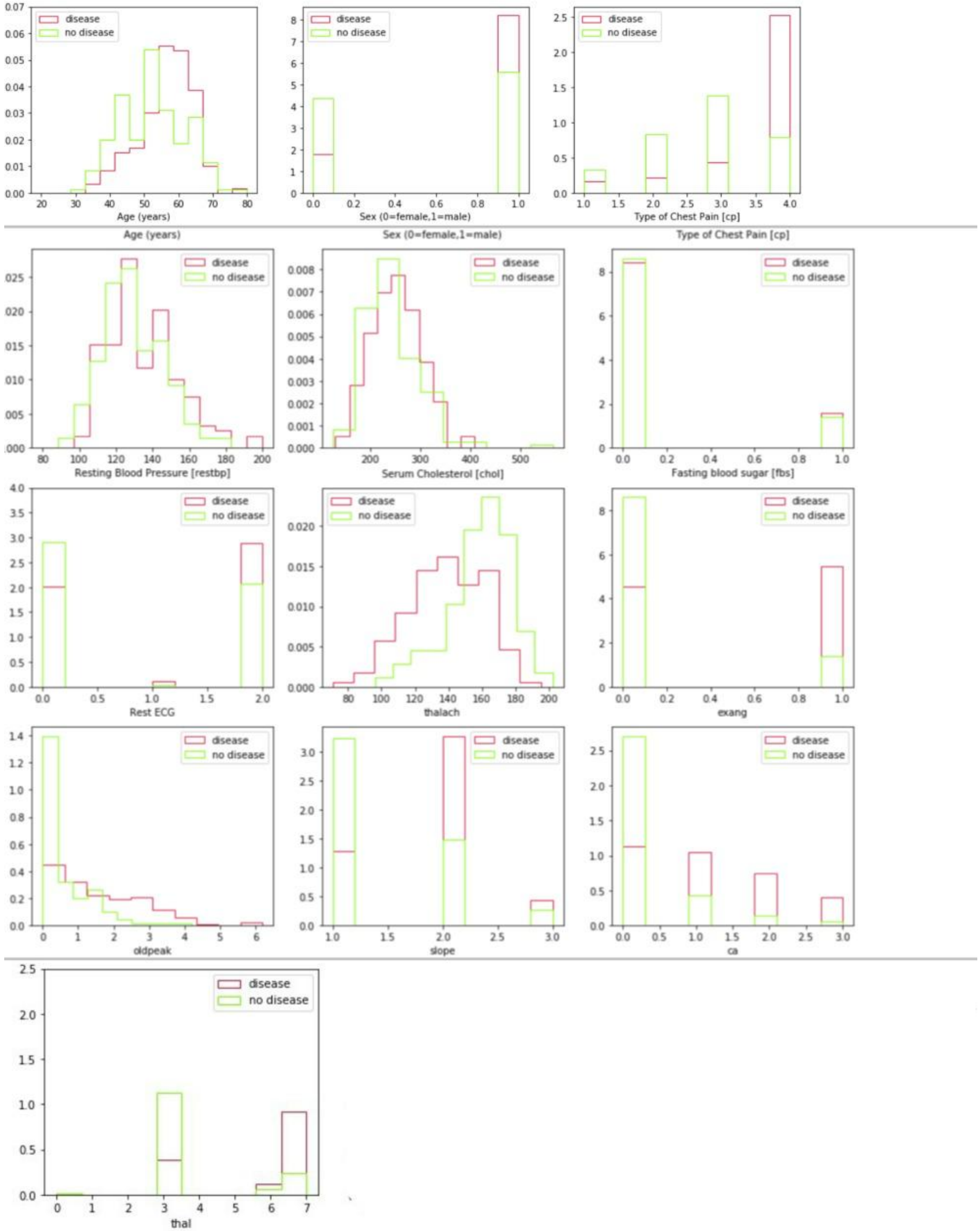

Figure 4. Graphical analysis-2 of data within the dataset

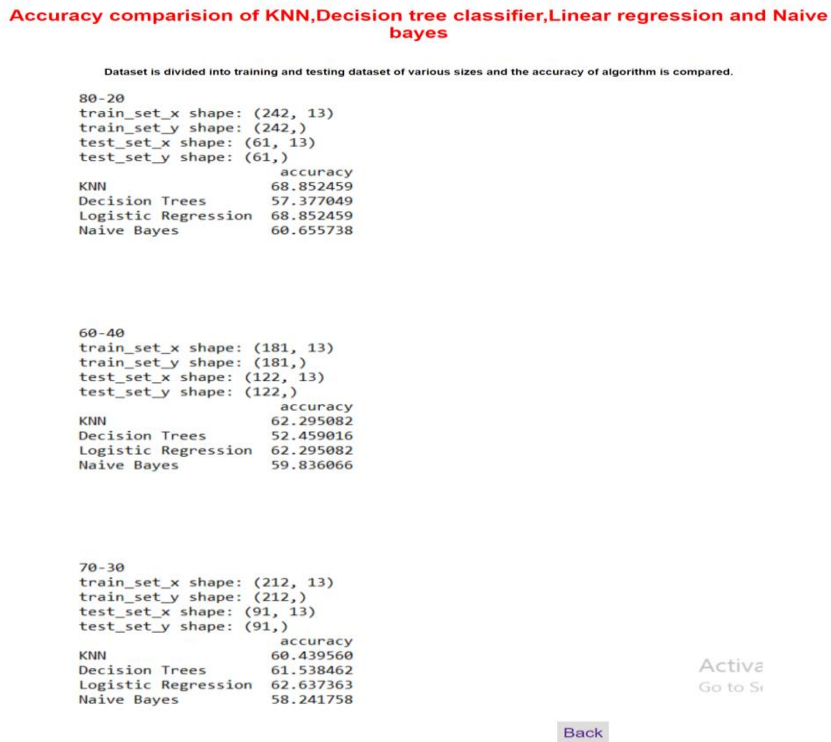

Figure 5 Accuracy comparison of KNN, Decision tree classifier, linear regression and Naïve bayes.

IV.CONCLUSION

Heart disease prediction system provides the deep insight into machine learning techniques for classification of heart diseases. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. The role of classifier is crucial in healthcare industry so that the results can be used for predicting the treatment which can be provided to patients. The existing techniques are studied and compared for finding the efficient and accurate systems. Machine learning techniques significantly improves accuracy of cardiovascular risk prediction through which patients can be identified during an early stage of disease and can be benefitted by preventive treatment. In this work naïve Bayes classifier is used which is more efficient and accurate than other classifiers like KNN, Decision tree and logistic regression algorithms. When a patient is predicted as positive for heart disease, then the medical data for the patient can be closely analyzed by the doctors. When the early symptoms of heart diseases are ignored, the patient might end up with drastic consequences in a short span of time.

V. REFERENCE

[1] JyotiSoni, Uzma Ansari, Dipesh Sharma, and SunitaSoni,June 2011, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers", International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 6, pp. 2385-2392.

