International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Comparative Study of Data Mining Techniques on Heart Disease Prediction System: a case study for the "Republic of Chad"

Alladoumbaye Ngueilbaye[1], Lin Lei[2], Hongzhi Wang[3]

[1,2,3] Harbin Institute of Technology, School of Computer Science and Technology, 92 Western Dazhi, Harbin 150001, China

Abstract: ***Nowadays, the healthcare sector is one of the areas where huge data are daily generated. However, most of the generated data are not properly exploited. Important encapsulated data are currently in the data sets. Therefore, the encapsulated data can be analyzed and put into useful data. Data mining is a very challenging task for the researchers to make diseases prediction from the huge medical databases. To succeed in dealing with this issue, researchers apply data mining techniques such as classification, clustering, association rules and so on. The main objective of this research is to predict heart diseases by the use of classification algorithms namely Naïve Bayes and Support Vector Machine in order to compare them on the basis of the performance factors i.e. probabilities and classification accuracy. In this paper, we also developed a computer-based clinical Decision support system that can assist medical professionals to predict heart disease status based on the clinical data of the patients using Naïve Bayes Algorithm. It is a web-based user-friendly system implemented on ASP.NET platform with C# and python for the data analysis. From the experimental results, it is observed that the performance of Naïve Bayes is better than the other Algorithm.***

Keywords: Data mining, Heart Disease, Naïve Bayes, Support Vector Machine

1. Introduction

Data mining or knowledge discovery is an important technique of extracting the useful and explicit information from huge data in several types of databases. Presently, data mining obviously plays an important role in the various types of fields such as Medical, Science, Business, the web, and Government etc. The most notorious used data mining techniques are Association rule, Classification, Neural Networks and Clustering.

In recent days, The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to heart diseases and Heart Disease remains one of the major causes of death in the world. In 2008, 17.3 million people died due to Heart Disease. The World Health Organization Statistics 2012 reports have shown enlightens the fact that one in three adults worldwide has raised blood pressure a condition that causes around half of all deaths from stroke and heart disease. In May 2014, WHO also estimated the rate of 93.49 of heart disease in the Republic of Chad and by 2030, almost 23.6 million worldwide will die due to Heart disease [1].

Heart disease can be also known as (CVD) cardiovascular disease, contains a number of conditions that affect the heart including the heart attacks. In addition, Heart diseases possess some functional problems of the heart such as infections of the heart muscles like myocarditis (inflammatory heart diseases), heart-valve abnormalities or irregular heart rhythms etc. are the reasons that can be led to heart failure.

2. Literature Review

In this paperwork [2] there are three different data mining techniques such as Naïve Bayes, K-NN, Decision were addressed to analyze the dataset in which Tanagra tool was also used for the classification and evaluation of the data using 10-fold cross validation and the results were compared. Tanagra is a data mining tool used for graphical user interface algorithms with the main objective of providing an easy environment to researchers and student to use data mining tool and to analyze either the data are categorized as real or synthetic. This powerful tool abounds several data mining techniques such as Supervised Learning Assessment, Meta Supervised Learning, Clustering, supervised learning, feature selection, data visualization, construction algorithms, feature selection, and statistics. In this paperwork, the experiment has been performed by the use of 3000 instances training dataset with 14 different attributes. The data set is classified into two categories in which we have 70% of the data were used for training while 30% were used for testing. Considering these experimental results, it is shown that the classification accuracy of Naïve Bayes algorithms is better compared to other algorithms.

Dhamodharan et.al [3] made use of different symptoms with Naïve Bayes and FT Tree algorithms to predict three major liver diseases such as liver cancer, cirrhosis, and Hepatitis. Then, the results of the two algorithms were compared based on their classification accuracy measure with the experimental results which Naïve Bayes algorithm predicted the results with high classification accuracy than the other algorithm.

Rosalina et al [4] used Support Vector machine (SVM) and Wrapper Methods to predict a hepatitis prognosis disease. Prior to the classification process, they implemented denoising features using wrapper methods. Firstly, SVM performed feature selection to get better accuracy and then Features selection were implemented to decrease the noise or irrelevance data. From the experimental results, they

observed the increment of the accuracy rate in the clinical laboratory test with less execution time with the target achievement by the combination of Wrappers Method and SVM techniques.

