Analysis of Association between Caesarean Delivery and Gestational Diabetes Mellitus Using Machine Learning

Nisana Siddegowda Prema[1,*], Mullur Puttabuddi Pushpalatha[2]

[1]Department of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, India

[2]Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, India

Received 16 September 2019; received in revised form 23 December 2019; accepted 09 January 2020

DOI: https://doi.org/10.46604/peti.2020.4740

Abstract

The study aims to analyze the association between gestational diabetes mellitus (GDM) and other risk factors of cesarean delivery using machine learning (ML). The dataset used for the analysis is from the pregnancy risk assessment survey (PRAMS), considered in two scenarios, i.e., all the data is taken, and all the data of the women who developed GDM. Further, the data is developed in two groups Data-I and Data-II by considering multiparous and primiparous women details, respectively. The correlation analysis and major classification algorithms are applied to the data. It is founded that the top risk factors for the first time cesarean delivery are the age, height, weight, race of the women, presence of hypertension and gestational diabetes mellitus. The major risk factor for repeated cesarean delivery is the previous cesarean delivery. The presence of GDM is also one of the risk factors for cesarean delivery.

Keywords: C-section, cesarean delivery, GDM, machine learning

1. Introduction

A cesarean (C-section) delivery is a surgical procedure in which a fetus delivered through an incision in the mother's abdomen and uterus. Over the past 30 years, there is an increase in C-section delivery. The CDC reported that the highest ever national cesarean birth rate was 29.1% in November, 2005 [1]. The factors associated with deciding the type of deliveries are social-demographic, medical, obstetric history, antenatal history, etc. It is very important to know the presence of any of the above-mentioned risk factors at earliest which will help the doctors and the patients to make the right decision at the right time.

1.1. Complications associated with C-section delivery

The potential complications to mothers are infection and bleeding which can lead to anemia. Further complications are pregnancies, pain, medicine, anesthesia negative reactions, etc. The rate of maternal mortality is higher in C-section deliveries when compared to vaginal deliveries. Complications for the baby are low birth weight and respiratory problems. There is a possibility of injury for the baby during incision.

1.2. Machine learning in healthcare

Machine learning (ML) is a technique of training a machine to recognize patterns using data and an algorithm. The prediction accuracy of the machine increases with the data and the complexity of the rules fed into the machine. Healthcare is one of the leading applications of ML which has massive and enormous data. ML assists healthcare professionals in analyzing the data and helps them in the right decision making [2].

* Corresponding author. E-mail address: premans@vvce.ac.in

The purpose of this work is to apply ML and statistical techniques for the analysis of the association between cesarean delivery with GDM and other risk factors, as well as to apply the predictions model for the prediction of C-section delivery for the first time delivery and repeated deliveries.

2. Literature Survey

Sunantha Sodsee [3] proposed a cephalopelvic disproportion (CPD) based on the nearest neighbors (NN) algorithm for the prediction of cesarean delivery. In the proposed NN model, two determined threshold distances were used to identify the nearest and farthest neighbors. The highest performance was achieved by the proposed model [3]. Obstetric and pregnancy factors were used in predicting the type of delivery using data mining models by Sonia Pereira et al. [4], and applied different data mining algorithms in different scenarios; they have achieved the highest accuracy of 84%. Farhad Soleimanjan et al. [5] used decision tree C4.5 for the prediction of cesarean delivery, the attributes considered are age, blood pressure, and heart problems. The obtained accuracy by the classifier is 86%.

Decision tree and artificial neural network were used to classify the births into normal and caesarian by Ayesha Sana et al. [6], the classification accuracy obtained was 80% and 82%, respectively. The authors also used association rules for the identification of caesarian birth patterns. In the study, the authors showed that high blood pressure, lack of education, and pulse rate are associated with caesarian birth. Alisha Kamat et al. [7] proposed a prediction model using Naive Bayes and ID3 classifiers to determine the type of delivery based on ultrasonography, urine and blood reports of pregnant women. ML techniques like a decision tree and Naïve Bayes were applied for the prediction of pregnancy-related risk factors [8] and to predict normal or abnormal stages of pregnancy [9]. A classification model was proposed, and it allows an estimation of the interval for the value of the Apgar score depending on mother and newborn data [10].

