I J C T A, 9(28) 2016, pp. 301-306

HYBRID DATA MINING TECHNIQUES FOR ACCURATE DIABETIC PREDICION

S. Sathya* A. Rajesh and K. Bhuvaneswari*****

Abstract: This research focus on producing a hybrid data mining technique by combining different data mining algorithm features in to one. Many researchers implemented different data mining algorithms to predict the accuracy of diabetes, but the level of accuracy is not up to the expected level. In this paper the two data mining algorithms Random tree and ID3 were implemented with the diabetic dataset with 13 attributes and 642 instances. The implementation has been done with the popular data mining tool Weka. It is found that Random tree yields the accuracy of 94.7867 % and ID3 gives the accuracy of 96.3665 %. The proposed hybrid model combines the features of Random Tree and ID3 algorithm and produces the improved accuracy of 99.0521 %. This paper concludes that Hybrid model produces improved rate of accuracy to predict the diabetic.

Keywords: Data Mining, ID3, Random Tree, Hybrid Model and Diabetes Prediction.

1. INTRODUCTION

The World Health Organization estimates that nearly 250 million people all over the world suffer from diabetes and this number is likely to be doubled by 2030 and 85% of the diabetes deaths occur in middle-income countries. In India, there are nearly 70 million diabetics, according to the statistics of the International Diabetes Federation. As the incidence of diabetes is on the rise, doctors say, there is a proportionate rise in the complications that are associated with diabetes. The disease has been named the fifth deadliest disease in the United States with no imminent cure in sight. This disease has many side effects such as higher risk of eye disease, higher risk of kidney failure, and other complications. However, early detection of the disease and proper care management can make a difference.

According to the American Diabetes Association, 20.9 million children and adults in the United States (i.e., approximately 8% of the population) were diagnosed with diabetes. Thus, the ability to diagnose diabetes early plays an important role for the patient's treatment process. By using data mining techniques it takes less time for the prediction of the disease with more accuracy [3].This causes sugar to build up in your blood leading to complications like heart disease, stroke, and neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage, and death. General Symptoms of Diabetes are increased thirst, increased urination, Weight loss, increased appetite, Fatigue, Nausea and/or vomiting -Blurred vision, Slow-healing infections and Impotence in men. Diabetic results in Multi organ failure in a human body and it is necessary to predict and prevent earlier.

1.1 Diabetes Prediction

Diabetes is a major health problem in Saudi Arabia. Diabetes is the most common endocrine disease across all population and age groups [1]. Diabetic is the most common form of eye problem affecting people with diabetes, usually only affects people who have had diabetes for a long time period and can result in blindness. Diabetes mellitus, or simply diabetes, is a set of related diseases in which the body cannot regulate the amount of sugar in the blood. It is a group of metabolic diseases in which a person

* Research Scholar, Dept of CSE, St.Peter's University, Chennai.
Assistant Professor, Dept of CSE, C.Abdul Hakeem College of Engineering & Technology, TamilNadu, India
** Professor & Head, Department Of CSE, C.Abdul Hakeem College of Engineering & Technology, Tamilnadu, India
*** Assistant Professor, Department of CSE, C.Abdul Hakeem College of Engineering & Technology, TamilNadu, India

has high blood sugar, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced. Diabetic retinopathy the most common diabetic eye disease, is caused by complications that occurs when blood vessels in the retina weakens or distracted [2]. Diabetes complications can be prevented or delayed by early identification of people at risk [7].

There are two types of Diabetes. Type 1 - Diabetes also called as Insulin Dependent Diabetes Mellitus (IDDM), or Juvenile Onset Diabetes Mellitus is commonly seen in children and young adults however, older patients do present with this form of diabetes on occasion. Type II - Diabetes is also called as Non-Insulin Dependent Diabetes Mellitus (NIDDM), or Adult Onset Diabetes Mellitus. Preventing the disease of diabetes is an ongoing area of interest to the healthcare community. Diabetes is one of the high prevalence diseases worldwide with increased number of complications, with retinopathy as one of the most common one. Diabetes is a major chronic disorder which has no cure. The diagnosis of diabetes is a significant and tedious task in medicine[4].

1.3 Weka Tool

Data mining is an approach which can help in decision making [9]. Weka is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. Weka is a state-of-the-art facility for developing machine learning (ml) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks.

The algorithms are applied directly to a dataset. Data mining approach helps to diagnose patient's diseases. Diabetes Mellitus is a chronic disease to affect various organs of the human body [5]. Weka implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools. The new machine learning schemes can also be developed with this package. Weka is open source software issued under the GNU general public license.

