Available online at www.ijiere.com
International Journal of Innovative and Emerging Research in Engineering
e-ISSN: 2394 – 3343 p-ISSN: 2394 – 5494

A Survey on Naive Bayes Based Prediction of Heart Disease Using Risk Factors

Sohana Saiyed[1], Nikita Bhatt[2] and Dr. Amit P. Ganatra[3]
[1]M.tech Student at CSPIT Changa, Gujrat-India
[2]Assistant Professor at CSPIT Changa, Gujrat-India
[3]Head Of Department at CSPIT Changa, Gujrat-India

ABSTRACT:
Data mining techniques have been used in clinical decision support systems for prediction and diagnosis of various diseases with excellent accuracy. The statistics show that heart diseases (or cardiovascular disease) are one of the leading causes of deaths all over the world. Many researchers are using statistical and data mining techniques for the diagnosis of heart disease. Naive Bayes is one of the simple data mining technique has shown better result and accuracy.None of the system predicts heart diseases based on risk factors such as age, family history, diabetes, hypertension, high cholesterol, tobacco smoking, alcohol intake, obesity or physical inactivity, etc. System based on such risk factors would not only help to reduce the mortality rate of Heart Disease patient in the rural areas but it would also give patients a warning and suggestion about the probable presence of heart disease even before he visits a hospital or goes for costly medical checkups. . In this paper, data mining methods namely, Naive Bayes, algorithm is analyzed on dataset based on risk factors.

Keywords: Data mining, Naive Bayes, Heart Disease, Risk Factors

I. INTRODUCTION

Data mining is the process of extracting knowledgeable information from huge repositories. Medical data mining extracts knowledgeable data for effective medical diagnosis. Classification is a supervised learning used to discover hidden patterns from existing medical data. Large numbers of data mining tools are used for heart disease prediction [13]. Heart Disease (or CVDs) is the number one cause of death globally. More people die annually from Heart Disease than from any other diseases. An estimated 17.5 million people died from Heart Diseases in 2012, representing 31% of all global deaths. An estimated 7.4 million people died due to coronary heart disease and 6.7 million were due to stroke [1].Most of the cardiovascular diseases can be prevented and diagnosed by addressing behavioral risk factors such as, unhealthy diet and obesity(fatness), physical inactivity and additional use of alcohol and smoking using population-wide strategies. People with cardiovascular disease or who are at high risk (due to the presence risk factors such as hypertension, diabetes or already established disease) need early detection and management using treatment and medicines, as appropriate. High blood pressure and High Cholesterol are the major risk factor for heart disease. Hypertension and diabetes mellitus have been appropriately highlighted as recognized predictors of cardiovascular disease. These risk factors have become targets for influencing cardiovascular risk; their assessment, treatment, and monitoring are major accents of clinical care. Notably, lifestyle risk factors, including dietary habits, physical inactivity, smoking, strongly influence the established cardiovascular risk factors and also affect paths of risk such as inflammation, endothelial function, thrombosis/coagulation, and arrhythmia [2]. There are many studies and researches on the prevention of heart disease risk. Information from readings of population has helped in prediction of heart diseases, based on blood pressure, smoking habit and alcohol, cholesterol and blood pressure levels, diabetes etc. The Framingham Risk Score (FRS) is a popular risk prediction criterion which is used in algorithms for heart disease prediction [4]. This study aimed at developing the risk prediction factors categories.

II. DATA MINING TECHNIQUES

Each data mining technique works a different purpose depending on the modeling objective. Data mining techniques are used to explore and extract medical data using complex algorithms in order to discover unknown patterns. Many researchers are using different data mining techniques for the diagnosis of many diseases such as heart disease, diabetes, smoking, stroke and cancer and many data mining techniques have been used in the diagnosis of heart disease with good accuracy. Researchers have been used multiple data mining techniques such as naïve bayes, neural network, decision tree, and support vector machine for prediction and diagnosis of heart diseases [5][6].

