Prediction of Diabetes Disease Using Classification Data Mining Techniques

R. Manimaran[1] and Dr. M.Vanitha[2]

[1]Research Scholar, PG and Research Department of Computer Science,
J.J.College of Arts and Science (Autonomous), Pudukkottai,Tamilnadu,India.
Email: rmanimaran_jjc@rediffmail.com

[2]Assistant professor, Department of Computer Applications,
Alagappa University,Karaikudi, India.
Email: mvanitharavi@gmail.com

Abstract- Data mining uses important techniques and classification is one of them. Classification is also an accepted technique in analyzing huge databases. It is used for solutions in several fields of Science, Business and Industry. Classification is implemented by finding rules that classify data. There are several classification and Statistical methods. This paper demonstrates the use of Decision Tree algorithm for classification and predict Diabetes in patients. Further, it attempts to develop a Decision Tree Algorithm for diabetes prediction in patients

Keywords: Data mining, Classification, Decision Tree, Prediction, Training set.

1. INTRODUCTION

Diabetes is a dangerous disease and greatly affects human life periods. Diabetes is transmittable from mothers to unborn children. Its effects include premature Death, Strokes, Heart diseases, blindness, and kidney failure. This objective of this paper is to propose an easy technique for Prediction of Diabetic patients. A diabetic person is identified with low levels of blood sugars and insulin in their body. Diabetes can be classified into three types namely Type 1, Type 2 and Gestational [1].

Type 1 – Diabetes was previously known as "**Insulin-Dependent Diabetes Mellitus**". Diabetes can be formed at any age, but needs to be diagnosed below 20 years. This type of diabetes is formed insulin producing cells or beta cells in the pancreas get destroyed

Type 2 - Diabetes was previously known as **non–insulin-dependent diabetes** as it was diagnosed in patients above 20 years of age.

Gestational – Diabetes, can occur in pregnant women, when pancreas does not create required amount of insulin in the body. These three types of diabetes need treatment and when detected and treated early, complications associated with them can be avoided [2].

Data mining is a powerful tool for data analysis in its process of discovering interesting pattern from huge amounts of data like massive datasets or data warehouses. This work uses data mining techniques for the prediction of the above listed diseases in patients databases.

2. OVERVIEW OF DATA MINING

Data mining is used for data analysis of large data sets like World Wide Web, external sources, databases and warehouses. Interesting patterns when discovered through data mining are easy to understand, and potentially useful [3]. Sorting techniques are also used for discovering hidden patterns. The aims of data mining are knowledge Discovery of patters, extraction, reducing complexity, and saving processing time [4]. Data mining ends in Knowledge Discovery in Database (KDD) [5] . KDD is a continuous process and consist of thefollowing steps [6]

1. **Selection:** used to select the data.
2. **Preprocessing:** lessen repetitive data.
3. **Transformation:** transform data for processing.
4. **Data mining:** identify desired results.
5. **Interpretation: giving** meaningful information

2.1 DATA MINING TECHNIQUES

The aim of any predictive model can be achieved using a number of data mining techniques [7].

2.1.1 Classification: Classification is based on categories and this technique depends on a supervised learning. It is based on training set and values.

2.1.2 Regression: Regression is used to map a real data item into a valued prediction [8]. Regressions are used for predicting results.

2.1.3 Clustering: Clustering is grouping similar data objects. Finds similarities in data.

2.1.4 Association Rule: Association is an important data mining tool and finds most frequent item sets. It discovers patterns in the database based on relationships between transactions [8].

2.1.5 Prediction: It is a technique which discovers relationships between an independent and dependent variables [6]

2.1.6 Time Series Analysis: Time series analysis is a statistical technique used to model and explain dependency of data points based on time [9].

2.1.7 Summarization: Summarization are abstractions of data. They provide an overview of data and a set of relevant tasks.

