Indian Journal of Science and Technology, Vol 9(29), DOI: 10.17485/ijst/2016/v9i29/93880, August 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease

K. R. Anantha Padmanaban* and G. Parthiban

Department of Computer Science and Applications, SRM Arts and Science College, Kattankulathur - 603203, Tamil Nadu, India; toananths@yahoo.com, trgparthi@gmail.com

Abstract

Objective: This paper aims at predicting the early detection of chronic kidney disease also known as chronic renal disease for diabetic patients with the help of machine learning methods and finally suggests a decision tree to arrive at concrete results with desirable accuracy by measuring its performance to its specification and sensitiveness. **Methods:** The behaviour of learning algorithms determined on a set of data mining indicators has a proportionate effect on the resulting models. Discovering the knowledge from wide databases is termed as Data mining. Besides studying the existing available Clinic Foundation Heart Disease dataset, 600 clinical records collected by us from a leading Chennai based diabetes research centre. We have tested the dataset for classification using Naïve Bayes and Decision tree method. **Findings:** On comparing the classification algorithms with respect to Naïve Bayes and Decision tree, we came to conclusion that the accuracy is up to 91% for Decision tree classification. **Applications/Improvement:** In order to increase the accuracy of the prediction result, we have utilized algorithms such as neural network and clustering data which greatly helped in our mission and also gave scope for future research.

Keywords: Chronic Kidney Disease (CKD), Data Mining, Decision Tree, Diabetes, Naïve Bayes Method

1. Introduction

Data mining is becoming more popular nowadays in healthcare, as also in fraud, abuse detection etc. In[1] classification is a more useful data mining function to handle items in a collection to target categories or classes. In[2] kidney failure falls one among several classes viz heart disease, blindness etc which results due to chronic Diabetes[3]. Dialysis is the only method to keep the kidneys function artificially and it is also painful and expensive process. According to World Health Organization about millions of people around the world are suffering from severe kidney disorder and its number is increasing every year[4,5]. Therefore, an early diagnosing technique is immediately required so that precautions or controls can be taken before hand in time.

For obtaining essential information from medical databases Data mining technique was found very much useful. By combining machine learning and statistical analysis intelligently, very useful information can be drawn from medical databases. Machine learning methods which coordinates various statistical analyses and databases helps us to extract hidden patterns and relationships from huge and multiple variable data. In order to ensure the chosen classifier's accuracy the available test phases are verified. Moreover, these attributes like Specificity, sensitivity, and accuracy are common for disease detection. By applying Naïve Bayes and Decision tree techniques for our desired classification methods, we were able to achieve the result of identification of Kidney disorder at the early stages.

2. Methods

Data mining techniques were earlier used by many researchers for prediction purpose. Identification of several diseases are done by combindly utilising and analysing machine learning, statistical and medical database exercises[6,7].

*Author for correspondence

The two different methods viz classification and evaluation which comes under Data mining technique, helps to build training data, classified predictive model and also testing the classification efficiency[8]. The guidance for future activities are also derived and attained from the above classification data[9,10]. Several corporate alsoapply Rapid miner software for theirresearch, training etc., for their good governance. Rapid miner software is also being used ineducation and rapid prototyping etc., which very much help in coordinating the activities in machine learning, data mining, text mining, predictive analytics and business analytics etc[11]. The two data mining tools such as WEKA (Waikato Environment for Knowledge Analysis) and YALE (Yet Another Learning Environment) both written in Java, developed at the University of Waikato, is useful in contributing to commerce and industry, knowledge bases, scientific and clinical research.

The proposed architecture is shown in Figure 1.

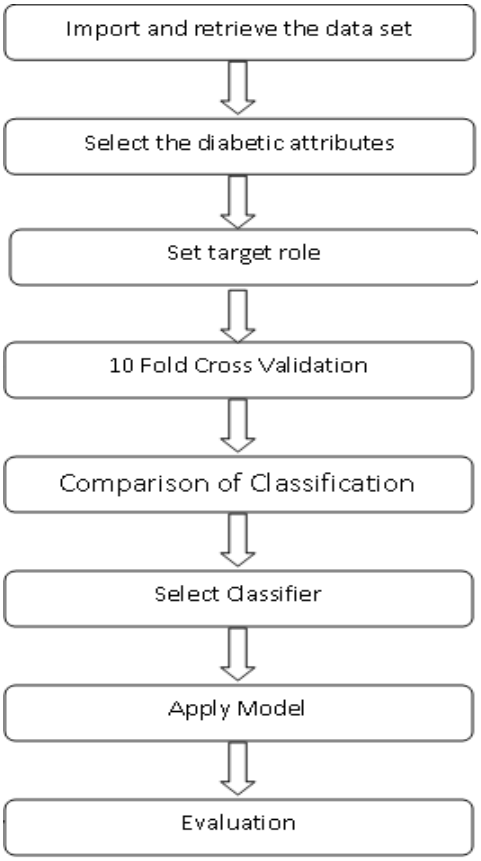

Figure 1. Proposed architecture.