Chaitrali S. Dangare et al.,[5] used the Multilayer Perception Neural Network(MLPNN) with Backpropagation (BP) Algorithm and Weka tool in the developed Heart Disease System (HDS) and the performance results of these techniques were compared based on the accuracy. In the existing system, 13 types of Medical terms were used for the prediction process, which 2 new terms as Obesity and Smoking were included. The analysis results showed that the classification Neural Network has predicted the heart disease with the highest accuracy.

Ashish Kumar Sen et al., [6] predicted Heart Disease by implementing two layered approaches using Neural Networks and fuzzy rules. These layers consist of different parameters each, and they also designed an automated system which could analyze the possibilities of Heart Disease occurrence by the use of the Matlab tool.

Dhanashree S. Medhekar et al.,[7] designed the Classification Approach with Naive Bayes algorithm which results in categorizing the Medical data into different categories such as No, Low, Average, High and Very High.

Shadab Adam Pattekari et al.,[8] applied Decision Trees, Naive Bayes, and Neural Networks to develop a prototype Heart Disease Prediction System (HDPS) and it is implemented in a web application in which the user answers the predefined questions. Then it will retrieve and compare the hidden data from stored database with the user's values in the trained data set.

Mr.Pankaj S. Kulkarni et al., [9] used the horoscope approach for the purpose of predicting the Heart Disease. There are twelve different types of planets are available in the horoscope approach. Each planet having different types of behaviors. These twelve planets take an important role in this horoscope approach.

Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques such as Decision Trees, Naïve Bayes, and Neural Network is implemented in [10] using .NET platform. IHDPS is the Web-based, user-friendly, scalable, reliable and expandable system. It can also answer complex "what if" queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease. As a Data source, a total of 909 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database. Naïve Bayes appears to be most effective as it has the highest percentage of correct predictions for patients with heart disease, followed by Neural Network (with a difference of less than 1%) and Decision Trees. Decision Trees, however, appears to be most effective for predicting patients with no heart disease compared to the other two models.

3. Problem Definition

The diagnostic of heart disease remains more or less the most difficult and tedious task in the medical field and it various factors and symptoms of prediction which is involved in several layered issue that could engender the negative presumptions and unpredictable effects. Wu et al proposed that the integration of clinical decision support with relation to the computer-based system of the patient record could reduce the rate of errors in medical predictions, low the unwanted practice variation, enhance safety for patients, and the improvement of patient outcome [21]. This knowledge provides a useful environment which can help to significantly improve the quality of clinical decisions.

Many of hospital information in recent days are designed to implement patient billing, patient data storing, inventory management and generation of simple statistics computation. Most of the hospitals use decision support systems but they are still in most cases bounded. The majority of doctors are predicting heart disease symptoms based on their learning and working experience. In this case, prediction system should be implemented so that to reduce the risk of Heart Disease.

4. Methodology

Data Source
In this paper, the dataset accommodates information about heart disease and the data were collected from the hospitals (Clinique providence and Hôpital de Reference Nationale) databases. We have in total 315 instances of which 150 instances are assigned to the healthy and 165 instances of the heart disease and 14 clinical features have been recorded for each instance.

Features Description

Parameters with related information

Predictable attribute
1. Diagnosis (0 and 1 are used as codes for patient having heart disease and patient without heart disease. 1 (one) indicates Patient with heart disease and 0 (zero) indicates Patient without heart disease)

Key Attribute
Patient ID- patient's identification number

Input Attributes
P1-Age (1 :< =40, value 2 :< =60 and>40, value 3 :> 60) in year
P2- Gender (value 1: Male, Value 2: Female)
P3- Chest Pain Type (value 1: low, value 2: Medium, value 3: High, value 4: Very High)
P4 – Blood pressure (value 1 :< =80, value 2: <= 120 and >80, value 3 :> 120)
P5 – Serum Cholesterol (value 1 :< =180, value 2: <= 400 and >180, value 3 :> 400)
P6- Blood sugar >120? Yes=1, no=0
P7-restecg: resting electrocardiographic output 0, 1, 2

P8-thalach: the maximum heart rate achieved.
P9-exang: exercise-induced angina states (1=yes; 0=no)
P10-oldpeak=ST depression induced by exercise relative to the resting time.
P11-slope: the slope of the peak exercises ST segment interval.
P12-ca: number of major blood vessels ranges (0 to 3) colored by fluoroscopy
P13-thal: 3=normal, 6=fixed defect, 7=reversible defect
P14-diagnostic of heart disease.

