A Medical Document Classification System for Heart Disease Diagnosis Using Naïve Bayesian Classifier

D. J. S. Sako & J. Palimote
Department of Computer Science
Rivers State University,
Port Harcourt,
Nigeria
sunday.sako@ust.edu.ng

Abstract
One of the most challenging projects in information systems is extracting information from unstructured texts, including medical document classification. The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. The paper describes the process of extracting knowledge from information stored in dataset in order to generate clear and understandable description patterns. The Statlog (Heart) dataset, obtained from the UCI database, was used for the experiments with the simulations carried out using weka tool. Naïve Bayesian classifier was used to predict the absence or presence of heart disease in a patient. To evaluate the performance of the system, a confusion matrix, prediction accuracy, recall, and precision were used. The experimental results show that the system achieves very promising results in classifying heart disease patients.

Keywords: *Naïve Bayes, heart disease, weka, document classification, medical records, heart disease, and data mining.*

1. Introduction

Nowadays modern hospitals are well equipped with monitoring and other data collection devices resulting in enormous data which are collected continuously through health examination and medical treatment. All this led to the fact that medical area produces increasingly voluminous amount of electronic data which are becoming more complicated. The produced medical data have certain characteristics that make their analysis very challenging and attractive (Al-Aidaroos, Bakar & Othman, 2012). This overwhelmed amount of medical information available in the research literature, makes the use of automated information classification methods essential for both medical experts and novice users (Lakiotaki, Hliaoutakis, Koutsos, & Petrakis, 2013).

In the past, various statistical methods have been used for modelling in the area of disease diagnosis. These methods require prior assumptions and are less capable of dealing with massive and complicated nonlinear and dependent data (Lin, 2009). However data mining (and machine learning) has proven to be more powerful and effective and it provides processes for discovering useful patterns from large data sets (Thongkam, Xu, Zhang & Huang, 2008).

Data mining uses a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases (Rohanizadeha & Moghadama, 2009). Data mining must also be considered as an iterative process that requires goals and objectives to be specified (Han & Kamber, 2001). These data mining are generally classified into supervised and unsupervised models.

Heart disease is a term covering any disorder of the heart and is a primary cause of death. An estimated 17.7 million people died from cardiovascular diseases (CVDs) in 2015, representing 31% of all global deaths[1]. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects that one is born with (congenital heart defects), among others[2].

Numerous factors are involved in the diagnosis of heart disease, which complicates a physician's task (Liu, et al, 2017). To help physicians make quick decisions and minimize errors in diagnosis, classification systems enable physicians to rapidly examine medical data in considerable detail (Temurtas, Yumusak & Temurtas, 2009). These systems are implemented by developing a model that can classify existing records using sample data. Various classification algorithms have been developed and used as classifiers to assist doctors in diagnosing heart disease patients (Liu, et al, 2017).

In this paper, we are concerned with the supervised (ie classification) method which requires the data to include a special response attribute, known as the class attribute and therefore known as classification models. We describe the process of extracting knowledge from information stored in dataset in order to generate clear and understandable description patterns. Bayesian classifier was used to classify the absence or presence of heart disease in patients. The experiments and Simulation were performed with weka tool using the Statlog (Heart) dataset[3] from the UCI machine learning database. The performances of system are evaluated using a confusion matrix, prediction accuracy, and recall. The experimental results show that the system achieves very promising results.

2. Theoretical Background

2.1 Naïve Bayesian Classifications

Naïve Bayesian classifier, or simply put Naïve Bayes (NB) is one of the most effective and efficient classification algorithms for supervised learning. It is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions. The elegant simplicity and apparent accuracy of Naïve Bayes even when the independence assumption is violated, fosters the on-going interest in the model (Al-Aidaroos, Bakar & Othman, 2012).

Data classification with naïve Bayes is the task of predicting the class of an instance from a set of attributes describing that instance and assume that all the attributes are conditionally independent given the class. It has proven its effective application, in text classification, medical diagnosis and systems performance management.

Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $x = (x_1, \dots, x_n)$ representing some n features (independent variables), it assigns to this instance probabilities

$$p(C_k | x_1, \dots, x_n)$$

for each of K possible outcomes or *classes* C_k

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables

[1] http://www.who.int/mediacentre/factsheets/fs317/en/

[2] https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

[3] UCI Repository of Machine Learning Databases,http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29.

is infeasible. The model is, therefore, reformulated to make it more tractable. Using Bayes theorem, the conditional probability can be decomposed as[4]:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \qquad (1)$$

Where $p(C_k|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*), $p(C_k)$ is the prior probability of *class*, $p(x|C_k)$ is the likelihood which is the probability of *predictor* given *class* and $p(x)$ is the prior probability of *predictor*.

To construct a naïve Bayes classifier from the naive Bayes probability model, we combine this independent feature model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the *maximum a posteriori* or *MAP* decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label *E* for some *k*.

Therefore, given a set of training instances with class labels and a test case *E* represented by n attribute values ($x_1, x_2 \ldots x_n$), Naïve Bayes classifier uses the following equation to classify *E*:

$$C_{NB}(E) = argMax\, p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \qquad (2)$$

Where, $C_{NB}(E)$ denotes the classification given by Naïve Bayes (NB) on test case *E*. Based on this equation, each of the probabilities can be determined directly from the training data.

Compared to other classifiers, Naïve Bayes is simple, computationally efficient, requires relatively little data for training, do not have lot of parameters, is very transparent and is naturally robust to missing and noise data (Al-Aidaroos, Bakar, & Othman, 2010).

2.1 Medical Document Classification

In document classification or categorization, which is example of Machine Learning (ML), the task is to assign a document to one or more classes or categories. Broadly speaking, there are two classes of ML techniques: supervised and unsupervised. In supervised methods, a model is created based on a training set. Categories are predefined and documents within the training dataset are manually tagged with one or more category labels. A classifier is then trained on the dataset which means it can predict a new document's category from then on (Ghaffari, 2015).

Given a set of class, classifier determines which class a given object belongs to. This may be done manually or algorithmically. Classification is done mainly based on attributes, behavior or subjects. Classification techniques have been applied to spam filtering, email routing, language identification, etc. The Classification problem can be stated as a training data set consisting of records. Each record is identified by a unique record id, and consists of fields corresponding to the attributes. An attribute with a continuous domain is called a continuous attribute. An attribute with a finite domain of discrete values is called a categorical attribute. One of the categorical attribute is the classifying attribute or class and the value in its domain are called class labels. The objective is to use the training data set to build a model of the

[4] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

class label based on the other attributes such that the model can be used to classify new data not from the training data set attributes (Purohit, Atre, Jaswani & Asawara, 2015).

Kononenko (2001) considered NB as a benchmark algorithm that in any medical domain has to be tried before any other advanced method. While Abraham et al (2006) argued, based on their study, that simple methods are better in medical data mining and classification that is makes NB performs well for such data. Compared to other classifiers, NB is simple, computationally efficient, requires, relatively little data for training, do not have lot of parameters and is naturally robust to missing and noise data (Ai-Aidaroos et al, 2010).

One of the main advantages of NB approach which is appealing to physicians is that all the available information is used to explain the decision. This explanation seems to be “natural” for medical diagnosis and prognosis i.e. is close to the way how physicians diagnose patients (Zelic, Kononenko, Lavrac & Vug, 1997). When dealing with medical data, naïve bayes classifier takes into account evidence from many attributes to make the final prediction and provides transparent explanations of its decisions and therefore is considered as one of the most useful classifiers to support physicians’ decisions (Al-Aidaroos, Bakar, & Othman, 2012).

2.3 WEKA Tool

WEKA, a data mining and machine learning software tool, is used in this experiment. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. Weka is a collection of machine learning algorithms for solving real-world data mining problems. The algorithms can either be applied directly to a dataset or called from your own Java code (Holmes, Donkin &Witten, 2010).

3. Statement of Problem

To assign a medical document containing records of patients to one of the two classes or categories: identify whether a given person in a dataset will have heart disease or not based on the attribute values in the dataset.

The dataset contains the details of person like age, sex, chest pain (type 4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (with values 0,1,2), maximum heart rate achieved, exercise induced angina, old peak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy, that (3 = normal; 6 = fixed defect; 7 = reversible defect).

The general structure of the system is given in figure 1.

