International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Cardiac Disease Prediction Using Naïve Bayes Machine Learning Algorithm

B.Soumya, V.Sushma, V.Divya, S.Lalith Kumar, B.Venkateswarulu

Abstract: ***With the increasing population and comforts many of the people with white collar jobs will not have enough time or don't like to spend time on doing exercises or doing physically stress causing works which leads to a lot of serious problems that include blood pressure, heart diseases and etc. If any of these problems are left unattended or not recognized prior, these may become deadly. For example, some heart diseases like heart stroke or heart attacks can be easily prevented if the patients are treated with better medication. From this idea of predicting the heart diseases prior gave a spark to this research area of using machine learning techniques to predict the possibility of a person that may get heart diseases soon or later based on some simple answers to the questions that are given as the attributes to the naive Bayes classification that will in turn provide the best results possible. Naive Bayes is not only an easy algorithm but also a powerful one which can easily handle big data sets. With the help of this system, there can be a good an outbreak of the people getting self-conscious about the current condition they are in and get health conscious, which can pretty well help them in getting themselves out of the health troubles. Prediction and classification all add up to data mining, but the giant data sets and the accuracy for it can only be done right by Naive Bayes, which is a well known algorithm which is being used from many decades and still works better than the algorithms that are invented after it.***

Index Terms: ***Data Mining, Data-Sets, Heart Disease, Naïve Bayes.***

I. INTRODUCTION

Now a days heart diseases have become a very serious health issue in everyone's life, in previous days humans use to do a lot of physical work and also they used to eat organic foods. As they do not have any chemicals or artificial elements that promote the desired qualities of the natural food, they used to stay healthy. But it's not the same now, we have artificially modified natural food and also we are not doing necessary physical work which results in many of the threatening diseases. Heart diseases are one of the most dangerous diseases among all.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
B.SOUMYA, Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada, India.
V.SUSHMA, Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Guntur, India.
V.DIVYA, Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Guntur, India.
S. LALITH KUMAR, Computer Science Engineering, Koneru Lakshmaiah Education Foundation , Guntur, India.
Mr.B.VENKATESWARULU, Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada, India.

As we know that heart is the main part of the blood circulation and also the main thing of our body to function, slightest difference in the working might affect our health a lot. Having some other diseases might also have some effect on the heart as well. For example, having blood pressure can also be a serious problem for heart as the blood of blood pressure patient will get thickened and that in turn results in difficulty in flow of blood and also thereby might cause trouble for heart which may lead in heart diseases.We proposed a system that which helps a patient get to know if he is ever going to be diagnosed with heart diseases by knowing cholesterol, ECG etc. Data mining might help a lot in this kind of cases as we can possibly predict what common symptoms or habits that are observed with the previous patients of heart diseases. We used Bayesian Probabilistic method to predict the probability of the person to getting diagnosed with heart diseases.

II. LITERATURE SURVEY

We as a team surveyed few research papers which made a base to our developed project.

[1] Article- in this the authors aimed to develop an Intelligent System using data mining modeling technique, namely Naïve Bayes and the performance metrices are Accuracy, Complex queries, search iterations, confusion matrix. his has addressed complex questions, each with its very own quality no sweat of model understanding, access to definite data and precision. Choice Support in Heart Disease Prediction System is developed utilizing Naive Bayesian Classification technique.

[2] Article- In this the authors conveyed detecting the heart disease using naïve bayes and finding the accuracy of it and the performance matrices are Weka tool, accuracy, attribute value distribution, data analysis and the results showed incorrect instances of 13.5%. So they were working on other data mining applications for best accuracy and prediction.

[3] Article- In this the writers described about A system architecture which is built as a web application that diagnoses the heart condition of a patient using naïve bayes. So It is completely automated online software and Tasks are conducted more efficiently. They tend to use some intelligent data mining techniques to predict the foremost correct illness or problem that might be related to patient's details. Based on result, system automatically shows the result specific doctors for more treatment. The system permits user to look at doctor's details. The system can be use in case of emergency.

[4] Article-In this the authors scripted about Accuracy, optimization, precision, recall, matthwes correlation coefficient (MCC) and they have given The generalized system that was developed. It can be served as a training tool for medical students and that Predicts the risk of the heart disease if unknown sample is given as an input.

[5] Article- In this it was given about Answering complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique.