These 14 above-mentioned attributes in which include 8 symbolic and 6 numeric such as: age (age in years), sex (male, female), Chest pain type (typical angina, atypical angina, non-angina pain, asymptomatic), Trestbps (resting blood pressure in mm Hg), Cholesterol (serum cholesterol in mg/dl), fasting blood sugar < 120 mg/dl (true or false), resting electrocardiographic output (normal, having ST-T wave abnormality, showing probable or definite left ventricular hypertrophy by Estes' criteria), max heart rate, exercise-induced angina (true or false), old peak (ST depression induced by exercise relative to rest), Slope (up, flat, down), number of vessels colored by fluoroscopy (0-3), Thal (normal, fixed defect, reversible defect), and class (healthy, not healthy) were used in this system.

4.1 Data mining technical terms

Association *rule*
Association rule is the process of extracting the useful correlations, usual patterns between the variables or sets of items in the transaction database or a particular data repository. It is one of the most important and well-researched techniques of data mining and was first introduced in [Agrawal et al. 1993].

Classification
Classification is data mining or machine learning techniques of identifying which set of categories belongs to the basis of a training set of data for its most effective and efficient application. For example, classification can be used to prophesy (predict) whether the patient has heart Disease or the patient has no heart disease[5].

Clustering
Clustering is data mining techniques used to analyze data objects without referring back to a known label. It is the most useful and important techniques used for the process of discovery of data distribution. Clustering has two major types available. Portioning clustering which is the process of portioning the database into the limited number of clusters and Hierarchical clustering do a sequence of partitions by assigning each item to it cluster [5].

Supervised learning
Supervised learning can be explained as machine learning task which allows deciding that a certain function is true from the grouped training data. In supervised learning, every example is composed of an input object and an expected output value. In this case, the supervised learning will analyze the data and bring out an inferring function called classifier which will predict the accurate output for any input object by requiring the learning algorithm to make the general view from the training data to hidden situations in a satisfactory way [2, 5].

Unsupervised learning
Unsupervised in machine learning makes reference to the process of finding hidden structure in ungrouped data, once the example given to the learner are ungrouped, no error or a reward signal will occur to rate a potential solution [11].

Reinforcement learning
In this algorithm, the machine is trained to make specific decisions by exposing to an environment where it trains itself regularly by using trial and error. It learns from the previous experience and tries to catch the important possible knowledge in which will lead to the accurate business decisions. Markov Decision can be one of the examples of Reinforcement Learning [11].

Decision Tree
A decision tree is a non-parametric supervised learning method used for classification. It is one of the most commonly used classification techniques in the machine learning; recent surveys demonstrates that it is the most commonly used techniques that you don't have to know much about machine learning to understand how it works[12].

Table1: Dataset used to build Decision Tree

Age	Dataset		
	Gender	Intensity Symptoms	Disease prediction
41	M	Medium	Yes
39	M	Low	Yes
39	M	Low	Yes
42	M	High	Yes
49	F	High	Yes
40	F	Medium	Yes
54	M	Low	No
52	F	Medium	No
60	F	Medium	No
55	F	Low	No
50	M	Low	No

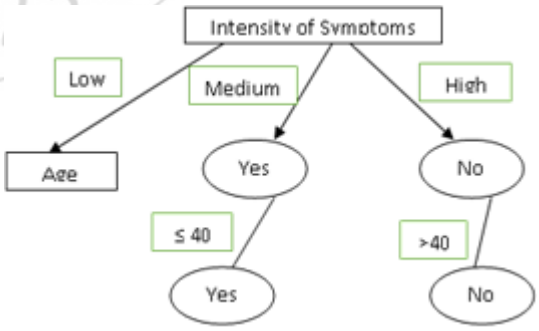

Figure 1: Decision representation

A decision tree can be used here as the classification to find a model that describe and distinguish data classes to predict an unknown class of data instance with the help of the above data set given in Table 1. The main purpose behind is to move the instance down to the tree by following the branches whose attributes values match the instances attribute values until the instance get a destination to the specific leaf node, whose class label is therefore assigned to the instance [13].