Sellappan Palaniappan, Rafiah Awang has developed a model Intelligent Heart Disease Prediction System (IHDPS) which uses data mining techniques such as Decision Trees, Naive Bayes and Neural Network. IHDPS can help by serving a training tool to train nurses and medical students to detect patients with heart disease. It can also provide decision support to help doctors to make better clinical decisions [7].

Punam Bajaj, Preeti Gupta has proposed a system in which they have investigate the effect of hybridizing more than one technique which shows better result in the diagnosis of heart disease. In this research the neural network is trained with selected pattern for the diagnosis of heart disease and Genetic Algorithm has been used for optimizing the neural network [8].

Nidhi Bhatla et al. proposed that observations reveal that Neural networks with 15 attributes has exceed over all other data mining techniques. Another conclusion from the examination is that decision tree has also shown good accuracy (89%) with the help of genetic algorithm and feature subset selection[9].

Latha Parthiban et al. proposed a model on basis of Coactive Neuro-Fuzzy Inference System (CANFIS) for prediction and diagnoses of heart disease. The CANFIS model diagnosed the presence of disease by integrating different data mining techniques that includes the neural network, fuzzy logic and further merging with genetic algorithm.[10]

M.A.Nishara Banu et al. proposed that preprocessed data is clustered using clustering algorithms like K-means to cluster related data in database. Maximal Frequent Itemset Algorithm (MAFIA) is used for removing maximal frequent patterns in heart disease database. The frequent patterns further can be classified using C4.5 algorithm as training algorithm using the concept of information entropy. The results showed that the considered prediction system is capable of predicting the heart attack successfully [16].

Algorithm selection is a time consuming task which includes experimentation with different classifiers and analyzing the performance of those classifiers. So while analyzing the different techniques discussed, this paper proposes a novel system using Naive Bayes algorithm for predicting the risk of heart diseases. A naive Bayes classifier is a term dealing with a probabilistic classification based on applying Bayes' theorem. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. What is even more different in this paper is that it is the first time that such a technique is applied on risk factors for the accurate prediction of heart disease. Naive Bayes classifiers frequently work much better in many complex real-world situations than one might believe.

III. STUDY OF RISK FACTORS

The problem with risk factors related to heart disease is that there are many risk factors involved like age, usage of cigarette, blood cholesterol, person's fitness, blood pressure, stress and etc. and understanding and categorizing each one according to its importance is a difficult task. Also a heart disease is often detected when a patient reaches advanced stage of the disease. Hence the risk factors were analyzed from various sources [11]-[12]. The dataset was composed of 12 important risk factors which were sex, age, family history blood pressure, Smoking Habit, alcohol consumption, physical inactivity, diabetes, blood cholesterol, poor diet, obesity .The system indicated whether the patient had risk of heart disease or not. Most of the heart disease patients had many similarities in the risk factors. The TABLE I below show the identified important risk factors and the corresponding values.

TABLE I **Risk Factors Values and their Encodings**

	Risk Factor	Values
1	Sex	Male or Female
2	Age in Years	Age in Numeric
3	Blood Cholesterol	Below 120 mm Hg- Low 120 to 139 mm Hg- Normal Above 139 mm Hg- High
4	Blood Pressure	Below 120 mm Hg- Low 120 to 139 mm Hg- Normal Above 139 mm Hg- High
5	Hereditary	Family_Member diagnosed with HD -Yes Otherwise –No
6	Smoking	Yes or No
7	Alcohol Intake	Yes or No
8	Physical Activity	Low , Normal or High
9	Diabetes	Yes or No
10	Diet	Poor , Normal or Good
11	Obesity	Yes or No
12	Stress	Yes or No
Output	Class	Yes or No

IV. **NAIVE BAYES CLASSIFIER**

Naive Bayes or Bayes' Rule is a very efficient data mining technique which calculates the various probabilities and gives a result on the basis of these probabilities.