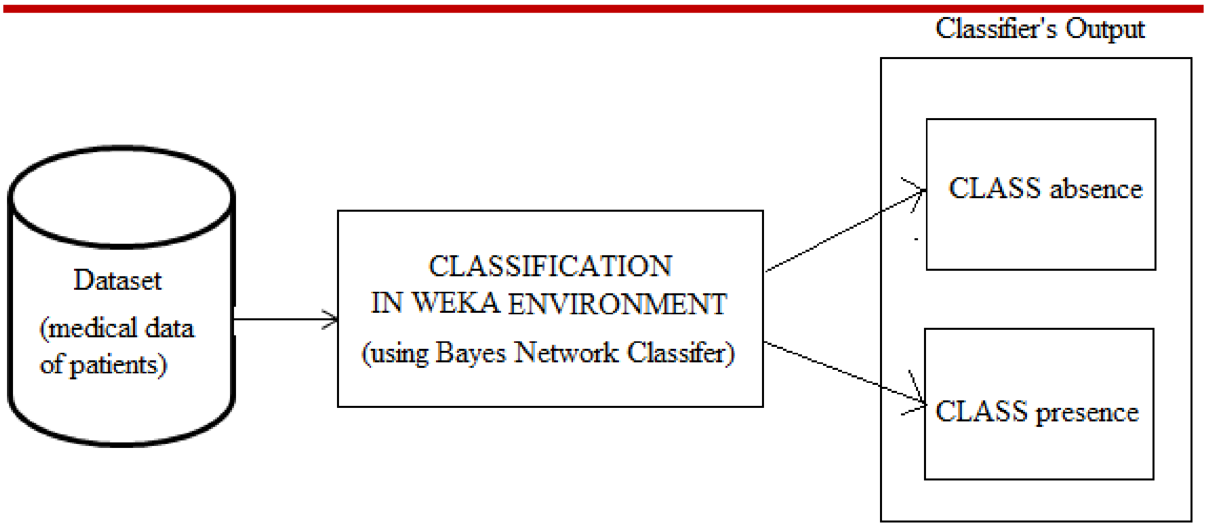


Figure 1: The classification System.

4. Experimental Results and Discussions

4.1 Dataset

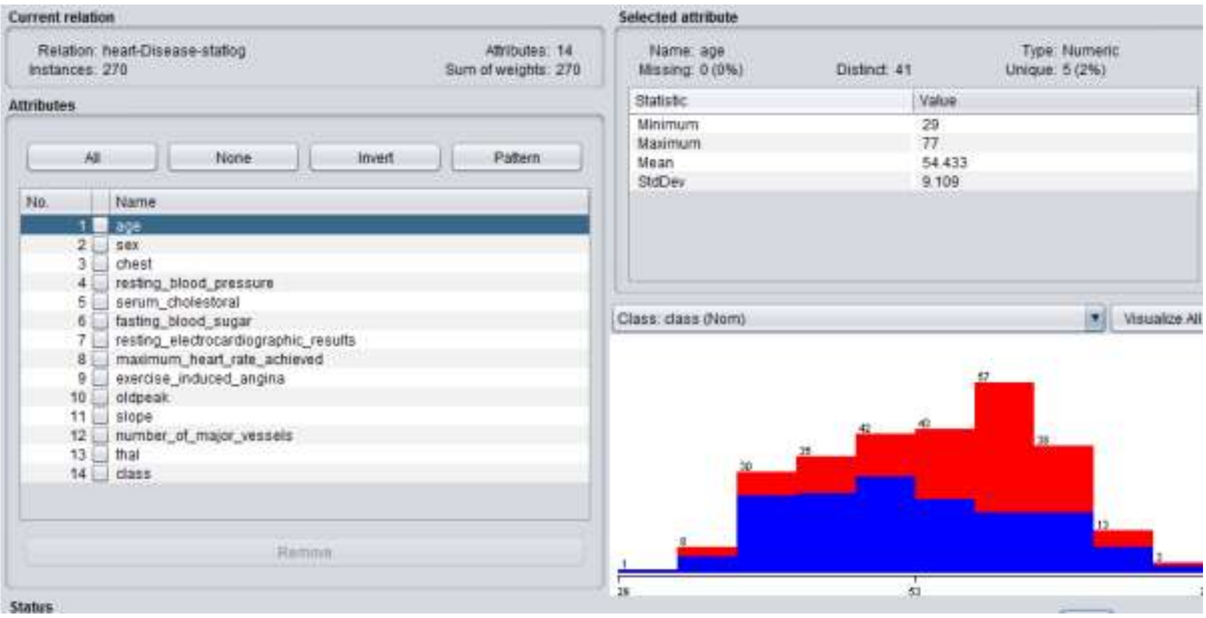
The Statlog (Heart) dataset used in our work was obtained from the UCI machine learning database. This dataset contains 270 observations and 2 classes: the presence and absence of heart disease given the results of various medical tests carried out on the patients. The samples include 13 condition features (attributes) which have been extracted from a larger set of 75, presented in Table 1. We denote the 13 features as *X1* to *X13* and one output attribute (Y) with two classes; absence or presence of heart disease.