Abbas et al. [11] proposed a decision support system using ML techniques to assist the physicians in adopting the correct decisions. They had used neural networks, kNN, Naïve Bayes, and SVM classifiers for the analysis of risk factors of C-section delivery. Md Rafiul Hassan et al. [12] proposed a feature selection algorithm coupled with automated classification using ML techniques to analyze and predict IVF pregnancy in greater accuracy. They had used five attributes to asses the prediction ability of IVF pregnancy and five different ML models, namely MLP, SVM, C4.5, CART, and random forest. Prema Pushpalatha [13] used SVM and logistic regression for the diagnosis of preterm birth in the pregnant woman having either diabetes mellitus or GDM; the highest accuracy obtained is 86%. Prema Pushpalatha [13] also conducted a review on applications of data mining techniques in preterm birth prediction. Furthermore, the most commonly used data mining technique which was founded is classification in that the usual techniques are support vector machine (SVM) and logistic regression. The most commonly considered risk factors are socio-demographic, behavioral (lifestyle), and pregnancy history.

Further, the application of deep neural networks for the prediction of diabetes was proposed, for five-fold cross-validation, and achieved good results [15]. A prediction of the model using support vector machine and the nearest neighbor was proposed for the prediction of breast cancer, they have obtained the highest accuracy of 99.68% for support vector machine in the training phase [16].

3. Materials and Methods

3.1. Dataset

The data used for this work is the pregnancy risk assessment survey (PRAMS) collected from the centers for control and prevention (CDC). The PRAMS data will be helpful to identify the group of women and infants with high-risk health issues. The data set contains about 41000 instances with more than 350 attributes related to maternal and child healthcare. The data set used for this work is reduced to 13550 instances after removing missing values instances and the outliers. Outliers are detected

using the interquartile range. The PRAMS dataset has about 350 attributes, in that few are redundant and many are irrelevant to the present study, hence only the required attributes are considered. About 20 attributes (as risk factors) are considered, which are related to the following factors:

- Maternal Age
- Height and weight of the mother
- Obesity
- Diabetes Mellitus(DM & GDM)
- Hypertension
- Lifestyle
- Obstetric history
- Educational level
- Maternal race
- Urban or rural

Most of the values are all categorical, few attributes are of type ordinal. The analysis of the data is done by considering complete data and the details of the women who developed GDM, respectively. Further, the data is developed in Data-I and Data-II groups by considering multiparous and primiparous women details, respectively. In the data set containing the details of women having GDM about 9% and 6% cases are C-section deliveries in Data-I and Data-II respectively as shown in Fig. 1.

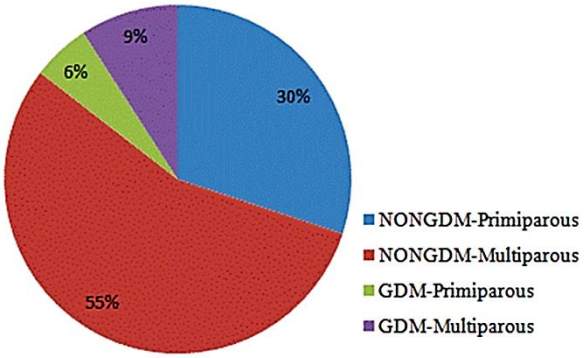

Fig. 1 C-section Delivery distribution in GDM and NON-GDM data

3.2. Association between the features

Feature association refers to the statistical relationship(s) between the feature variables. Nominal variables are measured at the nominal level and have no inherent ranking. There are many statistical measures available to measure the dependency between the variables, for our data the following measures are used:

(1) *The correlation coefficient*: The strength of the relationship between two variables is measured using the correlation coefficient. The correlation may be negative or positive is decided by the correlation coefficient values, -1 and 1, respectively; the value 0 indicates there is no relationship.

