Journal of Advances in Mathematics and Computer Science

24(2): 1-11, 2017; Article no.JAMCS.35610

Previously known as British Journal of Mathematics & Computer Science
ISSN: 2231-0851

Hypertension Prediction System Using Naive Bayes Classifier

Babajide O. Afeni[1*], Thomas I. Aruleba[1] and Iyanuoluwa A. Oloyede[1]

[1]Department of Computer Science, Joseph Ayo Babalola University, Ikeji - Arakeji, Nigeria.

Authors' contributions

This work was carried out in collaboration between all authors. Author BOA designed the study, wrote the protocol and the first draft of the manuscript. Author TIA did the statistical analysis of the study. Author IAO managed the literature searches of the study. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMCS/2017/35610
Editor(s):
(1) Huchang Liao, Business School, Sichuan University, China.
Reviewers:
(1) D. Senthilkumar, Anna University, India.
(2) Aurora Martínez Romero, Universidad Juárez del Estado de Durango, Mexico.
(3) M. Ilayaraja, Kalasalingam University, India.
Complete Peer review History: http://www.sciencedomain.org/review-history/20775

Original Research Article

Received: 21st July 2017
Accepted: 22nd August 2017
Published: 1st September 2017

Abstract

Hypertension is an illness that often leads to severe and life-threatening diseases such as heart failure, coronary artery disease, heart attack and other severe conditions if not promptly diagnosed and treated. Data Mining the use of a variety of techniques to smoothen information discovery or decision-making knowledge in the database and extracting these in a way that they can put to use in areas such as predictions, forecasting and estimation. This research has developed hypertension predictive system using data mining modelling technique, namely, Naïve Bayes. Medical profiles such as age, sex, blood pressure, chest pain and blood sugar it can predict the likelihood of patients getting a hypertension. This work was implemented in WEKA environment as an application which takes medical test's parameter as an input. The 10-fold cross validation method was used to train the developed predictive model and the performance of the models evaluated. This paper presents a model for predicting hypertension with 83.69%. The naïve Bayes' classifier proved to be an effective algorithm for predicting the diagnosis of hypertension in Nigerian patients. It can serve a training tool to train nurses and medical students to diagnose patients with hypertension.

Keywords: Data mining; Naive Bayes; hypertension; prediction.

**Corresponding author: E-mail: babajideafeni@gmail.com;*

1 Introduction

Hypertension disease is a significant health problem, and patients may not be able to recognize this disease for years. As a result, it may damage the patient's kidney, heart and veins. Hence, early diagnosis and hypertension treatment is very important. The reason for this importance is the damage caused by hypertension on organs, and high treatment costs and loss of labour that occur as a result [1]. A major challenge facing health care institutions is that the provision of quality services at reasonable prices. Quality service entails diagnosing patients properly and administering treatments that are effective [2]. Poor clinical choices will result in devastating consequences. Hospitals should also reduce the price of clinical tests. This can be achieved by using computer-based information or decision support systems. Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but are largely limited in the sense that they can't answer complex queries. However, they cannot answer complex queries like such as "Given patient records, predict the probability of patients getting hypertension." [3]. Most of the time, 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 unwarranted medical costs which affects the quality of service provided to patients. The anticipated system is aimed at reducing medical errors, enhance patient safety, decrease unwanted practice disparity, and improve patient outcome. This research work is presents data mining as a viable tool for generating a knowledge rich environment which can help to significantly improve the quality of clinical decisions. Data mining combines applied mathematics analysis, machine learning and information technology to extract hidden patterns and relationships from huge databases. In this, work Naive Bayes algorithm is employed to make a model with predictive capabilities. It provides new ways that of exploring and understanding knowledge. It learns from the "evidence" by calculating the correlation between the target (dependent) and different (independent) variables.

1.1 Research Aim and Objectives

The aim of this study is to develop a predictive model developed using Naïve Bayes classifier.

The specific objectives are to

a. identify key patterns from the dataset and select attributes that are more relevant in relation to hypertension prediction;
b. formulate the predictive model using Naïve Bayes model based on the attributes in (a);
c. simulate and validate the model in (b)

2 Literature Review

The various studies were conducted regarding the diagnosis of heart disease and some are given below.

