IJBPAS, January, 2023, 12(1): 475-483

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

PREDICTION OF ISCHEMIC HEART DISEASES WITH NAÏVE BAYES CLASSIFIER IN R STUDIO

CHAITHRA N[1*], MADHU B[2] AND ASHA SRINIVASAN[3]

1: Division of Medical Statistics, School of Life Sciences, JSS Academy of Higher Education & Research, India

2: Department of Community Medicine, JSS Medical College, JSS Academy of Higher Education & Research, India

3: School of Life Sciences, JSS Academy of Higher Education & Research, India

***Corresponding Author: Dr. Chaithra N: E Mail: chaithra.mstats@jssuni.edu.in**

Received 26th Feb. 2022; Revised 25th March 2022; Accepted 22nd June 2022; Available online 1st Jan. 2023

https://doi.org/10.31032/IJBPAS/2023/12.1.6797

ABSTRACT

Background: Cardiovascular diseases are the leading causes of death in the world, caused by disorders of the heart and blood vessels. Among all the heart related diseases, IHD is more prevalent in Indian population. Naïve Bayes algorithms refers to a classification technique that depends on the application of Bayes theorem and which is the best model in machine learning. It is relatively simple to build a model to obtain the estimated probability for a prediction and capable of handling extremely large datasets.

Methodology: A retrospective study was designed to access the 7304 echocardiography records of patients who underwent transthoracic echocardiography at Department Cardiology, JSS Hospital in the year 2016. A model was developed with ECHO database using Naive Bayes classifier. The dataset consists of 6191 patients without IHD and 1113 patients with IHD along with their ECHO parameters.

Results: The model can be trained using naivebayes, e1071 and caret packages, which allows us to perform Naïve Bayes in a powerful and scalable architecture. The final output displays that a Naive Bayes classifier was built which has the capability of predicting whether a person suffers from IHD or not, with an accuracy of approximately 95%. The value of Kappa Statistc (0.810), Precision (0.838), F – Score (0.839) and ROC (0.969).

Conclusion: The Naïve Bayes model is implemented in R-studio as an application, which takes ECHO parameter as an input and it was tested for its accuracy (95 %) in predict disease risk.

Keywords: ECHO database, Ischemic Heart Disease, Naïve Bayes Classifier, R studio

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the largest cause of mortality, accounting for around half of all deaths resulting from Noncommunicable Diseases (NCDs) and are the leading causes of death in the world **[1]**. They are caused by disorders of the heart and blood vessels which includes Ischemic Heart Disease (IHD), Rheumatic Heart Disease, Congenital Heart Disease, Cardiomyopathy, Valvular Heart Disease, Aortic Valve Sclerosis & Stenosis and Atherosclerosis **[2]**. Among all the heart related diseases, IHD is more prevalent in Indian population. IHD is defined as inadequate blood circulation to a local area due to blockage of the blood vessels supplying the heart muscle and it can be diagnosed in several ways **[3]**. An Echocardiography (ECHO**)**, is an ultrasound test used to view moving pictures of the heart on a screen**.** It is used to detect and evaluate a variety of conditions, including heart valve problems, abnormal heart rhythms, congenital heart disease, heart murmurs or infections involving the heart **[4, 5]**.

The Naïve Bayes is a family of probabilistic models that utilize Bayes theorem under the assumption of conditional probability, which refers to the probability of an event occurs depending on the past events information **[6]**. The algorithm explains a simple method to apply Bayes theorem to classification problems. Bayesian Classifiers makes use of training data in order to compute an observed probability of each outcome on the evidence provided by feature values. Bayesian methods were applied to problems in which the information from numerous attributes should be considered simultaneously in order to estimate the overall probability of an outcome **[7]**. Naïve Bayes classifier is the best model in machine learning method that utilizes Bayesian methods were applied to problems in which the information from numerous attributes should be considered simultaneously in order to estimate the overall probability of an outcome. It is relatively simple to build a model to obtain the estimated probability for a prediction and capable of handling extremely large datasets **[8, 9]**.