Table 1: Attribute information of Statlog (Heart) dataset

Code	Attribute	Description	Type
X1	AGE	Age of Person (years)	Numeric
X2	Sex	Male or Female	Numeric
X3	chest pain type	chest pain type (4 values)	
X4	resting blood pressure	resting blood pressure	Numeric
X5	serum cholesterol in mg/dl	serum cholesterol in mg/dl	Numeric
X6	fasting blood sugar	fasting blood sugar > 120 mg/dl	Numeric
X7	resting electrocardiographic results	resting electrocardiographic results (values 0,1,2)	Numeric
X8	maximum heart rate achieved	maximum heart rate achieved	Numeric
X9	exercise induced angina	exercise induced angina	Numeric
X10	Old peak	Old peak = ST depression induced by exercise relative to rest	Numeric
X11	the slope of the peak exercise ST segment	the slope of the peak exercise ST segment	Numeric
X12	number of major vessels (0-3) colored by fluoroscopy	number of major vessels (0-3) colored by fluoroscopy	Numeric
X13	thal	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect	Numeric
Y	Class	Absence (1) or presence (2) of heart disease	Nominal

4.2 Experimental setup.

In this paper, WEKA[5] version 3.8.2 was used for the evaluation of the performance of Naïve Bayes in the classification of the medical document and was employed based its default parameters of the WEKA application. We used 10-fold cross-validation to minimize the bias associated with the random sampling of the training and holdout data samples (Kohavi, 1995). The dataset was classified into two classes: absence or presence of heart disease. Figures 2 and 3 show the main WEKA Explorer interface with the data file loaded for classification and the classifier output respectively.



Current relation

Relation: heart-Disease-statlog
Instances: 270
Attributes: 14
Sum of weights: 270

Attributes

No.	Name
1	age
2	sex
3	chest
4	resting_blood_pressure
5	serum_cholestoral
6	fasting_blood_sugar
7	resting_electrocardiographic_results
8	maximum_heart_rate_achieved
9	exercise_induced_angina
10	oldpeak
11	slope
12	number_of_major_vessels
13	thal
14	class