With the help of the above tool an activity design with 10-Cross validation a Machine Learning algorithm is developed to enhance the performance level. This validation is done by dividing the data into 10 components each consisting 90% of the original data. Cross-validation will be very much useful to create a set of training data with validation folds. We compared the expected error on the results using the training data.

The attributes data view of each record is shown in Table 1. Here M denotes Male, F denotes Female, Y denotes Yes and N denotes No.

Table 1. Attributes used in our experimentation

Name	Type	Description
Sex	Binomial	M / F
Age	Integer	Age of the patient
Heredity	Polynomial	Father, Mother, Both
Weight	Numeric	Weight of the patient
Smoking	Numeric	Y / N
BP	Polynomial	Blood pressure of the patient
Fasting	Integer	Fasting Blood Sugar
PP	Integer	Post prondial Blood Glucose
A1C	Numeric	Glycosylated Hemoglobin Test
LDL	Integer	Low Density Lipoprotein
VLDL HDL	Integer Numeric	Very Low Density Lipoprotein High Density Lipoprotein
Risk Class	Polynomial	Vulnerability of the patients to Kidney disease.

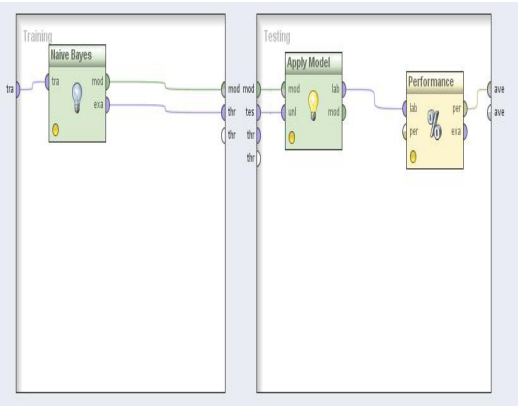

Figure 2. Naïve Bayes performance screen.

2.1 Naives Bayes Method

The Bayes theorem is applied using Naïve Bayes Classifier with assumptions to deal with conditional simple prob-

abilities. Using Bayes theorem, the probability of an event occurred is found[12,13]. If 'A' denotes prior event and 'B' denotes dependent event, Bayes' theorem can be given as

Prob(BgivenA) = Prob(AandB)/Prob(A)

To compute the framework required for classifying and determining the variances of variables for each class only a few items of training data are required, since independent variables are assumed.

The Naïve Bayes activity screen is shown in Figure 2.

2.2 Decision Tree

It is a prediction method used to construct classification or regression model in a tree structured type[14]. While it disintegrate in to smaller subsets also simultaneously develops decision nodes and leaf nodes. The two types of Decision trees used in data mining are classification tree analysis and Regression tree analysis for different predicted outcome such as belonging to particular data class or real number.

A statistical classifier C4.5 an algorithm developed by Ross Quinlan (1996) used to produce a Decision tree[15] is taken in our review and its working is given in Figure 3.

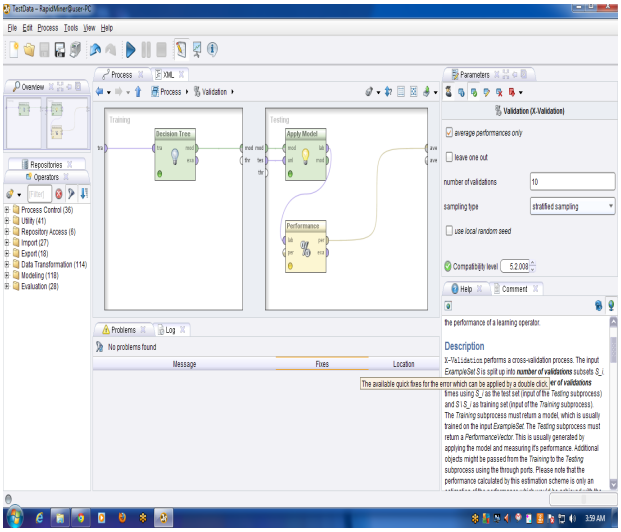

Figure 3. Decision tree performance screen.

3. Results and Discussion

A model with machine learning algorithm with validation is prepared with utmost accuracy. One trained sub-set is tested out of 10 classifier and cross validation is arrived. Thus the percentage classification data is arrived. With the help of Naïve Bayes classification technique normal distribution probabilities are assumed. The prime attributes identified for evaluating through Naïve Bayes method areage, sex, smoking, alcohol, cholesterol HDL etc.

