Comparative Analysis of ID3 and Naïve Bayes for Heart Disease Prediction

Syed Ahmed Yasin
Assistant Professor,
JSPM's JSCOE,
Pune (M.S)

Nikhil Kumar B.S.
Assistant Professor,
JSPM's JSCOE,
Pune (M.S)

Ravindra P. Bachate
Assistant Professor,
JSPM's JSCOE,
Pune (M.S)

ABSTRACT
Improper working of blood vessels within heart causes heart disease .Hospitals are using medical application software for their day to day operation for billing and generation of simple satictcs.Multispecilaity hospitals are using expert system but they have some limitations. Heart dieses predication is difficult task because we need lot of patient historical data, medical history and it also depend on knowledge and experience of doctors. In this paper decision support systems made by two data mining techniques decision tree and naive bayes. Performance analysis is performed on both the methods.

General Terms
Heart Disease, Expert System, Decision Support System, Data Mining, Naive Bayes, and Decision Tree

Keywords
Heart Disease, Expert System, Decision Support System, Data Mining, Naive Bayes, and Decision Tree

1. INTRODUCTION
Data mining and disease predication has strong relationship as data mining will extract knowledge e from large amount of medical data which may be hidden that can be acquire through hospital paper based data base. Alternative term for data mining is knowledge discovery which consist of cleaning of dirty data, integration of different kind of data, selection and extraction of useful data and finally knowledge presentation. Survey shows that majority of deaths occurred during past decade is due to heart diseases [1].In Asia majority of deaths are due heart related problem.

2. SYSTEM ARCHITECTURE
Fig 1 shows complete architecture of system, starting from data collection to final stage of deployment of the system.Intially data preprocessing is performed on UCI and actual data set. On actual data set two additional attributes, smoking and obesity are included and system is designed for Naive bays and decision tree are created and in the last results are noted.Konwlege discovery steps are carried out, such as data preparation, data cleaning, data integration and knowledge presentation in user understandable format [2]Doctors can used this system for predication of heart disease.

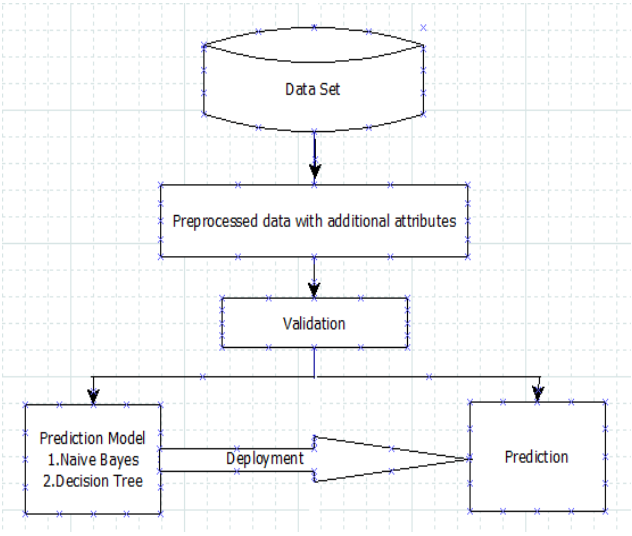


Fig.1 Architecture of the System

3. METHODOLOGY USED
Decision support system consists of data management, and knowledge management. For heart disease predication data is patient history and knowledge will be the result of predication. For this work two data mining techniques have been used and result is analyzed in following manner.

Fig.2 System Development

3.1 Naïve Bays Classifier
Naïve bays are a classification method works on bays theorem which assumes independence among attributes. The basis assumption is that presence of a particular feature in a class unrelated to presence of any other .Naive bays requires less amount of data as compared to other method for estimation of parameters for classification. For this reason mean and variance of all 15 attribute is calculated [6].