A. Classifier Module

Naive Bayesian classifiers have proven to be powerful for solving classification problems in a variety of fields. A Naive Bayesian classifier is a model of a joint probability distribution over a set of stochastic variables [6]. It is composed of a single class variable, showing the likely outcomes or classes for the problem under study, and a set of feature variables, modeling the features that provide for distinguishing between the numbers of classes; the feature variables are assumed to be equally independent given the class variable [11]. Instances of the classification problem under study are presented to the classifier as a combination of values for the feature variables; the classifier then returns a posterior probability distribution over the class variable, that is, it produce a probabilistic summary for each class[6]. Naive Bayesian classifiers have been successfully applied in the medical domain where they are used for solving diagnostic problems. Naive Bayesian classifiers are typically learned from data. Learning such a classifier amounts to establishing the prior probabilities of the different classes and estimating the conditional probabilities of the various features given each of the classes. Not for every classification problem, however, will a dataset be available that is rich enough to allow for reasonable probability estimates. According to the Bayes theorem of probability theory-

$P(C_k|\{A\}) = P(\{A\}|C_k)P(C_k) / P(\{A\})$

For most domains, estimating $P(\{A\}|C_k)$ needs an impractically large training set, as we need to consider all possible attributes A1, A2, A3…..AM. Moreover, it needs a lot of computation, which, in turn, can require a lot of time. So it is usually assumed that the attributes A1 to AM are class conditionally independent, which means it is often assumed that

$P(\{A\}|C_k) = \prod_{i=1}^{N} P(Ai)=V_j|Ck)$

After making the above assumption, the classifier is called the Naïve Bayes classifier.

B. How Naive Bayes algorithm works

Let's understand this with an example. Here I have a training data set of weather and consistent target variable 'Play' (possibilities of playing). Here, we need to classify whether players will play or not based on weather condition. Let's follow the below steps to perform [14].

Step 1:- Convert the data set into a frequency table

Step 2:- Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing based on weather is 0.64

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Figure 1 **Example of Naive Bayes[14]**

Step 3:- Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability will be the outcome of prediction.

Problem: Players will play if weather is sunny. Is this statement is correct?

We can solve it using above discussed method of posterior probability [14].

P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability [14].

Naive Bayes uses a related method to predict the probability of different class based on different attributes. This algorithm is generally used in text classification and with problems having various classes.

Pros and Cons of Naive Bayes:

Pros:

- It is very simple and fast to predict class of test data set. It also perform well in multi class prediction[14]
- When assumption of independence holds, a Naive Bayes classifier performs well compare to other models like logistic regression[14]
- Fast to train and Fast to classify[15]
- It perform better in case of categorical input variables compared to numerical variables[14]
- Handles real and discrete data[15]

Cons:

- If categorical variable has a category in test data set, which was not observed in training data set, then model will allocate a 0 (zero) probability and will be not be able to make a prediction. This is also known as "Zero Frequency" [14]
- Loss of accuracy[15]
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost difficult to get a set of predictors which are completely independent [14]

Applications of Naive Bayes Algorithms:

- Real time Prediction: Naive Bayes is a keen learning classifier and it is fast. Thus, it could be used for making predictions in real time[14]
- Multi class Prediction: This algorithm is also known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable[14]
- Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers mostly used in text classification (due to better result in multi class difficulties and independence rule) have greater success rate as compared to other algorithms. As a result, it is broadly used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)[14]
- Recommendation System: Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter hidden information and predict whether a user would like a given resource or not[14]

Table III. **Dataset Characteristics**

Each dataset has different characteristics and classifier performance depends on the dataset characteristics. The experiment is performed by using Heart-Disease dataset.