Selected attribute

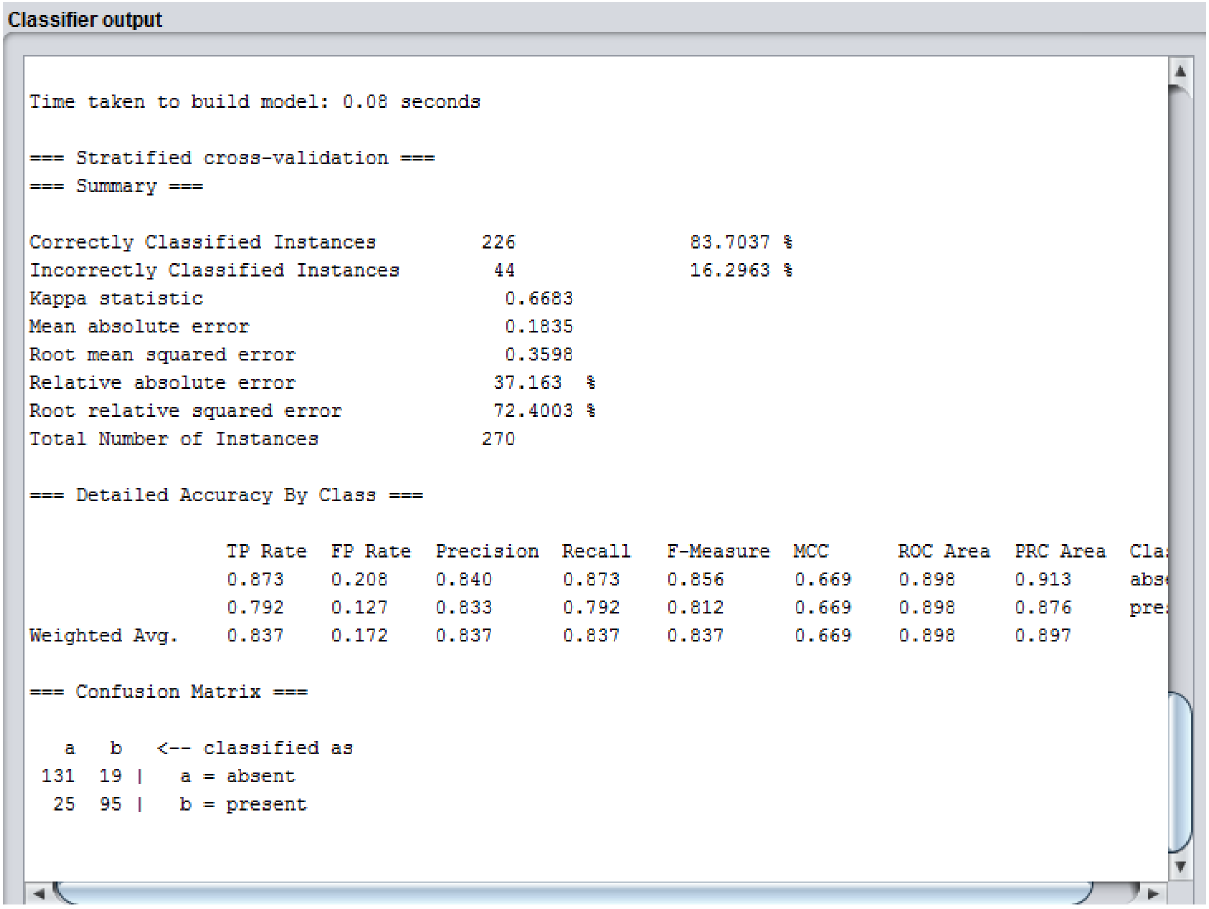
Name: age
Missing: 0 (0%)
Distinct: 41
Type: Numeric
Unique: 5 (2%)

Statistic	Value
Minimum	29
Maximum	77
Mean	54.433
StdDev	9.109

Class: class (Nom)

Figure 2: The main WEKA Explorer interface with the data file loaded

[5] https://www.cs.waikato.ac.nz/ml/weka/



```
Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         226               83.7037 %
Incorrectly Classified Instances        44               16.2963 %
Kappa statistic                          0.6683
Mean absolute error                      0.1835
Root mean squared error                  0.3598
Relative absolute error                 37.163  %
Root relative squared error             72.4003 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.873    0.208    0.840      0.873    0.856      0.669    0.898     0.913     abs
                 0.792    0.127    0.833      0.792    0.812      0.669    0.898     0.876     pre
Weighted Avg.    0.837    0.172    0.837      0.837    0.837      0.669    0.898     0.897

=== Confusion Matrix ===

   a   b   <-- classified as
 131  19 |   a = absent
  25  95 |   b = present
```

Figure 3: Classifier Output

Table 2 shows the results obtained for each of the classes: the predicted healthy persons (absent of heart disease) and the predicted patients with heart disease (present of heart disease)

Table 2: Results Obtained for each of the classes

Class	**Entries**	**% of Persons (Patients)**
Predicted healthy persons (ABSENT)	226	83.7037
Predicted patients with heart disease (PRESENT)	44	16.2963

4.3 Performance Evaluation

4.3.1 Confusion Matrix. We used the confusion matrix to describe the performance of our classification model (classifier). A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix (Kumari, Vohra & Arora, 2014). Table 3 shows the confusion matrix for a two class classifier

Table 3: Confusion matrix for our binary classifier

	Predicted healthy persons	Predicted patients with heart disease
Actual healthy persons	TP (=131)	FN (=19)
Actual patients with heart disease	FP (=25)	TN (=95)

Therefore, the entries in the confusion matrix have the following meaning in the context of our study. TP is the number of true positives; the positive tuples that were correctly labelled by the classifier, representing the cases with absent of heart disease that is correctly classified into the healthy class. FN is the number of false negatives; the positive tuples that were mislabelled as negative, representing cases with no heart disease that are classified into heart disease class. TN is the number of true negatives; the negative tuples that were correctly labelled by the classifier, representing heart disease cases that are correctly classified into the heart disease class. Finally, FP is the number of false positives; the negative tuples that were incorrectly labelled as positive, representing the heart disease cases that are incorrectly classified into the healthy class.

