Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method

G. Parthiban
Research Scholar,
Dr. MGR Educational
Research and Institute,
Maduravoyal,
Chennai, India.

A. Rajesh
Professor, Dept of CSE
C.Abdul Hakkeem College
of Engineering and
Technology,
Melvishram,
Vellore, India.

S.K.Srivatsa
Sr. Professor, Dept of
E & I,
St.Joseph's College of
Engineering,
Chennai, India.

ABSTRACT
The objective of our paper is to predict the chances of diabetic patient getting heart disease. In this study, we are applying Naïve Bayes data mining classifier technique which produces an optimal prediction model using minimum training set. Data mining is the analysis step of the Knowledge Discovery in Databases process (KDD). Data mining involves use of techniques to find underlying structures and relationships in a large database. Diabetes is a set of related diseases in which body cannot regulate the amount of sugar specifically glucose (hyperglycemia) in the blood. The diagnosis of diseases is a vital role in medical field. Using diabetic's diagnosis, the proposed system predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease.

Keywords: Knowledge Discovery**,** Data Mining, Diabetes, Heart disease, Naïve Bayes Method.

1. INTRODUCTION
Knowledge discovery in medical databases is well-defined process and data mining is an essential step. Data mining is the non trivial extraction of potential useful information about data**.** **[**1][2] Thus data mining should have been more appropriately named "knowledge mining from data". [3] Diabetes Mellitus is a chronic disease which causes serious health complications including renal (kidney) failure, heart disease, stroke, and blindness [4].People with diabetes either do not produce enough insulin (type 1 diabetes) or cannot use insulin properly (type 2 diabetes), or both. Type1 diabetes was also called Insulin Dependent Diabetes Mellitus (IDDM) or Childhood-onset diabetes. Type2 diabetes was also referred to as Non-Insulin Dependent Diabetes Mellitus (NIDDM) or Adult-onset diabetes [5]. Type1 diabetes is typically recognized in childhood or adolescence. At least 90% of patients with diabetes have type2 diabetes and it is typically recognized in adulthood where the body cannot effectively use the insulin produced [6] [13]. The causes of diabetes mellitus are unclear, however, there seem to be both hereditary (genetic factors passed on in families), and environmental factors involved.

The risk factors for type 2 diabetes are being 45 years of age or older, being overweight, having a parent or sibling with diabetes (family heredity), having high blood pressure (140/90 or higher), having high cholesterol (HDL 35 or lower; triglycerides 250 or higher) and acute stress**.** [7] Over 80 per cent of people with type 2 diabetes are overweight and it is treated with diet and exercise, the blood sugar level is lowered with drugs. [8] [15] A family history of diabetes research has shown that people are more at risk if there is a history of diabetes in close family members. The physical inactivity research has shown that people who do not lead an active life are more at risk of developing type 2 diabetes [9][14].

Diabetes also increases the risk of micro-vascular damage and macro-vascular complications. People with diabetes are two to four times more likely to get cardio vascular diseases. Thus diabetes is found to be one of the leading causes of global death by disease. There are several methods in the literature individually to diagnosis diabetes or heart disease. There is no automated diagnosis method to diagnose Heart disease for diabetic patient based on diabetes diagnosis attributes to our knowledge.

In this paper, we propose a Naïve Bayes based method to diagnose heart disease for diabetic patients. It should be noted that the attributes used in our proposed method are those used for diagnosis of diabetes and are not direct indicators of heart disease.

2. BACKGROUND
Naïve Bayes Classifier is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature [10].

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

Prob (B given A) = Prob(A and B)/Prob(A)

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone.

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. Since independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire. It can be used for both binary and multi class classification problems.

3. EXPERIMENTAL METHODOLOGY

3.1 Data set and used variables

The data set used in this work are clinical data set collected from one of the leading diabetic research institute in Chennai and contain records of about 500 patients. The clinical data set specification provides concise, unambiguous definition for items related to diabetes.

The diabetes data set is developed to ensure people with diabetes have up to date records of their risk factors, current management, treatment target achievements and arrangements and outcomes of regular surveillance for complications, to help them monitor their care and make informed choices about their management. It will also ensure that when people with diabetes meet health care professionals the consultation is fully informed by comprehensive, up to date and accurate information.

The diabetes attributes used in our proposed system and their descriptions are shown in Table 1.

