

Comparison of Prediction Models for Gestational Diabetes Mellitus: Logistic Model and Random Forest Model

Yanru Chen *

Sdu-Anu Joint Science College, Shandong University, Weihai, Shandong, 264209, China

* Corresponding author: 202100700004@mail.sdu.edu.cn

Abstract. This study aimed to develop and evaluate gestational diabetes mellitus (GDM) prediction models based on the logistic regression approach and Random Forest algorithm. GDM, defined as intolerance for glucose when first diagnosed during pregnancy, influences 5-12% of pregnancies worldwide and is a serious threat to the mothers' and infants' health. Therefore, doctors and mothers must predict and intervene in GDM as early as possible. Data from 1012 women who were pregnant from laboratories in the Kurdistan Region was collected for this study, including age, number of deliveries, weight, height, body mass index (BMI), genetic information, and GDM diagnosis. Among them, 21.4% of pregnant women were diagnosed with GDM. By data preprocessing, division of training and test sets (80% training set, 20% test set), and using the SMOTE-Tomek Links sampling method to deal with data imbalance, it constructed logistic regression (LR) and random forest (RF) models, respectively. A grid search algorithm optimized the random forest model with parameters ($n_estimators=300$, $max_depth=20$, $min_samples_split=10$) and performed five-fold cross-validation. These two models were evaluated on the test set and the outcomes showed that the RF model had a good performance in terms of accuracy (0.875, 95% confidence interval (CI) [0.8261, 0.9239]), precision (0.874, 95% CI [0.7199, 1.0000]), specificity (0.979, 95% CI [0.9536, 1.0000]), and area under curve (AUC) value (0.740, 95% CI [0.6611, 0.8153]), which were superior to the LR model. According to the study, the random forest algorithm shows better performance in GDM prediction.

Keywords: Gestational diabetes, Maternal health, Prediction, Random Forest model.

1. Introduction

Poor glucose tolerance measured for a pregnant woman is the definition of gestational diabetes mellitus (GDM), which is a common condition. [1]. It affects 5-12% of pregnancies worldwide, and it is anticipated that this figure will increase to 1 in 6 by 2045 [2, 3]. Both mothers and newborns are adversely affected by GDM. GDM may lead to adverse pregnancy outcomes in the short run such as pregnancy-induced hypertension (PIH) and preeclampsia, increasing the first caesarean section rate, neonatal hypoglycemia, Jaundice-related, neonatal intensive care unit (NICU) admissions, and respiratory morbidity [4-6]. In the long term, it is associated with the progression of type 2 diabetes (T2DM) which is related to cardiovascular disease (CVD) in the mother, as well as later rates of obesity and diabetes in the newborn [7]. The risk of T2DM is increased 4.62-fold and 8.43-fold in singleton and multiple pregnancies respectively by GDM [8]. The adverse impacts of gestational diabetes mellitus (GDM) can be effectively reduced by implementing early intervention strategies during pregnancy, as demonstrated [9]. Therefore, it is important to be able to predict and intervene in GDM earlier.

Machine learning is a method that relies on data as the basis for input and training to make predictions or classifications by learning patterns and regularities in the data. In prognostic assessment, feature selection, and statistical testing methods, Random Forest, a machine learning algorithm that incorporates multiple factors and can accommodate multiple interactions in data, is commonly employed [10].

Early risk prediction studies of GDM have mainly used traditional logistic regression models, for example, HCJ and his team used logistic regression (LR) to identify relevant factors for early poor glucose tolerance in Korean people through some prenatal factors, which are β -cell function and pre-pregnancy obesity. Teede's Monash GDM risk prediction model identified six risk factors for GDM [11, 12]. In recent years, many researchers have begun to experiment with machine learning

algorithms to solve the problem of disease prediction. For instance, Parkhi et al. compared sparse logistic regression and the tree-based algorithms in machine learning in the postnatal pre-diabetes prediction. Belsti and his team evaluated different machine learning methods with logistic regression in gestational diabetes and concluded that the Catboost classifier (CBC) performs optimally [13, 14]. Zhou et al. attempted to predict gestational diabetes using a stepwise prediction method to be developed and tested [15]. However, there is no definitive conclusion about the superiority of traditional logistic regression models or machine learning methods for GDM prediction. Therefore, it is necessary to study and compare the traditional logistic regression and machine learning methods' model in gestational diabetes mellitus prediction models

This paper aims to develop the GDM prediction model alongside the random forest algorithm and logistic regression method based on 1012 pregnant women and evaluate them with different criteria to recognize the model with better performance. The information is collected from Kurdistan region laboratories.

