Sensors and Materials, Vol. 33, No. 9 (2021) 3011–3025
MYU Tokyo

S & M 2670

Prognosis Model for Gestational Diabetes Using Machine Learning Techniques

Sumathi Amarnath,[1*] Meganathan Selvamani,[2] and Vijayakumar Varadarajan[3]

[1]SRC, SASTRA Deemed to be University, Kumbakonam, Tamil Nadu 612001, India
[2]SRC, SASTRA Deemed to be University, Kumbakonam, Tamil Nadu 612001, India
[3]School of Computing Science and Engineering, The University of New South Wales, Sydney, Australia

(Received September 27, 2020; accepted May 17, 2021)

Keywords: gestational diabetes mellitus, data mining, classification, prediction

Gestational diabetes mellitus (GDM) is a syndrome that occurs among women during pregnancy and is characterized by lack of insulin hormone secretion. GDM occurs in about 4% of all pregnancies and is diagnosed at later stages of pregnancy. It can occur in women with no known history of diabetes. Since no signs or symptoms occur at the onset of GDM, it is possible to diagnose it only through screening tests. GDM poses some major health risks such as hormonal imbalance, delivery risks, and the development of Type 2 diabetes (T2D) after delivery. The condition can be diagnosed from the blood sugar level. Those diagnosed with GDM are likely to be obese, have a weak constitution, and be undergoing a stressful life or living in a stressful environment, eating unhealthy food, and living an unhealthy lifestyle. Other risk factors to be considered are family history, heredity, and the occurrence of diabetes in the past. Apart from diagnosis, the most crucial stage in managing GDM is its prognosis. If the disease is diagnosed at earlier stages, one can avoid its complications. Advanced technologies such as IoT and wearable sensors can help healthcare professionals in identifying the early signs and symptoms of GDM. In this scenario, data mining techniques are recommended for the prognosis of GDM using existing medical reports and risk factors related to women. A patient's medical history and their family history should be correlated with each other to find the likelihood of GDM occurrence. Classification is a technique in which a training dataset is used to predict the importance of related factors using an inference function. Our aim is to develop a prognosis model for GDM using a classification technique. A GDM prognosis model is developed using a training set of disease parameters along with an individual's risk factors. From the results of our experiments, it is inferred that the proposed model can be used for predicting the likelihood of GDM in its earlier stages.

1. Introduction

Gestational diabetes mellitus (GDM)[(1–3)] is a syndrome that occurs among women during their pregnancy. World Health Organization (WHO) stated that the prevalence of GDM is

*Corresponding author: e-mail: sumathi@src.sastra.edu
https://doi.org/10.18494/SAM.2021.3119

ISSN 0914-4935

increasing every year owing to lifestyle changes and the high number of Type 2 diabetes (T2D) patients. GDM has pre- and post-implications for both the mother and the infant. After birth, the mother may have the possibility of T2D or Type 1 diabetes (T1D). The infant may experience the problem of poor nutrition and be prone to diabetes in the future. WHO[4] revised the treatment regimen for diabetes based on race, country, and the individual. Research is ongoing to prognosticate GDM and diagnose the condition. Innovative biomarkers that can be used to identify the disease with normal tests have recently been introduced.

The emergence of sensor devices in recent years has led to rapid advances in a wide range of applications. In the healthcare domain, smart patient assistance is a notable field that provides intensive care to patients at remote locations. Hospitals have undergone drastic changes in providing 24 × 7 lifeline support to patients across the globe. In this scenario, it is important to promote research on GDM using advanced technology. Various studies that focus on applying data mining and machine learning concepts to the maintenance and analysis of patient records and disease biomarkers have been conducted. Researchers have recently identified new and highly helpful biomarkers that can be used to periodically check patients for symptoms of GDM. Data mining classifiers[5] are widely used in the prediction of GDM. Our current research aims to improve the accuracy of diagnosis by enhancing the quality of data and finding suitable classifiers such as the support vector machine (SVM) and k-nearest neighbors (KNN) for GDM prediction. The classifiers, which are mostly used in diabetic research, are compared in terms of their accuracy rate.

2. Related Studies

Schoenaker *et al.*[6] proposed an important prediction model for GDM based on electronic health records. The model tracks the history of previous pregnancies and compares it with current pregnancy data. It also selects the features of a diabetic dataset based on correlation. This model achieved its maximum accuracy with the use of data mining classifiers. Earlier, Iyer *et al.*[7] developed a framework to predict GDM using multiclassifier techniques. They focused on developing an autonomous decision-making model to diagnose GDM using an ensemble classifier approach with higher accuracy than previous models. Milewski *et al.*[8] proposed a prediction model using principal component analysis (PCA), K-means clustering, and the logistic regression (LR) classifier.[9] Kumar and Umatejaswi[10] proposed a new model to solve the basic diagnosis problem, with which they analyzed and identified the severity of diabetes. They developed guidelines for doctors and hospital management to predict and diagnose diabetes as well as its risk levels. Kavakiotis *et al.*[11] conducted a systematic review of studies on diabetes research that have been conducted with biological tools, machine learning, and data mining. Nagarajan *et al.*[12] and Omiotek *et al.*[13] proposed a new algorithm to improve the diagnosis of GDM using data mining techniques.

