Prediction of the Development of Gestational Diabetes Mellitus in Pregnant Women Using Machine Learning Methods

M. R. Basarab, ORCID 0000-0002-3260-674X
K. O. Ivanko, PhD Assoc.Prof., ORCID 0000-0002-3842-2423
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute» ROR 00syn5v21
Kyiv, Ukraine

Vishwesh Kulkarni, Assoc.Prof., ORCID 0000-0002-2285-8652
School of Engineering, University of Warwick ROR 01a77tt86
Warwick, United Kingdom of Great Britain and Northern Ireland

Abstract — The paper is devoted to the application of machine learning methods to the prediction of the development of gestational diabetes mellitus in early pregnancy. Based on two publicly available databases, study assesses influence of such features as body mass index, thickness of triceps skin folds, ultrasound measurements of maternal visceral fat, first measured fasting glucose, and others a predictors of gestational diabetes mellitus. The supervised machine learning methods based on decision trees, support vector machines, logistic regression, k-nearest neighbors classifier, ensemble learning, Naive Bayes classifier, and neural networks were implemented to determine the best classification models for computerized gestational diabetes mellitus disease prediction. The accuracy of the different classifiers was determined and compared. Support vector machine classifier demonstrated the highest accuracy (83.0% of total correctly prognosed cases, 87.9% for healthy class, and 78.1% for gestational diabetes mellitus) in predicting the development of gestational diabetes based on features from Pima Indians Diabetes Database. Extreme gradient boosting classifier performed the best, comparing to other supervised machine learning methods, for Visceral Adipose Tissue Measurements during Pregnancy Database. It showed 87.9% of total correctly prognosed cases, 82.2% for healthy class, and 93.6% for gestational diabetes mellitus).

Key words – gestational diabetes mellitus, diabetic fetopathy, machine learning, prediction.

I. INTRODUCTION

Gestational diabetes mellitus (GDM) is a disease characterized by hyperglycemia (increased blood glucose levels), which was first detected during pregnancy. Most often, women's blood glucose levels return to normal after childbirth, but there is a high risk of developing diabetes in the subsequent pregnancy or in the future life. Disorders of carbohydrate metabolism can develop in any pregnant woman, taking into account those hormonal and metabolic changes that occur sequentially at different stages of pregnancy. But the highest risk of developing gestational diabetes is observed in pregnant women with overweight or obese, presence of diabetes in close relatives and disorders of carbohydrate metabolism detected before pregnancy (e.g. impaired glucose tolerance, increased fasting blood glucose, gestational diabetes in previous pregnancies) [1-5].

To overcome the physiological insulin resistance and maintain normal pregnancy glucose levels in the blood, there is a compensatory increase in insulin secretion by the pancreas of the healthy pregnant women. However, in pregnant women with a hereditary predisposition to diabetes or obesity (body mass index greater than 30 kg/m²), the existing insulin secretion does not always overcome the physiological insulin resistance that develops in the second half of pregnancy. This leads to the increase in blood glucose levels and the development of gestational diabetes [1-5]. With the flow of blood, glucose is immediately transported unhindered through the placenta to the fetus, helping it to produce its own insulin. Insulin of the fetus stimulates the growth of its internal organs on the background of slowing down their functional development. The excess glucose coming from the mother with the help of fetal insulin is deposited in the form of subcutaneous fat. As a result, the mother's chronic hyperglycemia damages the development of the fetus.

These are a number of fetus diseases and risks that can occur as a result of prolonged uncompensated hyperglycemia in pregnant women [4-8]. Among them, heavy weight of the fetus, violation of body proportions, swelling of fetus tissues, chronic hypoxia of the fetus caused by impaired blood flow in the placenta, delay in the formation of lung tissue of fetus, injuries during childbirth and high risk of perinatal mortality.