Dataset	No. of Instances	No. of Attributes	Nominal Count	Numeric Count	Class unit	Missing Value
Heart Disease	1028	13	12	1	2	0

C. Impact of Dataset on Classifier Performance

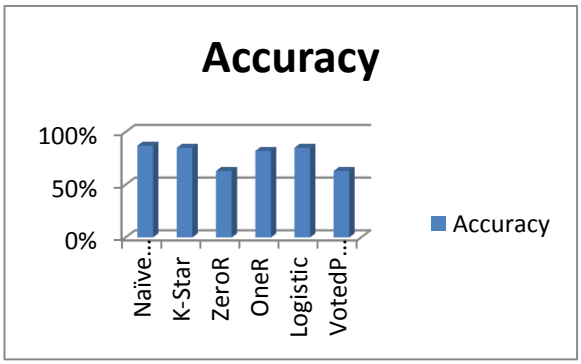
According to Bayes theorem, different algorithms are equivalent when their performance is averaged across all possible problems. The experiment is carried out in Weka 3.6 tool. Here dataset is selected with different characteristics and classifier is selected from different class. For evaluating performance of the classifier, accuracy is considered.

Table IV**. Performance of Different Classifiers based on Accuracy**

Classifiers	**Accuracy**
Naïve Bayes	86.38%
K-Star	85%
ZeroR	63%
OneR	82%
Logistic	85%
VotedPerceptron	63%

The above table shows the Performance of different Classifiers on the basis of accuracy.

In below Graph 1, Accuracy is considered as parameter to measure the performance in which Naïve Bayes gives better performance.



Graph 1 **Performance of dataset on different classifier according to column chart.**

V. CONCLUSION

In this paper we presented major research accomplishments and techniques that immerged in the field of Heart Disease prediction. The overall objective is to study the various data mining techniques available to predict the heart disease and to compare them to find the best method of prediction. This paper has delivered the summary of Naïve Bayes used for Heart Disease they classified. By using Naïve Bayes as Data Mining technique the chance of getting Heart Disease can be predicted which is helpful for timely detection of the disease.

VI. REFERENCES

[1] "Global atlas on cardiovascular disease prevention and control", WHO, 2011
[2] Mozaffarian D, Wilson PW, Kannel WB, Beyond established and novel risk factors: lifestyle risk factors for cardiovascular disease. Circulation 117: pp:3031–3038, 2008.
[3] Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. Am Heart J., 121: pp: 293–298, 1991.
[4] Kannel WB, An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol 110: pp:281–290, 1979.
[5] K. Thenmozhi , P.Deepika, "Heart Disease Prediction Using Classification with Different Decision Tree Techniques", International Journal of Engineering Research and General Science Volume 2, Issue 6, ISSN 2091-2730, October-November, 2014.
[6] Binal A. Thakkar, Mosin I. Hasan,Mansi A.Desai," HEALTH CARE DECISION SUPPORT SYSTEM FOR SWINE FLU PREDICTION USING NAÏVE BAYES CLASSIFIER" 2010 International Conference on Advances in Recent Technologies in Communication and Computing,IEEE,pp:101-105, 2010.
[7] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, IEEE,pp:108-115,2008
[8] Punam Bajaj, Preeti Gupta," Review on Heart Disease Diagnosis Based on Data Mining Techniques", International Journal of Science and Research (IJSR), Volume 3 Issue 5, May 2014
[9] Nidhi Bhatla,Kiran Jyoti," An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 8, October – 2012.
[10] Latha Parthiban, R.Subramanian,"Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm."International Journal of Biological and Life Sciences 3:3 2007.
[11] Centre for Disease Control and Prevention,," http://www.cdc.gov/heartdisease/risk_factors.htm"
[12] American Heart Association, "http://www.heart.org/HEARTORG/Conditions".
[13] M.A.Jabbar, B.L Deekshatulu, Priti Chandra," Computational Intelligence Technique for early Diagnosis of Heart Disease" International Conference on Engineering and Technology,IEEE,(ICETECH), 2015
[14] http://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained
[15] Monika Gandhi,Dr.Shailendra Narayan Singh," Predictions in Heart Disease Using Techniques of Data Mining", 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management, IEEE ,2015
[16] M.A.Nishara , Banu,B. Gomathy," Disease Forecasting System Using Data Mining Methods", 2014 International Conference on Intelligent Computing Applications, IEEE, 2014.