(2) *Cramer's V*: By using Cramer's V to calculate the strength of association, if the chi-square test is significant, the V value is between 1 and 0 which tells strong and little association between the variables respectively.

(3) *Feature selection*: The following features selection techniques are applied to the data to figure out the most important features.

- Correlation-based feature-subset selection: This technique is applied for the selection of the best subset of features. In this method subset of attributes is evaluated by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred.
- CfsSubsetEval: Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.
- GainRatioAttributeEval: Evaluates the worth of an attribute by measuring the gain ratio concerning the class. Gaines(Class, Attribute)=(H(Class)-H(Class|Attribute))/H(Attribute).
- InfoGainAttributeEval: Evaluates the worth of an attribute by measuring the information gain concerning the class. InfoGain(Class,Attribute)=H(Class)-H(Class|Attribute).
- OneRAttributeEval: Evaluates the worth of an attribute by using the OneR classifier.
- SymmetricalUncertAttributeEval: Evaluates the worth of an attribute by measuring the symmetrical uncertainty concerning the class.
- SymmU(Class, Attribute)=2*(H(Class)-H(Class|Attribute))/H(Class)+H(Attribute)

3.3. Classification

It is a supervised technique of grouping objects into different labeled classes. The classifiers used for this work are random forest, Naive Bayes, logistic regression and the nearest neighbor classifier(kNN).

(1) *NaiveBayes*: This is a classifier based on Bayes theorem with independence assumptions between the features, which uses the maximum likelihood method.

(2) *Random forest*: A random forest is an ensemble approach that can also be thought of as a form of the nearest neighbor predictor. Each decision tree is constructed by using a random subset of the training data. Ensembles are a divide-and-conquer approach used to improve performance.

(3) *Logistic regression*: Logistics is the regression analysis when the dependent variable is binary. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

(4) *kNN*: It is a non-parametric, lazy learning algorithm, which can be used for both classification and regression. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

4. Result and Discussion

4.1. The correlation coefficient

The correlations between all the individual features are calculated for Data-I and Data II as shown in Fig. 2 and Fig. 3, respectively. It can be observed from the matrix that if a mother is diabetic then there is a strong chance of getting, GDM in the current pregnancy. Features like hypertension, age, weight, and GDM are associated with C-section delivery. For repeated C-section, a previous C-section feature has a strong correlation.