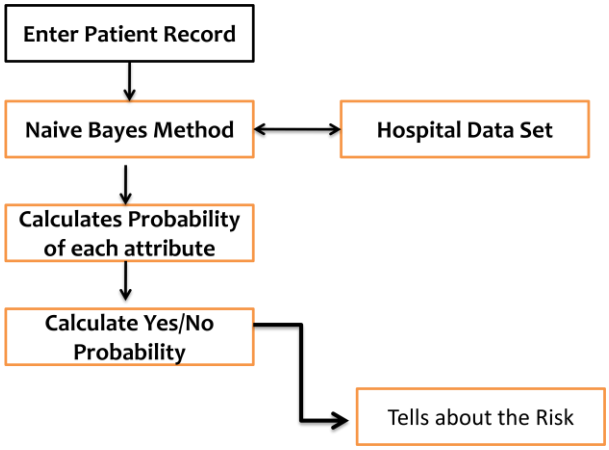
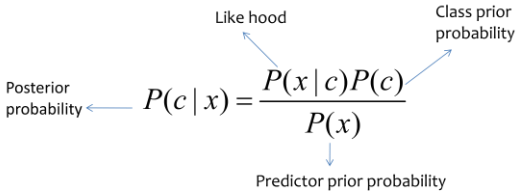


Fig.3 Naive bays Implementation

Bayes Theorem :



$$P(c \mid x) = P(x_1 \mid c) \times P(x_2 \mid c) \times \ldots\ldots\ldots P(x_n \mid c) \times P(c)$$

$$P(disease \mid symptom) = \frac{P(symptom \mid disease)P(disease)}{P(symtom)}$$

The Naive bayes classifier is based on Bayes theorem with independent assumptions between predictors

Continuous values associated with each class are distributed according to a Normal distribution [6].

Let the training data contain a continuous attribute x,

Probability Density of some value given a class can be find out by

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e - \frac{(v-\mu)^2}{2\sigma^2}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} xi \qquad \mu - Mean$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(xi-\mu)^2\right] \qquad \sigma - S\,\text{tandard deviation}$$

3.2 ID3 Decision Tree Algorithm

Iterative Dectomizer 3 is one of the algorithms of decision tree invented by Ross Qualanian which is used to build a decision tree from dataset.Id3 uses greedy method by selecting the best attribute to divide the dataset on each iteration.ID3 can handle over fitting of data for the small decision tree should prefer larger data set. One improvement we can make in ID3 to be use in backtracking during the search for optimal decision tree. Performance of ID3 and naive bays is compared based on accuracy, as per performance resulted is noted as decision tree has accuracy of 76.09% and for Naïve bays it is 81.14%.Above comparison shows that Naïve bays has more accuracy then ID3 algorithm.

ID3 Algorithm:

1. Establish Classification Attribute
2. Compute Classification Entropy.
3. For each attribute in R, calculate Information Gain using classification attribute.
4. Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).
5. Remove Node Attribute, creating reduced table.
6. Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows in the reduced table[7].

The formula for entropy is:

$$H(S) = -\sum_{x \in X} p(x)\log^2 p(x)$$

Where,

S = current (data) set for which entropy is being calculated,

p(x) =Proportion of the number of elements in class x, to the number of elements in set [7].

Information Gain :

In decision trees, nodes are created by singling out an attribute. ID3aim is to create the leaf nodes with homogenous data. That means it has to choose the attribute that fulfils this requirement the most. ID3 calculates the "Gain" of the individual attributes. The attribute with the highest gain results in nodes with the smallest entropy.

$$IG(A) = H(S) - \sum_{t \in T} p(t)H(t)$$

Where,

H(S) = Entropy of set *S;T*= The subsets created from splitting set *S;* by attribute *A* such that, *p(t)* = the proportion of the number of elements in *t* to the number of elements in set *S*,

H(t) = Entropy of subset *t*[7].