Let's consider this example of the data instances to be classified in the following features such as Age=38, Gender=female, Intensity of symptoms equal to medium, Goal =? ,where "?" represents our aim which is the unknown class label of the goal instance. In this example, Gender attribute is not relevant to a particular classification task. The tree tests the intensity of symptom value in the instance and checks if the requested answer is medium; then, the instance is pushed down through the relevant branch and reaches the Age node. And then the tree tests the Age value in the instance and checks if the answer is 23, the instance is again moved down through the relevant branch according to the condition statement. Finally, the instance gets the final destination of the leaf node where it is classified as yes which is our aim to be reached.

Neural networks Algorithm

Neural network algorithm is an artificial neural network that imitates the human brain activity of functioning when presented with a problem. It examined all the possible inputs and outputs combination and assigns importance to their relationships. It also considers the combination of the correlated inputs and outputs to an output even though the inputs alone do not. The inputs are not directly correlated to an output because of the hidden layer of nodes between the inputs and the outputs[14].

Prediction

The prediction is one of a data mining techniques used to discover the relationship between dependent and independent variables. The prediction here is to analyze the data and show the unknown, missing values or what will happen. In this study, the technique used for the research work is mostly based on heart disease prediction. In this case, the system will predict whether the patient has heart disease or does not have heart disease. Based on this prediction, an algorithm is implemented[15].

Heart disease

Heart disease is the type of disease that deals with the operations of the heart by narrowing or blockage of the arteries and vessels that supply oxygen and nutrient-rich. It is caused by the following factors:

- **The family history of heart disease (Heredity):** people should know that the heart disease can be inherited from the family when one of them is heart disease contractor.
- **Smoking:** Approximately 40% of persons die from tobacco due to heart attack and blood vessel diseases and a smoker's risk of heart attack can rapidly reduce within twelve (12) months of tobacco abstinence.
- **Cholesterol:** The increment of fats in the blood is a risk factor for heart diseases. Cholesterol is a substance consist of lipids in the bloodstream and all over the body's cells. High level of the fat in the body with high levels of LDL (low-density lipoprotein) cholesterol can rise up atherosclerosis which will lead to the risk increment of heart diseases.
- **High blood pressure:** High blood pressure also known as HBP or hypertension is a widely misunderstood medical condition. High blood pressure increases the risk of the walls of our blood vessels walls becoming overstretched and injured. Also, increase the risk of having a heart attack or stroke and of developing heart failure, kidney failure, and peripheral vascular disease.
- **Obesity:** the term obesity is used to describe the health condition of anyone significantly above his or her ideal healthy weight. Being obese puts anybody at a higher risk for health problems such as heart disease, stroke, high blood pressure, diabetes and more.
- **Lack of physical exercise:** lack of exercise is a risk factor for developing coronary artery disease (CAD). Lack of physical exercise increases the risk of CAD because it also increases the risk for diabetes and high blood pressure.

Heart diseases are generally provoked by the above-mentioned factors and the world health organization survey shows in 2012, an estimated 56 million people died worldwide, every year because of Heart Diseases [1].

4.2 Data Mining Techniques

Bayesian Theorem

The Bayesian Classification is a type of supervised learning and statistical method used for classification. The principle way of capturing uncertainty model is made by probabilistic classifications with the aim of determining probabilities of outcomes. Nearest Neighbor approaches are called lazy learners because when we give them a set of training data, they just basically save or remember the set while Bayesian methods are called eager learners. When given a training set eager learners immediately analyze the data and build a model. Eager learners tend to classify instances faster than lazy learners. The ability to make probabilistic classifications and the fact that they are eager learners are two advantages of Bayesian methods [16]. It can solve diagnostic and predictive problems.

Bayes Theorem is classification techniques, based on the relationship between P(h), P(h/D), P(D), and P(D/h) can be given as

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)}$$, where P(h/D) is the posterior probability of class (h, target) given predictor (D, attributes), P (h) is the prior probability of the class, P(D/h) is the likelihood which is the probability of predictor given class and P(D) is the prior probability of predictor. This theorem is the most important part of all Bayesian methods usually used in data mining to decide among alternative hypothesis.

