Journal of Applied Information Science
9 (1) 2021, 22-25
http://www.publishingindia.com/jais

An Efficient Heart Disease Detection System utilizing Naive Bayes Classification

Cherakuru Prasad
Seshachala Degree & P.G. College, Puttur, Andhra Pradesh, India.

Abstract: Coronary illness is one of the most basic human infections on the planet and influences human life seriously. Heart-related maladies or Cardiovascular Diseases (CVDs) are the primary explanation behind countless passings on the planet in the course of the most recent couple of decades and has risen as the most perilous infection, in India as well as in the entire world. In coronary illness, the heart cannot push the necessary measure of blood to different pieces of the body. Exact and on-time analysis of coronary illness is significant for cardiovascular breakdown anticipation and treatment. The analysis of coronary illness through customary clinical history has been considered as not solid in numerous angles. Along these lines, there is a need for a solid, exact and practical framework to analyze such illnesses in an ideal opportunity for appropriate treatment. The proposed Naïve Bayes grouping framework can without much of a stretch distinguish and arrange individuals with coronary illness from solid individuals. The proposed Naïve Bayes order-based choice emotionally supportive network will help the specialists to conclusion heart patients productively. A significant test in Data Mining is to manufacture exact and computationally productive classifiers for clinical application. In this paper we considered order rule digging for information disclosure and produced the guidelines by applying our created approach on Heart expire databases [1, 2, 3].

Keywords: Classification, Data mining, Heart disease, Naive Bayes.

I. Introduction

The enthusiasm for breaking down clinical information has developed colossally as of late, as clinical associations have found the capability of utilizing the patient information dissipated in different clinical frameworks as one rational entire for better understanding and the board of the clinical databases. To dissect information a large number of advances is required, to be specific advances from the territories of Data Mining, Machine Learning, Artificial insight and Data Visualization [2].

We see as of late different clinical associations are delivering tremendous measures of information which are hard to deal with. Emergency clinics have aggregated huge amounts of data about patients and their clinical accounts. Information digging is scanning for connections and examples that could give valuable information to powerful dynamic. Clinical information mining is one of the key issues to get helpful clinical information from clinical databases [3].

So, it is important to utilize information mining apparatuses to investigate the clinical records, and mine the covered up and significant information to deciding. Numerous strategies have been utilized for information mining. We consider characterization is one of the most helpful methods. A regulated AI task includes developing planning from input information (ordinarily depicted by a few highlights) to the suitable yields. In an arrangement learning task, each yield is at least one class of which the information has a place. The objective of characterization learning is to build up a model that isolates the information into various classes, with the point of arranging new models in future.

This is the mother purpose behind some related clinical issues like heart ambush, liver disillusionment, kidney dissatisfactions, nerves damages and vision setback. One of the significant true clinical issues is the recognition of coronary illness at its beginning phase.

The principle goal of this paper is the expectation of coronary illness utilizing Naïve Bayes characterization calculation. The objective is to separate the shrouded designs by applying information mining procedures on the dataset, which are significant to heart illnesses and to anticipate the nearness of coronary illness in patients where the nearness is esteemed on a scale. The expectation of coronary illness requires an enormous size of information which is excessively perplexing and monstrous to process and break down by customary procedures.

II. Related Study

Heart Disease

Heart is the most crucial organ in the human body in the event that that organ gets affected; at that point, it similarly impacts the other key pieces of the body. Thusly it is imperative for people to go for a coronary ailment investigation [1]. The most significant organ of the human body is the heart. The capacity of the heart is to siphon the blood and courses the whole body [5]. The coronary illness (HD) has been considered as one of the complex and life deadliest human infections on the planet. In this malady, as a rule, the heart cannot push the necessary measure of blood to different pieces of the body to satisfy the typical functionalities of the body, and because of this, at last, the cardiovascular breakdown happens. As per the World Health Organization (WHO), an expected 17 million individuals bite the dust every year from cardiovascular illness, especially respiratory failures and strokes [9].

The manifestations of coronary illness incorporate brevity of breath, a shortcoming of the physical body, swollen feet, and exhaustion with related signs, for instance, raised jugular venous weight and fringe edema brought about by utilitarian heart or non-cardiovascular irregularities [8]. The coronary illness finding and treatment are intricate, particularly in the creating nations, because of the uncommon accessibility of symptomatic mechanical assembly and lack of doctors and others assets which influence legitimate expectation and treatment of heart patients. The exact and appropriate determination of the coronary illness hazard in patients is fundamental for decreasing their related dangers of extreme heart issues and improving the security of heart [9].

