International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 9, Issue 6, June 2022)

Heart Attack Prediction System using Data Mining Techniques

Prof. Shadab Adam Pattekari
Computer Science and Enginnering
Principal, R.S.I.E.T., Pedhambe
Chiplun, India
shadabpattekari@gmail.com

Mr. Ajinkya Yadav
Computer Science and Enginnering
Lecturer, BSIET. Kolhapur
Kolhapur, India
yadhav.ajinkya008@gmail.com

***Abstract*— The main objective of this research is to develop an Intelligent System using data mining modeling technique, namely, Naive Bayes. It is implemented as web based application in this user answers the predefined questions. It retrieves hidden data from stored database and compares the user values with trained data set. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs.**

***Index Terms*— Data mining, heart disease, healthcare, Navie bayes**

I. INTRODUCTION

In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease. It is a world known fact that heart is the most essential organ in human body if that organ gets affected then it also affects the other vital parts of the body. Therefore it is very important for people to go for a heart disease diagnosis.

As a result of this people go to healthcare practitioners but the prediction made by them is not 100% accurate. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Heart Attack Prediction System (HAPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. HAPS can answer complex what if queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. HAPS are Web-based, user-friendly, scalable, reliable and expandable.

II. RESEARCH OBJECTIVE

The main objective of this research is to develop a prototype Health Care Prediction System using, Naive Bayes .The System can discover and extract hidden knowledge associated with diseases (heart attack, cancer and diabetes) from a historical heart disease database. It can answer complex queries for diagnosing disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results in tabular and PDF forms.

III. SCOPE OF THE PROJECT

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome[1].This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions .The main objective of this research is to develop a prototype Heart Attack Prediction System (HAPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network. So it provides effective treatments, it also helps to reduce treatment costs and also enhances visualization and ease of interpretation [8].

With immense knowledge and accurate data in that field. Large corporations invest heavily in this kind of activity

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 9, Issue 6, June 2022)

to help focus attention on possible events and risks that are involved. Such work brings together all available past and current data, as a basis on which to develop reasonable expectations about the future [2] [3].

IV. DATA SOURCES

Questionnaires have advantages over some other types of medical symptoms that they are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have[3][4]. Here our questionnaire is based on the attribute given in the data set, so the questionnaire contains:

A) Input attributes

Sr.No	Attribute	Description
1	Sex	value 1: Male; value 0: Female
2	Chest Pain Type	value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic
3	Fasting Blood Sugar	value 1: > 120 mg/dl; value 0:< 120 mg/dl
4	RestECG	resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5	Exang	exercise induced angina (value 1: yes; value 0: no)
6	Slope	the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7	CA	number of major vessels colored by floursopy (value 0 – 3)
8	Thal	value 3: normal; value 6: fixed defect; value 7: reversible defect
9	Trest Blood Pressure	mm Hg on admission to the hospital
10	Serum Cholesterol	mg/dl
11	Thalach	maximum heart rate achieved
12	Oldpeak	ST depression induced by exercise relative to rest
13	Age	In year
14	Height	In cms
15	Weight	In kgs

Table 1 Data set description

V. IMPLEMENTATION OF NAÏVE BAYES CLASSIFIER

A) Process of data mining Architecture

In the case of data mining there are many algorithms in existence such as decision tree, neural networks and naïve bayes out of these naïve bayes gives outstanding performance therefore we implement the naïve bayes theorem in order to achive a reasonably accurate result. Before getting to know about the naïve bayes classifier in detail, let us first come across some basic concepts such as

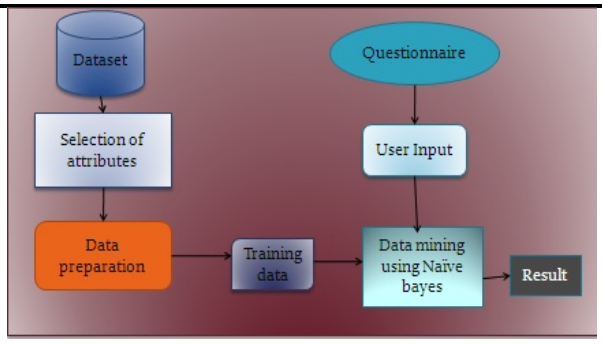

Figure.1Process of Data mining

B) Classifier

A classifier is a process of mapping from a (discrete or continuous) feature space *X* to a discrete set of labels *Y*. Here we are dealing about learning classifiers, and learning classifiers are divided into supervised and unsupervised learning classifiers. The applications of classifiers are wide-ranging. They find use in medicine, finance, mobile phones, computer vision (face recognition, target tracking), voice recognition, data mining and uncountable other areas [4].
An example is a classifier that accepts a person's details, such as age, marital status, home address and medical history and classifies the person with respect to the conditions of the project.