Naïve Bayes

Naive Bayes is the machine learning techniques used to predict very large data sets. It is based on Bayes' theorem with the assumption of independence between predictors. In simple ways, Naïve Bayes classifier is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions[16]. It is favored when data is high, when the attributes are independent of each other and for more efficient output as compared to other methods output.

Algorithm

Given the hospital training data set

1. We need to maximize or estimate the prior probability P(cj) of each class which can be computed based on the training tuples.
2. For each attribute, xi finds P (xi) by computing the number of events occur in each tuple.
3. Find the probability P(Ci/xi) by computing how many times each value occur in the class in the training data set.
4. We have to apply this to all attributes and values of these attributes to classify a target tuple estimate P (ti/cj) = ∏Pk = 1 P(xik/cj).
5. We can compute P (ti) by finding the likelihood that this tuple is in each class and then adding all these values.
6. Find posterior probability P(cj/ti) for each class. This is the product of conditional probabilities for each attribute value.
7. The highest probability value of P(cj/ti) will be the value for test tuple.

Mathematical Expressions

P (Heart Disease Yes) = Number of Records with Result Yes / Total number of Records.

P (Heart Disease No) = Number of Records with Result No / Total number of Records

P (t/yes) = P (Age (low) yes) * P (Sex (Male) yes) * P (BP (High) yes) * P (Chol (High) yes) * P (Heart_Rate (High) yes) *P (Ca (High) yes) *P (Chest_Pain (High) yes) *P (ECG (High) yes) *P (Exer_angina (High) yes) *P(old_peak(High)yes)*P(Thal(High)yes)*P(Blood_sugar (High) yes) * P (Slope_peak (High) yes)

P (t/no) = P (Age (low) no) * P (Sex (Male) no) * P (BP (High) no)* P (Chol (High) no) * P (Heart_Rate (High) no) * P (Ca (High) no) * P (Chest_Pain (High) no) * P (ECG (High) no) * P (Exer_angina (High) no) * P (old_peak (High) no) * P (Thal (High) no) * P (Blood_sugar (High) no) * P (Slope_peak (High) no)

P (Likelihood of yes) = P (t/yes) * P (Heart_Disease yes)
P (Likelihood of no) = P (t/no) * P (Heart_Disease no)

Now we find the total probability,
P (yes/t) = P (t/yes) * P (Heart_Disease yes) / P (T)
P (no/t) = P (t/no) * P (Heart_Disease no) / P (T)
If P (yes/t) >= P (no/t) then input query is classified as Heart Disease category
Else No Heart Disease category

Accuracy Calculation

Accuracy can be determined by the percentage of correct predictions made by the model compared with actual classifications in the test data.
Accuracy = Total number of Correctly Predicted Record / Total number of training Record.

4.3 Support Vector Machines

Support vector machines (SVMs) can be described as the set of supervised learning techniques applied for classification, regression, outlier's detection and it is the current binary classification technique. SVM reduces simultaneously with the verifiable classification error and increases the geometric margin by notifying as the Maximum Margin classifiers with the use of kernel trick can be efficiently executed the non-linear classification. [17, 18].

The RBF (Radial Basis Function) kernel of SVM can be used as classifier due to the functionality of RBF kernel which can analyze high dimensional data, can map samples into a higher dimensional space and also due to the minority of numerical difficulties [18, 19]. The values are normalized from the initial stage in the SVM classifier to improve the accuracy and the Test data sets were used to access the performance of the SVM model.

The application of test data sets in the sense of validation is to avoid potential bias of the performance, estimate due to the over-fitting of the model to training data sets. Indeed, the Support Vector Machine classifier and the Radial Basis Function kernel are put into used for classification. The automated classifier system for the heart disease prediction between the patient with heart disease and without heart disease has been conceived using supervised learning algorithm called support vector machine which can provide the overall performance without any need of adding a prior knowledge. The main purpose of this algorithm is to discover the best classification function to rise up the disagreement between members of the two classes in the training data [20].