Ture et al. [4] in their work compared the performances of classification techniques in order to estimate the risk of hypertension disease. Retrospective analysis was carried out on 694 data. 3 Decision Trees, 4 Statistical Algorithms and 2 Artificial Neural Networks (ANN) were compared. In their result, a Multilayer Perceptron (MLP) having an Artificial Neural Network showed a better performance in hypertension estimation than other methods. Rani [5] analyzed a heart disease dataset using neural network approach. Increase in the efficiency of the classification process parallel approach was also adopted in the training phase. Dangare et al. [6] developed a Heart Disease Prediction system using Neural network. The system predicts the likelihood of patient getting a Heart disease. For prediction, the system uses sex, blood pressure, cholesterol like 13 medical parameters. It added two more parameters are added i.e. obesity and smoking for better accuracy. From the results, it has been seen that neural network predicts heart disease with nearly 80% accuracy. A classification approach was introduced by Jabbar et al. [7] which uses ANN and feature subset selection for the classification of heart disease. The approach was applied on Andhra Pradesh heart disease data base. The results show that accuracy improved over traditional classification techniques. Waghulde and Patil [8] developed a Heart Disease Prediction System using Neural Network and Genetic Algorithm. The system calculates the number of hidden nodes for neural network which train the network with proper selection of neural network architecture and uses the global optimization of genetic algorithm for initialization of neural network. For prediction, the system uses 12 parameters such as sex, age, blood cholesterol etc. From the result, it is found that genetic neural approach predicts the heart disease better. Florence et al. [9], proposes the system which uses neural network and Decision tree (ID3) to predict the heart attacks. Here the dataset with 6 attributes is used to diagnose the heart attacks. The dataset used is acute heart attack dataset provided by UCI machine learning repository. The results of the prediction give more accurate output than the other techniques. Aljumah and Siddiqui [10] in their work computed the probability and prediction of hypertension using data mining techniques and concluded that smoking cessation is the best intervention followed by exercise, diet, weight and drug for the hypertension intervention in Saudi Arabia. Hence, all hypertension patients are unambiguously advised to stop smoking.

Awang and Siraj [11], assessed the application of artificial neural network in predicting the presence of heart disease, mainly the angina in patients. The prediction and detection of angina are significant in determining the most appropriate form of treatment for these patients. The best network model produced prediction accuracy of 88.89 percent. As the pilot project, the application developed could be used as the starting point in building a medical decision support system, particularly in diagnosing the heart disease. In [12], Kokyer in his work created hypertension database belonging to patients who arrived at hospital in different times which includes: age, sex, body mass index, HDL, LDL, triglyceride, uric acid, smoking and whether that person has hypertension or not; and the data were analyzed through Decision Table and Random Forest algorithms, which are data mining classification algorithms. In this way, a system able to predict whether or not hypertension patient candidates are hypertension was developed.

3 Methods

The methodological approach of this study is composed of: identification of the required variables for the diagnosis of hypertension, the collection of historical datasets about hypertension risk cases about patients, formulation of the predictive models using the supervised machine learning algorithm, the simulation of the predictive models using the WEKA simulation environment and the performance evaluation metrics applied during model validation for the evaluation of the performance of the predictive models.

3.1 Data Collection

For the purpose of this study, data was collected from 52 patients undergoing treatment at a hospital located in the south-western part of Nigeria from hospital case files following the processing of health records' ethical clearance. The information collected from the hospital was collected and stored in a spreadsheet application – Microsoft Excel of the Microsoft Office 2013. Information collected from the patients contained the explanatory variables for the diagnosis of hypertension as proposed by the cardiologist for each patient. A description of the attributes contained in the dataset is presented in Table 1.

3.2 Data-preprocessing

Following the collection of data from the 52 patients alongside the attributes (10 risk factors) alongside the diagnosis of hypertension, the data collected was checked for the presence of error in data entry including misspellings and missing data. Following this process, there was no error in misspellings but there were missing data in the cells describing the some records for the attributes chest pain and exang.