Table 1 Diabetes attributes used in the experimentation

Attribute	Description
Sex	A classification of the sex of the person
Age	Age of the patient
Family Heredity	Previous history (Father / Mother)
Weight	Patient's weight
BP	Blood pressure
Fasting	Sugar level after fasting
PP	Post Prandial blood glucose level
A1C	HbA1c level Glycosylated Last 4 months sugar level
LP Tot Cholesterol	Total cholesterol level

3.2 Preprocessing and Sampling

Except for the attributes sex and family heredity all the other attributes listed in Table 1 have numeric values. The attribute sex takes on values 'M' or 'F' to denote male or female respectively. The attribute family heredity takes on values 'Father', 'Mother' or 'Both'. In case there is no previous diabetes history for the patient the attribute is left empty.

Since no attribute value should be left empty for the mining algorithm to work properly, we have used the value 'No' for patients without any previous diabetes history. Likewise, we need to have a categorical attribute based on which the data sets are to be classified. The aim of our work is to predict the chances of a diabetic patient getting heart disease. Hence, we have taken the 'LP Tot Y/N' attribute as the class attribute. Since the 'LP Tot Y/N' attribute is a numeric attribute, we have categorized the attribute values into high cholesterol value ('Yes') or low cholesterol value ('No').

This categorization has been done based on the fact that a cholesterol value of 180 or more is taken to be high cholesterol for Indians.

3.3. Data Analysis

The distribution of the attribute values with respect to the class attribute 'LP Tot Y/N' is shown in Figure 1.