3. RELATED WORKS

Mohd Fauzi bin Othman and Thomas Moh Shan Yau [10], examined the performances of different classification and clustering methods with a large set of data. Kawsar Ahmed, et.al [11], in their work proposed a system to detect Lung cancer in patients. Kaur[12], reviewed six clustering techniques of data mining namely K-means clustering, Hierarchical clustering, DBSCAN clustering, OPTICS, and STING. Bharat Chaudharil, Manan Parikh [13], analyzed the performances of clustering algorithms are K-Means, Hierarchical clustering and Density based clustering algorithm using weka tools. Manish Verma, et. al [14], analyzed six clustering techniques k-Means Clustering, Hierarchical Clustering, DBSCAN clustering, Density Based Clustering, Optics and EM Algorithm, in their study. Shraddha K.Popat, et.al [15], surveyed different clustering techniques in their work. P. Thangaraju and B.Deepa [16], surveyed on preclusion and discovery of skin melanoma in patients by using clustering techniques.Khaled Hammouda, Prof. Fakhreddine Karray [17], reviewed four off-line clustering algorithms in their study. Pradeep Rai and Shubha Singh [18], in their study surveyed and provided a comprehensive review of different clustering techniques of data mining. Amandeep Kaur Mann and Navneet Kaur [19], also reviewed different clustering techniques in data mining. Dr.N. Rajalingam, K. Ranjini [20compared implementations of Hierarchical clustering algorithms, agglomerative and divisive clustering for various attributes.

4. DATA SET DESCRIPTION

MV dataset, collected from various districts is used to predict diabetes Disease using Data Mining Classification Techniques. It contains 1024 complete instances with 26 Parameters. The data was gathered from answers to Questionnaires given during the research work. The main objective of the questionnaire was to converse on a set of parameters for diagnosis of diabetes in patients. Table 1 lists the characteristics of MV data sets, while Table 2. Describes the attributes used in the study.

Table 1. Characteristics of MV data sets.

Dataset	No. of Attributes	No. of Instances
MV	26	1024

Table 2. Attribute Description.

Attribute	Values
Age	1.20-30,2.31-40,3.41-50,4.Above 50
Sex	1. Male, 2. Female
Marital status	1.Married,2.Single,3.Widowed, 4.Divorced
Nature of works	1.Physical,2.Mental,3.Both
Occupation	1.Govt.2.Private,3.Business,4.Student, 5.Other
Working Environment	1. City, 2. Village
Working hour	1.6hrs, 2.8hrs, 3.>8hrs
Job Satisfaction	1. Yes, 2. No
Smoking	1. Yes, 2. No
Alcohol	1. Yes, 2. No
Tobacco	1. Yes, 2. No
Food	1. Veg., 2.NV, 3. Both
Water	1. Normal, 2. R.O, 3. Well, 4. All
Affected Age	1.20-30,2.31-40,3.41-50,4.Above 50
Diabetes	1. Yes, 2. No
Heart disease	1. Yes, 2. No
Kidney failure	1. Yes, 2. No
Giddiness	1. Yes, 2. No
Stroke	1. Yes, 2. No
Exercise	1. Yes, 2. No
Treatment	1. Yes, 2. No
High BP	1. Yes, 2. No
Blood Urea	1.20-40mg, 2.Less than, 3. Greater than
Blood Sugar After Food	1.80-140mg, 2.Less than, 3. Greater than
Blood Sugar Before Food	1.70-110mg, 2.Less than, 3. Greater than
Weight	1.30-50,2.51-70,3.71-90,4.Above 90

5. METHODOLOGY

5.1.1 Classification

Classification is categorical and data is classified based on the training set and values, for prediction of Diabetes.

- **Multilayer Perceptron (MLP):** This is a commonly used neural network classification algorithm. Architectures that use MLP during simulations on PIDD dataset consist of three layer feed-forward neural network namely input, hidden, and output layer.
- **BayesNet** - Bayesian networks are used with the presumptions that attributes are nominal with no missing values.
- **JRip** - JRip to RIPPER is a basic and important algorithm in data mining classification.
- **C4.5 –** Is a Decision tree classifier to classify a new item and needs to create a decision tree using the training data.
- **Fuzzy Lattice Reasoning (FLR)** – This Classifier is used descriptive and decision-making.

6. PERFORMENCE ANALYSIS OF ALGORITHMS

MV dataset with 1024 instances found 272 persons with Diabetes. The data set was split into Training and Test Data sets as listed in Table 3. Table 4 lists the results of the classification algorithms performances on multiple factors.

Table 3. Number of instances.