After reviewing the literature about data mining techniques, we propose a novel approach to handling decision-making in the prediction of GDM that uses data mining and technological improvements.

3. Methodology

The objective of our study is to predict GDM through data mining and machine learning algorithms. The Pima Indian diabetes dataset, sourced from the UCI repository, is used in this study.[14] First, the data is preprocessed, during which the missing data is handled effectively to improve the accuracy of the classifier. Normalization is used to scale the data of an attribute so that it falls within a small range (0–1). A predictive model is developed with the Random Forest (RF) classifier[15] and cross-validated[16] using part of the dataset. The model is tested for its effectiveness in predicting GDM using patient health data.

Data mining has different stages, among which preprocessing is the first step. The input data should be preprocessed prior to the application of a data mining technique to remove the noise and increase the accuracy and output of the process. During preprocessing, data cleaning and transformation are applied as preliminary steps. To predict GDM for the given dataset with higher accuracy, a set of significant classifiers was selected and compared as performance measures in this study.

3.1 Data cleaning and transformation

Data cleaning and transformation are important steps since the dataset should be refined and developed for application in data mining and machine learning approaches. Real-time datasets mostly have a few missing values, encoded as blanks, NaNs, or other placeholders. These missing values should be handled prior to the actual processing. However, it is challenging to manage these values and use them in the development of strong models. Different approaches should be used to handle such missing values. In the current study, various methods, such as the drop-down of the entire tuple, a mean imputation method that replaces the missing value with the mean[17] of each column, and a grouping-based mean imputation method, were tested to overcome missing data values. Among these methods, the grouping-based mean imputation method is used here since it has excellent performance in replacing the missing values based on grouping, thus increasing the classification accuracy. The age attribute is considered in the current study.

3.2 Group-based mean (GBM)

The mean is a suitable method for handling missing and inconsistent values in a dataset. In this method, the value of an attribute is replaced by the mean of its group. The mean can be used to approximate some attributes. For instance, the blood pressure (BP) values of patients differ according to their age. However, when calculating the mean to replace the missing data in the BP attribute, age must also be considered. Therefore, a GBM method is proposed in the current study, in which the missing values are filled with categorized group-based values. To achieve this, the dataset is grouped according to the age of the patients, then the GBM is applied. The results attained from the GBM increase the quality of the dataset.

3.2.1 Normalization

Normalization[18] is a stage in the data preprocessing technique[19] in machine learning. In this data transformation technique, attribute values are allowed to repeat until they lie within a relevant range with the help of a common scale. The common value of each attribute may differ. Normalization reduces the distortion and increases the quality of data. Thus, the data is normalized so that the values of all the attributes are between 0 and 1. We analyzed the organization of the data with and without normalization in this study.

Min-max normalization is a way to normalize data using feature values and transformations. This method guarantees the same scale for all the features. Min-max normalization is one of the most common ways to normalize data. For every feature, the minimum value of that feature is transformed into 0, the maximum value is transformed into 1, and every other value is transformed into a decimal between 0 and 1 using Eq. (1).

$$V = \frac{V' - MIN(A)}{MAX(A) - MIN(A)}\left(newMAX(A) - newMIN(A)\right) + newMIN(A) \tag{1}$$

Here, A is the attribute, and $MIN(A)$ and $MAX(A)$ are the minimum and maximum absolute values of A, respectively, which define the range of A. V' represents the new data entry and V represents the old data entry. Finally, $newMAX(A)$ and $newMIN(A)$ are the maximum and minimum values of A, respectively.

3.3 Classification

Classification is a supervised[20] and self-automated machine learning algorithm that is used to test the unknown data from sampled data. In contrast to other algorithms in data mining, classifiers are used to handle both continuous and discrete attributes. Classification techniques are evolving and are now able to cope with medical datasets consisting of both continuous and discrete attributes, and are highly useful for classifying data into ranges. Classifiers that have been used in earlier studies on GDM were considered in this study. Figure 1 shows different classifiers. In diabetic research,[21] a few selected classifiers such as LR, SVM, Gaussian naïve Bayes (NB), KNN, and RF[22,23] are widely used. These classifiers were used for performance evaluation in this study. We analyzed different classification algorithms for their performance, accuracy, and output with the given dataset. Among the classifiers, RF achieved the highest accuracy rate. Another advantage of RF is its scalability;[24] when the dataset is large and dynamic in nature, the performance of RF is increased.