Among the children experienced diabetic fetopathy, the more common consequences are macrosomia (newborn weight ≥4000 g), violation of adaptation to extraterine life manifested by immaturity of the newborn even in full-term pregnancy and its large size, respiratory disorders, hypoglycemia of the newborn, enlargement of the internal organs, cardiomyopathy, disturbances in the blood coagulation system, and metabolic disorders. In
children born by mothers with undiagnosed, uncompensated gestational diabetes, neurological diseases such as cerebral palsy and epilepsy can occur with birth trauma [4-8].

Pregnant women with gestational diabetes are more commonly suffer from polyhydramnios, urinary tract infections, toxocosis of the second half of pregnancy, premature birth; anomalies of patrimonial activity, injuries during childbirth, and delivery by cesarean section [6-9].

Gestational diabetes mellitus has no clinical manifestations associated with hyperglycemia as dry mouth, premature birth; anomalies of patrimonial activity, injuries during childbirth, and delivery by cesarean section, infections, toxicosis of the second half of pregnancy, diabetes mellitus [6-9].

The data base contains measured features of 768 pregnant women, healthy and diabetic. 268 cases among 768 have diagnosed gestational diabetes, the rest data correspond to patients not suffering from GDM. Since diabetic and nondiabetic classes are imbalanced in the considered database, we generated additional data for diabetic class using the distributions of initial prognostic features in GDM class. After generation of additional data the number of samples in each of classes has become equal.

The following features demonstrated the highest correlation with gestational diabetes development: number of pregnancies, thickness of triceps skin folds, body mass index, and age of the pregnant woman. Thickness of triceps skin folds is a measure of subcutaneous fat and gives information about the fat reserves of the body. In a similar way, body mass index (BMI) plays the role of assessment of body fat and serves as a measure of patients’ risk for diseases that can occur accompanying overweight and obesity.

The second analyzed dataset is Visceral Adipose Tissue Measurements during Pregnancy Database [23] provided by Physionet resource [24]. A distinctive feature of this database is the study of visceral fat measurements during pregnancy, which is reported to be quite important for predicting metabolic risk in pregnant women [25-28]. Traditional methods of fat assessment as waist-to-thigh ratio and high cost abdominal magnetic resonance imaging are not convenient to use during pregnancy. Consequently, the use of ultrasound to assess visceral adipose tissue is a very promising area in pregnancy screening. Recent studies have shown an increased risk of gestational diabetes due to increased visceral fat [25-28].

The “Visceral adipose tissue measurements during pregnancy” database contains data from studies of 133 pregnant women up to 20 weeks of pregnancy and their observation until delivery. Women with pre-existing type 1 or 2 diabetes mellitus were excluded from the study. Evaluation of the thickness of the mother's visceral adipose tissue was carried out during the ultrasound examination with an ultrasound sensor located in the sagittal position (Fig.3). The mean of two measurements during maternal inhalation and exhalation was used to calculate visceral adipose tissue thickness. The first measurement of the mother's weight before the 12th week of pregnancy and her height were used to calculate the body mass index. The data provided in the database were collected as part of a study aimed at estimation whether the value of maternal visceral adipose tissue thickness could be used as one of the early predictors of the development of gestational diabetes mellitus [23, 28].
To build predictive models for the development of gestational diabetes using machine learning algorithms, we used the following prognostic variables from the dataset: age of women in years; ethnicity (0 = white; 1 = not white); mean diastolic blood pressure in mmHg; mean systolic blood pressure in mmHg; maternal visceral adipose tissue measurement in mm; number of pregnancies; first measured fasting glucose; pregestational body mass index in kg/m².

As outcomes of pregnancy, the following target variables were available: gestational age at birth; type of delivery (vaginal birth or Cesarian section); child birth weight in grams (more than 4000 g corresponds to macrosomia); gestational diabetes (healthy women or suffering from GDM).

