



Analysis of Gestational Diabetes using Data Mining Algorithms

J. Sharmila¹, G. Seenuvasan²

Research Scholar, P.G. and Research Dept of Computer Science, Swami Vivekananda Arts and Science College,
Villupuram, Tamil Nadu, India¹

Assistant Professor, P.G. and Research Dept of Computer Science, Swami Vivekananda Arts and Science College,
Villupuram, Tamil Nadu, India²

Abstract: Data Mining is used to discover new patterns from large volume of data. In data mining the classification is very important task. Gestational diabetes is a condition characterized by high blood sugar levels that is first predictable during pregnancy period of a woman. Diabetes is a disease in which levels of blood glucose, also called blood sugar, are above normal. People with diabetes have problems converting food to energy. Generally, after a meal, the body breaks food down into glucose, which the blood carries to cells throughout the body. Cells use insulin, a hormone made in the pancreas, to help them convert blood glucose into energy. Through the second and third trimester, a mother's diabetes can lead to over-nutrition and excess growth of the baby. Having a large baby rises risks during labour and delivery. In addition, when foetal over-nutrition occurs and hyper insulinemia results, the baby's blood sugar can drop very low after birth, since it won't be receiving the high blood sugar from the mother. However, with proper treatment, a gestational diabetic mother can deliver a healthy baby despite having diabetes. In this paper, many classification algorithms like J48, simple CART and Naïve bayes algorithm are used to diagnose the diabetes in pregnant women and they are compared for their accuracy levels.

Keywords: Data Mining, Classification, Gestational Diabetes, Blood glucose, J48, CART, Naïve bayes.

I. INTRODUCTION

Pregnant women who have never had diabetes before but who have high blood sugar levels during pregnancy are said to have gestational diabetes. Based on recently announced diagnostic criteria for gestational diabetes, it is estimated that gestational diabetes affects 18% of pregnancies. Gestational diabetes is a condition characterized by high blood sugar (glucose) levels that is first recognized during pregnancy [4]. During pregnancy, increased levels of certain hormones made in the placenta help shift nutrients from the mother to the developing foetus. Other hormones are produced by the placenta to help prevent the mother from developing low blood sugar [11].

They work by resisting the actions of insulin. Over the course of the pregnancy, these hormones lead to progressive impaired glucose intolerance (higher blood sugar levels). To try to decrease blood sugar levels, the body makes more insulin to get glucose into cells to be used for energy [13]. Usually the mother's pancreas is able to produce more insulin (about three times the normal amount) to overcome the effect of the pregnancy hormones on blood sugar levels. If, however, the pancreas cannot produce enough insulin to overcome the effect of the increased hormones during pregnancy, blood sugar levels will rise, resulting in gestational diabetes. Diabetes can affect the developing foetus throughout the pregnancy. In early pregnancy, a mother's diabetes can result in birth defects and an increased rate of miscarriage. Many of the birth defects that occur affect major organs such as the brain and heart [5].

During the second and third trimester, a mother's diabetes can lead to over-nutrition and excess growth of the baby. Having a large baby increases risks during labour and delivery. For example, large babies often require caesarean deliveries and if he or she is delivered vaginally, they are at increased risk for trauma to their shoulder. In addition, when fetal over-nutrition occurs and hyper insulinemia results, the baby's blood sugar can drop very low after birth, since it won't be receiving the high blood sugar from the mother [12].

However, with proper treatment, a gestational diabetic mother can deliver a healthy baby despite having diabetes. In this paper, many classification algorithms like J48, simple CART, Naïve bayes algorithm, Multilayer perception, SMO and simple logistic algorithm are used to diagnose the diabetes in pregnant women and they are compared for their accuracy levels.



International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2017

II. RELATED WORKS

Pedro J.Caraballo et. al [1], proposed the method of distributional association rule mining to identify sets of risk factors and the corresponding patient subpopulations that are significantly increased risk of progressing to diabetes. And to discover sets of risk factor, here uses bottom up summarization algorithm which produces most suitable summary that describes subpopulations at high risk of diabetes. The Subpopulation identified by this summary covered most high risk of patients, had low overlap and were at very high risk. This method is used for when the patient having high risk.

Dr. Zuber khan et. al [2], worked on the concept of Diabetes Mellitus using kNearest Neighbor algorithm which is most Important technique of Artificial Intelligence. The accuracy rate is showing that how many outputs of the data of the test dataset are same as the output of the data of different features of the trained dataset. The error rate is sighting that how many outputs of the data of the test dataset are not same as the output of the data of different features of the training dataset.

