International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Evaluation of Risk Factors of Gestational Diabetes Mellitus (GDM) Using Data Mining

Prema N S, Pushpalatha M P

Abstract: ***Diabetes is one of the major chronic diseases in all population which is a major health challenge. The diabetes developed during pregnancy is called Gestational diabetes mellitus (GDM). The identification of GDM at early stages of the pregnancy is very important otherwise it will lead to major health issues both in mother and the baby. We developed a Data mining (DM) model to analyze the risk factors of GDM using different DM techniques. Dataset used for analysis contains the details of the pregnant women collected from the local hospital of Mysuru, India. The clustering and classification techniques used are k-means clustering, J48 Decision Tree, Random-Forest and Naive-Bayes classifier. Classification accuracy is enhanced by using feature subset selection wrapper approach. A balanced dataset is developed by using Synthetic Minority Over-sampling Technique (SMOTE).Using accuracy the performances of classifiers are compared.***

Index Terms: Data mining, Gestational Diabetes Mellitus, SMOTE, K-means.

I. INTRODUCTION

It is estimated that 1 out of every 200 pregnancies is complicated by diabetes mellitus and in every 200 pregnant women 5 will develop GDM. Globally prevalence of diabetes is increasing and India is no exception. Four million women have GDM in India alone so, it is important to identify the GDM at the earliest; else it threatens lives of mother and baby [1].

The factors for increasing prevalence of gestational diabetes in India are;

- The age of the women
- Obesity
- lack of physical activity
- Modern lifestyles, smoking, alcohol consumption etc.

Diagnosing a pregnant woman with Gestational Diabetes Mellitus (GDM) is very important because diabetes mellitus is associated with significant metabolic alterations, increased perinatal mortality and morbidity, maternal morbidity and exaggerated long span illness among the mothers and their off springs. Since there is a scarce of healthcare resources and very low doctor to population ratio – the medical specialist remain overcome with the workload; therefore, it becomes a challenging task for them to provide value care to the patients. To meet the trial of saving lives we provide a combined e-health solution – using Clinical Decision Support System (CDSS) which is based on data mining which analyzes the patient's data and classifies as either normal or high risk. Extracting useful knowledge from large repository of data is data mining. Medical data mining is an application of data mining, where data mining techniques are used for the analysis of medical data. In Medical data mining approaches are applied for the following tasks: diagnosis, treatment, prognosis, monitoring and management. The aim of medical data mining is to help and assist physicians, improve public health and support patients.

The main two tasks of data mining are prediction and description, prediction includes classification, regression and description includes clustering and association analysis applications of both in the field of healthcare can be found in literature. In most of the work referred in the paper,

literature. In most of the work referred in the paper, the data set used is the Pima Indian diabetes data set taken from UCI machine learning repository which contains data about female patients [2]. Many classification algorithms are applied on Pima diabetes data set and their objective is to classify the data into either diabetic or non-diabetic and they have considered only Type-I and Type-II diabetes they have not taken gestational diabetes into consideration.

To mention few, Alexis *et al*., proposed diagnosis of type II diabetes by applying artificial metaplasticity on multilayer perceptron, the data set used is Pima Indians diabetes [3]. B. M. Patil *et al*., have prepared hybrid prediction model proposed for the prediction of Type II diabetes which uses k-means and C4.5 algorithm [4]. Similarly Nahla H Barakat and his team have used support vector machines (SVMs) for the diagnosis of diabetes [5].

In literature we could find many works related to maternal healthcare data, to mention few; M. Jamal Afridi and Muddassar Farooq presented a combination of data mining techniques for effective classification of high risk pregnant women [6]. The model classifies four major risk factors of mortality – hypertension, hemorrhage, septicemia and obstructed labor - in a reliable, autonomous and accurate fashion. Aparna Gorthi *et al*., proposed a machine learning approach for early identification of the risk category of pregnancy based on patterns taken from profiles of known clinical parameter.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Prema N S*, Department of Information Science and Engineering , Vidyavardhaka College of Engineering, Mysuru, India.
Pushpalatha M P, Department of Computer Science and Engineering ,Sri Jayachamarajendra College of Engineering, Mysuru, India.