2. MATERIALS AND METHODS

2.1 Diabetic Dataset Used

The data collected by medical and healthcare industry is not turned into useful information for effective decision making [6]. The data set is obtained from a Local Health Center. The data set includes 13 essential attributes and 642 instances needed for diabetic prediction. The availability of huge amount of patient's data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in diagnosis of diseases[10]. This work has been implemented to change data in to information. The attributes involved is tabulated below for reference in Table 1.

Table 1.
Data set Attributes Used

S.No	*Attribute*	*Abbreviations*
1.	PID	Patient ID
2.	Sex	Sex
3.	Age	Age
4.	Weight	Weight
5.	BP	Blood Pressure
6.	Type	Type of Diabetic
7.	Fasting	Fasting (Empty Stomach)
8.	HDl	High Density Lipo Protein
9.	LDl	Low Density Lipo Protein
10.	HDl/LDl	Ratio of Cholesterol
11.	Height	Height of Patient
12.	Heriditory	Heriditory of Patient
13.	Category	Category (Normal/Diabetes/Prediabetes)

2.2 Random Tree Implementation

Random Tree Implementation with the diabetic data set has been done with weka tool and the results obtained were tabulated below for reference. The number of correctly classified instances was 600 out of 642 yielding 94.7867 % accuracy. The other measures were tabulated in the Table 2.

Table 2.
Measures obtained from Random Tree

Measure	*Value*
Correctly Classified Instances	600 (94.7867 %)
Incorrectly Classified Instances	33 (5.2133 %)
Kappa statistic	0.9174
Mean absolute error	0.0346
Root mean squared error	0.1775
Relative absolute error	8.2251 %
Root relative squared error	38.6994 %
Coverage of cases (0.95 level)	95.8926 %
Mean rel. region size (0.95 level)	34.0706 %

Table 3.
Other Measures of Random Tree

	TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*
N	0.987	0.012	0.987	0.987	0.987	0.975	0.987	0.980
D	0.911	0.028	0.922	0.911	0.917	0.887	0.951	0.897
P	0.910	0.034	0.899	0.910	0.904	0.873	0.955	0.870
Weighted Avg.	0.948	0.022	0.948	0.948	0.948	0.926	0.970	0.931

The other measures TP Rate, FP Rate, Precision, Recall, F-measure, and ROC Area with respective weighted average are given in the above Table-3.

2.3 Naïve Bayes Implementation

Naive Bayes classifier is a well known type of classifiers. A set of programs that assign a class of predefined set to an object under construction based on the descriptive attributes[8]. Naïve Bayes Implementation with the diabetic data set has been done with weka tool and the results obtained were tabulated below for reference. The number of correctly classified instances was 610 out of 642 yielding 96.3665% of accuracy. The other measures were tabulated in the Table 4 Measures Obtained from Naïve Bayes below for reference.

Table 4.
Measures Obtained from Naïve Bayes

Measure	*Value*
Correctly Classified Instances	610(96.3665 %)
Incorrectly Classified Instances	23(3.6335 %)
Kappa statistic	0.9426
Mean absolute error	0.0304
Root mean squared error	0.1375
Relative absolute error	7.233 %
Root relative squared error	29.9828 %
Coverage of cases (0.95 level)	98.7362 %
Mean rel. region size (0.95 level)	36.3876 %

Table 5.
Other Measures of Naïve Byes

	TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*
N	0.990	0.003	0.997	0.990	0.993	0.987	0.996	0.997
D	0.893	0.002	0.993	0.893	0.941	0.923	0.984	0.982
P	0.987	0.044	0.880	0.987	0.931	0.909	0.990	0.934
Weighted Avg.	0.964	0.013	0.967	0.964	0.964	0.951	0.991	0.977

The other measures TP Rate, FP Rate, Precision, Recall, F-measure, and ROC Area with respective weighted average are given in the above Table-5.

2.4 Implementing Hybrid Data mining Technique

The Hybrid Proposed Model Implementation with the diabetic data set has been done with weka tool and the results obtained were tabulated below for reference. The number of correctly classified instances was 627 out of 642 yielding 99.0521% accuracy. The other measures were tabulated in the Table 6 Measures Obtained from Proposed Hybrid Model below for reference.

Table 6.
Measures obtained from Proposed Hybrid Model

Measure	*Value*
Correctly Classified Instances	627(99.0521 %)
Incorrectly Classified Instances	6 (0.9479 %) 0.986
Kappa statistic	0.985
Mean absolute error	0.0107
Root mean squared error	0.0794
Relative absolute error	2.5334 %
Root relative squared error	17.3032 %
Coverage of cases (0.95 level)	99.0521 %
Mean rel. region size (0.95 level)	33.3333 %

Table 7.
Other Measures of Proposed hybrid Model

	TP Rate	*FP Rate*	*Precision*	*Recall*	*F-Measure*	*MCC*	*ROC Area*	*PRC Area*
N	0.997	0.006	0.994	0.997	0.995	0.991	0.997	0.997
D	0.988	0.002	0.994	0.988	0.991	0.988	0.994	0.981
P	0.981	0.006	0.981	0.981	0.981	0.974	0.985	0.954
Weighted Avg.	0.991	0.005	0.991	0.991	0.991	0.986	0.993	0.982

The other measures TP Rate, FP Rate, Precision, Recall, F-measure, and ROC Area with respective weighted average are given in the above Table-7 for reference.