The result they showed that as the value of k increases, accuracy rate and error rate will increase. K-Nearest Neighbour algorithm is one of the most important techniques of AI which is used widely for diagnostic purposes. Through KNN more Accurate results can be obtain. This method is very effective for the training data set which is very large.

Mukesh kumari et. al [3], worked on the concept of data mining is to extract knowledge from large volume of data stored in dataset and produce clear and understandable description of patterns. The techniques are attributes selection, data normalization and then classifier is applied on data set to construct Bayesian model.

Bayesian network classifier was proposed for the prediction of person whether diabetic or not. By using Bayesian classifier patient is undergoing categorized in classes of Pre-diabetic, Non-diabetic, Diabetic according to the attributes selected. The techniques they applied as pre-processing attribute identification and selection, data normalization. Then classifier is applied to the modified data set to build the Bayesian model. The Bayesian network has an advantage of it encodes all variables and missing data entries can be handled successfully.

Dr.Pramanand Perumal et. al [6], proposed an idea about diabetes mellitus its diagnosis using data mining with minimum number of attributes applied to classification algorithms. They worked on Apriori and FP-growth techniques. In FP-growth the novel data structure frequent pattern tree is being implemented for storing compressed crucial information about frequent pattern.

It is observed that both of the techniques generate the same number of frequent sets as an importance same number of rules for the same known dataset under the same constraints. With the help of data Apriori and FP-growth algorithms, the computation cost decreases and also the classification performance increases.

Satyanarayana Gandi et. al [7], worked on the initial training data set to the optimal process to extract the optimal data set, on that optimal dataset they applied classification with Bayesian classifier. Bayesian classifier methods is uses getting training data set and convert it into classified data. Initially they extract the optimal feature set from existing training data and calculates the positive and negative probability, until the new data set if formed with same size and forwards the current generated dataset for classification .It classifies the testing dataset with new feature.

Ramkrishnan Shrikant et al [8], proposed a systematic framework of building a risk prediction model for type-2 diabetes disease. The GBRE algorithm identifies the best set of indicators that can predict risk level of diabetes and then multiple classifiers are trained and their accuracy are measured.

Sanchita paul et. al [9], proposed an approach for feature selection, classification and used Genetic Algorithm, Multilayer Perceptron Neural network on diabetes data set. With features selection methodology using Genetic algorithm they improve the accuracy but achieved slightly less ROC. With feature Selection methodology genetic algorithm improved accuracy but achieved less ROC by applying GA, MLP NN methodology classification ROC is also improved.

Alan J. Garber,MD et. al [10], proposed case study includes evaluation for Complications and staging, Lifestyle Modifications, Algorithm for adding/Intensifying insulin, CVD Risk factor algorithm, Profiles of anti-diabetic Medications. Principles of the ACE Algorithm for the treatment of type 2 diabetes.



International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2017

III. PROPOSED WORK

Classification is an important task in data mining. There are several classification algorithms available to mine the data and these algorithms are used in several disciplines. The classification techniques also play a vital role in analysing the data and to predict information [14]. Some of the classification algorithms used to predict gestational diabetes is J48 algorithm, simple CART, and Naïve Bayes algorithm. They are being used depending upon the problem specificity; these techniques have their own advantages and drawbacks [15]. The experiment carried out using WEKA environment which is open source data mining tools. These experiments have used thyroid data set which is collected from UCI repository data source. This data set applied in different data mining techniques for classification of pregnancy causing diabetes diseases. This experiment is applied to different partitions of data set in different data mining techniques like J48, Navie Bayes and Simple CART for classification of diabetes data. First here applied this data set into various individuals' data mining techniques and calculated the accuracy of models. Second to ensemble the two models for classification of diabetes data. And also ensemble J48, Navie Bayes and Simple CART for classification of this data which gives higher accuracy compared to each individual's models. Partitions of data plays very important role for accuracy of model. From one partition to other partitions accuracy is varying and the proposed ensemble Simple CART gives high classification testing accuracy 99.41% in case of 90-10% as training- testing partitions. Table 1 depicts the comparison or classification algorithms.