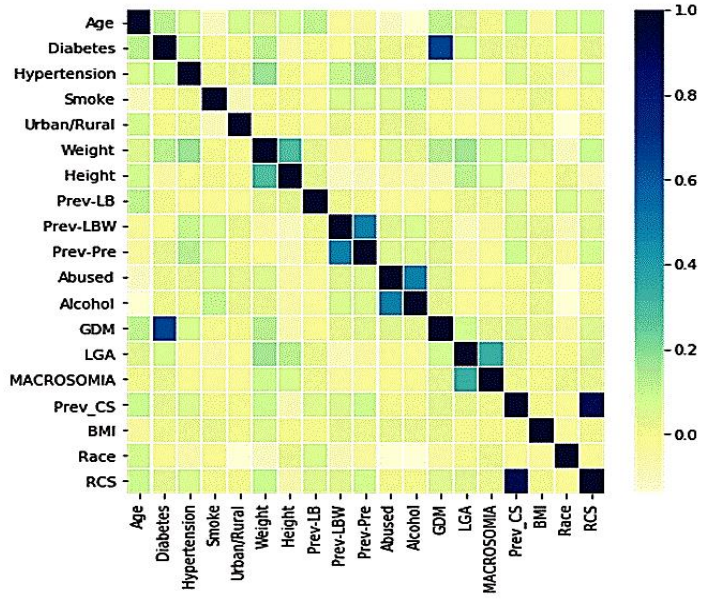

Fig. 2 Correlation matrix of Data-I

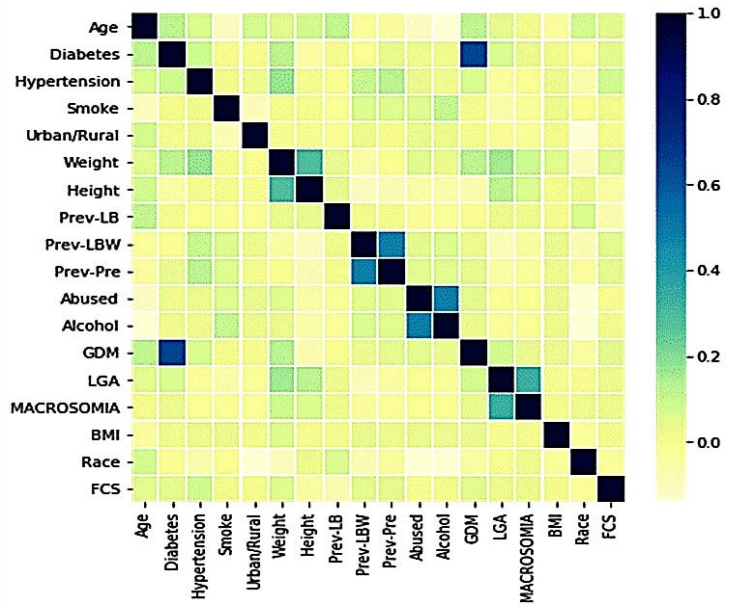

Fig. 3 Correlation matrix of Data-II

Cramer's V test is done on the Data-I and Data-II separately, and the values are tabulated in Table 1. From the V values, it can be observed that there is a strong association between previous C-section delivery and repeated C-section delivery in Data-I. There is a moderate association between C-section delivery and the age, weight, height of the women in both the data sets. The maternal chronic condition like diabetes and hypertension are also associated with C-section delivery.

Table 1 Top features after Cramer's V test

Data set used	Feature	Cramer' V
Data-I (Repeated C-section)	• Previous C-section • GDM • Hypertension • Mothers Weight • Mothers Height	0.9038 0.0469 0.0602 0.1695 0.0900
Data-II (First C-section)	• Mothers Age • GDM • Mothers Weight • Mothers Height • Hypertension	0.0581 0.0449 0.1454 0.0720 0.0799

4.2. Feature selection

The top-ranked attributes by the above-mentioned features selection techniques are shown in Table. 2, It can be observed in Data-II that, the most ranked feature is hypertension, body mass index (BMI), the height of the women, age as well as weight. And, in Data-I, the attributes previous C-section is the top-ranked attribute by all the techniques. Along with previous C-section, the other top-ranked attributes for Data-I is height, weight, BMI, and age of the women.

Table 2 The Ranked features using various features selection techniques

Sl..no	Method	Selected features: For Data-II	Selected features: For Data-I
1	CfsSubsetEval: Bestfit	• Hypertension • Mothers height • Previous live birth	• Previous C-section • Mothers BMI
2	CorrelationAttributeEval	• Hypertension • Mother weight • GDM	• Previous C-section • Mothers weight • Mothers Age
3	GainRatioAttributeEval	• Hypertension • GDM • Diabetes	• Previous C-section • Mothers weight • Hypertension
4	InfoGainAttributeEval	• Hypertension • Mothers BMI • Mothers Age	• Previous C-section • Mothers BMI • Mothers Age
5	OneRAttributeEval	• Mothers Race • Maternal Education • Mothers height	• Previous C-section • Mothers Race • Smoking
6	SymmetricalUncertAttributeEval	• Hypertension • Mothers BMI • Mothers Age	• Previous C-section • Mothers BMI • Mothers weight
7	ReliefFAttributeEval	• Mothers Race • Mothers BMI • Maternal Education	• Previous C-section • Mothers BMI • Maternal Education