The data was transformed into the attribute file format (.arff) for the purpose of the development of the predictive model for hypertension risk using the simulation environment. Fig. 1 shows a screenshot of the format of the .arff used for model development in the Waikato Environment for Knowledge Analysis (WEKA) – a light-weight java application composed of a suite of supervised and unsupervised machine learning tools. The dataset collected for the purpose of the development of the predictive model for the diagnosis of hypertension was stored in .arff in the name *hypertensionData.arff* while the number of attributes listed in the attribute section were 11 including the target attribute. Following this, the values of the risk factors for the record of the 52 patients considered for this study was provided.

Table 1. Identified variables for diagnosis hypertension

S/N	Variable names	Labels
1.	Age	Numeric
2.	Sex	Male, Female
3.	Chest Pain	NAP, TTA, T1, Asymptotic
4.	Systolic Blood Pressure	Numeric
5.	Diastolic Blood Pressure	Numeric
6.	Cholesterol	Numeric
7.	Fasting Blood Sugar	Numeric
8.	Thalach	Numeric
9.	Exang	Yes, No
10.	Old peak	Numeric
11.	Diagnosis of hypertension	Yes, No

```
@relation hypertensionData

@attribute Age numeric
@attribute Sex {M,F}
@attribute Chest_Pain {NAP,TTA,Nil,T1,Asymptomtic}
@attribute Systolic_BP numeric
@attribute Diastolic_BP numeric
@attribute Cholesterol numeric
@attribute fasting_blood_sugar numeric
@attribute thalach numeric
@attribute exang {Yes,No}
@attribute old_peak numeric
@attribute Diagnosis {Yes,No}

@data
39,F,NAP,220,150,5.8,204.4,90,No,32,Yes
81,M,TTA,160,90,3.2,144,80,Yes,18,No
60,F,Nil,170,120,4.1,304.4,150,No,60,Yes
71,M,Nil,110,80,4.6,182.2,106,No,34,Yes
50,F,T1,220,110,4.8,?,120,Yes,42,Yes
90,F,TTA,140,90,?,?,80,Yes,25,No
34,F,T1,180,120,2.8,224,100,No,60,No
75,M,NAP,160,80,?,?,62,No,24,Yes
55,M,Nil,120,100,?,124.4,72,No,94.5,Yes
59,M,T1,90,70,5.7,290,25.8,No,32,Yes
85,F,Asymptomtic,160,80,4.3,208,80,Yes,30,No
65,F,?,150,90,3.8,304.2,100,?,24,Yes
68,F,NAP,160,100,4.4,292,60,Yes,40,No
36,F,T1,210,110,4.2,190,56,No,34,Yes
73,F,?,120,75,4.72,142,80,No,28,No
33,M,Nil,90,60,4.4,100,100,No,42,Yes
42,F,Asymptomtic,180,100,4.8,180,78,Yes,64.5,Yes
65,F,?,140,80,?,132,132,?,62,Yes
75,M,Nil,160,100,5.5,112,90,No,23,Yes
38,F,Asymptomtic,130,80,?,300,110,Yes,18,Yes
60,F,T1,90,40,2.3,100,100,No,40,Yes
69,M,TTA,80,60,2.6,105,108,Yes,36,No
86,F,Nil,120,90,4,175,100,No,64,Yes
72,M,T1,110,90,4.2,185,80,No,34,Yes
76,M,?,150,90,?,160,62,?,?,Yes
91,M,TTA,120,100,5.4,145,?,Yes,?,No
53,F,Nil,110,90,4.6,144,?,No,?,No

Normal text file    length : 2,279    lines : 65    Ln : 38    Col : 36    Sel : 0 | 0    Windows (CR LF)    ANSI    INS
```

Fig. 1. arff file containing identified attributes

3.3 Model formulation

Supervised machine learning algorithms are Black-boxed models, thus it is not possible to give an exact description of the mathematical relationship existing among the independent variables (input variables) with respect to the target variable (output variable – diagnosis of hypertension). Cost functions are used by supervised machine learning algorithms to estimate the error in prediction during the training of data for model development. Although, the decision trees algorithm is a white-boxed model owing to its ability of been interpreted as a tree-structure.

3.3.1 Naïve Bayes' classifier

Naive Bayes' Classifier is a probabilistic model based on Baye's theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data.