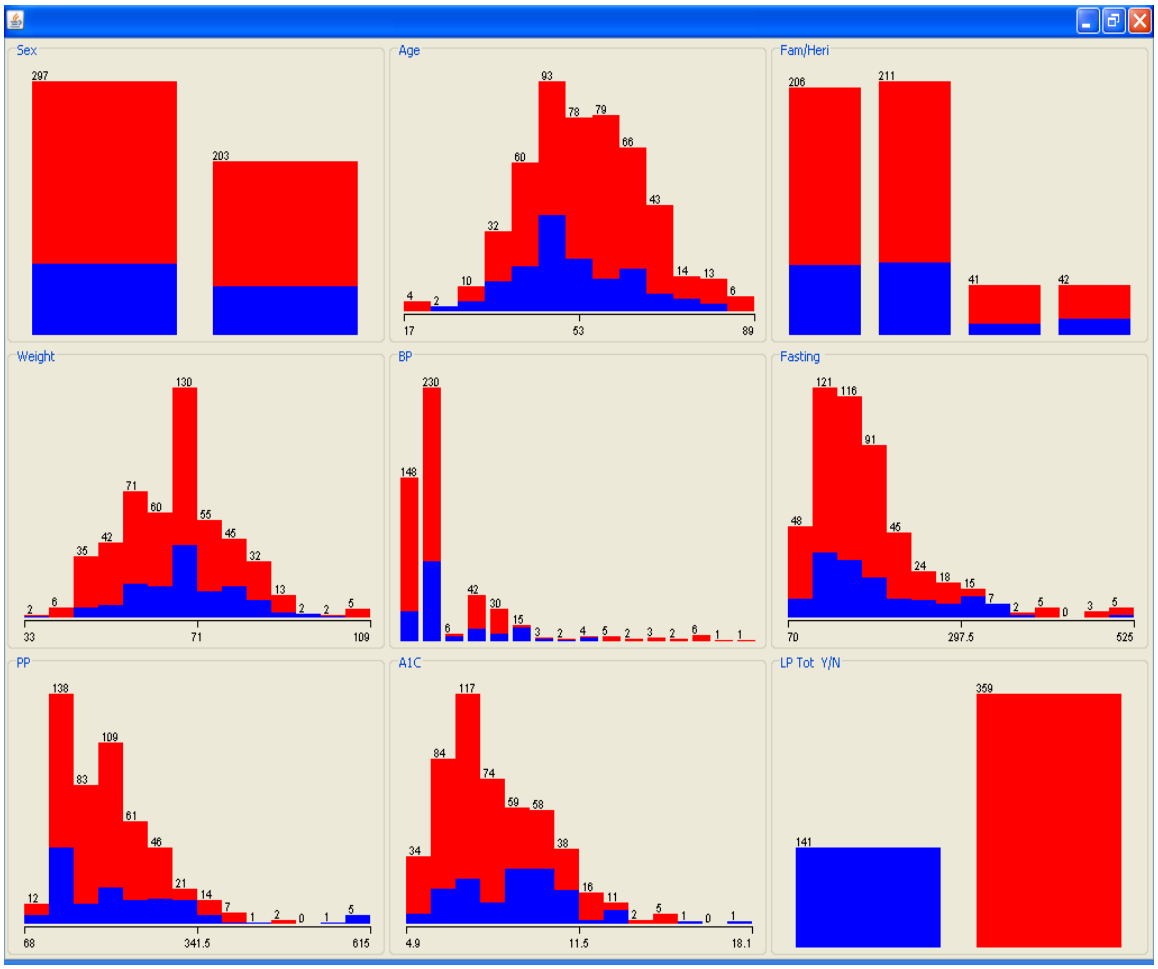

Figure 1 Attribute value distributions with respect class attribute LP Tot Y/N

The blue colored regions in the graphs in Figure 1 denote high cholesterol values. From the graphs we can see that, most of the diabetic patients with high cholesterol values are in the age group of 45 – 55, have a body weight in the range of 60 – 71, have BP value of 148 or 230, have a Fasting value in the range of 102 – 135, have a PP value in the range of 88 – 107, and have a A1C value in the range of 7.7 – 9.6.

3.4. Using Data Mining in data set

The WEKA ("Waikato Environment for Knowledge Analysis") tool is used for Data mining. Data mining finds valuable information hidden in large volumes of data. Weka is a collection of machine learning algorithms for data mining tasks, written in Java and it contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. [11] The key features of Weka are it is open source and platform independent. It provides many different algorithms for data mining and machine learning [12]. We have used Naïve bayes method to perform the mining and classification process. We have used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process.

4. RESULTS AND DISCUSSION

The results of our experimentation are shown in Figure 2.