4.3. Classification

Table 3 Accuracy obtained for the classifiers

Data set used	Classifier	Accuracy in %		
		For All features	For Top -7 features	For GDM data
Data-I (Repeated C-section)	KNN(k=15)	96	96	95
	Naive Bayes	96.5	96.5	96.9
	Logistic Regression	96.5	96.5	96.8
	Random-Forest	95	96.5	95
Data-II (First C-Section)	KNN(k=15)	88.9	88.8	84.8
	Naive Bayes	88.6	88.9	84.6
	Logistic Regression	88.9	88.9	84.8
	Random-Forest	88.9	88.9	84.9

The above-mentioned classifiers were applied for Data-I and Data-II in both the scenarios by considering all the attributes, and the accuracy obtained by the algorithms are shown in Table 3. For the prediction, different classifier are used firstly; random-forest adopted about ten trees and maximum instances for splitting is taken as five. For the k-nearest neighbor classifier, the value of k chosen is 15, and the measure of the distance used is Euclidean. The different k values have experimented with KNN. If the value k is 15, the highest accuracy is achieved. The classifiers were evaluated using 10 fold cross-validation which the original sample is randomly partitioned into 10 equal-sized subsamples. Of the 10 subsamples, a single subsample is retained

as the validation data for testing the model, and the remaining subsamples were used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. The accuracy of the classifiers is more for the Data-II compared to Data-I in both the scenarios. Logistic regression and Naïve Bayes classifiers are giving the highest accuracy for both the data sets. There is a decrease in the accuracy of the classifiers in the first C-section in the dataset containing GDM women's details when compared with the accuracy of the data. The same classifiers are applied for the top-selected features by using the above-mentioned feature selection methods but there is not much improvement in the accuracy of the classifiers. The ROC analysis curves of both data are shown in Fig. 4 and Fig. 5.

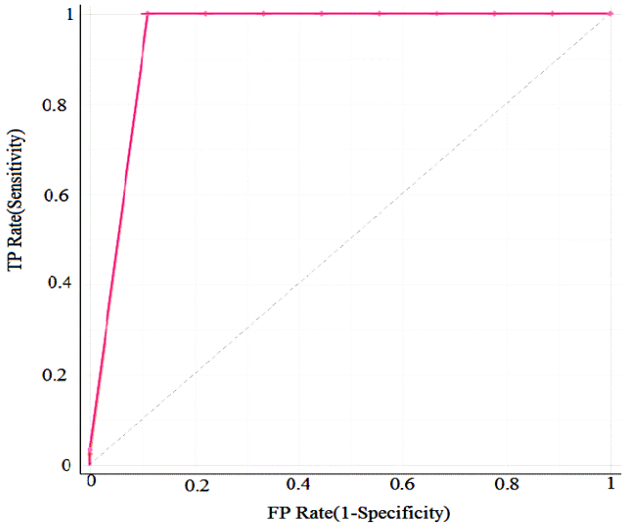

Fig. 4 ROC for Data-I

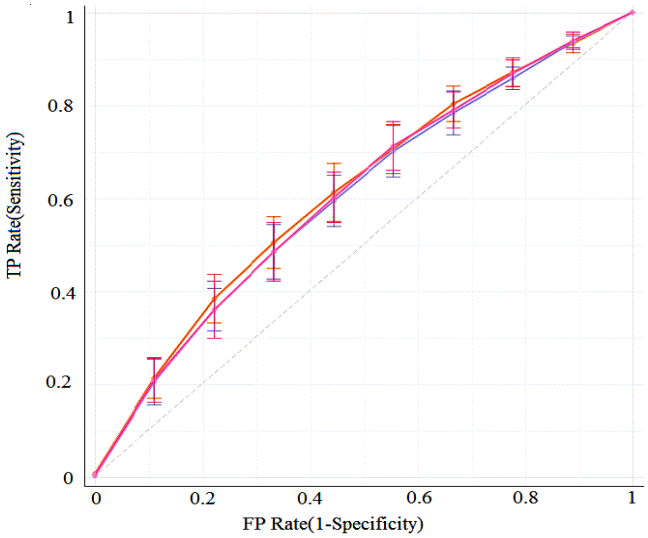

Fig. 5 ROC for Data-II

5. Conclusions

A machine learning-based decision support system has been presented, and the top risk factors for Cesarean delivery have been analyzed in both the first times and repeated pregnancy cases. The major risk factors for the first time Cesarean delivery are the age, height, weight, race of the mother, and the presence of hypertension. For repeated Cesarean, the delivery of previous Cesarean delivery is the major risk factor. There is an association between GDM and Cesarean delivery in the data set considered for the study. The classification model has achieved the highest accuracy of 96% for multiparous and 89% for primiparous mothers in predicting C-section delivery.