[6] Article- In this it has given classification evaluation measures, sensitivity, specificity, precision, accuracy and confusion matrix disease or not. In this paper, they designed a cloud-assisted privacy preserving mobile health monitoring system. They elaborated on both objective and subjective elements of information concerning the patient.

III. BASIC TERMINOLOGY

A. Machine learning: is a study of the behaviours and the basic working methodologies of several algorithms. Machine learning is basically divided into three main streams they are: Supervised, Unsupervised and Reinforcement. Where Supervised as the name suggests a human supervision should be there. Where as Unsupervised human supervision may not be needed. Reinforcement is that the algorithm adapts itself with the comfort of the environment.

***B.* Decision tree:** Decision tree is basically a flow diagram representation of algorithm which covers different possiblities as the decision tree itself is a tree with branches, in which branches are the possiblities.

***C.* Artificial neural networks:** Artificial neural networks are inspired and designed from the observaton of a human brain. How human brain thinks in different ways. It have three layers namely input layer, hidden layer and output layer. In hidden layer we have layers that are composed of neurons.

***D.* K-Means Algorithm**: K means algorithm is good with databases that have massive data. This algorithm will learn about the data itself without human supervision. so it is semi supervised algorithm.

E. Bolster Vector Machine: or SVM is a method for implication of both straight and non-direct information. It applies a non-direct mapping strategy with the goal that it can change the preparation information into a higher measurement. A hyperplane is a sort of line which isolates the information variable space in SVM. The hyperplane can isolate the focuses in the info variable space containing their class that is either 0 or 1.

F. Naïve Bayes: Naive Bayes algorithm is a probablity based Algorithm that predicts the further or future values based on the rule set or instructions in bayes the predictors work individually in predicting without interdepending.

These are few algorithms which are also used to predict or estimate the heart infected conditions.

IV .PROPOSED SYSTEM:

A. Data Set:

The dataset of UCI statlog coronary illness has been utilized in this investigation. The dataset contains fourteen highlights which are essential in determination of heart illnesses.

The main factors that to be noted are "Age", " Sex"," ChestPain", "RestBp", "Chol", "FBS" ,"RestECG", "MaxHR", "ExANG", "Oldpeak" ,"Slope", "Ca", "Thal", "AND".

Names Of Attributes	Illustration
Age	The patient's Age
Sex	0-female 1-male
ChestPain	Chest pain.These are of 3 categories 1.Typical 2.Non-typical 3.Asymptomatic
RestBp	Value of Resting blood pressure (single digit)
Chol	Cholestral Content
FBS	Fasting Blood sugar
RestECG	Resting electro cardio graphic results 0,1,2
MaxHR	The Maximum Heart Rate

ExANG	Exercise Induced Angina (unbearable pain that spreads throught the body)
Oldpeak	ST- Severe Disease. ST depression induced by exercise relative tp rest.
Slope	The slope of the peak exercise ST segment
Ca	Number of major vessels(0 to 3) Colored fluoroscopy(Barium X- Rays, authography).
Thal	Thalassemia (a blood disorder) Which causes our body to make less Hemoglobin. The 3 states are 1.Normal 2.Fixed 3. .Reversable
AHD	Analysis/Estimation of Heart Disease.

B. Naïve Bayes Detailed Explanation:

In likelihood hypothesis, Bayes' hypothesis (regularly called Bayes' law after Thomas Bayes) relates the restrictive and minimal probabilities of two irregular occasions. Usually used to process back probabilities given perceptions. For instance, a patient might be seen to have certain side effects. Bayes' hypothesis can be utilized to register the likelihood that a proposed finding is right, given that perception. A credulous Bayes classifier is a term managing a basic probabilistic arrangement dependent on applying Bayes' theorem. It is a simple technique for constructing classifiers. It is a probabilistic classifier dependent on Bayes' hypothesis.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

All Naive Bayes classifiers expect that the estimation of a particular element is independent of the estimation of some other element, given the class variable. Bayes hypothesis is given as pursues:

$$P(C|X) = P(X|C) * P(C)/P(X)$$

Where X is the information tuple and C is the class such that P(X) is constant for all classes. In spite of the fact that it accept an unrealistic condition that trait esteems are conditionally independent, it performs surprisingly well on substantial datasets where this condition is expected and holds. The naïve bayes classifier works on prior probability and conditional probability.