2. Material and Methodology

2.1. Data Source

Information from 1,012 pregnant women from the Kurdistan Regional Laboratory was selected for analysis. The data set included information on age, number of births, weight (in kilograms), height (in meters), body mass index (BMI), genetics, and the results of gestational diabetes mellitus (GDM) in mothers. Of the mothers in the data set, 21.4% were diagnosed with GDM.

2.2. Data Preprocessing

Random forests are integrated algorithms based on decision trees. In a decision tree, each node is based on a segmentation of a feature, and this segmentation usually relies on a comparison of the magnitude of the feature values. Therefore, random forests are usually less sensitive to the scale of features, which means that different magnitudes of features do not significantly affect the performance of the model. Therefore, the variables were not normalized in this study.

2.2.1. Split the train and test set

This study split the dataset into two parts which is the training set and the test set. The random seed parameter was set to 42 to ensure consistency in splitting and fairness in model comparisons. A training set was made up of 80% of the dataset for model training, and the model's function was tested by dividing the rest of the dataset into a test set. In this, the training set was subjected to SMOTE Tomek comprehensive sampling followed by five-fold cross-validation for parameter optimization. To compute the model function, the test set is subjected to 1000 bootstrap resampling.

2.2.2. SMOTE-Tomek links sampling

In the dataset, there were 2 instances of positive cases and 8 instances of negative cases. The imbalance in the dataset may affect the effectiveness of model training. Hence Synthetic Minority Over-Sampling Technique and Tomek links method (SMOTE-Tomek Links) were used to diminish the impact of imbalance in the number of healthy cases and cases with GDM in the dataset. The SMOTE Tomek sampling method is a data processing method that incorporates SMOTE oversampling and Tomek link under sampling techniques. Firstly, the SMOTE technique is applied to oversample the minority class samples. In this step, the minority class samples' number is increased by synthesizing new minority class samples and generating new sample points according to the feature differences between existing minority class samples, thus making the dataset more balanced in terms of class distribution. Subsequently, under sampling is performed using Tomek linking, a technique for detecting and removing noisy or outlier points, which identifies those samples in nearest-neighbor pairs of samples that belong to different classes and removes the majority class samples in these pairs as noise or outliers. This step helps to clean up the dataset and reduce the

impact of noise on model training. This sampling method combines the advantages of oversampling and under sampling to balance the dataset and reduce noise and outliers, thus improving the classification performance and generalization ability of the model.

2.3. Logistic regression modeling

A generalized linear regression model called logistic regression (LR) is used. In this model, a set of independent variables is used to estimate the probability of an event occurring. It is commonly applied in data mining, automatic disease diagnosis, economic forecasting, and other fields.

LR model uses a Sigmoid function to map the inputs to the interval $[0,1]$ and therefore to have probabilistic significance. The parameter estimation is then realized by maximum likelihood estimation (MLE).

2.4. Random forest modeling

The RF method, a Bagging algorithm in machine learning, utilizes decision trees as estimators. Randomly back-selecting the dataset and incorporating randomly selected features are part of multiple decision trees that are combined. Finally, the majority classification result is selected as the result.

It reduces overfitting and at the same time has a better tolerance to noise which leads to better performance.

2.5. Model evaluation

Finally, it evaluates the models' performance on the 20% test set by 1000 bootstrap resampling iterations, calculates the 95% confidence intervals, and obtains the sensitivity, accuracy, area under the curve (AUC) value, specificity, and precision rate as the metrics for evaluating the models.

3. Result

Through the grid search algorithm, the optimal parameter combinations of random forest are $n_estimators = 300$, $max_depth=20$, $min_samples_split=10$. A 3d heatmap is used to display the accuracy of the RF model with different parameters' combinations. (shown in figure 1).