The investigation of heart disease is a troublesome task, which can offer automated assumption regarding the heart condition of the patient with the objective that further treatment can be made feasible. Heart ailments join an alternate extent of messes: coronary hallway contaminations, stroke, hypertensive rheumatic coronary sickness, heart arrhythmia and various others. Thus, the acknowledgement of heart diseases from various components is an unpredictable issue. So coronary illness expectation is significant. This crisis has pushed the drive towards shield sedate, where the fundamental concern sees infirmity peril and making a move at the soonest sign [10].

III. Methodology

A. Naive Bayes Classifier

The Bayesian Classifier is equipped for ascertaining the most likely yield contingent upon the information. It is conceivable to include new crude information at runtime and have a superior probabilistic classifier. A Naive Bayes classifier is a term managing a straightforward probabilistic arrangement dependent on applying Bayes' hypothesis. In straightforward terms, a credulous Bayes classifier expects that the nearness (or nonattendance) of a specific component of a class is random to the nearness (or non-appearance) of some other element. Contingent upon the exact idea of the likelihood model, guileless Bayes classifiers can be prepared proficiently in a directed getting the hang of setting [4] [6]. Gullible Bayes classifiers frequently work much better in numerous mind-boggling certifiable circumstances than one may anticipate. Here free factors are considered with the end goal of expectation or event of the occasion.

In likelihood hypothesis, Bayes' hypothesis relates the restrictive and minimal probabilities of two arbitrary occasions. It is frequently used to figure back probabilities given perceptions [7]. For instance, a patient might be seen to have certain manifestations. Bayes' hypothesis can be utilized to register the likelihood that a proposed analysis is right, given that perception.

B. Bayesian Theorem

Given training data X, the posterior probability of a hypothesis H, P(H|X), follows the Bayes theorem $P(H|X) = \frac{P(X|H)P(H)}{P(X)}$

Leave X alone information tuple and H be some speculation with the end goal that the information tuple X has a place with a predefined class C. For arrangement issues, we need to decide P (H|X), the likelihood that the hypotheses H holds the given proof or watched information tuple X.

P (H|X) is the posterior probability of H conditioned on X

P (H) is the prior probability of H

P (X|H) is the posterior probability of X conditioned on H

P(X) is a prior probability of X

Algorithm

The Naive Bayes calculation depends on Bayesian hypothesis as given by the condition. Steps in the calculation are as per the following:

1. Each data sample is represented by an n dimensional feature vector, X = (x1, x2…..xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.
2. Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned if and only if: P (Ci/X)>P (Cj/X) for all 1< = j< = m and j! = i Thus we maximize P (Ci|X). The class Ci for which P (Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem.

3. As P(X) is constant for all classes, only P (X|Ci) P (Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P (C1) = P (C2) = ….= P (Cm), and we would therefore maximize P (X|Ci). Otherwise, we maximize P (X|Ci) P (Ci).

Note that the class prior probabilities may be estimated by P(Ci) = Si/s, where Si is the number of training samples of class Ci, and s is the total number of training samples on X. That is, the naive probability assigns an unknown sample X to the class Ci (2).

IV. Experimental Results

A. Data Set

We consider the Heart Disease Data of UCI Repository [10]. This Data set has 270 lines and 13 segments. So, in this information there are two class marks i.e., the absent class has 150 and present class has 120. Enlightening measurements can give us incredible knowledge into the state of each property of information. So, we can synopses Heart Disease information through clear measurements has exhibited Table I.

Table I: Heart Disease Data Attribute Information

Attribute ID	Attribute Definition
age	Age
sex	Sex
chest	Chest Pain Type
resting_blood_pressure	Resting Blood Pressure
serum_cholestoral	Serum Cholesterol in mg/dl
fasting_blood_sugar	Fasting Blood Sugar
resting_electrocardiographic_results	Resting electrocardiographic result
maximum_heart_rate_achieved	Maximum heart rate achieved
exercise_induced_angina	Exercised-induced angina
oldpeak	Old peak
slope	Slope
number_of_major_vessels	Number of major vessels
thal	Thal
class	Class label: absent, present

The assessments have been coordinated by using the Python programming language. It is an open-source programming language give astounding use of different data examination and Visualization procedures. It is a noteworthy library that gives numerous AI gathering figurings, capable contraptions for data mining and data assessment. The Python Scikit-learn is a pack for the data request, backslide, bundling and portrayal. The standard dataset is apportioned into two sets (70% and 30%), one for getting ready and another set for testing.