2. METHODOLOGY

2.1 Study Subject and Dataset

A retrospective study was designed to access the 7304 echocardiography records

of patients who underwent transthoracic echocardiography at Department Cardiology, JSS Hospital in the year 2016. A model was developed with ECHO database using Naive Bayes classifier. The dataset consists of 6191 patients without IHD and 1113 patients with IHD along with their ECHO parameters. Here is the list of the 29 independent variables that classifies the patient as either having heart disease or not: Sex, Age, Aortic Root (AO), Left Atrium (LA), Right Ventricle (RV), Left Ventricle Internal Diameter during Diastole (LVID_d), Left Ventricle Internal Diastole during Systole (LVID_s),Intact Ventricular Septum Diameter during Diastole(IVS_d),Intact Ventricular Septum Diastole during Systole(IVS_s),Left Ventricular Posterior Wall Diameter during Diastole(LVPW_d), Left Ventricular Posterior Wall Diastole during Systole (LVPW_s),End Diastolic Volume (EDV), End Systolic Volume (ESV),Stroke Volume (SV), Ejection Fraction (EF (%)), and fractional Short (FS(%)).Mitral valve - ratio of the early (E) to late (A) ventricular filling velocities(MV_E/A), Mitral regurgitation(MR), Tricuspid valve - ratio of the early (E) to late (A) ventricular filling velocities(TV_E/A), Tricuspid regurgitation (TR), Aortic Valve- The maximal aortic jet velocity(AV_VMAX), Pulmonary Vascular - The maximal Pulmonary jet velocity(PV_VMAX), Pulmonary regurgitation (PR), Left Atrium, Right Atrium, Pulmonary Artery and The dependent variable "Diagnosis" was identified as a predicted attribute with value is equal to "1" for patients suffering from IHD and value equal to "0" for patients not suffering from IHD. The records were split equally into two datasets: training dataset (5112) and testing dataset (2191).

2.2 Naive Bayes Algorithm

Naïve Bayes algorithms refers to a classification technique that depends on the application of Bayes theorem and it strongly assumes that predictors exhibit independency with each other [10]. A probability model characterising the explanatory observation X_1, X_2, X_3,…, X_d where it consists of d values, each being an outcome of a measurement of a different characteristic X_i. For this study, d=29 the characteristics X_1, X_2, X_3,…, X_{29} may represent Sex, Age, AO, LA, RV, LVID_s ,LVID_d , IVS_s , IVS_d , LVPW_d , LVPW_s , EDV , ESV , SV, EF FS, (%), MV_E, MV_A, MR, TV_E, TV_A, TR, AV_VMAX, AR, PV_VMAX, PR, Left Atrium, Right Atrium and Pulmonary Artery respectively, are their ECHO measurements of a particular patient. Furthermore, given X = x, which is a compact notation for $X_1= x_1$, $X_2 =x_2$, $X_3 = x_3$,…, $X_d = x_d$. We are interested in

predicting another characteristic Y, which can take on K possible values denoted by (C_1, C_2, C_3,…,C_k). In other words, we have a multi-class classification problem with K specifying the number of classes. If K = 2 the problem reduces to the binary classification. In the context of our attrition data, we are seeking the probability of an patients belonging to diagnosis class C_k (where C_{yes} = IHD and C_{No} = non-IHD) considering the values of predictors to be X_1, X_2, X_3,…, X_d which can be expressed as $P(C_k \setminus X_1, X_2, X_3,\ldots, X_d)$. Mathematically, the Bayes theorem for calculating this probability is representing as **[11, 12]**.

$$P(C\setminus X) = \frac{P(X\setminus C)\,P(C)}{P(X)}$$

In the above equation:

- C: It is referred to as the proposition and X is the evidence.
- P (C): The class prior probability of the outcome of event C occurring.
- P(X): The probability of occurrence of an event X.
- P (C\X): The posterior probability that an event C occurs, given the event X.
- P (X \ C): The likelihood probability or conditional probability of event X occurring, given the event C.

The Naïve Bayes is a simple form of classification technique and this classification problem is tackled first by applying the Bayes' theorem to the class-specific conditional probabilities $P(Y = C_k \setminus X = x)$, thus we have.

$$P(Y = C_k \setminus X = x) = \frac{P(Y = C_k)P(X = x \setminus Y = C_k)}{P(X = x)}$$

$$P(Y = C_k \setminus X = x) = \frac{P(Y = C_k)\prod_{i=1}^{d} P(X_i = x_i \setminus Y = C_k)}{P(X_1 = x_1, X_2 = x_{2,\ldots,} X_d = x_d)}$$

Since the denominator $P(X_1 = x_1, X_2 = x_{2,\ldots,} X_d = x_d)$ is a constant with respect to the class label C_k, the conditional probability $P(Y = C_k \setminus X = x)$ is proportional to the numerator. We can solve this classification task by applying the maximum posterior classification rule, chose the highest probability value of Y given X using this assumption for the prediction algorithm.