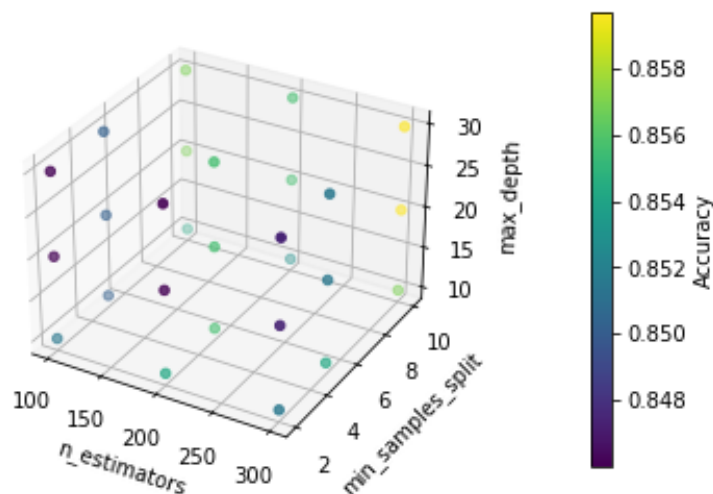


Figure 1. The accuracy of the random forest model with varying combinations of parameters
(Picture credit: Original)

After 1000 bootstrap resampling iterations, the results of the LR model together with RF as Table 1.

Table 1. Evaluation results of the LR and RF models' performance

	LR model	RF model
Accuracy	0.842 (95% CI [0.7880, 0.8913])	0.875 (95% CI [0.8261, 0.9239])
Precision	0.788 (95% CI [0.5788, 0.9524])	0.874 (95% CI [0.7199, 1.0000])
Sensitivity	0.374 (95% CI [0.2222, 0.5238]),	0.500 (95% CI [0.3415, 0.6522])
Specificity	0.973 (95% CI [0.9448, 0.9937])	0.979 (95% CI [0.9536, 1.0000])
AUC value	0.675 (95% CI [0.5996, 0.7521])	0.740 (95% CI [0.6611, 0.8153])

All model evaluation metrics show that the RF model outperforms the LR model.

4. Discussion

In this study, the RF model significantly outperformed the LR model in accuracy, precision, sensitivity, specificity, and AUC value, suggesting that it can better predict GDM. This is consistent with the better predictive performance of machine learning over traditional regression models in the study by Yitayeh Belsti et al [14].

Despite the exemplary performance of the model's metrics in this study, there are still limitations. The sensitivity of both models is not optimal and further research is required to enhance their performance. Furthermore, the dataset used in this study was derived from a single region, which limits the potential influence of factors such as race and environment on GDM. Additionally, other studies have indicated the potential involvement of additional factors in the development of GDM, beyond those considered in this investigation. For instance, Haymanot has proposed that pregnant women residing in rural areas may exhibit an elevated risk of GDM [16]. Furthermore, the ABO group has been proven to play a role in the cause of gestational diabetes mellitus by Shimodaira [17]. Therefore, it is evident that further research is required to enhance the precision of our model.

5. Conclusion

The objective of this study was to develop a predictive model for gestational diabetes mellitus (GDM) using logistic regression and the random forest algorithm within the field of machine learning. The optimal parameters of the random forest were established using a grid search algorithm, with the model trained on 80% of the dataset and the remaining dataset was for model validation. Ultimately, two models with commendable performance were obtained, with the RF model and the LR model exhibiting accuracy rates of 87.5% and 84.2%, respectively. A comparison of the accuracy, precision, sensitivity, specificity, and AUC values of the two models with 95% confidence intervals on this dataset shows that the RF model performs better in the prediction of GDM.

Clinical practice has significant implications from the findings. The developed random forest model can assist medical practitioners in early identifying women who may develop gestational diabetes (GDM), enabling timely intervention and management strategies to reduce the associated maternal and infant complications. The predictive capability of this model highlights the efficacy of machine learning in healthcare, particularly in disease prediction and risk assessment.