4.4 Systems Architectures

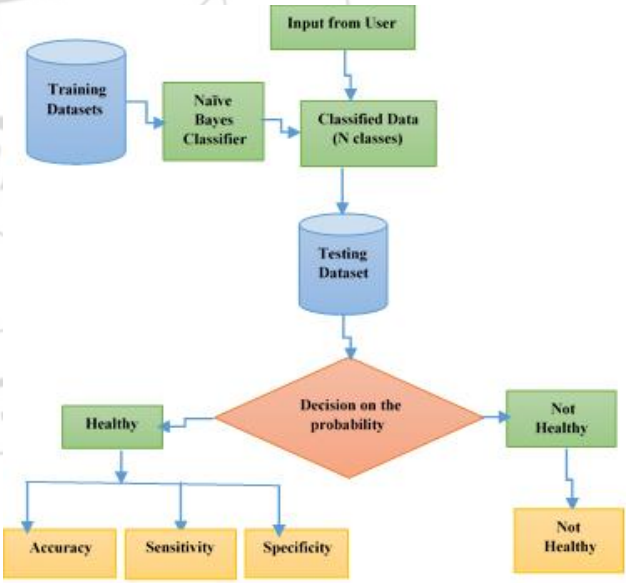

Figure 2: Decision support system with Naïve Bayes

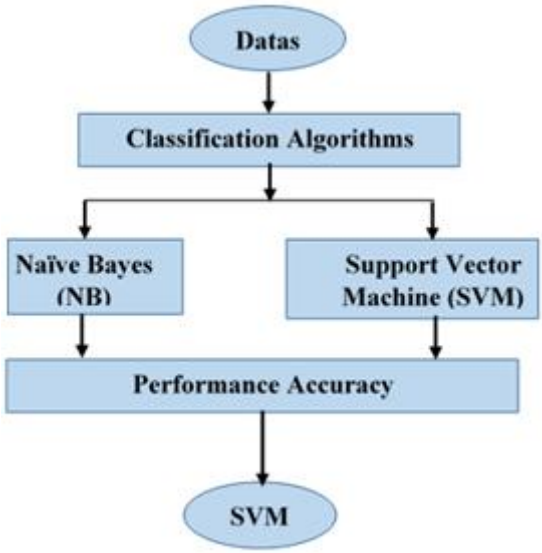

Figure 3: system architecture for Naïve Bayes and Support vector Machine (SVM)

5. Results and Discussion

The proposed web-based application system approach is implemented in Visual Studio 2012 tool developed by Microsoft. Visual studio is used to design the user interfaces with the programs written in C# language and the data analysis with python programming language.

Table2: Classification accuracy

Data Mining Techniques	Accuracy
Naïve Bayes	91.42%
Support Vector Machine	89.56%

The data in table 1 are the accuracy results of data analysis implemented in python.

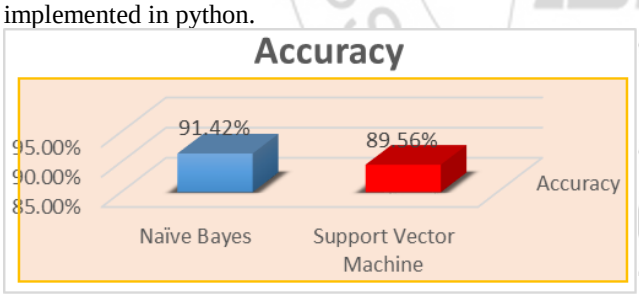

Figure 4: accuracy measure

This graph chart is formed by using the table 1 with Data Mining tools and Accuracy Level as illustrated in Fig. 1. In this chart, the prediction accuracy level of different data mining applications has been compared.