Let X_{ij} be a dataset sample containing records (or instances) of i number of risks factors (attributes/features) alongside their respective diagnosis of hypertension, C (target class) collected for j number of records/patients and $H_k = \{H_1 = Yes, H_2 = No\}$ be a hypothesis that X_{ij} belongs to class C. For the classification of the risk of hyperthension given the values of the risk factor of the jth record, Naïve Bayes' classification required the determination of the following:

- $P(H_k|X_{ij})$ – Posteriori probability: is the probability that the hypothesis, H_k holds given the observed data sample X_{ij} for $1 \leq k \leq 2$.
- $P(H_k)$ - Prior probability: is the initial probability of the target class $1 \leq k \leq 2$;
- $P(X_{ij})$ is the probability that the sample data is observed for each risk factor (or attribute), i; and
- $P(|X_{ij}|H_k)$ is the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X_{ij}.

Therefore, the posteriori probability of an hypothesis H_k is defined according to Bayes' theorem as follows:

$$P(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P\left(X_{ij}|H_k\right)P(X_i)}{P(H_k)} \quad for\ k = 1,2 \tag{1}$$

Hence, the risk of hypertension for a record is thus:

$$max.\left[P\left(H_1|X_{ij}\right), P\left(H_2|X_{ij}\right)\right] \tag{2}$$

3.4 Performance evaluation

In order to evaluate the performance of the Naïve Bayes algorithm used for the classification of the diagnosis of hypertension, there was the need to plot the results of the classification on a confusion matrix (Fig. 2). A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the vertical. All correct classifications lie along the diagonal from the north-west corner to the south-east corner also called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN). In this study, the likely cases are considered as the positive case while the unlikely and probable cases are the negative cases; the definitions are presented as follows:

a. True positives (TP) are correctly classified Yes cases;
b. False positives (FP) are incorrectly classified No cases;
c. True negatives (TN) are correctly classified No cases; and
d. False negatives (FN) are incorrectly classified Yes cases.

The true positive/negative and false positive/negative values recorded from the confusion matrix can then be used to evaluate the performance of the prediction model. A description of the definition and expressions of the metrics are presented as follows:

a. True Positive (TP) rates (sensitivity/recall) – proportion of positive cases correctly classified.

$$TP\ rate_{Yes} = \frac{TP}{TP + FN} \tag{3}$$

$$TP\ rate_{No} = \frac{TN}{FP + TN} \tag{4}$$

b. False Positive (FP) rates (1-specificity/false alarms) – proportion of negative cases incorrectly classified as positives.

$$FP\ rate_{Yes} = \frac{FP}{FP + TN} \tag{5}$$

$$FP\ rate_{No} = \frac{FN}{TP + FN} \tag{6}$$

c. Precision – proportion of predicted positive/negative cases that are correct.

$$Precision_{Yes} = \frac{TP}{TP + FN} \tag{7}$$

$$Precision_{No} = \frac{TN}{TN + FP} \tag{8}$$

d. Accuracy – proportion of the total predictions that was correct.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

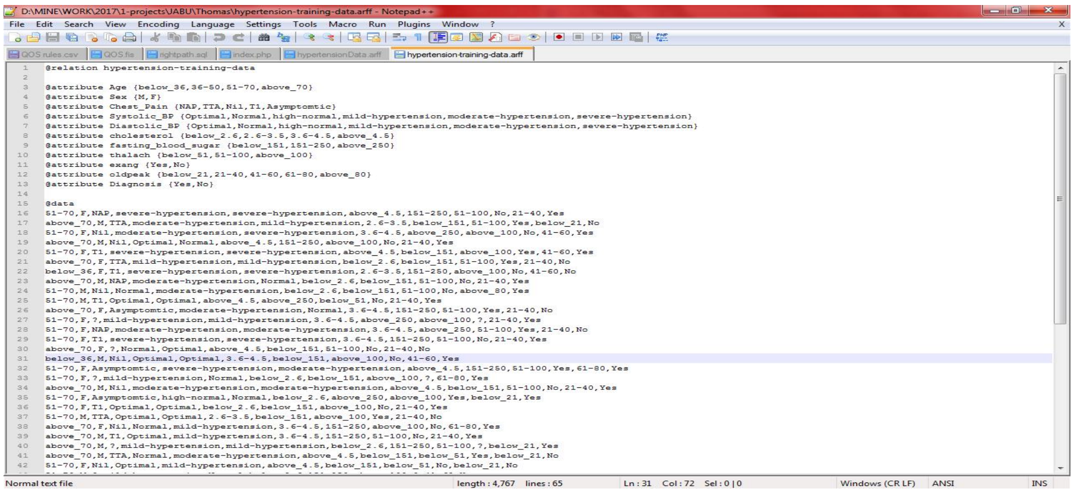
Yes	No	<- Predicted as
TP	FN	Yes
FP	TN	No