$$\begin{aligned}
\hat{y} &= g(x) \\
&= \arg_{y\in[0,1]} \max \hat{P}(Y \setminus X) \\
&= \arg_{y\in[0,1]} \max \hat{P}(Y)\,\hat{P}(X \setminus Y) \\
&= \arg_{y\in[0,1]} \max \hat{P}(Y = y)\prod_{j=1}^{d} \hat{P}(X_j = x_j \setminus Y = y) \qquad \text{Naive Bayes Assumption} \\
&= \arg_{y\in[0,1]} \max \left(\log \hat{P}(Y = y) + \sum_{j=1}^{d} \log \hat{P}(X_j = x_j \setminus Y = y) \right)
\end{aligned}$$

The class conditional probabilities $P(Y = C_k \setminus X = x)$ are the main interest which is equivalent to predict, then the log posterior probability is transferred back to the original space and then normalised.

$$For\ \ y=0\ \ \Rightarrow\ \ \hat{P}(Y=0)+\sum_{j=1}^{d}\hat{P}(X_j=x_j\setminus Y=0)$$

$$For\ \ y=1\ \ \Rightarrow\ \ \hat{P}(Y=1)+\sum_{j=1}^{d}\hat{P}(X_j=x_j\setminus Y=1)$$

Then the final equation becomes

$$\hat{y}=\arg_{y\in[0,1]}\max\left(\hat{P}(Y=y)+\sum_{j=1}^{d}\hat{P}(X_j=x_j\setminus Y=y)\right)$$

This algorithm is both fast and stable when training and making predictions. Although this above equation seems intimidating, the series of steps are fairly straightforward. Begin by building a frequency table, use this to build a likelihood table, and multiply the conditional probabilities according to the Naive Bayes rule. Finally, divide by the total likelihood to transform each class into a probability.

3. RESULTS AND DISCUSSION

3.1 Building a Model with Naïve Bayes Classifier using R Studio.

Despite the availability of wide set of R add-ons, installation is made by the package format and utilize a process that is virtually effortless. To start training a Naive Bayes classifier in R, we will install and load the naivebayes, e1071 and caret packages, which allows us to perform Naïve Bayes in a powerful and scalable architecture. R supports a package called e1071, contains the naivebayes function and it allows numeric and factor variables to be used in the model. In order to conserve memory, R does not load every installed package by default. Instead, users are required to load packages when needed by making use of the library () function **[13, 14]**.

Step 1: Install and load the required packages.

```
> install.packages("naivebayes")
> install.packages("e1071")
> install.packages("caret")
```

Step 2: Exploring and preparing the data.

The first step towards constructing our classifier involves processing the raw data for analysis. Before we study the dataset let's convert the output variable into a categorical variable. This is necessary because our output will be in the form of two classes, Yes or NO. Where Yes will denote that a patient has IHD and No denotes that a person is without IHD. Once the data has been converted into the CSV format then it is imported into R using the following command. Using the str() function, we see that the data frame includes 7304 records.

```
> data=read.csv(file.choose(),header=T)
# Reading data into R
> str(data)                # Studying the structure of the data
> summary(data)            # Print the Descriptive statistics
```

The summary of the model which was printed in e3071 package is stored in learner model and these summary calculations provide the mean, median,

25th and 75th quartiles, min, max for each class.

Step 3: Data preparation – creating training and test datasets

This stage begins with a process called Data Splicing, wherein the data set is split into two parts i.e. training set and testing set.

Training set: 70 % data used to build and train the Machine Learning model.

Testing set: 30% records used to evaluate the efficiency of the model.

```
> set.seed (2)
> id=sample(2,nrow(data),prob=c(0.7,0.3), replace=T)
> datatrain=data[id==1,]
> datatest=data[id==2,]
> nrow(datatest)
[1] 2191
> nrow(datatrain)
[1] 5113
```

Step 4: Data modelling

The model can be trained using e1071 package and this package contains function called naivebayes () is a simple, elegant implementation of the Naive Bayes classification.

naiveBayes(formula, data)

- The formula is $Y \square X_1 + X_2 + X_3,...,+X_n$
- The data is typically a data frame of numeric or matrix containing training data.

The function will return a Naïve Bayes model object that can be used to make predictions.

The naiveBayes () function assumed gaussian distributions for numeric variables. Also, the priori probabilities and the conditional probabilities are calculated from the proportion of the training data. The values are shown when the object is printed.

```
> data_nb=naiveBayes(Diagnosis~., data= datatrain)
> data_nb
```

Naive Bayes Classifier for Discrete Predict ors

Call: naiveBayes.default(x = X, y = Y, lapl ace = laplace)

Table 1: A-priori probabilities

P (No)	P (Yes)
0.85	0.15

Priori probabilities results of **Table 1** indicates that the diagnosis is No with probability of 0.85 and Yes with probability of 0.15. Naive Bayes algorithm is trained by constructing a likelihood table for the appearance of these Independent variables (Sex, Age , AO , LA , RV , LVID_d , LVID_s , IVS_d , IVS_s , LVPW_d , LVPW_s , EDV , ESV , SV, EF FS, (%), MV_E, MV_A, MR, TV_E, TV_A, TR, AV_VMAX, AR, PV_VMAX, PR, Left Atrium, Right Atrium, Pulmonary Artery).