Data set	No. of training data	No. of test data	Total
MV	712	312	1024

Table 4. Shows the Performance metrics**.**

Algorithm	Time Taken	Accuracy %	Positive Recall	Error Rate
MLP	530	75	0.3725	0.2733
BayesNet	515	85	0.5	0.36
JRip	531	86	0.119	0.36
FLP	530	75	0.3725	0.2733
C4.5	550	86	0.38	0.28

It is evident from the above table that BayesNet, MLP, and FLP has lowest computational on the MV dataset. A Confusion matrix was obtained to calculate the specificity, sensitivity, and accuracy, since confusion matrix is a representation of authenticity in results. The results of accuracy is depicted in Figure 1.

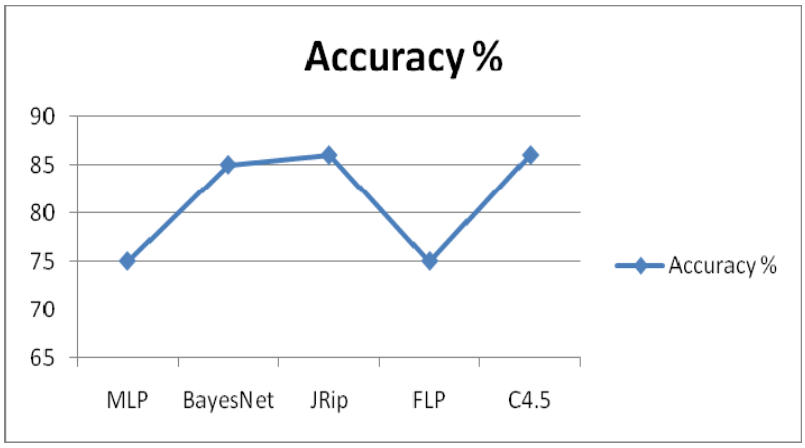

Fig. 1 - Predicted Accuracy

It is evident from Figure 1 that BayesNet, C4.5 and JRip result in 85% accuracy, while MLP and FLP are comparatively less accurate. Figure 2 depicts the positive recall values. On positive recall BayesNet performsthe highest, while MLP, FLP and C4.5 recall around the same range of values. JRip's performance on positive recall is the lowest.

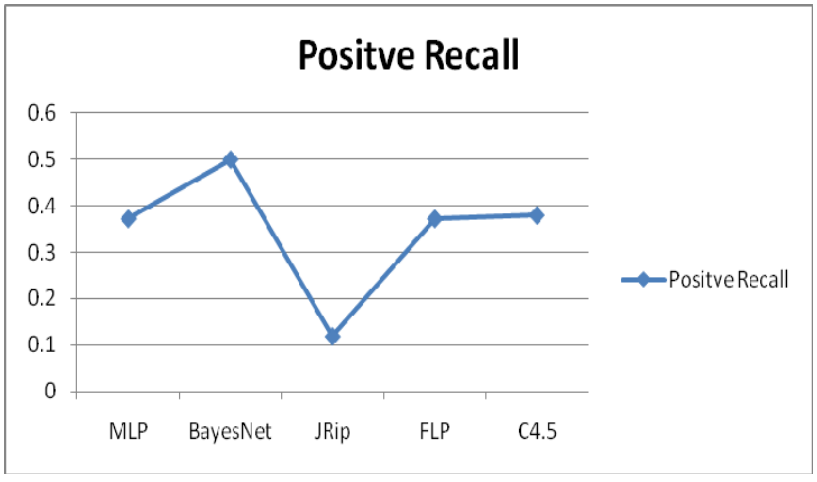

Fig. 2- Positive Recall

Figure 3 depicts the error rates of the compared algorithms, where all of them revolve around 0.25 to 0.35 percentage.

