A COMPARATIVE STUDY ON DECISION TREE AND BAYES NET CLASSIFIER FOR PREDICTING DEABETES TYPE 2

Nipjyoti Sarma[1], Sunil Kumar[2], Anupam Kr. Saini[3]
NITTTR, Chandigarh

ABSTRACT
Datamining provides efficient algorithm for implementing different classification problem related to various decision support system. Here the decision tree algorithm and naive bayes and bayes net classifier are important classification algorithm which gives better performance. These classifiers can be efficiently used in bioinformatics problem. One very important problem is the prediction of diabetes mellitus disease in a person by observing the symptoms and so that proper diagnostic could be done. In this paper we are simply trying to put forward the approach of decision tree and bayesnet classifier in predicting disease by surveying different paper in this area and mention the accuracy parameters to measure the performance. The comparision and other surveying ware done upon the Pima Indian Dataset.

Keywords: *decision tree, bayesnet classifier, specificity, accuracy, precision*

I. INTRODUCTION
Diabetes is a disease in which the body could not produce insulin or sometimes could not use the produced insulin properly. This leads to gathering of glucose particle in the blood instead of going into the body cell. The gathering of the glucose particle in the blood invites various kinds of instabilities in the body. Normally the diabetes disease can be considered in two classes, one is type to which is called insulin dependent and the other type which is called the insulin independent. The type 1 is normally seen in children of less age group whereas the type two which is also called insulin independent is normally seen in the adult people. Statistics showed that this disease is increasing day by day. According to the International Diabetes Federation, there are currently 246 million diabetic people worldwide, and this number is expected to rise to 380 million by 2025.[1]. Therefore it is of utmost important to diagnose the disease at its early stage, so that it can be prevented or delayed to reduce costs and to save human life. Since now a day the medical information systems in hospitals and medical institutions become larger and larger and process of extracting useful information becomes more difficult. Therefore the Traditional manual data analysis has become inefficient and methods for efficient computer based analysis are much needed. This leads to may approaches to computerized data analysis have been considered and examined. The datamining is the best suitable approach since it is the main analytical tool of this era .It is also showed that the the benefits of introducing data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to save human resources.[2]

In this context different data mining classifier algorithm are used to construct prediction model. Out of these two very basic classifiers are Decision tree classifier and Bayesnet classifier. These two are very basic classifier and are very efficient in predicting whether or not a person is going to develop diabetes or not. In this paper we are going to discuss some properties of this two classifier and their prediction model, accuracy in context of the PID dataset.

II. THE DATASET
This article presents the accuracy comparison of classifier using decision tree and naive bayes implementing on the Pima Indian Diabetic Dataset. This is a popular dataset from the National Institute of Diabetes and Digestive and Kidney Diseases [11]. Several constraints were placed on the selection of the instances of the dataset from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm; The attributes of the dataset are given below:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

III. METHODOLOGY USED BY BOTH THE CLASSIFIERS

The decision tree is a very efficient classifier for classification of binary class value attribute[3]. For this purpose the dataset should be properly pre-processed using various methods like descretization, normalization, transformation. The missing value present in the dataset should be appropriately removed by proper measure. For his the mean value for that attribute can be used o sometimes the whole instance is deleted otherwise it will make the classifier inefficient. In the PID dataset the missing values are present in certain fields and are processed accordingly. As specified in [3] the dataset is firstly pre-processed by using
attribute identification and selection process, then appropriate methods are used to remove the missing values present in some attribute like Pregnant, Triceps SFT, Serum–Insulin etc. Different methods are available for handling missing values like k nearest neighbourhood algorithm. The K-nearest neighbour method replaces missing values in data with the corresponding value from the nearest-neighbour column.[4] The nearest-neighbour column is the closest column in Euclidean distance. However, sometimes this technique can bias the dataset. The other task is descretization which is essential for constructing decision tree. The WEKA datamining tool cpuld be used for this purpose. After performing numerical descritization the decision tree could be constructed. WEKA is a very nice tool for implementing the decision tree algorithm [5]. Here the dataset could be classified by choosing the J48 algorithm which is a decision tree learner and is the implementation of Quinlan C4.5 in Weka software. The test method could be used as 10 fold cross validation which gives the best result by classifying the dataset into 10 different folds and considering one fold as testing and other folds as training. Normally it is the best method for performing j48 algorithm in WEKA[6].It can be drawn as given below[3]:

Similarly as we can see the bayes net classifier for classifying the diabetes data. As we know that the naive bayes classifier is [12, 13] is a popular classifier. Programs that assign a class from a predefined set to an object or case under consideration based on the values of descriptive attributes. They do so using a probabilistic approach, i.e., they try to compute conditional class probabilities and then predict the most probable class. The NaIvebayes classifier finds the , marginal probabilities of symptoms P(si) and diseases P(dj),from a training set of patient data and conditional probabilities of symptoms on all diseases P(sildj) . These can be calculated by counting frequencies in the data. Here to find out the posterior probability of a patient the following equation can be used to calculated by using a given set of symptoms (S {si}):[10]

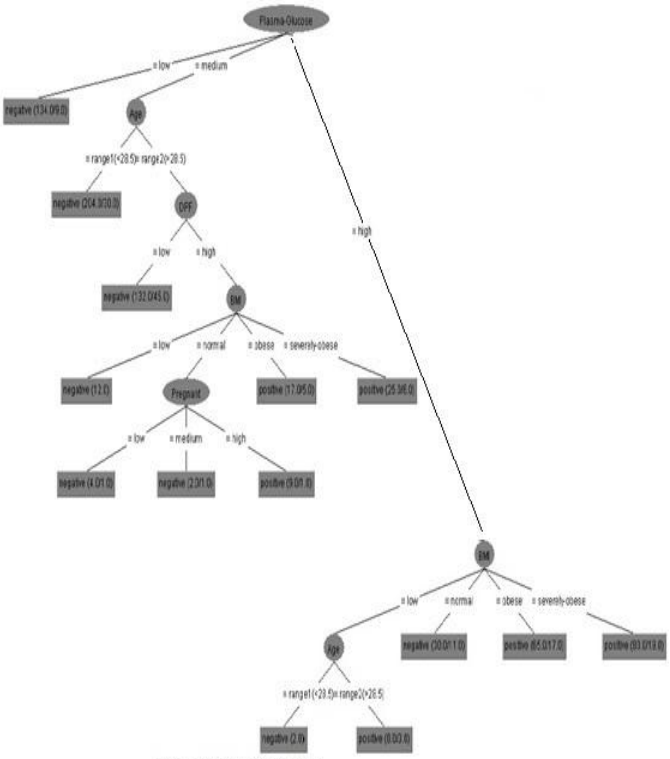

Fig 1 : A Sample Decision tree

$$p(d_i|s) = pd_j \prod s_i \in s \frac{p(s_i|d_j)}{p(s_i)}$$

The diagnostic value could be calculated by using the following equation:

$$p(dj|s) = p(d_j \prod_{s_i \epsilon s} p_{(s_i|d_j)}$$

This classifier learns from observed data of the conditional probability [12]of each variable Sj , where the class label is S.Then the classification is done by applying Bayes rule to compute the probability p(S Isi, ..., sn) and then predicting the class with the highest posterior probability. Where the variables ,S j are conditionally independent on class S.

The bayes net can be drawn with the help of the WEKA tool as shown below and appropriately can construct the Bayesian classifier model[10].

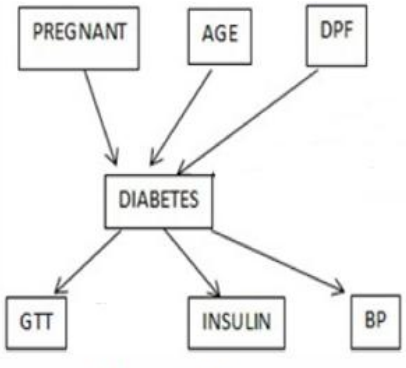

Fig 2 : A Bayes network

IV. PERFORMANCE MEASUREMENT

After performing the decision tree as well as the bayes net classifier, the confusion matrix is created, which is of the form given below[13]