Fig. 2. Diagram of a confusion matrix

4 Results

The analysis of the data containing information about the attributes for the 52 patients are shown in Tables 2 and 3. Table 2 shows the description of the nominal variables while Table 3 shows the distribution of the numeric variables. From the description shown in Table 2, there were more female than male respondents owing for a ratio of 1:1.33 for men to women. The number of records diagnosed of hypertension consisted of 67.3% of the dataset while the remaining consisting of those without hypertension. Chest pain and exang had missing values of 7 and 6 respectively representing about 14.3% and 12.2% of the data values for each variables respectively.

From the description shown in Table 3, the analysis of the numeric datasets is presented showing the values of the minimum, maximum, mean and standard deviation of each variable presented in the dataset. Following the description of the numeric dataset, the numeric dataset were discretized into nominal datasets by creating intervals to which classes were defined. Table 4 shows a description of the discretization of the numeric datasets into nominal datasets while Fig. 3 shows a diagram of the arff file for the new training data stored in the file *hypertension_training_data.arff.*



```
@relation hypertension-training-data

@attribute Age {below_36,36-50,51-70,above_70}
@attribute Sex {M,F}
@attribute Chest_Pain {NAP,TTA,Nil,T1,Asymptomtic}
@attribute Systolic_BP {Optimal,Normal,high-normal,mild-hypertension,moderate-hypertension,severe-hypertension}
@attribute Diastolic_BP {Optimal,Normal,high-normal,mild-hypertension,moderate-hypertension,severe-hypertension}
@attribute cholesterol {below_2.6,2.6-3.5,3.6-4.5,above_4.5}
@attribute fasting_blood_sugar {below_151,151-250,above_250}
@attribute thalach {below_51,51-100,above_100}
@attribute exang {Yes,No}
@attribute oldpeak {below_21,21-40,41-60,61-80,above_80}
@attribute Diagnosis {Yes,No}

@data
51-70,F,NAP,severe-hypertension,severe-hypertension,above_4.5,151-250,51-100,No,21-40,Yes
above_70,M,TTA,moderate-hypertension,mild-hypertension,2.6-3.5,below_151,51-100,Yes,below_21,No
51-70,F,Nil,moderate-hypertension,severe-hypertension,3.6-4.5,above_250,above_100,No,41-60,Yes
above_70,M,Nil,Optimal,Normal,above_4.5,151-250,above_100,No,21-40,Yes
51-70,F,T1,severe-hypertension,severe-hypertension,above_4.5,below_151,above_100,Yes,41-60,Yes
above_70,F,TTA,mild-hypertension,mild-hypertension,below_2.6,below_151,51-100,Yes,21-40,No
below_36,F,T1,severe-hypertension,severe-hypertension,2.6-3.5,151-250,above_100,No,41-60,No
above_70,M,NAP,moderate-hypertension,Normal,below_2.6,below_151,51-100,No,21-40,Yes
51-70,M,Nil,Normal,moderate-hypertension,below_2.6,below_151,51-100,No,above_80,Yes
51-70,M,T1,Optimal,Optimal,above_4.5,above_250,below_51,No,21-40,Yes
above_70,F,Asymptomtic,moderate-hypertension,Normal,3.6-4.5,151-250,51-100,Yes,21-40,No
51-70,F,?,mild-hypertension,mild-hypertension,3.6-4.5,above_250,above_100,?,21-40,Yes
51-70,F,NAP,moderate-hypertension,moderate-hypertension,3.6-4.5,above_250,51-100,Yes,21-40,No
51-70,F,T1,severe-hypertension,severe-hypertension,3.6-4.5,151-250,51-100,No,21-40,Yes
above_70,F,?,Normal,Optimal,above_4.5,below_151,51-100,No,21-40,No
below_36,M,Nil,Optimal,Optimal,3.6-4.5,below_151,above_100,No,41-60,Yes
51-70,F,Asymptomtic,severe-hypertension,moderate-hypertension,above_4.5,151-250,51-100,Yes,61-80,Yes
51-70,F,?,mild-hypertension,Normal,below_2.6,below_151,above_100,?,61-80,Yes
above_70,M,Nil,moderate-hypertension,moderate-hypertension,above_4.5,below_151,51-100,No,21-40,Yes
51-70,F,Asymptomtic,high-normal,Normal,below_2.6,above_250,above_100,Yes,below_21,Yes
51-70,F,T1,Optimal,Optimal,below_2.6,below_151,above_100,No,21-40,Yes
51-70,M,TTA,Optimal,Optimal,2.6-3.5,below_151,above_100,Yes,21-40,No
above_70,F,Nil,Normal,mild-hypertension,3.6-4.5,151-250,above_100,No,61-80,Yes
above_70,M,T1,Optimal,mild-hypertension,3.6-4.5,151-250,51-100,No,21-40,Yes
above_70,M,?,mild-hypertension,mild-hypertension,below_2.6,151-250,51-100,?,below_21,Yes
above_70,M,TTA,Normal,moderate-hypertension,above_4.5,below_151,below_51,Yes,below_21,No
51-70,F,Nil,Optimal,mild-hypertension,above_4.5,below_151,below_51,No,below_21,No
```