3.4 RF algorithm

The RF algorithm proposed by Ho[24] is based on the stochastic discrimination approach followed by Kleinberg.[25] RF is a modern ensemble classifier that has started gaining attention in recent years owing to its good classification capability. In this classifier, every single learner

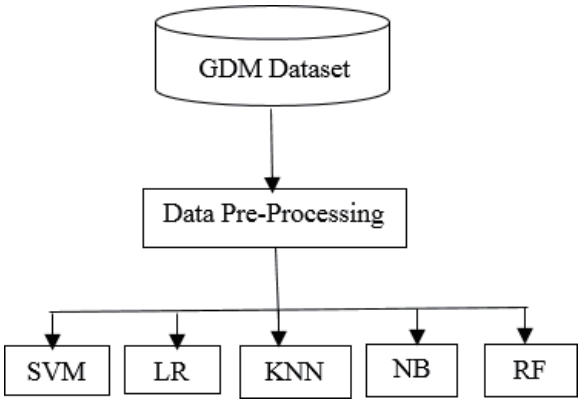

Fig. 1. Comparison of different classification techniques.

is a decision tree built on bagging data, while each node split is developed on the basis of a randomly selected feature subset. This is a supervised multilearning (ensemble) method that is based on the concept of decision trees. Compared with other classifiersclassifiers [26] such as ID3 and C4.5, RF is highly efficient since it can easily handle overfitted values. It creates multiple subsets of decision trees using regression, the mean, and the mode. The later versions of RF include bagging, boosting, and the control variance.[27,28] GDM data contains different sets of multidimensional attributes with continuous values. RF with tuned parameters can handle a GDM dataset with improved accuracy and lower error rate.

The pseudocode for RF is given as follows.

Input: Gestational dataset
Output: Predicted class variable
Procedure:

1. *Randomly select nodes.*
2. *Calculate node "d" using the best split point.*
3. *Split the nodes into sibling nodes using the best split.*

Repeat steps 1 to 3 until [the last node].

Figure 2 shows the test sample input of RF and the creation of subset trees. Figure 3 shows the workflow of the prediction model proposed in this study.

4. Experimentation and Results

The prediction model used in our investigation was developed using Python language[29,30] and R software.[31–33] The GDM dataset was input to the application. The data was preprocessed in the first step by using the GBM method. The result was used to develop a prediction model with promising attributes and ranges. After developing the training model, the preprocessed GDM data was tested using different classifiers. The results for each classifier were compared in terms of accuracy.

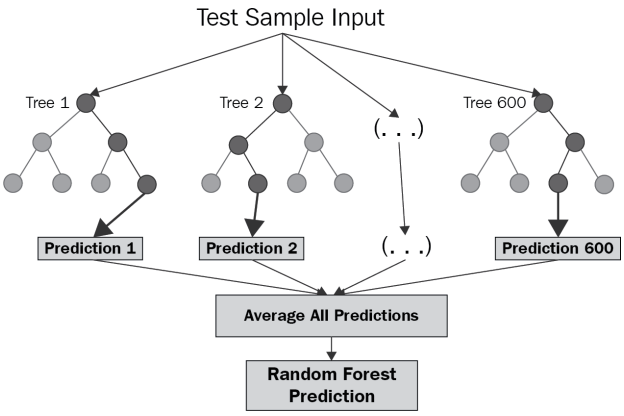

Fig. 2. RF tree construction.

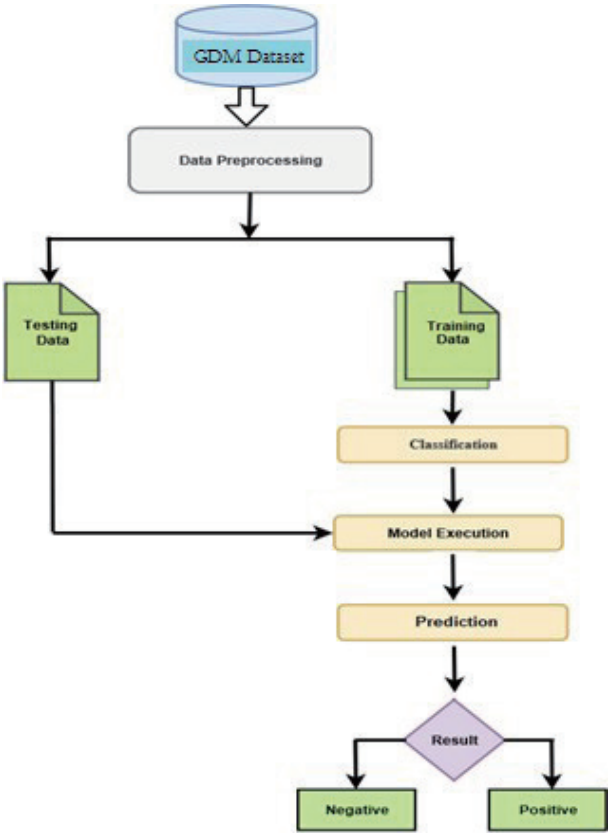

Fig. 3. (Color online) Prediction model for GDM.