C) Naïve Bayes

In probability theory, Bayes' theorem (often called Bayes' law after Thomas Bayes) relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations.
For example, a patient may be observed to have certain symptoms. Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation [8] [9].

A naive Bayes classifier is a term dealing with a simple 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. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple [5] [6].

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. Naive Bayes classifiers often work much better in many complex real-world situations than

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 9, Issue 6, June 2022)

one might expect. Here independent variables are considered for the purpose of prediction or occurrence of the event.
In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers [7] [6].

An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

D) Theorem

Bayesian Rule
The Bayesian Rule is:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (1)$$

Where we could say that:

$$Posterior = \frac{Likelihood \times Prior}{Evidence} \quad (2)$$

So for example, if we apply this to a Spam Filter, then P(C) would be the probability that the message is Spam, and P(X|C) is the probability that the given word (input) is Spam, given that the message is Spam. P(X) is just the probability of a word appearing in a message using the given training data.

$$P(C|X_1, X_2, ..., X_n) = \frac{P(X_1, X_2, ..., X_n|C)P(C)}{P(X_1, X_2, ..., X_n)} \quad (3)$$

For the Bayesian Rule above, we have to extend it so that we have:
Where, if we continued to use the spam filter idea, X_1,….,X_n would be the input, or the words from the training data. Naive Bayes is called so because it makes the assumption that all the input attributes are independent, such as one word doesn't affect the other in deciding whether or not a message is spam.

E) Example – Training

Let's say we have a table that decided if we should play tennis under certain circumstances. These could be the outlook of the weather; the temperature; the humidity and the strength of the wind:

Day	Outlook	Temp	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Table 2 Data set of weather report.

So here we have 4 attributes. What we need to do is to create "look-up tables" for each of these attributes, and write in the probability that a game of tennis will be played based on this attribute. In these tables we have to note that there are 5 cases of not being able to play a game, and 9 cases of being able to play a game.

OUTLOOK	Play = Yes	Play = No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

TEMPERATURE	Play = Yes	Play = No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

HUMIDITY	Play = Yes	Play = No
High	3/9	4/5
Normal	6/9	1/5

WIND	Play = Yes	Play = No
Strong	3/9	3/5
Weak	6/9	2/5

We also must note that the P(Play=Yes) = 9/14 and P(Play=No)=5/14

F) Testing

For this, say we were given a new instance, and we want to know if we can play a game or not, then we need to look up the results from the tables above. So, this new instance is:

International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 9, Issue 6, June 2022)

X = (Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

Firstly we look at the probability that we can play the game, so we use the lookup tables to get:

P(Outlook=Sunny | Play=Yes) = 2/9
P(Temperature=Cool | Play=Yes) = 3/9
P(Humidity=High | Play=Yes) = 3/9
P(Wind=Strong | Play=Yes) = 3/9
P(Play=Yes) = 9/14

Next we consider the fact that we cannot play a game. Then, using those results, you have to multiple the whole lot together. So you multiple all the probabilities for Play=Yes such as:
(2/9) * (3/9) * (3/9) * (3/9) * (9/14) = 0.0053

For those of you who could be bothered to lookup the Play=No probabilities, you should get the answer of 0.0206, or something along those lines

So, given the probabilities, can we play a game or not? Well:
P(Yes | X) < P(No | X) => 0.0053 < 0.0206

So the answer is no, you can't play tennis.

VI.BENEFITS AND LIMITATIONS

HAPS can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It can also provide decision support to assist doctors to make better clinical decisions or at least provide a "second opinion." The current version of HAPS is based on the 15 attributes listed in Table 1. This list may need to be expanded to provide a more comprehensive diagnosis system. Another limitation is that it only uses categorical data. For some diagnosis, the use of continuous data may be necessary. Another limitation is that it only uses three data mining techniques. Additional data mining techniques can be incorporated to provide better diagnosis. The size of the dataset used in this research is still quite small. A large dataset would definitely give better results. It is also necessary to test the system extensively with input from doctors, especially cardiologists, before it can be deployed in hospitals.

VII.CONCLUSION

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. This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. HAPS can be further enhanced and expanded. For, example it can incorporate other medical attributes besides the above list. It can also incorporate other data mining techniques. Continuous data can be used instead of just categorical data.

HAPS can be further enhanced and expanded. For example, it can incorporate other medical attributes besides the 15 listed in Figure 1. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining

REFERENCES

[1] Blake, C.L., Mertz, C.J.: "UCI Machine LearningDatabases", http://mlearn.ics.uci.edu/databases/heartdisease/, 2004.
[2] Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "CRISP-DM 1.0: Step by step data mining guide", SPSS, 1-78, 2000.
[3] Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.
[4] Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications for scientific databases", Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.
[5] Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2003.
[6] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.
[7] Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.
[8] Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE
[9] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.
[10] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management. 16(4), 50-55, 2002.