Fig. 3. arff file containing identified attributes after data pre-processing

Table 2. Description of the nominal variables in the dataset

Variables	Labels	Frequency (%)
Sex	Male	21 (42.9)
	Female	28 (57.1)
Chest Pain	NAP	3 (6.1)
	TTA	9 (18.3)
	Nil	9 (18.3)
	T1	14 (28.6)
	Asymptotic	7 (14.3)
	Missing	7 (14.3)
Exang	Yes	17 (34.7)
	No	26 (53.1)
	Missing	6 (12.2)
Risk of hypertension	Yes	33 (67.3)
	No	16 (32.7)

Table 3. Description of the numeric variables in the dataset

Variables	Minimum	Maximum	Mean	Standard deviation
Age	33.00	91.00	63.878	15.83
Systolic BP	80.00	2220.00	137.14	33.79
Diastolic BP	40.00	150.00	87.143	20.08
Cholesterol	2.30	5.80	4.20	0.92
Fasting Blood Sugar	100.00	305.00	182.26	64.33
Thalach	25.80	150.00	92.97	24.55
Old Peak	16.00	94.50	39.06	18.26

4.1 Simulation results and discussion

Naïve Bayes which is a supervised machine learning algorithm was used to formulate the predictive model for the diagnosis of hypertension. The simulation of the prediction models was done using the Waikato Environment for Knowledge Analysis (WEKA). The naïve Bayes' algorithm was implemented using the naïve Bayes' classifier available in the Bayes class all available on the WEKA environment of classification tools. The models were trained using the 10-fold cross validation method which splits the dataset into 10 subsets of data – while 9 parts are used for training the remaining one is used for testing; this process is repeated until the remaining 9 parts take their turn for testing the model.

Table 4. Description of the discretized numeric variables in the dataset

Variables	Labels	Frequency (%)
Age	Below 36	2 (4.1)
	36 – 50	0 (0.0)
	51 – 70	29 (59.1)
	Above 70	19 (38.8)
Systolic blood pressure	Optimal	12 (24.5)
	Normal	7 (14.3)
	High Normal	7 (14.3)
	Mild Hypertension	9 (18.4)
	Moderate Hypertension	7 (14.3)
	Severe Hypertension	7 (14.3)
Diastolic blood pressure	Optimal	10 (20.4)
	Normal	13 (26.5)
	High Normal	0 (0.0)
	Mild Hypertension	11 (22.4)
	Moderate Hypertension	7 (14.3)
	Severe Hypertension	8 (16.5)
Cholesterol	Below 2.6	13 (26.5)
	2.6 – 3.5	7 (14.3)
	3.6 – 4.5	17 (34.7)
	Above 4.5	12 (24.5)
Fasting blood sugar	Below 151	21 (42.9)
	151 – 250	20 (40.8)
	Above 250	8 (16.3)
Thalach	Below 51	5 (10.2)
	51 – 100	23 (46.9)
	Above 100	21 (42.9)
Old Peak	Below 21	14 (28.6)
	21 – 40	20 (40.8)
	41 – 60	10 (20.4)
	61 – 80	4 (8.2)
	Above 80	1 (2.1)