[7]. Here classification techniques are applied just to identify the severity of risk like low, medium and high. Mario W. L. et al., compared two Bayesian classifiers to classify hypertensive disorders in pregnancy care [8].

The first study on the application of machine learning techniques with EHRs to predict GDM was proposed by Hang Qui *et al*. In their work they developed prediction models capturing the future risk for the electronic medical records of women in West China Second Hospital; the average accuracy obtained is 62% [9].

The main purpose here is to apply data mining techniques in exploring the major risk factors of GDM which can be used for early detection of GDM.

The subsequent contents of the paper are organized as follows: In Section II, the data mining algorithms are described. The results of the algorithm are discussed in section III. The conclusions are given in Section IV.

II. MATERIALS AND METHOD

There is no dataset of pregnant women having gestational diabetes exists; therefore, we have made an attempt to create a new dataset which contains information about diabetes in pregnancy. The data used in this experiment are collected from the hospitals of Mysuru, Karnataka state, India. The medical records are taken after concealing the identity of patients in order to ensure confidentiality. The data set has been developed by keeping obstetrics and gynaecology consultants in a feedback loop. We have collected about 1352 pregnant women details. GDM dataset is developed by removing less relevant and irrelevant features with the help of doctors, then data cleaning and transformation is done, figure-1 show the steps followed in the proposed model

A. Removal of irrelevant Features: The data taken from hospital contains more than 20 attributes. The reduction of features is done manually by taking the help of gynaecologists. As a result, only 10 relevant features are retained that consultants use to detect gestational diabetes.

B. Data cleaning and transformation: This step is very important in developing a complete data set which can be used further in any machine learning techniques. It was very challenging task to extracts useful information from a manually entered medical record, as the entry was made manually there was lots ambiguity for in entering the values of some of the attributes for example for the attribute number of time pregnant (Gravida) some have entered in numbers and some have specified as multigravida etc. We applied data cleansing and transformation cycle to get a meaning data set. Once we have the meaningful attributes, the datasets is finalized on the basis of short listed 10 risk factors. Most of the attributes the values will be of type nominal and values will be either yes or no.

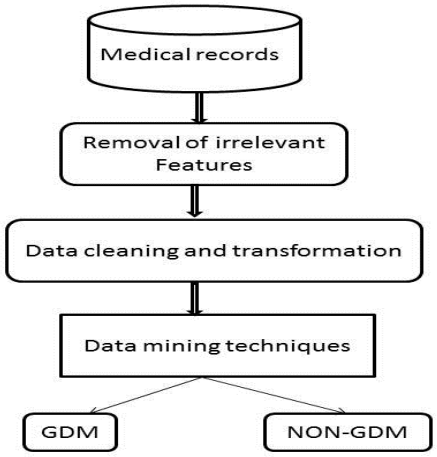

Figure-1 Steps in the model

GDM data set has totally 10 risk factors, they are Age

1. Past history of fetal loss(abortion or IUD)
2. Congenital anomalies in previous pregnancy
3. Macrosomia in previous pregnancy
4. Family history of Diabetes mellitus
5. Obesity
6. Past history of Pre-elampsia
7. Number of times pregnant
8. Unexplained neonatal loss
9. Previous history of GDM

Age is the major risk factor of GDM, older the age more chances of developing gestational diabetes. The figure 2 shows association between age and GDM. It is found that women with age more than 25 are having more chances of development of GDM.Many attribute are about previous pregnancy, they might be the cause for GDM, the selected attributes for this study are history of fetal loss by abortion or IUD, Congenital anomalies, GDM, unexplained neonatal loss, Macrosomia and Pre-elampsia.