The initial dataset consists of 133 patients, 18 of which appeared to have diabetes mellitus developed in pregnancy and the rest 115 women had no gestational diabetes. Since the available classes are significantly imbalanced and the total number of records in the Visceral Adipose Tissue Measurements during Pregnancy Database is low, we generated additional data separately for diabetic and non-diabetic patients using distribution of initial prognostic features. The obtained balanced dataset consisted of 200 samples (100 samples for each class) was used to apply machine learning methods for computerized early prediction of diabetes mellitus development in pregnant women. Histograms of the distribution of each prognostic feature for diabetic and nondiabetic patients from “Visceral adipose tissue measurements during pregnancy” database are shown in Fig.4. Correlation matrix heatmap for these data is presented in Fig.5.
We can see that GDM is more common among those women, who had multiple pregnancies. It means that the chance of having a GDM increases with the number of previous pregnancies. The thickness of central Armellini fat is also higher in the first half of pregnancy in patients who are suffering from GDM. That can be explained by the fact that visceral adipose tissue is a hormonally active component of total body fat, biochemical characteristics of which influence on the pathological processes development in the human body. The same conclusions can be drawn when analyzing pregestational BMI, distribution of which shows the increased number of women with BMI over the normal (25<BMI<29.9 corresponds to overweight, 30<BMI<34.9 reflects obesity, and BMI>35 is observed in people with extreme obesity).

First-trimester fasting plasma glucose level parameter showed the highest correlation with GDM outcome. This is consistent with a number of studies investigating this feature for prediction of GDM risk and unfavorable pregnancy outcomes [29-31].

III. MATERIALS AND METHODS

To determine the best classification models for predicting the development of gestational diabetes mellitus in pregnant women, we implemented supervised machine learning based on decision trees, support vector machines (SVM), logistic regression, k-nearest neighbors (KNN) classifiers, ensemble learning, Naive Bayes approach, and neural networks.

Decision Tree Classifier is a non-parametric supervised machine learning method. It can be used to create a model that predicts the value of a target variable by learning simple decision rules retrieved from the data features. The deeper the decision tree, the more complex the decision rules for it and model fitting.

Gradient Boosting Classifier (GBC) is an ensemble machine learning algorithm that uses decision trees. Boosting implies an ensemble approach, which sequentially adds models to the ensemble to correct the performance of prior models. Models are fit using differentiable loss function and gradient descent optimization algorithm. We used an ensemble of 100 decision trees for GBC.

Extreme Gradient Boosting (XGBC) Classifier algorithm is also based on using ensemble of decision trees. The algorithm builds short and simple decision trees, called “weak learners”, iteratively. The first simple tree demonstrates poor performance and the second decision tree is trained to define the weak points of the first tree. The algorithm works sequentially implementing more weak learners, each of which corrects the previous weak tree performance. The process continues while a stopping condition, such as the number of estimators, is not
reached. In this study we used ensemble of 40 decision trees for XGBC.

Logistic regression is a statistical method for predicting binary classes, which are in our case outcomes of pregnancy: type of delivery (vaginal birth or cesarean section) and presence or absence of gestational diabetes. Linear regression for continuous output $y$ and the sigmoid logistic function $p$ for dichotomous classification can be described by the following equations:

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n,$$

$$p = \frac{1}{(1 + e^{-y})}.$$

where $x_1, x_2, \ldots, x_n$ are the prognostic variables; $a_0, a_1, \ldots, a_n$ are the weight coefficients. Sigmoid logistic function $p$ presents a curve that can take any real value and convert it into a value between 0 and 1. If the curve tends to positive infinity, the predicted class is 1, and if the function goes to negative infinity, predicted class is 0. In case the output logistic function $p$ is more than 0.5, the outcome is classified as 1 (presence of gestational diabetes), and if it is less than 0.5, the outcome is classified as 0 (absence of gestational diabetes).

The k-nearest neighbors’ classifier is distance-based classifier implementing a simple supervised machine learning algorithm, which implicitly assumes that the smaller the distance between the points in the feature space, the more similarity between them. In order to make a prediction for a new sample, k-nearest neighbor algorithm finds the closest data from the training set, which are the nearest neighbors of the new data sample. Each new sample is assigned a class label based on the majority of votes of $k$ the closest learning samples from the train set. This classifier has two important parameters: the number of neighbors and the measure of the distance between the data samples.