In the future, the analysis will be done with paternal risk factors, including ethnicity, financial status, and stress-related factors. The data set is imbalanced, and the ratio of negative classes is more compared to the positive instances. Therefore, there is a lot of scopes to improve the classification accuracy handling imbalanced dataset problems.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] "Reasons For A Cesarean Birth," https://americanpregnancy.org/labor-and-birth/reasons-for-a-cesarean/, October 13, 2019

[2] P. Chowriappa, S. Dua, and Y. Todorov, Introduction to machine learning in healthcare informatics, Machine Learning in Healthcare Informatics, Springer, 2014, pp. 1-23.

[3] S. Sodsee, "Predicting caesarean section by applying nearest neighbor analysis," Procedia Computer Science, vol. 31, pp. 5-14, December 2014.

[4] S. Pereira, F. Portela, M. F. Santos, J. Machado, and A. Abelha, "Predicting type of delivery by identification of obstetric risk factors through data mining," Procedia Computer Science, vol. 64, pp. 601-609, December 2015.

[5] F. Soleimanian, P. Mohammadi, and P. Hakimi, "Application of decision tree algorithm for data mining in healthcare operations: A case study," International Journal of Computer Applications, vol. 52, no. 6, pp. 21-26, August 2012.
[6] A. Sana, S. Razzaq, and J. Ferzund, "Automated diagnosis and cause analysis of cesarean section using machine learning techniques," International Journal of Machine Learning and Computing, vol. 2, no. 5, p. 677-680, October 2012.
[7] A. Kamat, V. Oswal, and M. Datar, "Implementation of classification algorithms to predict mode of delivery," International Journal of Computer Science and Information Technologies, vol. 6, no.5, pp. 4531-4534, 2015.
[8] B. Lakshmi, T. Indumathi, and N. Ravi, "A comparative study of classification algorithms for predicting gestational risks in pregnant women," International Conference on Computers, Communications, and Systems, November 2015, pp. 42-46.
[9] R. Sawant and N. Gaikwad, "Hybrid prediction method for pregnancy data set," 2015 1st International Conference on Next Generation Computing Technologies, September 2015, pp. 918-920.
[10] R. Robu and Ş. Holban, "The analysis and classification of birth data," Acta Polytechnica Hungarica, vol. 12, no. 4, pp. 77-96, July 2015.
[11] S. A. Abbas, R. Riaz, S. Z. H. Kazmi, S. S. Rizvi, and S. J. Kwon, "Cause analysis of caesarian sections and application of machine learning methods for classification of birth data," IEEE Access, November 2018, pp. 67555-67561.
[12] M. R. Hassan, S. Al-Insaif, M. I. Hossain, and J. Kamruzzaman, "A machine learning approach for prediction of pregnancy outcome following IVF treatment," Neural Computing and Applications, September 2018, pp. 1-15.
[13] N. S. Prema and M. P. Pushpalatha, "Machine learning approach for preterm birth prediction based on maternal chronic conditions," Emerging Research in Electronics, Computer Science and Technology, Springer, 2019.
[14] N. S. Prema and M. P. Pushpalatha, "Prediction of preterm birth using data mining-a survey," IIOAB Journal, vol. 10, no. 2, pp. 13-17, January 2019.
[15] S. I. Ayon and M. Islam, "Diabetes prediction: A deep learning approach," International Journal of Information Engineering and Electronic Business, vol. 2, pp. 21-27, March 2019.
[16] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-nearest neighbors," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), December 2017, pp. 226-229.