V. WORKING SYSTEM

Out of many platforms we had opted **R Studio** to execute this. In R we have called on different packages and they are:

A. MICC package:

Missing data can be not so trivial problem when analysing a dataset and accounting for it is usually not so straight foward either. If the amount of data is very small relatively to the size of the dataset, then micc package is very useful to impute that values.
Example:

Before

Ozone	Solar	Wind	Temp
1.0	2	18	NA(missing Value)
4	NA	6	8

After

Ozone	Solar	Wind	Temp
1.0	2	18	6
4	7	6	8

B. VIM package:

VIM is a package which is also used to impute missing values. But it applicable also for large dataset

C. missForest package:

missforest is used to impute missing values particularly in the case of mixed type of data. It can be used to impute continuously and /or categorical data including complex interactions and non-linear relations. It yields an out - of - bag imputation error estimates. More over it can be run parallel to save computation time.

D. Impute Ts package:

This package specializes on time series imputation. It also provides plots and printing functions. It also provides statistics for missing data.

E. e1071:

It is used for functions for latent class analysis, short time faviour transform, fuzzy clustering, support vector machines, short path computation, bagged clustering.

F. Naïve Bayes Classifier:

Especially in this project we are going to use Naïve Bayes so this package is very important.

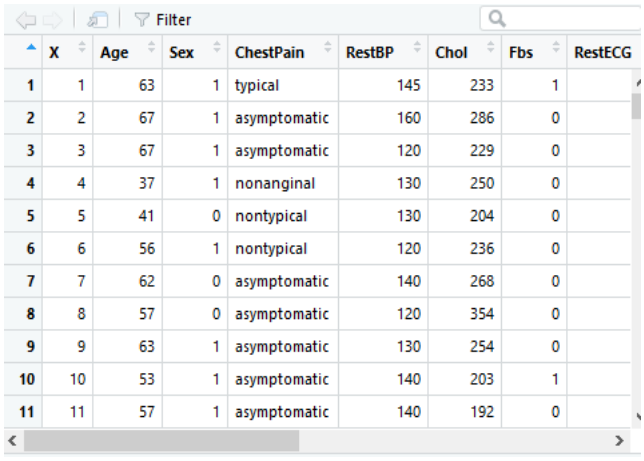
G. mlr:

It is a built-in function which returns the important variables from data. It also evaluates the performance of program this includes machine learning algorithms which we use frequently.

VI. TASK AND RESULT:

Firstly after setting path for your given dataset using setwd() (Shortcut→ ctrl+shift+H) getwd() Read the file using read.csv reads file to create data frames with cases with arguments as dataset and headers this is one out of all available formats this reads our heart dataset which is heart.csv format .

"X"	"Age"	"Sex"	"ChestPain"
"RestBP"	“Chol	“FBS”	“RestECG”
"MaxHR"	"ExAng"	"Oldpeak"	"Slope"



	X	Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG
1	1	63	1	typical	145	233	1	
2	2	67	1	asymptomatic	160	286	0	
3	3	67	1	asymptomatic	120	229	0	
4	4	37	1	nonanginal	130	250	0	
5	5	41	0	nontypical	130	204	0	
6	6	56	1	nontypical	120	236	0	
7	7	62	0	asymptomatic	140	268	0	
8	8	57	0	asymptomatic	120	354	0	
9	9	63	1	asymptomatic	130	254	0	
10	10	53	1	asymptomatic	140	203	1	
11	11	57	1	asymptomatic	140	192	0	

. *1.Figure shows the data set*

View is a function that pushes to new window to see the dataset in an r terminal which is read only mode of our heart dataset .

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Summary is a generic function that describes each attribute's mean median value and also any missing values that are present in our respective dataset.

```
       X                Age              Sex              ChestPain      RestBP
 Min.   :  1.0   Min.   :29.00   Min.   :0.0000   asymptomatic:144   Min.   : 94.0
 1st Qu.: 76.5   1st Qu.:48.00   1st Qu.:0.0000   nonanginal  : 86   1st Qu.:120.0
 Median :152.0   Median :56.00   Median :1.0000   nontypical  : 50   Median :130.0
 Mean   :152.0   Mean   :54.44   Mean   :0.6799   typical     : 23   Mean   :131.7
 3rd Qu.:227.5   3rd Qu.:61.00   3rd Qu.:1.0000                      3rd Qu.:140.0
 Max.   :303.0   Max.   :77.00   Max.   :1.0000                      Max.   :200.0

      Chol            Fbs             RestECG          MaxHR           ExAng
 Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0   Min.   :0.0000
 1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000
 Median :241.0   Median :0.0000   Median :1.0000   Median :153.0   Median :0.0000
 Mean   :246.7   Mean   :0.1485   Mean   :0.9901   Mean   :149.6   Mean   :0.3267
 3rd Qu.:275.0   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0   3rd Qu.:1.0000
 Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0   Max.   :1.0000
```