Figure 5: Application Home page

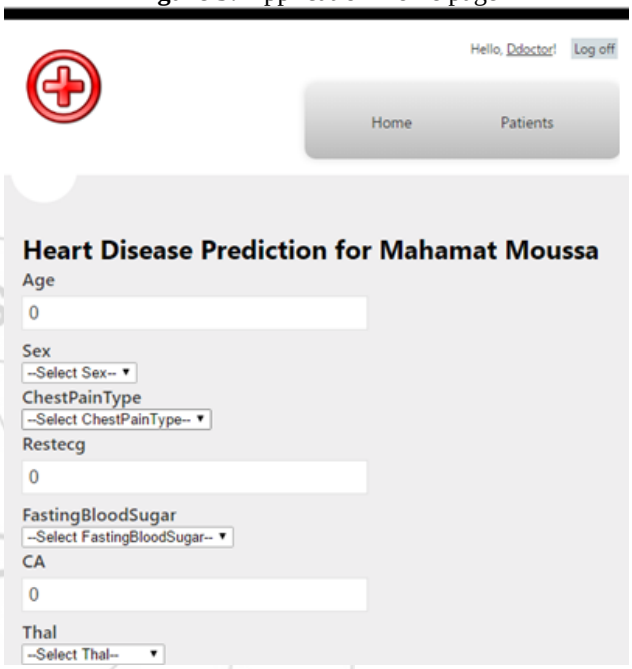

Figure 6: Data Input Page

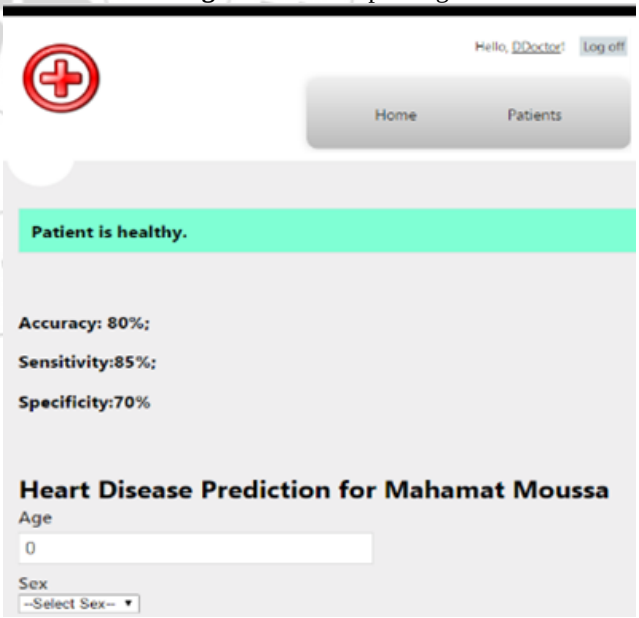

Figure 7: Sample of Heart Disease prediction showing that the patient is Healthy

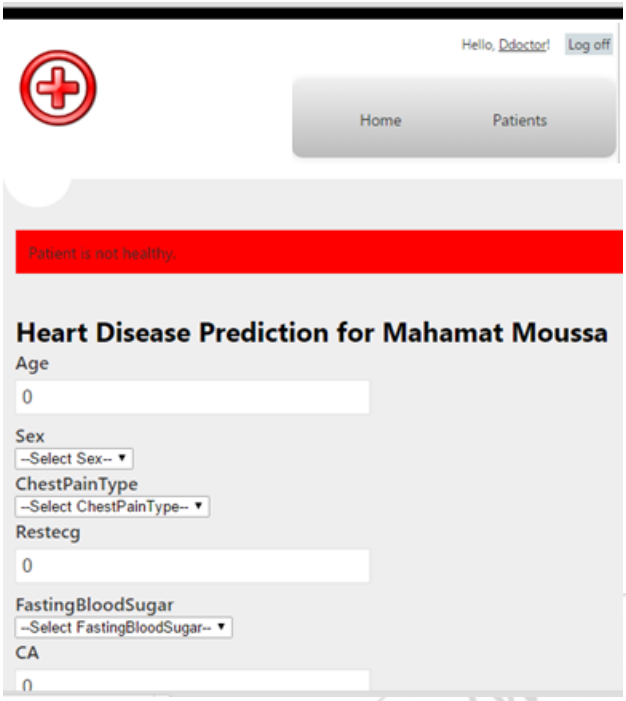

Figure 8: Sample of Heart Disease prediction showing that the patient is not healthy

6. Conclusions

This paper is conceived by the means of two data mining classification techniques in the sense to extract hidden knowledge from a historical heart disease database with respect to the application of data mining classification techniques by the mean of Naïve Bayes and Support Vector Machine. The training and validating models are tested against a test data set and the classification models are used to evaluate the effectiveness of the models. Besides that, we also developed a heart disease prediction system that helps medical professional to predict heart disease status based on the clinical of the patient using Naive Bayes.