4.1 Dataset

The Pima Indian diabetes dataset was sourced from the UCI online repository. It has the following attributes: pregnancy occurrences, oral glucose tolerance test (OGTT), diastolic BP, skinfold thickness, body mass index (BMI), plasma glucose level, diabetes pedigree function, age, and a class variable. There are nearly 750 observations taken with nine attributes. Figure 4 shows the dataset attributes with their values.

Figure 5 shows the correlation matrix that can be used to observe how the features are related to each other or to the target variable. It can be seen that the dataset is symmetrical about the leading axis and that each variable in the dataset is positively correlated with the others. Figure 6 shows the GDM dataset used to analyze the software prediction model.

4.2 Data preprocessing

The GDM dataset contained missing values and noisy data, which were preprocessed using the GBM method. In the table of values shown in Fig. 7, the empty values have been replaced with values during preprocessing. The missing values in the group concerning the age and BMI were replaced with the mean values as shown in the table.

	preg	plas	pres	skin	test	mass	pedi	age	class
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	0.264393	0.611491	0.593485	0.294479	0.183863	0.483718	0.194990	0.410381	0.348958
std	0.175023	0.152944	0.099150	0.088797	0.100498	0.102461	0.136913	0.145188	0.476951
min	0.058824	0.221106	0.196721	0.070707	0.016548	0.271237	0.032231	0.259259	0.000000
25%	0.117647	0.501256	0.524590	0.252525	0.143617	0.409836	0.100723	0.296296	0.000000
50%	0.264393	0.587940	0.591825	0.294479	0.183863	0.482861	0.153926	0.358025	0.000000
75%	0.352941	0.704774	0.655738	0.323232	0.183863	0.545455	0.258781	0.506173	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Fig. 4. Dataset model.

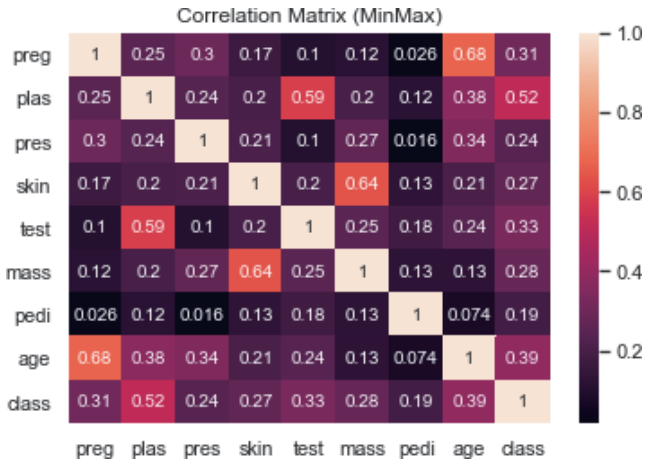

Fig. 5. (Color online) Correlation matrix of GDM.

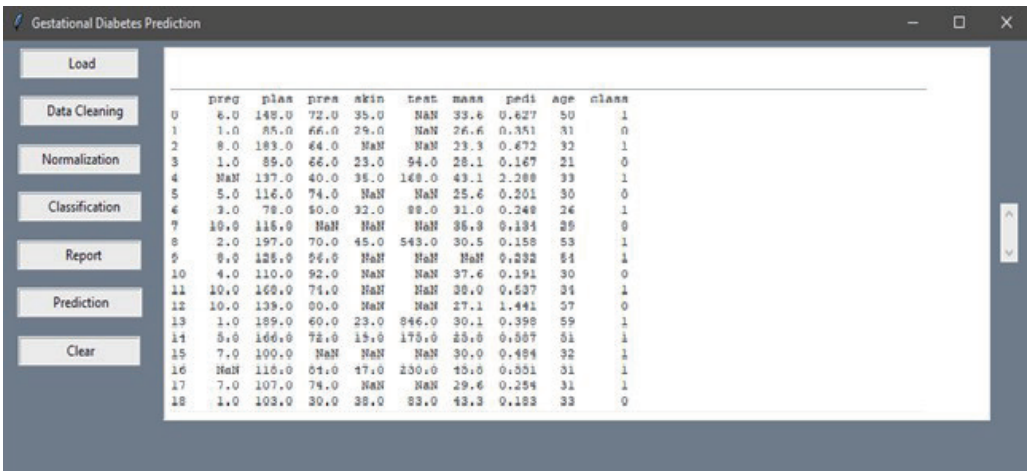

Fig. 6. (Color online) GDM dataset.