2.Figure shows each columns information

These are the following 14 attributes that are present and we need to check for any missing values present in the respecti-ve dataset using heart.csv. Now we will compute the missing values using mice package and we can artificially insert these values using prodNA() function with arguments as dat-a and proportion of missing values present in the dataset.

```
Console G:/
> heart.missing
    X Age Sex    ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope Ca
Thal  AHD
1   1  63   1      typical    145   NA  NA       2   150     0     2.3     3  0      f
ixed <NA>
2   2  67   1 asymptomatic    160  286  NA       2   108     1      NA     2  3     no
rmal  Yes
3   3  67   1 asymptomatic    120   NA   0       2   129     1     2.6     2  2 revers
able  Yes
4   4  37   1   nonanginal     NA  250   0       0   187     0     3.5     3  0
<NA>   No
5   5  41   0   nontypical    130  204  NA      NA    NA     0     1.4     1  0
<NA>   No
6   6  NA   1         <NA>    120  236  NA       0   178     0     0.8     1 NA     no
rmal   No
7   7  62   0 asymptomatic    140  268   0      NA   160     0     3.6     3  2     no
```

3.Figure shows missing(NA) values

These missing values can be imputed using many different methods but we have opted for mean method and after imputing all the missing values. And in another function it gives out the existing data as 1 and indicates empty or no value as 0.

Filter

MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	AHD
144	1	2.8	3	0.0000000	fixed	Yes
144	1	4.0	1	2.0000000	reversable	Yes
136	1	0.0	2	0.0000000	normal	Yes
182	0	0.0	1	0.0000000	normal	No
90	0	1.0	2	2.0000000	fixed	Yes
123	1	0.2	2	0.0000000	reversable	Yes
132	0	1.2	2	0.0000000	reversable	Yes
141	0	3.4	2	2.0000000	reversable	Yes
115	1	1.2	2	1.0000000	reversable	Yes
174	0	0.0	2	1.0000000	normal	Yes
173	0	0.0	1	0.6722408	normal	No

4.Figure shows NA values imputed with mean value

5.Figure Missing values

These missing values can be imputed using many different methods but we have opted for mean method and after imputing all the missing values. And in another function it gives out the existing data as 1 and indicates empty or no value as 0.

Filter

MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	AHD
144	1	2.8	3	0.0000000	fixed	Yes
144	1	4.0	1	2.0000000	reversable	Yes
136	1	0.0	2	0.0000000	normal	Yes
182	0	0.0	1	0.0000000	normal	No
90	0	1.0	2	2.0000000	fixed	Yes
123	1	0.2	2	0.0000000	reversable	Yes
132	0	1.2	2	0.0000000	reversable	Yes
141	0	3.4	2	2.0000000	reversable	Yes
115	1	1.2	2	1.0000000	reversable	Yes
174	0	0.0	2	1.0000000	normal	Yes
173	0	0.0	1	0.6722408	normal	No

6.Figure shows null space as 0 and occupied as 1

Now we need to split our data into test and train splits in the given dataset in the ratio of 70:30 and now we use data partition using sort function .70% to train and 30% to test.
Number of train samples :244

Number of test samples :59

Total number of samples: 303

Number of attributes(main):14

dim(heart)

303(rows) 15(columns)

VII. EVALUATION USING NAIVE BAYES METHOD

Naïve Bayes is a classification algorithm which in R to classify your given dataset which is related classifiers applying baye's theorem with strong independent assumptions between features.it is time efficient and obtain maximum-likelyhood hypothesis. This is loaded using a package(e1071) in R These take class label and training data to classify the dataset using NaiveBayes() method.After classification predict the examples that has been trained using the train dataset.