Yes	No	<- Predicted as
31	2	Yes
6	10	No

Fig. 4. Confusion matrix for the result of naïve Bayes' classifier

Using the naïve Bayes' classifier to train the predictive model developed using the training data via the 10-fold cross validation method, it was discovered that there were 41 (83.67%) correct classifications (31 for Yes and 10 for No – along the diagonal) and 8 (28.21%) incorrect classifications 6 for Yes and 2 for No – along the vertical) as shown in Fig. 4. Hence, the predictive model for the risk of hypertension using the naïve Bayes' classifier showed an accuracy of 83.67%.

From the information provided by the confusion matrix, it was discovered that out of the 33 Yes cases, 31 were correctly classified with 2 misclassified as No and out of the 16 No cases, 10 were correctly classified while 6 were misclassified as Yes cases.

Table 5 shows the results of the evaluation of the performance of the naïve Bayes' classifier using the metrics. Based on the results presented for the naive Bayes' classifier, the TP rate of the model was better for the Yes cases than for the No cases thus the model has the ability to predict the Yes better than the no cases (an average of 83.7% of actual cases); the FP rate for the No cases were better than that of the Yes cases since the model did not misclassify the Yes for No cases like it did for the No for Yes cases (an average of 27.2% of the actual cases) while for the precision, the model performed very well in predicting the Yes and No cases since most of the predictions made by the model were correct (at least 83% of the predicted cases).

Table 5. Performance evaluation of the results of the naïve Bayes' classifier

Class	TP rate	FP rate	Precision	Area under the ROC
Yes	0.939	0.375	0.838	0.900
No	0.625	0.061	0.833	0.900
Average	0.837	0.272	0.836	0.900

4.2 Results of the C4.5 decision trees classifier

Using the C4.5 decision trees classifier to train the predictive model developed using the training data via the 10-fold cross validation method, it was discovered that there were 38 (77.55%) correct classifications (32 for Yes and 6 for No – along the diagonal) and 11 (22.45%) incorrect classifications (10 for Yes and 6 for No – along the vertical). Hence, the predictive model for the risk of hypertension using the C4.5 decision trees classifier showed an accuracy of 77.55%. Using the decision tree the following rules were deduced and can be used to predict the likelihood of hypertension given the values of the four identified risk factors. The rule can be read as follows:

a. IF (Cholesterol = "below 2.6") THEN (Hypertension Diagnosis = **Yes**)
b. IF (Cholesterol = "2.6 – 3.5") THEN (Hypertension Diagnosis = **No**)
c. IF (Cholesterol = "3.6 – 4.5") THEN (Hypertension Diagnosis = **Yes**)
d. IF (Cholesterol = "above 4.5") THEN (Hypertension Diagnosis = **Yes**)

Table 6 shows the results of the evaluation of the performance of the C4.5 decision trees classifier using the metrics. Based on the results presented for the C4.5 decision trees classifier, the TP rate of the model was better for the Yes cases than for the No cases thus the model has the ability to predict the Yes better than the no cases (an average of 77.6% of actual cases); the FP rate for the No cases were better than that of the Yes cases since the model did not misclassify the Yes for No cases like it did for the No for Yes cases (an average of 43.1% of the actual cases) while for the precision, the model performed very well in predicting the Yes and No cases since most of the predictions made by the model were correct (at least 76% of the predicted cases).