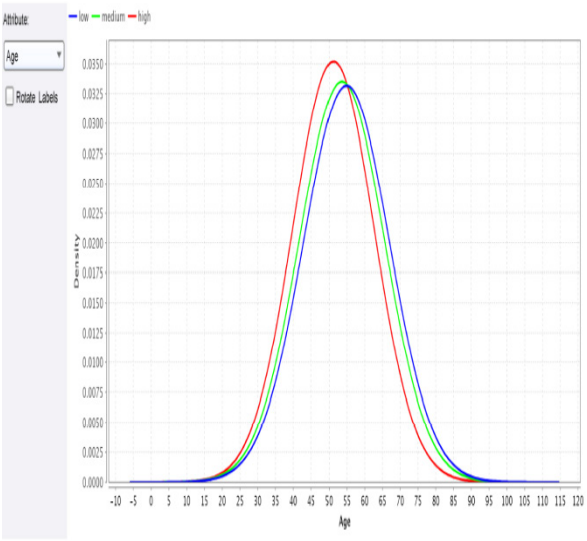

Figure 4. Bayes distribution plot by age attribute.

The risk is considered high, medium and low at the ages 51, 53 and 54 respectively which is denoted in X and Y axis as shown in Figure 4. Where X represents age and Y as density. With calculated value of above five attributes, age attributes are assigned and shown in the form of distribution Table 2. Here L, M, H denotes Low, Medium and High respectively.

Table 2. Naïve Bayes distribution table for age attribute

Attribute	Parameter	L	M	H
Age	Mean	53.92	52.62	50.78
Age	Standard deviation	11.03	10.80	10.42

With the help of True Positive (TP) and True Negative (TN) classifications the performance of the model is evaluated.

For a multiclass classification problem, we can define TP, FP, TN and FN for each class i. Similarly we define counts TPi, FPi, T,Ni, and FNi for the class i. Then certain parameters can be calculated to evaluate the multiclass classification results accordingly. For e.g., True Positive

Rate TPR, Precision and f-Measure value for each class and the overall accuracy.

TP_Rate = TPi / TPi + FNi

TP Rate is the parameter to measure how well the classifier can predict the true memebers for particular class. However, TP rate alone is not sufficient to fully measure performance of the classifier in a single class, therefore we compute Precision for class i as,

Precision i = TPi /TPi + FPi

F-score or F-measure is a measure of a test's accuracy and it is the harmonic mean of precision and recall which can be calculated as F = 2 * (Precision * Recall / Precision + Recall).

We can also compute F-Measure for class i as,

F=2 *(Precision i + TP_ratei / Precision i + TP_ratei

The performance of the Naïve Bayes classifier for the data sets given is shown in Table 3.

In Table 3 TL, TM, TH denotes True Low, Medium, High class values and PL, PM, PH denotes Predicted Low, Medium, High class values.

Table 3. Bayes distribution accuracy: 85.77%

Class Values	TL	TM	TH	Precision
PL	600	49	18	87.48%
PM	33	92	21	61.12%
PH	9	21	82	69.49%
Recall	91.46%	50.97%	62.86%	

Table 4. Accuracy by split methods using decision tree

Split method criteria	Accuracy in percentage	Classification error in percentage
Gain ratio	83.49	11.71
Information gain	90.69	9.21
Gini Index	76.69	26.31

The different levels of accuracy viz, Gain ratio, Information gain and Gini index in the form of decision tree using various split methods are shown in Table 4.

A classifier performance is predicted and measured based on error rate. The overall performance is measured based on the proportion of error and ratesin each set of instances. This is shown in Figure 5 and Figure 6 respectively.

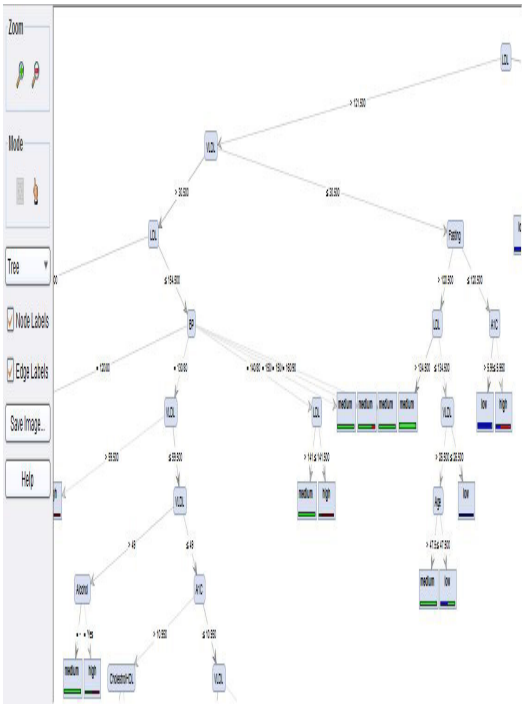

Figure 5. Decision tree diagram.