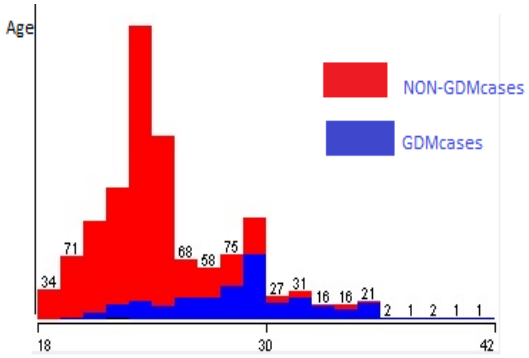

Figure-2 Distribution Of Age In GDM And Non-GDM Cases

Macrosomia is the situation where the birth-weight is over 4,000 g and is not depending on gestational age. Macrosomia affects about 3-15% pregnancies [10].Family history of diabetes can also be the reason for developing GDM, in the used dataset there about 138 cases where family history of diabetes is positive in that more than 90% cases we can see the development of GDM.

Obesity is the common risk factor for many deceases, for GDM also it can be considered as the major risk factors. The dataset contains about 95 obese women details where 95% of them have gestational diabetes.

C. Clustering: It is an unsupervised technique of grouping similar objects into disjoint groups using distance measures. K-means is a partition clustering techniques where it aims in partitioning the observation into k cluster with nearest centroid. The distance measure used find out the distance between the clusters is Euclidian distance.

D. Classification: The aim of study here is to find the application of the classification techniques for better classification. In our study, we use Decision Tree (J48), Random-Forest, Naive-Bayes classifiers.

Decision Tree (J48): It is a decision tree algorithm implemented for C4.5 from a set training data very similar to ID3, using Information entropy.

Random-Forest: It is an ensemble form of decision tree constructed using training data from a random subset. Random-Forest is of a collection of tree-structured classifiers.

For the GDM data set Random-Forest classifier is applied, 10 trees will be generated, each constructed while considering 4 random features with Out of bag error 0.1806.

Naive-Bayes: This is classifier based on Bayes theorem, which uses maximum likelihood method.

E. Class imbalance problem: Data contains very less ratio of positive class instances to negative class instances; so it is an imbalanced dataset. In medical datasets it can be observed that high risk patients instances will the minority class, so the cost of miss predicting the minority class will be more .Therefore, there is a need of a good sampling technique for medical datasets [11].

The technique used for oversampling is SMOTE, in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with duplicated real data entries. SMOTE blindly generates minority class samples without considering majority class samples and may thus cause over generalization [12].

III. RESULTS AND DISCUSSIONS

A. Clustering: The k means clustering technique is applied on the data set by taking k value as two, the two clusters formed are;

Cluster-1 with 90% of the data set.
Cluster-2 with 10% of the data set.

The cluster-2 contains the instances with GDM cases and the attributes values yes is considered *for Family history of diabetes, number of times pregnant, obesity and the average value for attribute Age, considered is 28.8*, where as in cluster-1 all the attributes values are no and they all belong to non-GDM cases.

B. Feature subset selection: The final 10 risk factors are given as input for the subset selection. Wrapper model approach is used for the feature subset selection which uses themethod of classification itself to measure the importance of features set; hence feature selected depends on the classifier model used. Evaluates attribute sets by using a learning scheme. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes. The selection of attributes is done using best fit approach which is a hybrid of DFS and BFS search

C. *Classifiers:* A Bayesian classifier Naive-Bayes and tree based classifier namely J48, Random-Forest are used for the feature selection purpose. Standard 10 fold cross validation techniques applied in all the experiments. This process allows the classifiers divides the data into 9 and 1 fold in which, 9 folds of data is used for training and 1 unused fold for testing.

Table I: Classification accuracy using wrapper feature selection approach for imbalanced dataset

Classification method	J48	Random -Forest	Naive-Bayes
Number of features selected	6	9	9
Accuracy for with selected features in %	93.8	95	90.7
Accuracy with all features in %	93.5	95	89

Table II: Classification accuracy using wrapper feature selection approach for balanced dataset

Classification method	J48	Random -Forest	Naive-Bayes
Number of features selected	5	7	6
Accuracy for with selected features in %	86.7	86	85.7
Accuracy with all features in %	86	87	84

The prediction of GDM is done by using same classifiers by using all the risk factors and selected risk factors Table I shows the accuracy of the classifiers for the selected features and the accuracy of classifiers with all features. The features selected by the classifiers are shown in Table II.