Support Vector Classifier finds a hyperplane in an N-dimensional feature space, which distinctly classifies the data points. Its goal is to find a plane that has the maximum margin, i.e., the maximum distance between data points of classes– norm and diabetes. We used the radial basis kernel function for mapping the data into a transformed higher dimensional space, where a hyperplane for separating the samples can be found.

We used Ridge Classifier based on ridge regression method, with leave-one-out cross-validation estimation. It converts the label data into [-1, 1] and solves the problem with regression method. In ridge regression, to reduce model complexity and prevent over-fitting, the cost function adds a penalty equivalent to square of the model coefficients. Shrinking the coefficients improves the prediction accuracy of the model.

Gaussian Naive Bayes Classifier is based on Bayes’ Theorem and uses the assumption that each class is normally distributed and the predictors are independent on each other.

In addition, in this study, to predict the development of gestational diabetes mellitus, we used 2 neural networks, conditionally called Artificial Neural Network 1 (ANN1) and Artificial Neural Network 2 (ANN2). Both utilized neural networks are sequential networks with a difference in the number of layers (five in first and three in second), neurons (16:8:1 and 64:64:1 structure respectively) and algorithms to perform optimization (adaptive moment estimation was used for ANN1 and adaptive learning rate method RMSprop was used for ANN2). Mixed input and first layer are common for both networks. The first one has two additional dropout layers with an exclusion rate of 0.2 - one after the first and one after the second dense layer. It is made to prevent the model from overfitting by randomly setting neurons on the hidden layers to zero at each training phase update. Both neural networks use binary cross-entropy as loss metrics and a combination of rectified linear unit and sigmoid activation functions for the layers.

IV. RESULTS

Feature importance values reflect which features have the biggest impact on the prediction that is performed by classification method. The chart in Fig. 6 shows the relative impact of each feature on the prediction determined using Gradient Boosting Classifier for Pima Indians Diabetes Database. This chart helps us to understand which features contribute to increasing the prediction probability. Importance is calculated for each decision tree by the amount that each feature split point improves the performance measure (i.e., the purity, Gini index or other error function) weighted by the number of observations the node is responsible for. The obtained feature importance is averaged for all decision trees within the model. Attention should be paid to the impact of BMI and diabetes pedigree function, which provides diabetes mellitus history in relatives and distinguished patients genetically predisposed to the diabetes mellitus disease.

Results of classification accuracy for models trained on predictors from Pima Indians Diabetes Database are presented in Fig. 7 and Table I. Support vector machine classifier demonstrated the highest obtained accuracy in predicting the development of gestational diabetes mellitus (83.0% of total correctly prognosed cases, 87.9% for healthy class, and 78.1% for gestational diabetes mellitus). Artificial Neural Network 1 performance reached 81.9% of total correctly progessed cases, 85.3% for healthy class, and 78.5 for gestational diabetes mellitus). Training and validation accuracy of Artificial Neural Network 1 for Pima Indians Diabetes Database depending on the number of epoch is presented in Fig. 8.

![Fig.6. Estimation of feature importance from a trained predictive model obtained with Gradient Boosting Classifier for Pima Indians Diabetes Database](image_url)
Results of classification accuracy for models trained with predictors from Visceral Adipose Tissue Measurements during Pregnancy Database are presented in Fig. 9 and Table II. Extreme gradient boosting classifier performed the best, comparing to other supervised machine learning methods, showing 87.9% of total correctly diagnosed cases, 82.2% for healthy class, and 93.6% for gestational diabetes mellitus.

One of the adverse consequences of gestational diabetes is the need for a caesarean section. Using the available data from “Visceral adipose tissue measurements during pregnancy database” we and ridge classifier for machine learning demonstrated 78.7% of correctly predicted delivery type, 82.0% for vaginal birth, and 75.4% for cesarean section). Of course, it is obvious that Cesarean section may also be necessary in other pathologies of pregnancy and labor, and this unfavorable outcome of gestational diabetes mellitus requires additional study.