4.3.2 Evaluation Metrics: The performance of the system was evaluated based on predictive accuracy test, recall (sensitivity), and precision which use the true positive (TP), true negative (TN), false negative (FN), and false positive (FP) terms. These criteria are calculated as follows[6] given that number of Positives (P) = TP+FN and number of Negatives (N) = FP+TN.

Accuracy metric: Accuracy is calculated as the total number of all correct predictions (TP + TN) divided by the total number of the dataset (P + N).

$$Accuracy = \frac{TP + TN}{P + N} \qquad (3)$$

$$Accuracy = \frac{131+95}{270} = 0.83703$$

Recall metric: Recall (sensitivity or True positive rate) measures how completely the medical document classification system classifies the heart disease document. It is calculated as the number of correct positive predictions (TP) divided by the total number of positives (P).

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{P} \qquad (4)$$

$$Recall = \frac{131}{131 + 19} = 0.8733$$

Precision metric: Precision measures how exactly the Medical document classification system classifies the document. It is a positive predictive value and is calculated as the

[6] https://classeval.wordpress.com/introduction/basic-evaluation-measures/

number of correct positive predictions (TP) divided by the total number of positive predictions (TP + FP)

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Precision = \frac{131}{131 + 25} = 0.8397$$

4.4 Results

According to experimental results, correctly classified instances for Naïve Bayes classifier is 226. Classification with Naïve Bayes shows an accuracy of 0.83703 (ie 83.703%) in the classification of the medical document, which is high. It also scores 87.33% and 83.97% for recall and precision respectively. The mean absolute error (MEA) is 0.1835 and the root mean squared error (MRES) is 0.3598. These results show that the classifier is a promising technique in classifying heart disease patients.

5. Conclusion

This paper described the process of extracting knowledge from information stored in dataset in order to generate clear and understandable description patterns. Naïve Bayes classification approach has been discussed and its main features are highlighted based on the medical data classification requirements. The Statlog (Heart) dataset, obtained from the UCI database, was used for experiments with the simulation carried out using weka tool. Naïve Bayes classifier was used to predict the absence or presence of heart disease in a patient. It is clear that when dealing with medical data, naïve Bayes classifier takes into account evidence from many attributes to make the final prediction. To evaluate the performance of the system, a confusion matrix, prediction accuracy, recall, and precision were used. The experimental results show that the system achieves very promising results and hence the suitability of the classifier to the medical domain problem to support physicians' decisions.

References

Abraham, R., Simha, J.B. & Iyengar, S.S. (2006). A Comparative Analysis of Discretization Methods for Medical data Minining with Naïve Bayesian Classifier. Proceeding of the 9th International Conference on Information Technology (ICIT 08). Bhubaneswar. Pp: 235-236

Alghoson, A.M.(2014). Medical Document Classification Based on MeSH. 47th Hawaii International Conference on System Science.

Al-Aidaroos, K. M., Bakar, A. A. & Othman, Z. (2010). Naive Bayes Variants in Classification Learning. Proceeding of the International Conference in Information Retrieval & Knowledge Management,(CAMP 2010). Selangor. pp. 276-281.

Al-Aidaroos, K. M., Bakar, A. A. & Othman, Z. (2012). Medical Data Classification with Naïve Bayes Approach. Information Technology Journal. Asian Network for Scientific Information, 11(9).

Dasari, D.B. & Venu, G.R.K. (2012). Text Categorization and Machine Learning Methods: Current State of the Art. Global Journal of Computer Science and Technology Software & Data Engineering. Volume 12 Issue 11 Version 1.0.

Frank, E., Hall, M.A., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I.H.(2005). Weka: A Machine Learning Workbench for Data Mining. In O. Maimon & L. Rokach (Eds), Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers (pp. 1305-1314). Berlin: Springer.

Friedman,N., Geiger, D & Goldszmidt, M. (1997). Bayesian Network Classifiers. Machine Learning. Vol. 29. pp. 131-163.