Performance of each classifier is measure in terms of the confusion matrix, sensitivity, specificity, precision, recall and accuracy. These metrics are traditionally defined for a binary classification task with positive and negative classes. That is:

Accuracy: Accuracy is a measure which determines the probability that how much results are accurately classified.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: Precision represents how precise the classifier predictions are since it shows the number of true positives that were predicted out of all positive labels assigned to the instances by the classifier. Precision is the proportion of positive predictions that are correct.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall: Recall is the proportion of positive samples that are correctly predicted positive. It shows the amount of truly predicted positive classes out of the amount of total actual positive classes.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

Where,

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

Table II: Confusion Matrix of Prediction Cases of Classification

		Predicted	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Confusion matrix is a visualization tool which is commonly used to present the accuracy of the classifiers in the classification that assist with performance evaluation purposes which consist of the concepts defined above measurements. This is illustrated in Table II. It is used to show the relationships between outcomes and predicted classes.

B. Results

The confusion matrix of Naive Bayes classification method is presented in Table III. The values to measure the performance of the methods (i.e. accuracy, precision, recall, and f1-score) are derived from the confusion matrix and shown in Table IV and same shown in the graphical representation in Fig. 1.

TABLE III: CONFUSION MATRIX OF HEART DISEASE DATA CLASSIFICATION

Testing Data (81)		
Desired Result	Output Result	
	Absent	Present
Absent	36	6
Present	9	30

TABLE IV: RESULTS OF HEART DISEASE PROPOSED NAIVE BAYES CLASSIFICATION

Accuracy	Precision	Recall	f1-score
81.48	82	81	81

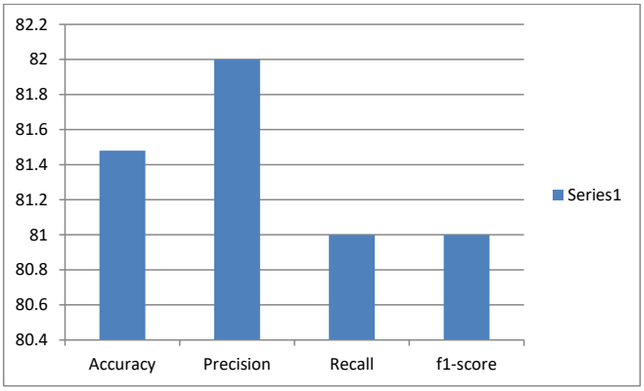

Fig. 1: Performance Metrics of Heart Disease Data

We observe in Fig. 1, The Naïve Bayes classifier algorithm gives a significant improvement in the accuracy. It is accomplished accuracy of 81.48%, precision got 82% and recall is achieved 81%.

V. CONCLUSION

In this paper, Naive Bayes classification of Data Mining has been discussed that can be used to predict the accuracy of Heart disease data. The accuracy or prediction rate of Naive Bayes is 81.48%. Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique. The system extracts hidden knowledge from a historical heart disease database. This is the most effective model to predict patients with heart disease. Hence, the proposed Naive Bayes Classifier approach will yield an effective method for both prediction and detection.

REFERENCES

[1] C. L. Blake, and C. J. Mertz, "UCI machine learning databases," 2004. [Online]. Available: http://mlearn.ics.uci.edu/databases/heartdisease/

[2] G. R. Kumar, G. A. Ramachandra, and K. Nagamani, "An efficient feature selection system for integrating SVM with genetic algorithm for large medical datasets," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 2, pp. 272-277, ISSN: 2277-128X, Feb. 2014.

[3] G. R. Kumar, V. S. Kongara, and G. A. Ramachandra, "An efficient ensemble based classification techniques for medical diagnosis," *International Journal of Latest Technology in Engineering, Management and Applied Sciences*, vol. 2, no. 8, pp. 5-9, ISSN: 2278-2540, Aug. 2013.

[4] I. H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., San Francisco: Morgan Kaufmann, 2005.

[5] H. G. Lee, K. Y. Noh, and K. H. Ryu, "Mining biosignal data: Coronary artery disease diagnosis using linear and nonlinear features of HRV," *LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining*, pp. 56-66, May 2007.

[6] J. Han, and M. Kamber, "Data mining concepts and techniques," The Morgan Kaufmann series in *Data Management Systems*, 2nd ed., San Mateo, CA: Morgan Kaufmann, 2006.

[7] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, A: Addision-Wesley, 2005.

[8] V. A. Sitar-Taut, et al., "Using machine learning algorithms in cardiovascular disease risk evaluation," *Journal of Applied Computer Science & Mathematics*, vol. 3, no. 5, 2009.

[9] The Atlas of Heart Disease and Stroke. [Online]. Available: http://www.who.int/cardiovascular_diseases/resources/atlas/en/

[10] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/