Comparison of Various Data Mining Algorithms in the Prediction of Risk for Gestational Diabetes

R. Catherine Stephina Mary
Research Scholar, Department of Computer Science
Bishop Heber College, Trichy, Tamil Nadu, India

B. Satheesh Kumar, M.Sc., M.Phil.,SET.,(Ph.d)
Assistant Professor, Department of Computer Science
Bishop Heber College, Trichy, Tamil Nadu, India

DOI: 10.23956/ijarcse/V7I7/0156

Abstract— Data Mining is a field of computer science which is used to discover new patterns for large data sets. Classification is an important task in data mining. In different areas of medicine, data mining has contributed to improve the results with other methodologies. Gestational diabetes is a condition characterized by high blood sugar (glucose) levels that is first recognized 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. Normally, 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. 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 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 Naive bayes algorithm are used to diagnose the diabetes in pregnant women and they are compared for their accuracy levels.

Keywords— Diabetes, glucose, insulin and diagnosis

I. INTRODUCTION

PREGNANT women who have never had diabetes before but who have high blood sugar (glucose) 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. During pregnancy, increased levels of certain hormones made in the placenta (the organ that connects the baby by the umbilical cord to the uterus) 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.

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. 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 fetus 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.

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. 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, Naive 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.

II. RELATED WORKS

Gyorgy J. Simon, 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. Zubair Khan, Shaifali Sing, et al. [2] worked on the concept of Diabetes Mellitus using k-Nearest 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 sighted 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 Kumar and Dr. Rajan Vohra [3] worked on the concept of data mining to extract knowledge from information stored in dataset and generate 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 weather diabetic or not. By using Bayesian classifier patient is undergoing classified in classes of Pre-diabetic, Non-diabetic, Diabetic according to the attributes selected. The techniques they applied as preprocessing attribute identification and selection, data normalization. And then classifier is applied to the modified data set to construct the Bayesian model. The Bayesian network has a benefit of it encodes all variables, missing data entries can be handled successfully.

Dr. Pramanand Perumal and Sankaranarayanan [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 a 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 Ganti and Amarendra Kothalanka [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 then it classifies the testing dataset with new feature.

Sanchita Paul and Dilip Kumar Choubey [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.

Ramkrishnan Shrikant and Rakesh Agrawal [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.


Rohit Prasad Bakshi and Sonali Agrawal [16] proposed a systematic framework of building a risk of prediction model for type-2 diabetes disease. The GBRE algorithm finds the best set of that can found risk level of diabetes and then multiple classifiers are trained and their accuracy are being measured. The classifier has been selected by voting policy technique. The suggested approach can be applied significantly in prediction modeling of other diseases.

S. Sapna and Dr. A. Tamilarasi[17] proposed a concept of Genetic Algorithm and Fuzzy system on chromosomes. To Obtained the accuracy of chromosome and to evaluate the diabetes in diabetic patient GA is implemented. The connection between fuzzy system and genetic algorithm is bidirectional. Genetic Algorithms are utilized to deal with various optimization problems involves fuzzy system. Using GA optimization of chromosome is obtained and based on the rate of old population diabetes can be retrained in new population to get chromosomal accuracy.

Srideivananai Nagarajan and R.M. Chandrasekaran[18] proposed a method for improvement of diagnosis of gestational diabetes with data mining techniques. Also they Analyse the performance of ID3, Naive Bayes, C4.5, and Random tree i.e. the algorithm for supervised Learning. They used the data set of Pregnant Womens. The results they found that Random tree served to be the best one with higher accuracy and least error rate. Veena

V. Vijayan and Aswathy Ravikumar[19] discussed the main data mining algorithm, K-Means Algorithm, Amalgam KNN algorithm and ANFIS algorithm. They proposed the study of Expectation Maximization algorithm used for sampling to determine and maximize the expectation in successive iteration cycles. K-Nearest Neighbor Algorithm is used for classification of objects and used for prediction of labels based on some closest training examples in the feature space. K-Means algorithm follows partition methods based on some input parameters on the dataset of n objects. They discussed about Amalgam Algorithm combines both the feature of K-Nearest Neighbor and K-Means with some additional processing and the Adaptive Neuro Fuzzy Inference System which combines the Features of Adaptive Neural Network and Fuzzy Inference System. They choose the dataset from PIMA Indian Diabetic Set from University of California.
K. Rajesh and V. Sangeeth [20] proposed that data mining relationship for efficient classification they applied data mining techniques to classify diabetes clinical data and predict the patient being affected with diabetes or not. They presented a system which gave training data on that data feature relevance analysis is done then comparison of classification algorithm. Selecting classifier then improved classification algorithm is applied and then found out the evaluation that compared with training data. They applied C4.5 Algorithm gave classification rate of 91%.

Dr. B. L. Shivkumar and S. Alby [21] presents a survey paper for data mining methods that have been commonly applied to diabetes data analysis and prediction of disease. They had done an analysis of various presentations and studies done by other researches. From the analysis of different research papers it is evident that the occurrence of diabetes is having strong relation with diseases like Wheeze Edema, Oral diseases, Female Pregnant and increase of age. Using data mining techniques the chance of diabetes can be predicted which is helpful for early detection of the disease.

Carlos Fernandez Llatas and Antonio Martinez Millanu [22] proposed the use of Interactive Pattern Recognition techniques for the iterative design of protocols and analyzing the problems of using process mining to infer care flows and how to cope the resulting spaghetti Effect.

### 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. 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.

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. In this experiment to applied different partitions of data set in different data mining techniques like J48, Navie Bayes and SimpleCART 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.

### Comparison of classification algorithms

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

SimpleCART is a well-known decision tree induction learning technique that has been successfully and extensively applied for medical data. SimpleCART 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 SimpleCART 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 SimpleCART 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 SimpleCART is cost efficient than the Naive Bayes classifier.
### IV. ARCHITECTURAL DESIGN

<table>
<thead>
<tr>
<th>Filtering Technique</th>
<th>No. of Attributes</th>
<th>Error Rate in Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before filtering</td>
<td>After filtering</td>
</tr>
<tr>
<td></td>
<td>After filtering</td>
<td>Before filtering</td>
</tr>
<tr>
<td>Fisher</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Runs</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>ReliefF</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Step Disc</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

### V. 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 un-control condition, may have 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.

### VI. FUTURE WORK

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