### Table 1

<table>
<thead>
<tr>
<th>#</th>
<th>Machine learning method</th>
<th>GDM (0 = no; 1 = yes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Extreme Gradient Boosting Classifier (with tune search cross validation)</td>
<td>77.8 (78.6, 77.0)</td>
</tr>
<tr>
<td>2</td>
<td>Logistic Regression Classifier</td>
<td>78.3 (80.2, 76.4)</td>
</tr>
<tr>
<td>3</td>
<td>K Neighbors Classifier</td>
<td>79.3 (84.5, 73.1)</td>
</tr>
<tr>
<td>4</td>
<td>Decision Tree Classifier</td>
<td>72.1 (88.2, 56.0)</td>
</tr>
<tr>
<td>5</td>
<td>Support Vector Classifier</td>
<td>84.0 (87.0, 78.1)</td>
</tr>
<tr>
<td>6</td>
<td>Gaussian Naive Bayes Classifier</td>
<td>80.7 (83.0, 78.4)</td>
</tr>
<tr>
<td>7</td>
<td>Stacking Classifier</td>
<td>82.4</td>
</tr>
<tr>
<td>8</td>
<td>Gradient Boosting Classifier</td>
<td>81.48 (86.14, 77.39)</td>
</tr>
<tr>
<td>9</td>
<td>Ridge Classifier (with cross validation)</td>
<td>76.38 (86.14, 67.83)</td>
</tr>
<tr>
<td>10</td>
<td>Artificial Neural Network 1</td>
<td>81.9 (85.3, 78.5)</td>
</tr>
<tr>
<td>11</td>
<td>Artificial Neural Network 2</td>
<td>75.24 (80.36, 76.0)</td>
</tr>
</tbody>
</table>

Fig. 7. Box-plot distributions of classification accuracy (number of correct classifications) obtained in 5-fold cross-validation procedure for Pima Indians Diabetes Database

Fig. 8. Training and validation accuracy of Artificial Neural Network 1 for Pima Indians Diabetes Database

Fig. 9. Estimation of feature importance from a trained predictive model obtained with Gradient Boosting Classifier for “Visceral adipose tissue measurements during pregnancy database”

Fig. 10. Box-plot distributions of classification accuracy (number of correct classifications) obtained in 5-fold cross-validation procedure for “Visceral adipose tissue measurements during pregnancy database”
The Pathophysiology of Gestational
Screening for gestational diabetes mellitus,
al Logistic Regressions for
Lecture
rd International
Sequential neural network model for diabetes prediction showed a performance of 81.9 % of total correctly prog-
nosed cases, 85.3% for healthy class, and 78.5% for ges-
tational diabetes mellitus. The worst result demonstrated
78.1% for gestational diabetes mellitus (390 cases).

<table>
<thead>
<tr>
<th>#</th>
<th>Machine learning method</th>
<th>Target</th>
<th>GDM (0 = no; 1 = yes)</th>
<th>Delivery type (0 = vaginal birth; 1 = cesarean section)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Extreme Gradient Boosting Classifier (tune search cross validation)</td>
<td>87.9 (82.2, 93.6)</td>
<td>75.7 (76.3, 75.1)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Logistic Regression Classifier</td>
<td>83.0 (84.6, 81.4)</td>
<td>76.7 (67.1, 86.3)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>K Neighbors Classifier</td>
<td>80.7 (79.5, 81.9)</td>
<td>74.7 (72.6, 72.8)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Decision Tree Classifier</td>
<td>83.2 (80.2, 84.35)</td>
<td>71.5 (58.3, 84.7)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Support Vector Classifier</td>
<td>85.0 (87.17, 82.83)</td>
<td>74.0 (71, 77)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Gaussian Naive Bayes Classifier</td>
<td>84.5 (80.5, 86.95)</td>
<td>78.5 (93.3, 63.7)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Stacking Classifier</td>
<td>85.9 (85.7, 86.1)</td>
<td>76.0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Gradient Boosting Classifier</td>
<td>81.0 (78.6, 83.4)</td>
<td>72.5 (74.0, 77.0)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Ridge Classifier (with cross validation)</td>
<td>84.0% (90.3, 77.7)</td>
<td>75.7 (82.0, 75.4)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Artificial Neural Network 1</td>
<td>81.0 (74.1, 87.9)</td>
<td>77.3 (73.0, 81.6)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Artificial Neural Network 2</td>
<td>83.0 (73.0, 93.0)</td>
<td>74.9 (67.0, 81.0)</td>
<td></td>
</tr>
</tbody>
</table>