From the experimental results, we can conclude that Naïve Bayes provides the accurate result compared to Support Vector Machine. In the future, this system can be further enhanced and expanded. It can also apply to analyze different datasets by just changing the name of the dataset file which is given for the training module.

7. Future Scope

The proposed web-based application system approach is a friendly user interface, scalable and trustful that can be adopted and implemented in remote areas to try to act like human diagnostic expertise for treatment of heart slight illness. The system can be enlarged in the sense that an acceptable number of records or attributes can be incorporated as well as the new significant rules can be generated using the applied techniques. For instance, the system has been using 14 attributes and 315 records from the hospitals (Clinique la Providence and Hôpital de Reference Nationale) databases. Considering the symptoms variation of a particular disease may vary according to the region, the system should be trained using local dataset collected from the clinic or hospital.

References

[1] A. N. Nowbar, J. P. Howard, J. A. Finegold, P. Asaria, and D. P. Francis, "2014 Global geographic analysis of mortality from ischaemic heart disease by country, age, and income: Statistics from World Health Organisation and United Nations," International journal of cardiology, vol. 174, pp. 293-298, 2014.
[2] A. Rajkumar and G. S. Reena, "Diagnosis of heart disease using data mining algorithm," Global journal of computer science and technology, vol. 10, pp. 38-43, 2010.
[3] S. Dhamodharan, "Liver Disease Prediction Using Bayesian Classification," 2014.
[4] A. Roslina and A. Noraziah, "Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method," in Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on, 2010, pp. 2209-2211.
[5] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," International Journal of Computer Applications, vol. 47, pp. 44-48, 2012.
[6] A. K. Sen, S. Patel, and D. Shukla, "A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level," International Journal Of Engineering And Computer Science ISSN, pp. 2319-7242, 2013.
[7] D. S. Medhekar, M. P. Bote, and S. D. Deshmukh, "Heart disease prediction system using naive Bayes," Int. J. Enhanced Res. Sci. Technol. Eng, vol. 2, 2013.
[8] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using Naïve Bayes," International Journal of Advanced Computer and Mathematical Sciences, vol. 3, pp. 290-294, 2012.
[9] M. P. S. Kulkarni, M. V. Belokar, S. Sane, and N. Bhale, "Heart Disease Prediction from Horoscope of a Person Using Data Mining."
[10] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, 2008, pp. 108-115.
[11] M. Hall, I. Witten, and E. Frank, "Data mining: Practical machine learning tools and techniques," Kaufmann, Burlington, 2011.
[12] R. Wu, W. Peters, and M. Morgan, "The next generation of clinical decision support: linking evidence to best practice," Journal of healthcare information management: JHIM, vol. 16, pp. 50-55, 2001.
[13] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques: Elsevier, 2011.
[14] J. MacLennan, Z. Tang, and B. Crivat, Data mining with Microsoft SQL server 2008: John Wiley & Sons, 2011.
[15] W. D. M. Primitives, "Data Mining: Concepts and Techniques."
[16] P. Harrington, Machine learning in action: Manning, 2012.

[17] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144-152.
[18] G. E. Güraksın, H. Haklı, and H. Uğuz, "Support vector machines classification based on particle swarm optimization for bone age determination," Applied Soft Computing, vol. 24, pp. 597-602, 2014.
[19] V. Vapnik, The nature of statistical learning theory: Springer Science & Business Media, 2013.
[20] C. J. Burges, "A tutorial on support vector machines for pattern recognition," Data mining and knowledge discovery, vol. 2, pp. 121-167, 1998.

Author Profile

Mr. **Allladoumbaye Ngueilbaye** received the B.Sc. degree in Computer Science from Ahmadu Bello University, Nigeria and currently an M.Sc. degree student in Computer Science and Technology at Harbin Institute of Technology, China in 2010 and July 2016 (Graduation), respectively. His areas of interest include Data Mining, Machine Learning, Natural Language Processing, and Databases Management System.

Lin Lei, Associate Professor, Department of Computer Science and Technology, Harbin Institute of Technology, China. His areas of interest include Data Mining, Machine Learning, Natural Language Processing and Computational Advertising.

Hongzhi Wang, Professor and Dr, Department of Computer Science and Technology, Harbin Institute of Technology, China. His areas of interest include Big Data, Data Mining, Machine Learning, Databases, and Information Retrieval