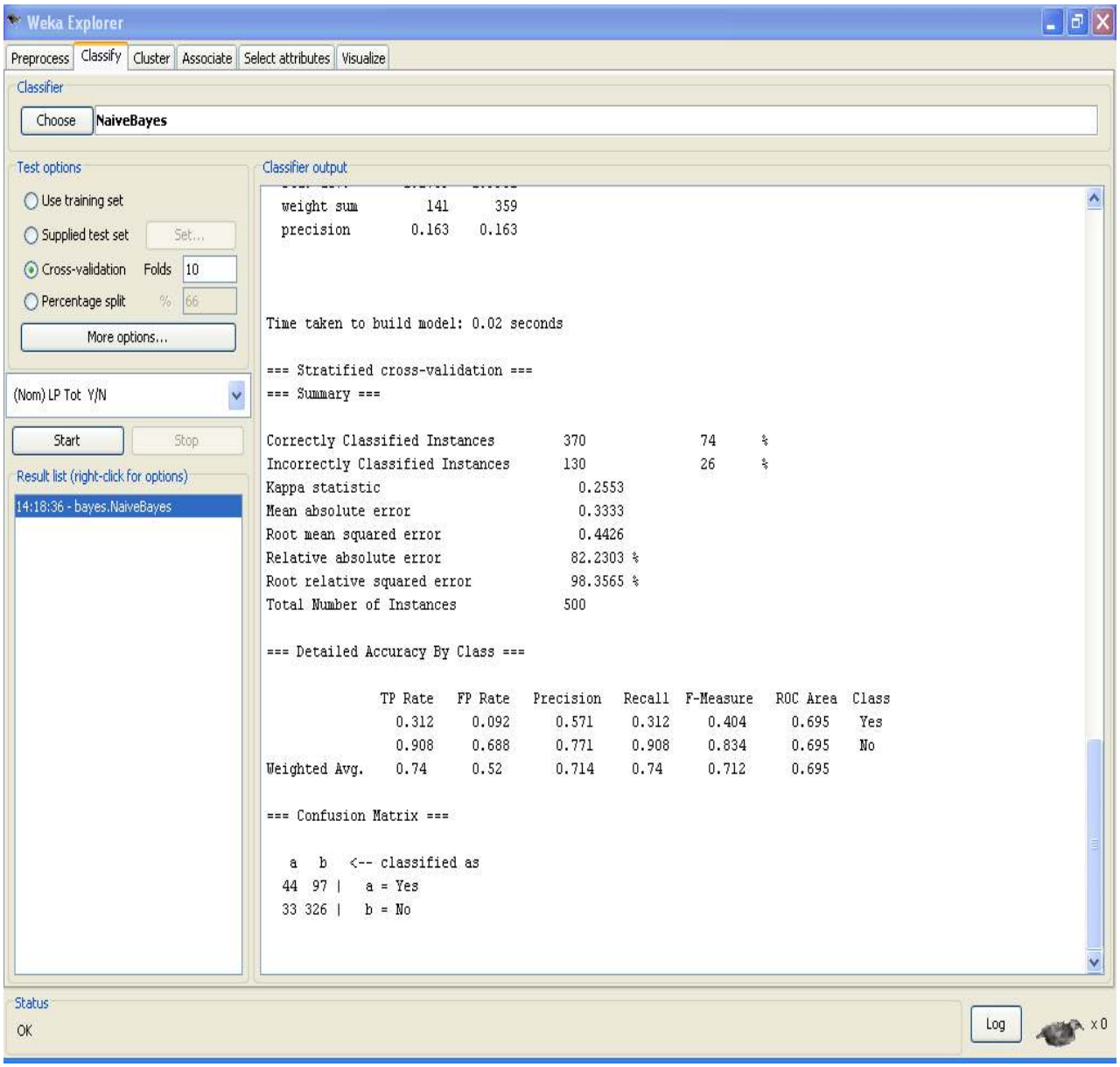

Figure 2 Result window of the data mining process

The proposed naïve bayes model was able to classify 74% of the input instances correctly. It exhibited a precision of 71% in average, recall of 74% in average, and F-measure of 71.2% in average. The results show clearly that the proposed method performs well compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

5. CONCLUSIONS AND FUTURE ENHANCEMENTS

Application of Data mining in analyzing the medical data is a good method for considering the existing relationships between variables. From our proposed approach we have shown that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict.

In our work we have tried to predict the chances of getting a heart disease using attributes from diabetic's diagnosis. This can be extended to predict other type of ailments which arise from diabetes, such as visual impairment in future. Further, the data analysis results can be used for further research in enhancing the accuracy of the prediction system in future.

6. ACKNOWLEDGEMENTS

We are grateful to Dr.V.Shesiah, Chairman and Managing director of Dr.V.Shesiah Diabetic Research Institute, Chennai for providing an access to medical diabetic data and for his involvement in this domain.

7. REFERENCES

[1] Frawley and Piatetsky-Shaprio, 1996. Knowledge Discovery in Databases – An Overview. The AAAI/MIT Press, Menlo Park,C.A.

[2] Cios, K. J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L. A. 2007. Data Mining: A Knowledge Discovery Approach, New York: Springer.

[3] Han, J., Kamber, M. 2006. Data Mining: Concepts and Techniques, 2nd ed. San Francisco: Morgan Kaufman.

[4] World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: http://www.who.int/topics/diabetes mellitus/en/

[5] Diabetes mellitus doctor's knowledge in MedicineNet : http://www.medicinenet.com/diabetes mellitus/page2.htm#toce.

[6] I. International Diabetes Federation, "Diabetes Atlas third edition", IDF 2007.

[7] M.Franciosi and M.Sacco, "Use of the diabetes risk score and impaired glucose tolerance", Diabetes care Vol.28,no.5, pp 1187-2005.

[8] Kelling, D.G. and J.A. Wentworth et al., 1997, Diabetes mellitus. Using a database to implement a systematic management program. NC.Med.J.,58:368-371.

[9]International Diabetes Federation(IDF), http://www.idf.org/about-diabetes

[10] Naïve bayes classifier based on applying bayes theorem: http://en.wikipedia.org/wiki/Naive bayes classifier

[11] Weka Data mining software http://www.cs.waikato.ac.nz/ml/weka

[12] An Introduction to the WEKA Data mining system - http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

[13] Jianchao Han, Juan C. Rodriguze, and Mohsen Beheshti, 2008. Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner. In Proceedings of the Second International Conference on Future Generation Communication and Networking.

[14] Asuncion, A., Newman, D. J. 2007. Pima Indians Diabetes Data Set, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabets, Irvine, CA: University of California, School of Information and Computer Science.

[15] Eleni Georga et al, 2009. Data Mining for Blood Glucose Prediction and Knowledge Discovery in Diabetic Patients: The METABO Diabetes Modeling and Management System. In Proceedings of the 31st Annual International Conference of the IEEE EMBS Minneapolis, Minnesota, USA.