Ghaffari, P. (2015). Text Analysis 101: Document Classification. https://www.kdnuggets.com/2015/01/text-analysis-101-document-classification.html [Accessed 15/12/2017]

Gope, H.L., Das, P.K. Islam, M.J. & Seddiqui, M.H. (2014)Medical Document Classification from OHSUMED Dataset. IJCSN International Journal of Computer Science and Network, Volume 3, Issue 4.

Han, J. & Kamber, M. (2001). Data Mining, Concepts and Techniques. Morgan Kaufmann Publishers,

Holmes, G., Donkin, A. & Witten, I.H. WEKA: A Machine Learning Workbench https://www.cs.waikato.ac.nz/ml/publications/1994/Holmes-ANZIIS-WEKA.pdf [Accessed 10/11/2017]

Hudson, D.L. & Cohen, M.E. (2002). Use of Intelligent Agents in the Diagnosis of Cardiac Disorders. Disorders Compu. Cardiol. 29:633-636.

Jain, S., Aalam, M.A. & Doja, M.N. (2010) K-Means Clustering using Weka Interface. Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, Bharati Vidyapeeth’s Institute of Computer Applications and Management.

Kohavi, R. (1995). A Study of Cross-validation and Bootstrap for Accuracy estimation and Model Selection. Proceeding of 14th International Joint Conference on Artificial Intelligence (IJCAI 95). Standfard. Pp. 1137-1143.

Kononenko, I. (2001). Machine Learning for Medical Diagnosis: History, State f the Art and Perspective. Artificial Intelligent Medicine. 23:89-109.

Kumara, M., Vohra, R. & Arora, A (2014). Prediction of Diabetes Using Bayesian Network. International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (4)

Lakiotaki, K., Hliaoutakis, A., Koutsos, S & Petrakis, E.G.M. (2013). Towards Personalized Medical Document Classification by Leveraging UMLS Semantic Network https://link.springer.com/chapter/10.1007/978-3-642-37899-7_8 [Accessed 12/11/17]

Li, Y.H. & Jain, A.K. (1998). Classification of Text Documents. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.100.7400&rep=rep1&type=pdf [Accessed 19/12/2017]

Lin, R.H. (2009). An Intelligent Model for Liver Disease Diagnosis. Artif. Intell. Med., 47: 53-62

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q. & Wang, Q. (2017). A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Computational and Mathematical Methods in Medicine. Hindawi Publishing Corporation http://dx.doi.org/10.1155/2017/8272091

Purohit, A., Atre, D., Jaswani, P. & Asawara, P. (2015). Text Classification in Data Mining. International Journal of Scientific and Research Publications, Volume 5, Issue 6.

Rohanizadeha, S.S. & Moghadama, M.B. (2009). A Proposed Data Mining Methodology and its Application to Industrial Procedures. Journal of Industrial Engineering 4 37-50 37

Thongkam, G., Xu, G., Zhang Y., & Huang, F. (2008). Breast Cancer Survivability via AdaBoost Algorithms. Proc. 2nd Aust. Workshop on Health Data Knowledge Management. 80:55-64.

Temurtas H., Yumusak N. & Temurtas F. (2009). A Comparative Study on Diabetes Disease Diagnosis using Neural Networks. Expert Systems with Applications. 36(4):8610–8615. doi: 10.1016/j.eswa.2008.10.032

Tomar, D. & Agarwal, S. (2014). Feature Selection Based Least Square Twin Support Vector

Machine for Diagnosis of Heart Disease. International Journal of Bio-Science and Bio-Technology, vol. 6, no. 2, pp. 69–82,

UCI Repository of Machine Learning Databases,http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29.

Zelic, I., Kononenko, I., Lavrac, N & Vug, V. (1997). Induction of Decision Tress and Bayesian Classification Applied to Diagnosis of Sport Injuries. J. Med. Syst., 21:429-444.

https://classeval.wordpress.com/introduction/basic-evaluation-measures/ [Accessed 12/01/2018]

https://www.cs.waikato.ac.nz/ml/weka/

https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118 [Accessed 12/01/2018]

https://en.wikipedia.org/wiki/Naive_Bayes_classifier [Accessed 12/01/2018]

http://www.who.int/mediacentre/factsheets/fs317/en/ [Accessed 12/01/2018]