In the second step of preprocessing the GDM dataset, normalization was performed. Box plot analysis[34] was used to estimate the amount of data that will be normalized. The results obtained before and after the normalization are respectively shown in Figs. 8 and 9.

	preg	plas	pres	skin	test	mass	pedi	age	class
0	6.0	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1.0	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	8.0	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	NaN	137.0	40.0	35.0	168.0	43.1	2.288	33	1

Mean()

	preg	plas	pres	skin	test	mass	pedi	age	class
0	6.000000	148.0	72.0	35.00000	155.548223	33.6	0.627	50	1
1	1.000000	85.0	66.0	29.00000	155.548223	26.6	0.351	31	0
2	8.000000	183.0	64.0	29.15342	155.548223	23.3	0.672	32	1
3	1.000000	89.0	66.0	23.00000	94.000000	28.1	0.167	21	0
4	4.494673	137.0	40.0	35.00000	168.000000	43.1	2.288	33	1

Fig. 7. Preprocessing using mean.

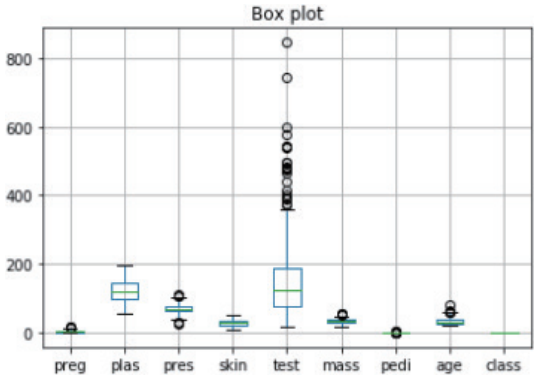

Fig. 8. (Color online) Box plot without normalization.

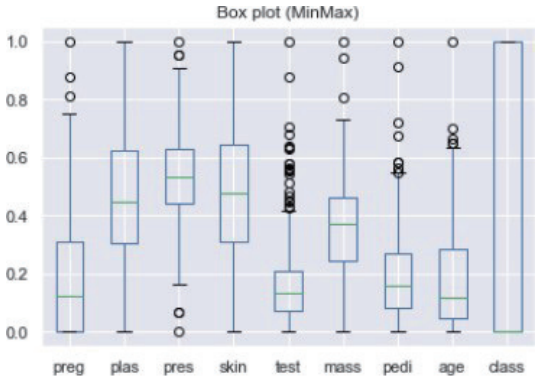

Fig. 9. (Color online) Box plot after min-max normalization.

Data visualization is an important step in data analysis. If the data is graphically visualized as box plots or histograms, it provides a better understanding of different feature values and their distribution. Figures 10 and 11 show the histograms obtained before and after the normalization of each attribute, respectively.

In the third step, the given GDM dataset and resampled dataset were analyzed using different classifiers. Both normalized and non-normalized data were fed as inputs to evaluate the performance of the classifiers. The result showed that RF achieved high accuracy for both normalized and non-normalized data. After min-max normalization was applied to the dataset, the RF classifier achieved much higher accuracy than the other classifiers.

Figure 12 shows an individual report of each classifier for relevant performance measures, where the accuracy rate of the classifier is presented using a confusion matrix. The dataset was used without normalization to determine the classifier's basic functionality.

Figure 13 shows the outputs attained using different classifiers along with their accuracy rates when using the max-abs normalization technique. Figure 14 shows the outputs attained using different classifiers along with their accuracy rates when using the mix-max normalization technique.

The above experimental results show that the normalization techniques perform well on the different classifiers. Min-max normalization performed well on most classifiers, whereas RF with tuned parameters and min-max normalization outperformed all the other classifiers for the given dataset.

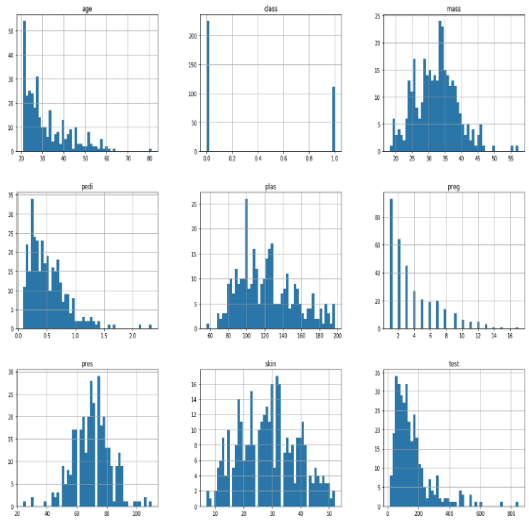

Fig. 10. (Color online) Histograms for different attributes.