Table 7 gives a summary of the simulation results by presenting the average value of each performance metrics that was evaluated for the machine learning techniques used. The True positive rate (recall/sensitivity), false positive rate (false alarm/1-specificity), precision, accuracy and the area under the receiver operating characteristics (ROC) curve were used. From the table, it was discovered that the naïve

Bayes' algorithms showed the highest accuracy due to the ability to predict 41 out of the 52 records correctly. The true positive rate was also highest for the naïve Bayes' classifier. The naïve Bayes' also showed the lowest value for the false positive rate. The naïve Bayes' classifier had the highest value for the precision alongside the highest value for receiver operating characteristics (ROC) curve – a graph of the TP rate against the FP rate. The area under the graph is used to identify the level of relevance that can be given to the machine learning algorithm at making predictions – thus, the higher the value then the lower the bias of the model. The naïve Bayes classifier showed the best performance in the development of the predictive model for diagnosing hypertension.

Table 6. Performance evaluation of the results of the C4.5 decision trees' classifier

Class	TP rate	FP rate	Precision	Area under the ROC
Yes	0.970	0.625	0.762	0.735
No	0.375	0.030	0.857	0.735
Average	0.776	0.431	0/793	0.735

Table 7. Summary of simulation results

Metrics	Accuracy (%)	TP rate (recall)	FP rate (False alarm)	Precision	Area under ROC Curve (AUC)
Naïve Bayes'	83.67	0.837	0.272	0.836	0.900
Decision Trees	77.55	0.776	0.431	0.793	0.735

5 Conclusions

In this paper, the development of a predictive model for hypertension given the values of risk factors was developed using dataset collected from patients in a hospital in the south-western part of Nigeria. 10 variables were identified by cardiologist to be necessary in predicting hypertension for which a dataset containing information of 52 patients alongside their respective hypertension diagnosis (Yes and No) was also provided with 10 attributes following the identification of the required variables. After the process of data collection and pre-processing, naïve bayes classifier algorithm was used to develop the predictive model for the diagnosis of hypertension using the historical dataset from which the training and testing dataset was collected. The 10-fold cross validation method was used to train the predictive model developed using the machine learning algorithm and the performance of the model evaluated. It can be concluded that naïve bayes' classifier is an efficient algorithm for predicting the diagnosis of hypertension in Nigerian patients.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Türk F, Barişçi N, Çiftçi A, Ekmekçi Y. Comparison of multi-layer perceptron and Jordan Elman neural networks for diagnosis of hypertension. Intelligent Automation & Soft Computing. 2015;21(1).

[2] Garima S, Kiran B, Shivani S, Shraddha S, Sulochana D. Heart disease prediction using Naïve Bayes. International Research Journal of Engineering and Technology (IRJET). 2017;04(03).

[3] Subbalakshmi G, Ramesh K, Chinna R. Decision support in heart disease prediction system using Naive Bayes. Indian Journal of Computer Science and Engineering (IJCSE). 2011;2(2).

[4] Ture M, Kurt I, Kurum T, Ozdamar K. Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications. 2005;29:583-588.

[5] Rani K. Analysis of heart diseases dataset using neural network approach. International Journal of Data Mining & Knowledge Management Process (IJDKP). 2011;1(5).

[6] Dangare S, Apte S. A data mining approach for prediction of heart disease using neural networks. International Journal of Computer Engineering and Technology (IJCET). 2012;3(3):30-40.

[7] Jabbar M, Deekshatulu B, Chandra P. Classification of heart disease using artificial neural network and features subset selection. Global Journal of Computer Science and Technology Neural & Artificial Intelligence. 2013;13(3).

[8] Waghulde N, Patil N. Genetic neural approach for heart disease prediction. International Journal of Advanced Computer Research. 2014;4(3).

[9] Florence S, Amma B, Annapoorani G, Malathi K. Predicting the risk of heart attacks using neural network and decision tree. International Journal of Innovative Research in Computer and Communication Engineering. 2014;2(11).

[10] Aljumah A, Siddiqui M. Hypertension interventions using classification based data mining. Research Journal of Applied Sciences, Engineering and Technology. 2014;7(17):3593-3602.

[11] Awang M, Siraj F. Utilization of an artificial neural network in the prediction of heart disease. International Journal of Bio-Science and Bio-Technology. 2013;5(4).

[12] Kokver Y., Barisci N., Unver H., Ciftci A., Data Mining Classification on Hypertension Database. Proceedings of The IRES 21st International Conference, Amsterdam, Netherland, 25th December 2015.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/20775