Tree

LDL > 121.500
| VLDL > 30.500
| | LDL > 154.500
| | | VLDL > 43.500: high {low=1, medium=0, high=62}
| | | VLDL 43.500
| | | | Age > 43: high {low=0, medium=1, high=23}
| | | | Age 43
| | | | | Age > 39.500: medium {low=0, medium=4, high=0}
| | | | | Age 39.500: high {low=0, medium=0, high=2}
| | LDL 154.500
| | | BP = 120/80
| | | | VLDL > 49.500
| | | | | Fasting > 133.500: high {low=0, medium=1, high=6}
| | | | | Fasting 133.500: medium {low=0, medium=8, high=0}
| | | | VLDL 49.500: medium {low=0, medium=26, high=0}
| | | BP = 130/80
| | | | VLDL > 59.500: high {low=0, medium=0, high=6}
| | | | VLDL 59.500

Figure 6. Decision tree text view.

The decision tree diagram shown in Table 4 gives the results of the distinguished data, the accuracy of which was found to be 90.69%.

In Table 5 TL, TM, TH denotes True Low, Medium, High class values and PL, PM, PH denotes Predicted Low, Medium, High class values.

Table 5. Performance of decision tree with an accuracy of 91% using information gain as split parameter

Class Values	TL	TM	TH	Precision
PL	650	19	8	93.81%
PM	12	135	20	79.97%
PH	3	9	99	83.00%
Recall	95.48%	81.00%	73.89%	

Table 6. Accuracy of NB and DT classification techniques

Technique	Accuracy
Naïve Bayes (NB)	86%
Decision Tree (DT)	91%

From Table 6, it can be arrived that compare to naïve Bayes method, Decision tree method adopted for the two models appear to be most effective since the correct predictions shown in 91% for patients with chronic kidney disease.

4. Conclusion

We arrived to a conclusion that application of Data mining technique for different analysis of medical data is a good method. The performance of Decision tree method was found to be 91% accurate compared to naive Bayes method. Classification algorithm on diabetes dataset performance was obtained as94% Specificity and 95% Sensitivity. We also found that mining helps to retrieve correlations from attributes which are not direct indicators of the class which we are trying to predict. We are also further working on enhancing the performance of prediction system accuracy in neural network and clustering algorithm data analysis.

5. References

1. Koh HC, Tan G. Data mining applications in healthcare. Journal of Healthcare Information Management. 2005; 19(2):64–72.
2. Purushotaman G, Krishnakumari P. A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology. 2015 Jun; 8(12):1–5.
3. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia. Available form: http://www.who. int/diabetes/en
4. Lakshmi KR, Nagesh Y. Performance comparison of three data mining techniques for predicting kidney dialysis survivability. International Journal of Advances in Engineering and Technology. 2014 Mar; 7(1):242–54.
5. National Kidney Foundation (NKF). The facts about Chronic Kidney Disease (CKD). Available from: https://www.kidney.org/kidneydisease/aboutckd
6. Jena L, Kamila NK. Distributed data mining classification algorithms for prediction of chronic kidney disease. International Journal of Emerging Research in Management and Technology. 2015 Nov; 4(11):110–8.
7. Vijayarani S, Dhayanand S. Data mining classification algorithms for kidney disease prediction. International Journal on Cybernetics and Informatics. 2015 Aug; 4(4):13–25.
8. Sinha P, Sinha P. Comparative study of chronic kidney disease prediction using KNN and SVM. International Journal of Engineering Research and Technology. 2015 Dec; 4(12):608–12.
9. Han J, Kamber M, Pei J. Data mining: Concepts and techniques. 2nd ed. San Francisco: Morgan Kaufman; 1996.
10. Fayyad U, Piatetsky-Shapiro, Smyth P. From data mining to knowledge discovery in databases. AI Magazine; 1996. p. 37–54.
11. Rapid Miner. Machine learning software getting started. Available from: http:// rapidminer.com/learning/getting-started/
12. Naïve bayes classifier based on applying bayes theorem. Available from: http://en.wikipedia.org/wiki/Naive bayes-classifier
13. Naïve bayes classifier. Bayes theorem. Available from: http://en.wikipedia.org/wiki/Naive bayes classifier
14. Sudha A, Gayathri P, Jaisankar N. Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications. 2012 Apr; 43(14).
15. Decision Tree, C4.5 Algorithm. Machine Learning Algorithm Description. Available from: http://en.wikipedia.org/wiki/C4.5_algorithm