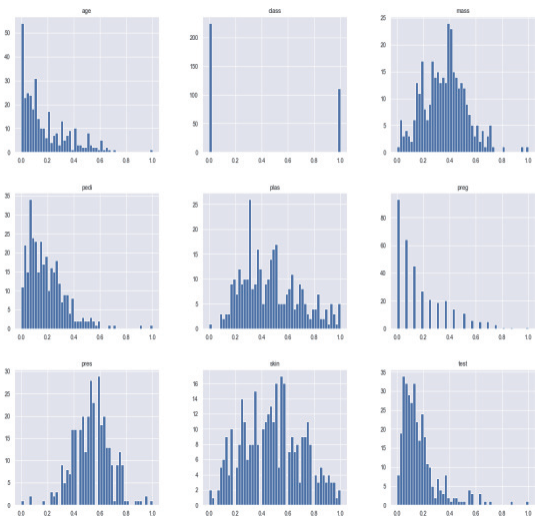

Fig. 11. (Color online) Normalized histograms for different attributes.

```
Logistic Regression Accuracy: 80.26315789473685
=== Confusion Matrix ===
[[63 19]
 [11 59]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.85      0.77      0.81        82
           1       0.76      0.84      0.80        70

   micro avg       0.80      0.80      0.80       152
   macro avg       0.80      0.81      0.80       152
weighted avg       0.81      0.80      0.80       152
```

```
SVM Accuracy: 73.68421052631578
=== Confusion Matrix ===
[[65 17]
 [23 47]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.74      0.79      0.76        82
           1       0.73      0.67      0.70        70

   micro avg       0.74      0.74      0.74       152
   macro avg       0.74      0.73      0.73       152
weighted avg       0.74      0.74      0.74       152
```

```
KNN Accuracy: 76.31578947368422
=== Confusion Matrix ===
[[68 14]
 [22 48]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.76      0.83      0.79        82
           1       0.77      0.69      0.73        70

   micro avg       0.76      0.76      0.76       152
   macro avg       0.76      0.76      0.76       152
weighted avg       0.76      0.76      0.76       152
```

```
Naive Bayes Accuracy: 76.97368421052632
=== Confusion Matrix ===
[[62 20]
 [15 55]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.81      0.76      0.78        82
           1       0.73      0.79      0.76        70

   micro avg       0.77      0.77      0.77       152
   macro avg       0.77      0.77      0.77       152
weighted avg       0.77      0.77      0.77       152
```

```
Random Forest Accuracy: 90.13157894736842
=== Confusion Matrix ===
[[72 10]
 [ 5 65]]


=== Classification Report ===
              precision    recall  f1-score   support

         0.0       0.94      0.88      0.91        82
         1.0       0.87      0.93      0.90        70

   micro avg       0.90      0.90      0.90       152
   macro avg       0.90      0.90      0.90       152
weighted avg       0.90      0.90      0.90       152
```

Fig. 12. Classification report without normalization.

```
Logistic Regression Accuracy: 80.26315789473685
=== Confusion Matrix ===
[[62 20]
 [10 60]]

=== Classification Report ===
              precision    recall  f1-score   support

         0.0       0.86      0.76      0.81        82
         1.0       0.75      0.86      0.80        70

   micro avg       0.80      0.80      0.80       152
   macro avg       0.81      0.81      0.80       152
weighted avg       0.81      0.80      0.80       152
```

```
SVM Accuracy: 80.26315789473685
=== Confusion Matrix ===
[[62 20]
 [10 60]]

=== Classification Report ===
              precision    recall  f1-score   support

         0.0       0.86      0.76      0.81        82
         1.0       0.75      0.86      0.80        70

   micro avg       0.80      0.80      0.80       152
   macro avg       0.81      0.81      0.80       152
weighted avg       0.81      0.80      0.80       152
```

```
KNN Accuracy: 80.92105263157895
=== Confusion Matrix ===
[[67 15]
 [14 56]]

=== Classification Report ===
              precision    recall  f1-score   support

         0.0       0.83      0.82      0.82        82
         1.0       0.79      0.80      0.79        70

   micro avg       0.81      0.81      0.81       152
   macro avg       0.81      0.81      0.81       152
weighted avg       0.81      0.81      0.81       152
```

```
Naive Bayes Accuracy: 76.97368421052632
=== Confusion Matrix ===
[[62 20]
 [15 55]]

=== Classification Report ===
              precision    recall  f1-score   support

         0.0       0.81      0.76      0.78        82
         1.0       0.73      0.79      0.76        70

   micro avg       0.77      0.77      0.77       152
   macro avg       0.77      0.77      0.77       152
weighted avg       0.77      0.77      0.77       152
```

```
Random Forest Accuracy: 91.44736842105263
=== Confusion Matrix ===
[[75  7]
 [ 6 64]]

=== Classification Report ===
              precision    recall  f1-score   support

           0       0.93      0.91      0.92        82
           1       0.90      0.91      0.91        70

   micro avg       0.91      0.91      0.91       152
   macro avg       0.91      0.91      0.91       152
weighted avg       0.91      0.91      0.91       152
```

Fig. 13. Classification report with max-abs normalization.