Table 1: Comparison of classification algorithms

S.No	Error Rate	Naive Bayes	J48	Simple CART
1	Kappa statistic	0.2916	0	0
2	Mean absolute error	0.3456	0.4544	0.4544
3	Root mean squared error	0.4663	0.4766	0.4766
4	Relative absolute error	76.0311	99.971	99.971
5	Root relative squared error	97.8216	100	100
6	Time (s)	0.01	0.02	21.5

Simple CART is a well-known decision tree induction learning technique that has been successfully and extensively applied for medical data. Simple CART is a software extension of the basic ID3 algorithm designed by Quinlan. It can be observed that CART algorithm gives a classification rate of ~ 99 % without feature relevance. However, when feature relevance technique is applied, the classification rate decreases to lesser than 95%. Error detection or Error control are techniques that enable reliable delivery of digital data over unreliable communication channels. Many communication channels are subject to channel noise, and thus errors may be introduced during transmission from the source to a receiver.

Error detection techniques allow detecting such errors, while error correction enables reconstruction of the original data in many cases. This proves that the, J48 and Simple CART is a simple classifier technique to make a decision tree. Efficient result has been taken from diabetes dataset using weka tool in the experiment. Naive Bayes classifier also showing good results. The experiments results shown in the study are about classification accuracy and cost analysis. J48 and Simple CART give more classification accuracy for class mortgage in bank dataset having two values Yes and No. Though here in this example, cost analysis valued same for both the classifier, with age attribute, here can prove that J48 and Simple CART is cost efficient than the Naïve Bayes classifier. Table 2 depicts the classification using different classification algorithms.

Table 2: Classification using Navie bayes, J48 and CART

Filtering Technique	No. of Attributes		Error Rate in Classification	
	Before filtering	After filtering	Before filtering	After filtering
Fisher	8	6	0.0938	0.1224
Runs	8	2	0.0938	0.1875
Relief	8	3	0.0938	0.1576
Step Disc	8	5	0.0938	0.1237