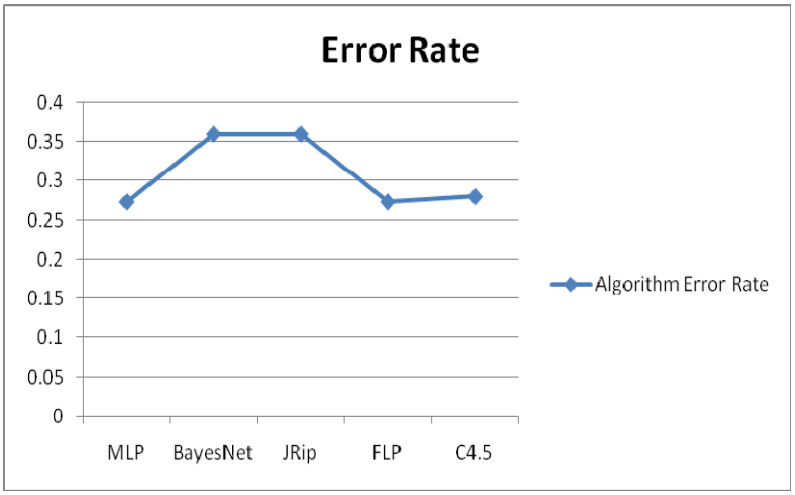

Fig. 3 - Error Rate

7. CONCLUSION

Five data mining classification techniques were compared on multiple factors on the same set of attributes in the MV database. They results were obtained for MLP, BayesNet, JRip, FLP, and C4.5 classification techniques. The techniques were compared on time, accuracy, recall and error rate. It was found that BayesNet, MLP, FLP had lower computation time. On accuracy, C4.5 and JRip had accuracy above 85%. Thus this work concludes that C4.5 and JRip are the most suited algorithms for prediction using classification on datasets with diseased kidney patients. Medical predictions need higher accuracy levels and accuracy above 85% is good for early detection/prediction of diabetes, thus helping doctors take preventive and early actions on treatment

8. REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2001.
[2] Global Diabetes Community, http://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html
[3] Yongjian Fu " data mining: task, techniques and application"
[4] Er. Rimmy Chuchra "Use of Data Mining Techniques for the Evaluation of Student Performance:A Case Study" International Journal of Computer Science and Management Research Vol 1 Issue 3 October 2012.
[5] J. Han and M. Kamber. "Data Mining, Concepts and Techniques", Morgan Kaufmann, 2000.
[6] Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar" Data Mining Techniques & Distinct Applications: A Literature Review" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November- 2012
[7] Dr. M.H.Dunham, "Data Mining, Introductory and Advanced Topics", Prentice Hall, 2002.
[8] Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology And Secured Transactions, 2012 International Conference For, pp. 471-472. IEEE, 2012.
[9] Time Series Analysis and Forecasting with Weka, http://wiki.pentaho.com/display/datamining
[10] Mohd Fauzi bin Othman, Thomas Moh Shan Yau, "Comparison of Different Classification Techniques using WEKA for Breast Cancer", F.Ibrahim, N.A. Abu Osman, J.Usman and N.A. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp.520-523, 2007
[11] Kawsar Ahmed, Abdullah-Al-Emran, Tasnuba Jesmin, Roushney Fatima Mukti,, Md Zamilur Rahman and Farzana Ahmed, "Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific Journal of Cancer Prevention, Volume 14, 2013.
[12] Aastha Joshi, Rajneet Kaur, "A Review: Comparative Study of Various Clustering Techniques in Data Mining", International journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.
[13] Bharat Chaudharil, Manan Parikh, "A Comparative Study of clustering algorithms using weka tools", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 1, Issue 2, October 2012 ISSN 2319-4847.
[14] Manish Verma, et. al, " A Comparative Study of Various Clustering Algorithms in Data Mining" International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.
[15] Shraddha K. Popat, et. al, "Review and Comparative Study of Clustering Techniques" International Journal of Computer Science and Information Technologies, Volume. 5 (1), 805-812, 2014.
[16] P.Thangaraju and B.Deepa, "A Case study on Perclusion and Discovery of Skin Melanoma Risk using Clustering Techniques", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 3, Issue 7, July 2014.
[17] Khaled Hammouda, Prof. Fakhreddine Karray, "A Comparative Study of Data Clustering Techniques", University of Waterloo,Ontario, Canada.
[18] Pradeep Rai and Shubha Singh , "A Survey of Clustering Techniques", International Journal of Computer Applications,Volume 7-No. 12, October 2010.
[19] Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on Clustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR)Volume2, Issue4, April 2013.
[20] Dr. N. Rajalingam, K. Ranjini, "Hierarchical Clustering Algorithm– A Comparative Study", International Journal of Computer Applications (0975-8887), Volume 19- No 3, April 2011.