5. Performance Evaluation

A method widely used for handling highly imbalanced datasets is called Resampling. It consists of adding or removing the samples from/to the training dataset. In such cases, the simplest approach involves adding or duplicating (Replicating, Reproducing the same) samples in the dataset. This type of resampling technique can be effective to have a better performance on the classification model. Table 1 illustrates the detailed results of the comparative analysis of the dataset.

The following are the metrics used to evaluate and compare the models discussed earlier:

- Recall – measures the ability of a classifier and its relevant distance,
- Precision – fraction of relevant instances among the retrieved instances,
- F1 score – combination of recall and precision using a harmonic mean,
- Confusion matrix – real, actual, and predicted labels from a classification problem, and
- Receiver operating characteristic (ROC) curve – differentiates values into true positive and false positive rates.

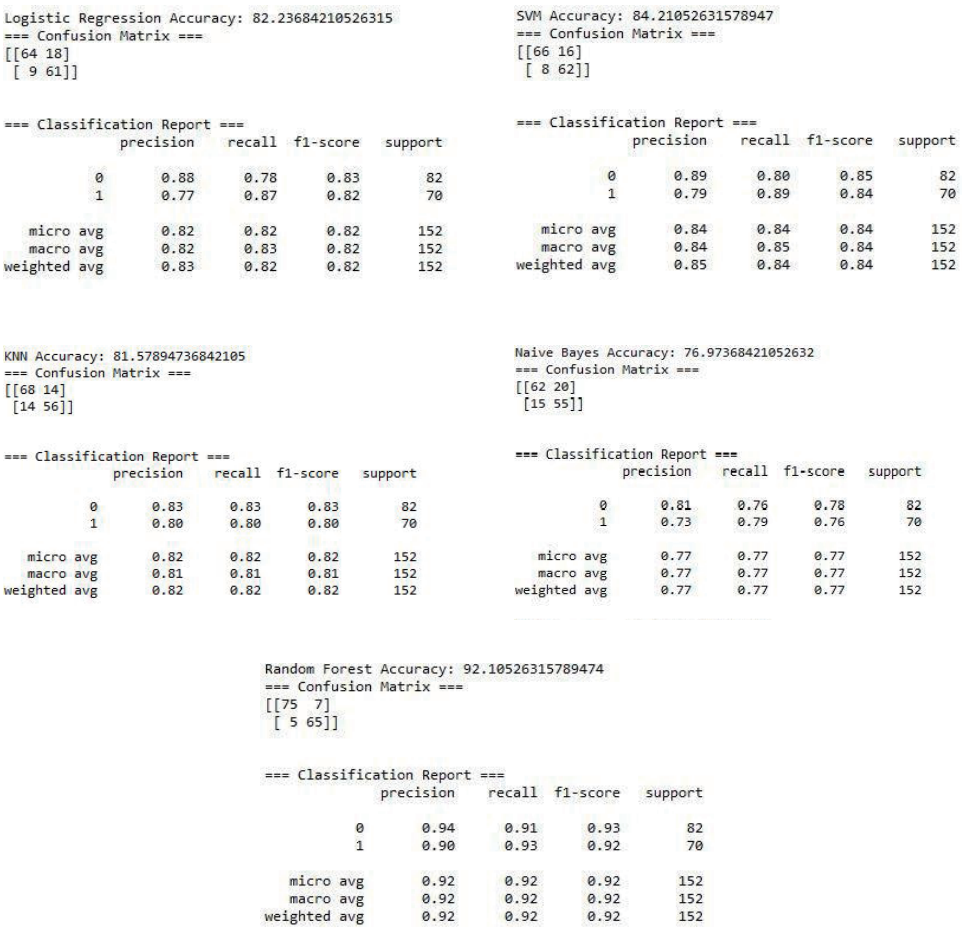

Fig. 14. Classification report with min-max normalization.

Table 1
Comparison of classification accuracies.