4. RESULTS

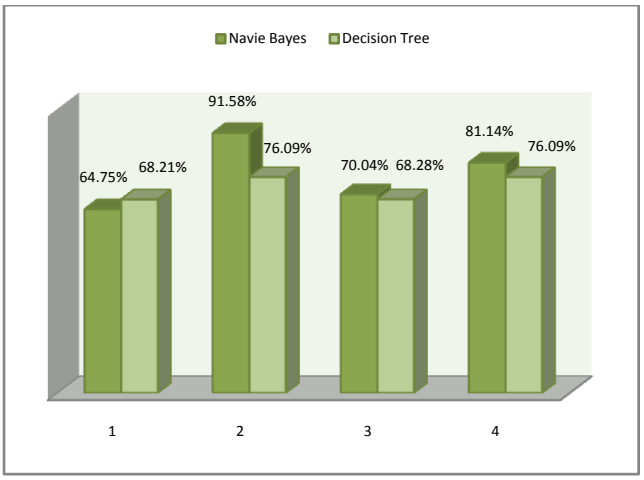
For both the techniques result noted, which is as follows

Techniques	Accuracy With		Accuracy With	
	13 Attributes of UCI Data	13 Attributes of Actual Data	13 Attributes of UCI Data	13 Attributes of Actual Data
Naive Bayes	64.75%	91.58%	70.04%	81.14%
Decision Tree	68.21%	76.09%	68.28%	76.09%

When 13 attributes of UCI data set is used for Naive Bayes then accuracy gain is 64.75%, and for decision tree it is 68.21%.Similaraly when 13 attributes and actual data set is used accuracy for Naive Bayes is 91.58% and for decision tree it is 76.09%. When 15 attributes of UCI data is used then accuracy gain is 70.04% and for decision tree it is 68.21%.Similaraly when 15 attributes of actual data set is used then accuracy gain for Naive Bayes is 81.14%, and for decision tree it is 76.09%. In this case accuracy of Naive Bayes algorithm is 81.14% which is 10% less then above case of 13 attributes of UCI data set, because a Naive Bayes classifier assumes that if a particular features is present or not present is unrelated to other features presence or absence. In

different case of 15 attributes of actual data set, number of attributes are increased as well as size of data set is also increased that's way system produces decrement in results by 10%.

Fig.2 Graphical Representation of Result



5. CONCLUSION

1. The system is able to extract and mine hidden information or knowledge from a heart disease dataset both models are trained and validated against test dataset.
2. Classification method is used to evaluate the effectiveness of result.
3. After applying both the techniques result is analyzed, both techniques has some advantages and disadvantages.

6. FUTURE WORK

1. The System can be further enhanced and expanded for future work. For example, it can incorporate other medical attributes such as height, weight family history, besides the 15 which are used in this paper.
2. It can also incorporate other data mining techniques, such as Clustering and Association Rules etc.
3. The size of the data set is also need to be expanded for better accuracy results as result is dependent of quantity of data set.

7. APPLICATIONS

1. Heart disease prediction system can be used as a training tool to train nurses and medical students
2. Can be used as a decision support for doctors to make better decisions and provide a second opinion
3. The system is user-friendly, scalable & reliable

8. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

9. REFERENCES

[1] K.Sudhakar, Dr. M. Manimekalai, "Study of Heart Disease Prediction Using Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277128, Volume 4, Issue 1, January 2014.

[2] Milan Kumar,Sunita Godara,"Comaparatuve Study of data mining classification methods in cardiovascular disese",International journal of engineering science and technology, Volume 2,Issue 3 ,pp 470-478 June 2012.

[3] N.Aditya Sundar,"Performance Analysis of classification data mining techniques over heart disease data base" international journal of engineering science and technology,Volume2,Issue 3 pp 470-478,June 2012.

[4] Joyti Soni,"Predication Data mining for medical Diagnosis", International Journal of Computer Application, pp 0975-887, Volume 17, No 8 March2011.

[5] M.A Jabbar,"Cluster Based Association Rule mining for heart attack predication", IJATIT, ISSN 1992-8645, 2011.

[6] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[7] https://en.wikipedia.org/wiki/ID3_algorithm