In addition to the Random Forest algorithm, there are many other common machine learning algorithms such as Gradient Boosting, Dimensional Reduction, K-Means algorithms, and so on. In future research, it will compare the performance of more different models and try to innovate the algorithms to find the best prediction model.

References

- [1] Sanjeeva R., Joanne E., Cheryce L. H., et al. Development, validation and clinical utility of a risk prediction model for adverse pregnancy outcomes in women with gestational diabetes: The PeRsonal GDM model. *eClinicalMedicine*, 2022, 52: 101637.

- [2] Sanjeeva R., Joanne E., Cheryce L. H., et al. External validation of risk prediction model for gestational diabetes: Individual participant data meta-analysis of randomized trials. *International Journal of Medical Informatics*, 2024, 190: 105533.
- [3] World Health Organization. Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy: a World Health Organization Guideline. *Diabetes Research and Clinical Practice*, 2014, 103 (3): 341 - 363.
- [4] E.P. T. Optimizing cardiometabolic risk factors in pregnancy: a review of risk prediction models targeting gestational diabetes and hypertensive disorders. *Journal of Cardiovascular Development and Disease*, 2022, 9 (2): 55.
- [5] Metzger B.E., Lowe L.P., Dyer A.R., et al. Hyperglycemia and adverse pregnancy outcomes. *New England Journal of Medicine*, 2008, 358 (19): 1991 - 2002.
- [6] Liran H., Howard B., Rania O., et al. Gestational diabetes mellitus is associated with adverse outcomes in twin pregnancies. *American Journal of Obstetrics and Gynecology*, 2019, 220 (1): 102.e1 - 102.e8.
- [7] Lowe W. L., Scholtens D. M., Lowe L.P., et al. Association of Gestational Diabetes with Maternal Disorders of Glucose Metabolism and Childhood Adiposity. *JAMA*, 2018, 320 (10): 1005 – 1016.
- [8] Amir N., Esther M.-S., Mordechai H., et al. Greater risk of type 2 diabetes progression in multifetal gestations with gestational diabetes: the impact of obesity. *American Journal of Obstetrics and Gynecology*, 2024, 231 (2): 259.e1 - 259.e10.
- [9] Arianne S., Wesley H., Helena B., et al. Epidemiology and management of gestational diabetes. *The Lancet*, 2024, 404 (10448): 175 - 192.
- [10] Ping S., Xiangwen W., Shenghai W., et al. bipolar disorder: Construction and analysis of a joint diagnostic model using random forest and feedforward neural networks. *IBRO Neuroscience Reports*, 2024, 17: 145 - 153.
- [11] Hak C. J., Chang-Hoon Y., Ki O. H., et al. Gestational diabetes mellitus in Korea: prevalence and prediction of glucose intolerance at early postpartum. *Diabetes Research and Clinical Practice*, 2003, 61 (2): 117 - 124.
- [12] Teede H. J., Harrison C. L., Teh W. T., et al. Gestational diabetes: development of an early risk prediction tool to facilitate opportunities for prevention. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, 2011, 51 (6): 499 - 504.
- [13] Durga P., Nishanthi P., Yonas G.-W., et al. Prediction of postpartum prediabetes by machine learning methods in women with gestational diabetes mellitus. *iScience*, 2023, 26 (10): 107846.
- [14] Yitayeh B., Lisa M., Lan D., et al. Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. *International Journal of Medical Informatics*, 2023, 179: 105228.
- [15] Fang Z., Xiao R., Fangliang S., et al. A stepwise prediction and interpretation of gestational diabetes mellitus: Foster the practical application of machine learning in clinical decision. *Heliyon*, 2024, 10 (12): e32709.
- [16] Haymanot T., Abebe Y., Asaye M., et al. Gestational diabetes mellitus and its association with ABO blood group type among pregnant women with pregnancy-induced hypertension in Northwest Ethiopia: A comparative study. *Clinical Epidemiology and Global Health*, 2024, 30: 101788.
- [17] Shimodaira M., Yamasaki T., Nakayama T. The association of maternal ABO blood group with gestational diabetes mellitus in Japanese pregnant women. *Diabetes and Metabolic Syndrome*, 2016, 10 (2 Suppl 1): S102 - 5.