CONCLUSIONS

Body mass index, visceral adipose tissue thickness, thickness of triceps skin folds, diabetes pedigree function, and first-trimester fasting plasma glucose level are the parameters, which should definitely be taken into account in predicting of gestational diabetes mellitus and its adverse consequences.

Support vector machine classifier demonstrated the highest obtained accuracy in predicting the development of gestational diabetes mellitus based on Pima Indians Diabetes Database (83.0% of total correctly diagnosed cases (830 cases), 87.9% for healthy class (440 cases), and 78.1% for gestational diabetes mellitus (390 cases)). Sequential neural network model for diabetes prediction showed a performance of 81.9 % of totally correctly diagnosed cases, 85.3% for healthy class, and 78.5% for gestational diabetes mellitus. The worst result demonstrated decision tree classifier.

Extreme gradient boosting classifier performed the best for “Visceral adipose tissue measurements during pregnancy database” predictors showing 87.9% of totally corrected diagnosed cases, 82.2% for healthy class, and 93.6% for gestational diabetes mellitus.

REFERENCES

Conference on Biological Information and Biomedical Engineering, Hangzhou, China, 2019, pp. 1-6.


[22] Kaggle repository URL: https://www.kaggle.com/


Copyright (c) 2021 Басараб М. Р., Іванько К. О., Вішвеш Кулкарні
Прогнозування розвитку гестаційного цукрового діабету у вагітних із використанням методів машинного навчання

Басараб М. Р., ORCID 0000-0002-3260-674X
Іванько К. О., к.т.н. доц., ORCID 0000-0002-3842-2423
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського" ROR 00syn5v21
Київ, Україна

Вішвеш Кулкарні, доц., ORCID 0000-0002-2285-8652
Інженерна школа, Університет Ворик ROR 01a77tt86
м. Ворик, Сполучене Королівство Великої Британії та Північної Ірландії

Анотація — Стаття присвячена застосуванню методів машинного навчання для прогнозування розвитку гестаційного цукрового діабету на ранніх термінах вагітності. На основі двох публічнодоступних баз даних оцінюється вплив таких показників, як індекс маси тіла, товщина шкірної складки трицепса, ультразвукове вимірювання вісцерального жиру у матері, перше визначення глюкози у плазмі венозної крові натщецерце та інших параметрів для прогнозування розвитку гестаційного цукрового діабету. Методи машинного навчання з вчителем, засновані на деревах рішень, методі опорних векторів, логістичній регресії, класифікаторі k-найближчих сусідів, ансамблевому навчанні, найвідомішому Байєсівському класифікаторі та нейронних мережах були реалізовані для визначення найкращих моделей класифікації для комп’ютеризованого прогнозування гестаційного діабету. В роботі визначено та порівняно точність різних класифікаторів. Метод опорних векторів продемонстрував найвищу точність класифікації у прогнозуванні розвитку гестаційного діабету на основі навчання з використанням показників з бази даних Pima Indians Diabetes Database (83,0% загальних вірно спрогнозованих випадків, 87,9% для класу здорових жінок та 78,1% для класу гестаційного цукрового діабету). Класифікатор з ансамблевого навчання дерев рішень показав найкращі результати порівняно з іншими методами машинного навчання на основі навчання з використанням показників з бази даних Visceral Adipose Tissue Measurements During Pregnancy - 87,9% загальних вірно прогнозованих випадків, 82,2% для класу здорових жінок та 93,6% для класу гестаційного цукрового діабету).

Ключові слова – гестаційний цукровий діабет, діабетична фетопатія, машинне навчання, прогнозування.