[2] VikasChaurasia, and Saurabh Pal, 2013, "Early Prediction of Heart Diseases Using DataMiningTechniques",Caribbean Journal of Science and Technology, ISSN: 0799-3757, Vol.1, pp. 208-218.

[3] Atul Kumar Pandey ,PrabhatPandey ,K.L. Jaiswal ,Ashish Kumar Sen ,2013, "A Heart Disease Prediction Model using Decision Tree", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 6 (Jul. - Aug. 2013), PP 83-86.

[4] NouraAjam, 2015, "Heart Diseases Diagnoses Using Artificial Neural Network", Network And Complex Systems, ISSN: 2224-610X (Paper), ISSN: 2225-0603(Online), Vol.5, No.4,pp. 7-11.

[5] S. Suganya, and P.TamijeSelvy, January 2016, "A Proficient Heart Disease Prediction Method using Fuzzy-Cart Algorithm", International Journal of Scientific Engineering and Applied Science (IJSEAS), Vol. 2, Issue1,ISSN: 2395-3470.

[6] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map",IEEE International Conference on Fuzzy Systems (FUZZ),pp. 1377-1382.

[7] Milan Kumari, SunilaGodara, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST Vol. (2), Issue (2), June 2016.

[8] Niti Guru, Anil Dahiya, NavinRajpal, Decision Support System for Heart Disease Diagnosis Using Neural Network, Delhi Business Review, Vol. 8, No. 1, January-June 2016.

[9] Jaymin Patel, Prof. TejalUpadhyay, and Dr. Samir Patel, Sep 2015-Mar 2016, "Heart Disease Prediction using Machine Learning and Data Mining Technique", Vol. 7, No.1, pp. 129-137.