IV. CONCLUSION

In summary, we have applied different DM technique for the identification of risk factors for predicting diabetes in pregnancy using 10 important attributes. If we consider the complete data set there is 24% GDM cases but K-means algorithm can group only 10% by considering limited attributes. Here the studies conclude that the classifiers achieve higher accuracy of 86% for the imbalanced data set and 93% for balanced data. The classification accuracy has been increased by 1 to 2% after selecting best attributes by applying wrapper approach of feature subset selection and except for the attributes Age and Obesity no attributes are

selected in common by the applied algorithms as major risk factors for GDM for the taken data set. Hence all 10 risk factors will be helpful in the prediction of GDM further the application can be developed which will help the pregnant women in primary diagnosis of gestational diabetes mellitus.Further we plan to consider more and different types of risk factors for gestational diabetes prediction and develop a model for large data set

REFERENCES

1. Martin Silink, "IDF Diabetes Atlas" 4th ed, of International Diabetes Federation (2009), Ala Alwan, Brussels, Belgium.
2. UCI Machine Learning Repository: Pima Indians Diabetes; (2007),https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
3. M. C. Alexis, Data mining for the diagnosis of type 2 diabetes, World Automation Congress, Mexico, 2012 (Vallarta, Mexico 2012), pp. 1-6
4. B. M. Patil, R. C. Joshi and Durga Toshniwal, Hybrid prediction model for Type-2 diabetic patients, Expert Syst. Appl. 37(12) (2010) 8102-8108.
5. N. Barakat, A. P. Bradley and M. N. H. Barakat, Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus, IEEE Trans. Inf. Technol. Biomed.14(4) 1114-1120
6. M. Jamal Afridi, Muddassar Farooq, OG-Miner: An Intelligent Health Tool for Achieving Millennium Development Goals (MDGs) in m-Health Environments, 44th Hawaii International Conference on System Sciences, USA, 2011 (Kauai, HI. USA, 2011), pp. 1-10.
7. Gorthi, C. Firtion, J. Vepa, Automated risk assessment tool for pregnancy care, International Conference of the IEEE Engineering in Medicine and Biology Society, USA 2009 (Minneapolis, MN, USA, 2009), pp. 6222-6225.
8. M. W. L. Moreira, J. J. P. C. Rodrigues, A. M. B. Oliveira, K. Saleem, A. V. Neto, An inference mechanism using Bayes-based classifiers in pregnancy care, IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), Germany 2016 (Munich, Germany 2016), pp. 1-6.
9. H. Y. Qiu, L. Y. Wang, Q. Yao, S. N. Wu, C. Yin, Bo-Fu, Zhu Xiao-Juan, Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy, Sci. Rep.-UK 7(2017) 16417.
10. Mohammadbeigi, F. Farhadifar, N. Soufizadeh, N. Mohammadsalehi, M. Rezaiee, M. Aghaei, Fetal Macrosomia, Risk Factors, Maternal, and Perinatal Outcome. Ann. Med. Health Sci Res. 3(2013) 546-550.
11. R. Laza, R. Pavon, M. Reboiro-Jata, F. Fdez-Riverola, Evaluating the effect of unbalanced data in biomedical document classification. J. Integr. Bioinform. 8(2016) 105-117.
12. S. J. Yen and Y. S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions, Expert. Syst. App. 36(2009) 5718-5727.

AUTHORS PROFILE

Prema N S
B.E M.Tech
Lif time member of ISTE and IETE,
Has published about 6 articles
Her area of research is Machine learning

Dr. M P Pushpalatha
BE, M.Tech, Ph. D.
Has published more than 20 articles
Her area of research is Machine learning and health informatics.