Sample output for Naïve bayes :

```
Console G:/

Naive Bayes Classifier for Discrete Predictors

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

A-priori probabilities:
Y
       No       Yes
0.5188679 0.4811321

Conditional probabilities:
     X
Y          [,1]     [,2]
  No  154.3273 83.35787
  Yes 165.5294 90.40390
```


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

```
Console G:/
  Yes 1.862745 0.5635674

     Ca
Y          [,1]      [,2]
  No  0.2752294 0.6364188
  Yes 1.1386139 1.0773087

     Thal
Y          fixed     normal reversable
  No  0.02727273 0.80909091 0.16363636
  Yes 0.09900990 0.24752475 0.65346535
```

7.Figure shows the probability of each attribute column. (only 2 attributes are shown)

Predict Syntax

```
predict(object, newdata, se.fit = FALSE,
scale = NULL, df = Inf,
        interval = c("none", "confidence",
"prediction"),
        level = 0.95, type = c("response",
"terms"),
        terms = NULL, na.action = na.pass,
        pred.var = res.var/weights,
weights = 1, ...)
```

`Levels: No Yes`

`predict.lm` produces predicted values, obtained by evaluating the regression function in the frame `newdata` (which defaults to `model.frame (object))`. If the logical `se.fit` is `TRUE`, standard errors of the predictions are calculated. If the numeric argument scale is set (with optional df), it is used as the residual standard deviation in the computation of the standard errors, otherwise this is extracted from the model fit. Setting intervals specifies computation of confidence or prediction (tolerance) intervals at the specified level, sometimes referred to as narrow vs. wide intervals. This prediction is our final output.

VIII. CONCLUSION:

The prediction is done being after classified using naïve Bayes classifier with 82 percent accuracy we have predicted this using naïve bayes classifier and from the construction we identify the accuracy using confusion matrix and by this we predict the accuracy of heart dataset.

```
> table(nbpredict1,htrain$AHD)

nbpredict1  No Yes
       No  102  17
       Yes  14  79
> table(nbpredict,htest$AHD)

nbpredict No Yes
      No  41  12
      Yes  7  31
```

This is the final confusion matrix we got. A system is said to be accurate when the diagonal matrix is zero, but in this result in "nbpredict" if there are 41 cases diagonised with "NO Disease" the system is giving a wrong prediction of 12 cases "have disease". In visa versa accurately there are 31 cases diagonised "Yes Diseased" but the remaining 7 even though they are diseased the system is giving a wrong prediction of having "No Disease". We are still working on the accurate result and to achieve the highest correct output.

REFERENCES

1. Ms.Rupali R.Patil," Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 5, May 2014. ISSN (Online) : 2278-1021 ISSN (Print) : 2319-5940.
2. K.VembandasamyPRR R.SasipriyaP Pand E.DeepaP, "Heart Diseases Detection Using Naive Bayes Algorithm", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015. ISSN 2348 – 7968.
3. Garima Singh, Kiran Bagwe, Shivani Shanbhag, ShraddhaSingh, Sulochana Devi." Heart disease prediction using Naïve Bayes", International Research Journal of Engineering and Technology(IRJET)e-ISSN: 2395.-0056, p-ISSN: 2395-0072 Volume: 04 Issue: 03| Mar-2017
4. Dhanashree S. Medhekar, Mayur P. Bote, Shruti D. Deshmukh," Heart Disease Prediction System using Naive Bayes", International Journal of Enhanced Research In Science Technology and Engineer-ing VOL. 2 ISSUE 3, MARCH.-2013 ISSN NO: 2319-7463.
5. Shadab Adam Pattekari and Asma Parveen, "Predicion System for Heart Disease using Naïve Bayes" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294
6. Arun.R, UG Scholar, Saveetha School of Engineering, SIMATS. Mrs. N.Deepa, Asst Professor, Saveetha School of Engineering, SIMATS. "Heart Disease Prediction Using Naïve Bayes" International Journal of Pure and Applied Mathematics,Volume 119 No. 16 2018, 3053-3065,ISSN: 1314-3395 (on-line version) url: http://www.acadpubl.eu/hub/

AUTHORS PROFILE:

Bonela Soumya, is a student pursuing study in the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. She is doing her research work in knowledge engineering.

V.Sushma, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. She is doing her research work in knowledge engineering.

V.Divya, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. She is doing her research work in knowledge engineering.

S.Lalith Kumar, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. He is doing her research work in knowledge engineering and IOT.

Mr.B.Venkateswarulu,. is an Assistant Professor in the Computer Science & Engineering Department at KLUniversity. His research areas include data mining, knowledge Engineering.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)