IV. ARCHITECTURAL DESIGN

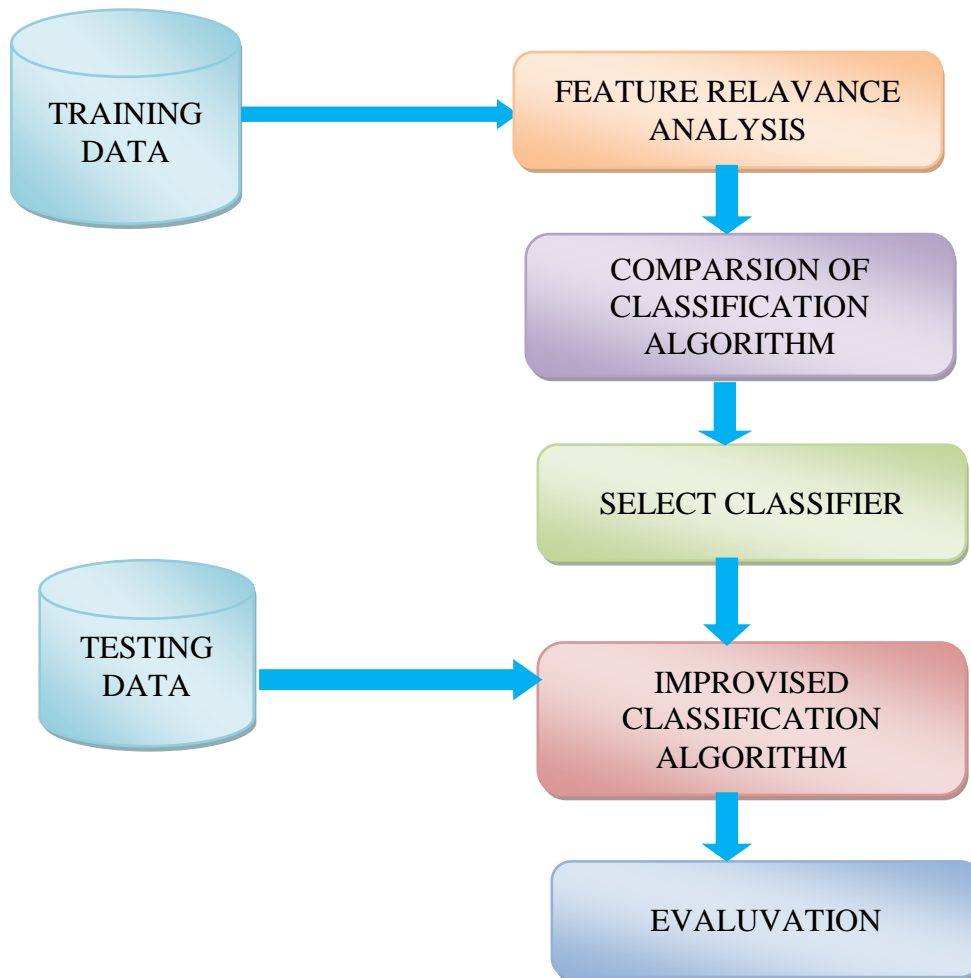


Figure 1: Architecture of Proposed System

IV. CONCLUSION

The Amount of Research work has been done for Prediction of diabetes using data mining technique. The bottom up summarization technique uses when patient has high risk of diabetes. The Naïve Bayesian Classifier, J48 and Simple CART all methods used for prediction of diabetes which gives patient's condition of Normal, Pre-diabetes, and Diabetes. All above methods used to predict diabetes. But if Patient is detected as diabetes firstly there is a need of finding Control and Un-control condition of diabetes. Because if Patient has diabetes in uncontrolled condition, may be the patient has severe effect on Patient's Organ like Heart, Eye, Kidney etc. So there is need of finding early Severity which may be help patient for reducing the Severity on Organ or Halting the Severe Effect on Organ. Though here in this example, cost analysis valued same for both the classifier, with age attribute, here can prove that J48 and Simple CART is cost efficient than the Naïve Bayes classifier. In Future will be use new technologies such as data mining should be explored and utilized to support medical decision, which improves in diagnosing the risk for pregnant diabetes.

REFERENCES

- [1] Gyorgy J.Simon,Pedro J.Caraballo,Terry M. Therneau,Steven S. Cha, M. Regina Castro and Peter W.Li "Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus", IEEE Transactions on Knowledge and Data Engineering ,vol 27,No.1,January 2015.
- [2] Dr.Zuber khan, shaifali singh and Krati Sexena, "Diagnosis of Diabetes Mellitus using K- Nearest Neighbor Algorithm", In proceeding of International Journal of Computer Science Trends and Technology, vol.2 , July-Aug 2014.
- [3] Mukesh kumari and Dr. Rajan Vohra, "Prediction of Diabetes Using Bayesian Network", In proceeding of International Journal of Computer Science and Information Technologies, vol. 5 , 2014.



**International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering**

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2017

- [4] Jianchao Han, Juan C. Rodriguze and Mohsen Beheshti , “Diabetes Data Analysis and Prediction Model Discovery Using Rapid Miner” , In proceeding of Second International Conference on Future Generation Communication and Networking, vol.2,2008 .
- [5] Wang ZuoCheng and XUE Li Xia ,“A Fast Algorithm for Mining Association Rules in Image”, In proceeding of International Conference on Data Engineering, vol.5, 2008 .
- [6] Dr. Pramanand Perumal and Sankaranarayanan, “Diabetic prognosis through Data Mining Methods and Techniques”, In proceeding of International Conference on Intelligent Computing Applications, vol. 2, 2014 .
- [7] Satyanarayana Gandhi and Amarendra Kothalanka,“An Efficient Expert System For Diabetes By Naïve Bayesian Classifier”, In proceeding of International Journal of Engineering Trends and Technology ,vol. 4 ,Issue 10 , Oct 2013.
- [8] Ramkrishnan Shrikant and Rakesh Agrawal,“Fast Algorithms for mining association rule,” in proceeding of IEEE International Conference on Data Engineering,vol.16,2007
- [9] Dilip Kumar Choubey and Sanchita Paul, “GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis”, In proceedings of IJ.Intelligent System and Applications, vol.1,pp.49-59,2016.
- [10] H. S. Kim, A. M. Shin, M. K. Kim, and N. Kim, “Comorbidity study on type 2 diabetes mellitus using data mining”, In proceedings of Korean J. Intern. Med., vol. 27, no. 2, pp. 197– 202, Jun. 2012 .
- [11] Kawita Rawat and Kawita Bhurse, “A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network”, In proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol.4, Nov. 2014.
- [12] G. S Collins, S. Mallett, O. Omar, and L.-M. Yu, “Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting”, In proceedings of BMC Med., 9:103,September 2011 .
- [13] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules”, In Proceedings of 20th VLDB, Santiago, Chile, 1994
- [14] M. A. Hasan, “Summarization in pattern mining” , In proceedings of Encyclopaedia of Data Warehousing and Mining, 2nd ed. Hershey, PA, USA:Information Science Reference, 2008 .
- [15] R.P.Bakshi and S.Agrawal, “Modeling Risk of Prediction of Diabetes - a preventive Measure”, In proceedings of BMC Med.,2012.