3. RESULTS AND DISCUSSION

The results obtained in implementing Random Tree, Naïve Bayes and Hybrid proposed models were shown in the Table .8 below.

Table 8.
Comparison of Accuracies

S.No	*Data Mining Tecnique*	*Accuracy*
1	Random Tree	94.79%
2	Naïve Bayes	96.37%
3	Hybrid	99.05%

The results obtained were charted in the figure shown below. It is proved that Hybrid Model provides high accuracy in predicting diabetes.

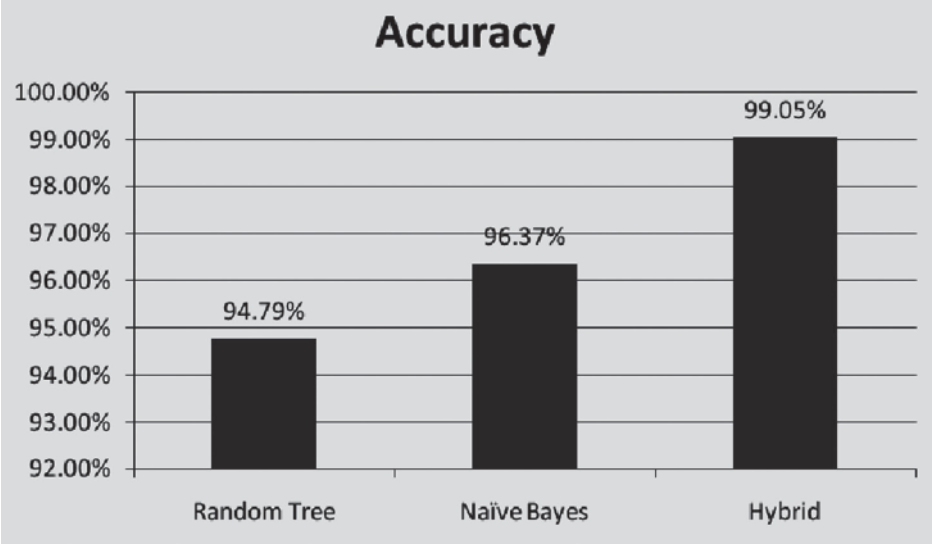

Figure 1. Comparing Accuracies

4. CONCLUSION

It is found that Random tree yields the accuracy of 94.7867 % and ID3 gives the accuracy of 96.3665 %. The proposed Hybrid model combines the features of Random Tree and ID3 algorithm and produces the improved accuracy of 99.0521 %. This paper concludes that Hybrid model produces improved rate of accuracy to predict the diabetic. The future enhancement required is the data set used consist of impure and inconsistent data and data cleaning may be implemented in future to obtain 100 percent accurate diabetic prediction results.

References

1. Aljumah, A. A., Ahamad, M. G., &Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*, *25*(2), 127-136.

2. Ananthapadmanabhan, K. R., &Parthiban, G. (2014).Prediction of chances-diabetic retinopathy using data mining classification techniques.*Indian Journal of Science and Technology*, *7*(10), 1498-1503.

3. Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. *International Journal on Recent and Innovation Trends in Computing and Communication ISSN*, 2321-8169.

4. Kumar, v., &velide, l. (2014).a data mining approach for prediction and treatment ofdiabetes disease.

5. Devi, M. R., &Shyla, J. M. (2016).Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus.*International Journal of Applied Engineering Research*, *11*(1), 727-730.

6. Bagdi, R., &Patil, P. (2012).Diagnosis of diabetes using OLAP and data mining integration.*International Journal of Computer Science & Communication Networks*, *2*(3).

7. Senthilkumar, D., &Paulraj, S. (2013). Diabetes Disease Diagnosis Using Multivariate Adaptive Regression Splines.*AGE*, *768*, 52.

8. Thirumal, P. C., and N. Nagarajan. "Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus - A Case Study", ARPN Journal of Engineering and Applied Sciences Vol. 10, No.1, January 2015.

9. Pandey, M. D. Prediction system to support medical information system using data mining approach.*International Journal of Engineering Research and Applications (IJERA) ISSN*, 2248-9622.

10. Evirgen, H., &Çerkezi, M. (2014).Prediction and Diagnosis of Diabetic Retinopathy using Data Mining Technique. *Turkish Online Journal of Science & Technology*, *4*(3).