Step 5: Model evaluation.

For checking the model efficiency, the testing data is initially executed on the model, and later we will evaluate the accuracy of the model by making use of a confusion matrix.

> pre=predict(data_nb,datatest)

- The predict () function is used to make the predictions.
- data_nb is a model train by the naiveBayes () function.
- datatest is a data frame containing test data with the same features the training data used to build the classifier.

This function will return a vector of predicted class values or raw predicted probabilities.

> confusionMatrix(table(pre,datatest$Diagnosis)

To compare the predictions to the true values, use the confusionMatrix () function.

Table 2: Confusion matrix

Actual value	Predicted value	
	No	Yes
NO	1797	55
Yes	54	285

Table 3: Prediction performance measures

True Positive rate (Sensitivity)	0.841 (95% CI = 0.808 - 0.868)
True Negative rate (Specificity)	0.970 (95% CI = 0.964 - 0.975)
Negative Predicted Value	0.971 (95% CI = 0.965 - 0.976)
Positive Predicted Value	0.838 (95% CI = 0.806 - 0.866)
Accuracy	0.950 (95% CI = 0.940 - 0.959)
Kappa	0.810 (95% CI =0.772 - 0.843)
Precision	0.838
F - Score	0.839
Correctly classified instances	2082 (95%)
Incorrectly classified instances	109 (4.97%)
Area under ROC	0.969

Confusion matrix is useful in comparing the predicted and the observed results is show in **Table 2**. True Positive Rate reflects the 285 cases of correctly identified IHD while True Negative Rate reflects the 1797 cases of correctly recognized healthy individuals. However, the false negative and false positive rate are reflected by their complements respectively. **Table 3** display the performance of each classification model is evaluated using statistical measures. The final output displays that a Naive Bayes classifier was built which has the capability of predicting whether a person suffers from IHD or not, with an accuracy of approximately 95%. The value of Kappa Statistc (0.810), Precision (0.838), F – Score (0.839) and ROC (0.969) are displayed in **Table 3**.

4. CONCLUSION

This research work has developed a Naïve Bayes model to predicting the likelihood of IHD. Which is implemented in R-studio as an application which takes echo parameter as an input and it was tested for its accuracy (95 %) in predict disease risk. From the above statistics it is clear that the model is highly specific than sensitive i.e. the negative values are predicted more accurately than the positives. This model contributes to medical sciences for supporting medical analysis as well as detecting in relation to cardiovascular disease.

REFERENCES

[1] Roman WP, Martin HD, Sauli E. Cardiovascular diseases in Tanzania: the burden of modifiable and intermediate risk factors. *Journal of Xiangya Medicine.* 2019; 4(33):1-13.

[2] World Health Organization [Internet]. 2020. Available from: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

[3] IOM (Institute of Medicine). 2010. Cardiovascular disability: updating the social security listings. Washington, DC: The National Academies Press.

[4] Nishimura RA et al. 2014 AHA/ACC guideline for the management of patients with valvular heart disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*. 2014;63(22):2438-88.

[5] Kaddoura S. Echo made easy E-Book. 2nd ed. Toronto. Churchill Livingstone; 2016.

[6] Majka M. Introduction to naïve bayes package main functions. 2020; 1–15.

[7] Lantz B. Machine learning with R. 2nd ed. Mumbai: PACKT Publishing; 2015.

[8] Singh G, Bagwe K, Shanbhag S, Singh S, Devi S. Heart disease prediction using naïve bayes. *Internation Research Journal of Engineering and Technology*. 2017; 4(3):1-3.

[9] Manjusha KK, Shankaranarayanan K, Seena P. Prediction of different dermatological conditions using naïve bayesian classification. *International Journal of Advanced Research in Computer Science and Software Engineering.* 2014; 4(1): 864-868.

[10] Selvamuthu D and Das D. Introduction to statistical methods, design of experiments

and statistical quality control. Springer; 2018.

[11] Majka M. Introduction to naïve bayes package main functions. 2020; 1–15.

[12] Monroe W. Naïve bayes. Lecture notes #21 based on chapter by Chris Piech; 2017.

[13] Zhang Z. Naïve bayes classification in R. *Annals of Translational Medicine*. 2016; 4(12):241-245.

[14] Lewis ND. Machine learning made easy with R: an intuitive step by step blueprint for beginners. 2017.