Predicted class

	yes	no	Total
Yes	TP	FN	P
No	FP	TN	N
total	P'	N'	**P+N**

and the accuracy of the model is measured s given below:

Accuracy =TP+TN/(TP+TN+FP+FN)

Precision=TP/(TP+FN)

Specificity=TN/(TN+FP)

Where TP=True Positive, FP=False Positive, TN=True Negative, FN=False Negative

These measures showed that The decision tree can give accuracy of 78% to 80% [3][7], Which is the best accuracy without implementing any neural network structure.
Mostly the accuracy result of the bayes net classifier is placed between 71% to 74% depending upon the number of cross validation applied on the dataset when performing the test.

V. DISCUSSION AND CONCLUSION

From the above it is observed that the methodologies of data preprocessing applied in case of the decision tree algorithm as well as in the case of naive bayes classification is some what similar , like implementation od Attribute selection and identification, normalization, descretization etc. Butthe dissimilarity here is the implementation of the algorithm approach. In case of the decision tree algorithm implementation the bining or descretization gives a better accuracy in the overall performance of the classifier, it happens because due to the bining method the classification on the basis of the value of the bin form different branches of the tree.The attribute selection method of this algorithm is responsible for determining the splitting criteria of nodes . Here the the nodes arethe condition and the edges are the implications that if the condition (that is the parent node) holds then the children the conditional parent would be the result if it is a terminal node otherwise the children would be another condition . In this way the whole decision tree is constructed. The accuracy of this approach is better compared to the naive bayes classifier for predicting whether a patient goes to develop diabetes or not .The naive bayes uses the probability model of symptoms and disease and ultimately gives the probability of number of instances having both symptoms and disease. This is the accuracy of naive bayes classifier. However its predictive accuracy is small compared to the decision tree method. Other methods like neural network, fuzzy systems gives higher accuracy while building classifiers.

REFERENCES

[1] International Diabetes Federation, Diabetes Atlas, 3rd ed. Brussels,Belgium: International Diabetes Federation, 2007
[2] R. Bellazzi, "Telemedicine and diabetes management: Currentchallenges and future research directions," J. Diabetes Sci. Technol.,vol. 2,no.l,pp. 98-104,2008
[3] Decision Tree Discovery for the Diagnosis of Type II Diabetes BY Asma A. AlJarullah
[4] Jayalskshmi, T.; Santhakumaran, A.; , "Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks,"Machine Learning and Computing (ICMLC), 2010 SecondInternational Conference on , vol., no., pp.109-112, 9-11 Feb. 2010
[5] I. H. Witten and E. Frank, "Data mining," Practical Machine Learning Tools and Techniques, 2000.
[6] Andrew Roberts, "AI32 —Guide toWeka", 2005.
[7] Jianchao Han; Rodriguez, J.C.; Beheshti, M.; , "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner," Future Generation Communication and Networking, 2008. FGCN '08. Second International Conference on , vol.3, no., pp.96-99, 13-15 Dec.2008
[8] P. Langley, W. Iba, and K. Thompson. An Analysis of Bayesian Classifiers. Proc. 10th Nat. Coni. on Artificial Iritelliyence (AAAI'92, San Jose, CA, USA), 223-228. AAAI Press and MIT Press, Menlo Park and Cambridge, CA, USA 1992
[9] P. Langley and S. Sage. Iiiductiori of Selective Bayesian Classifiers. Proc. 10th Corif. u7r Wricertozrrty zrr Arlsjiciul Irrlelliyence (UAI '94, Seattle, W A, USA), 399-406. Morgan Kaufinarl, Sail Mateo, CA, USA 1994
[10] Using Bayes Network for Prediction of Type-2 Diabetes by Yang Guo Guohua Bai and Yan Hu
[11]https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
[12]http://en.wikipedia.org/wiki/Naive_Bayes_classifier#Probabilistic_model
[13] Data mining Concepts and Techniques by Jiawei Han , Micheline Kamber , Jian Pei by Moorgan Kaufman publishers