Algorithm	Classification accuracy		
	Without normalization	Max-abs normalization	Min-max normalization
LR	80.26	80.26	82.23
SVM	73.68	80.26	84.21
NB	76.97	76.97	76.97
KNN	76.31	80.26	81.57
RF	90.13	91.44	92.10

The ROC curve is an evaluation measure used to analyze the performance of classifiers. It is a probability curve that is deployed to estimate the capability of an algorithm or model. It is plotted with the true positive rate on the *y*-axis and the false positive rate on the *x*-axis. The two cutoff points are the sensitivity and specificity with a threshold. In this curve, each point represents a sensitivity and specificity pair corresponding to a particular decision threshold. Figures 15 and 16 show the ROC curves used to differentiate whether the data from a patient falls under a disease or normal category.

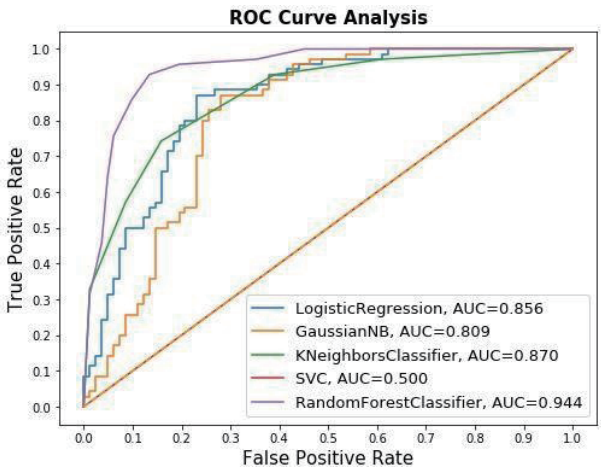

Fig. 15. (Color online) ROC curve analysis for max-abs normalization.

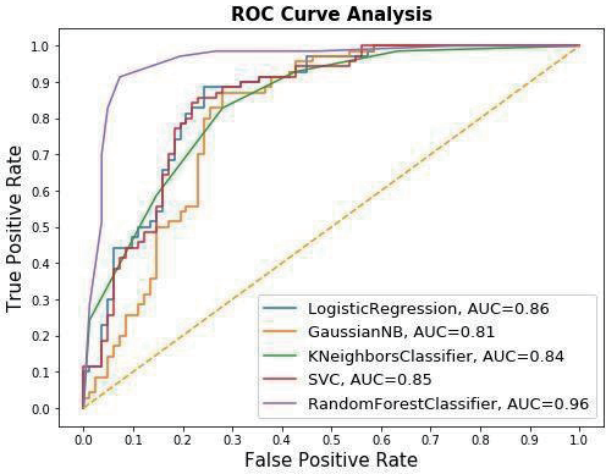

Fig. 16. (Color online) ROC curve for proposed normalization model.

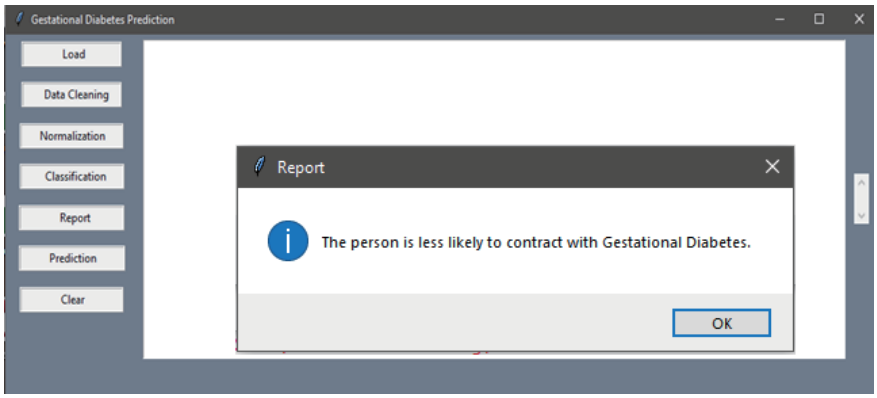

Fig. 17. (Color online) Prediction report (negative).

Classification algorithms that optimize the overall accuracy or class distribution purity often suffer from the classification of imbalanced data. In most scenarios, the testing set is classified under the majority class. However, such imbalanced data classification strongly considers accuracy in identifying the minority class (e.g., diseased samples). Thus, low sensitivity is highly undesirable. When numerous data features are collected and engineered along with appropriate estimator selection, it is possible to increase the performance. The ROC curve is a two-dimensional graph in which sensitivity is plotted against specificity, i.e., accuracy, in identifying the majority class (e.g., normal samples). The ROC curve is deemed to be an accurate means of evaluating the performance of a classification. In general, RF not only improves the classification accuracy but also gives a highly balanced classification result compared with other classification algorithms. Figures 17 and 18 show the prediction reports of gestational diabetes.

The proposed model has three steps: preprocessing, classification, and prediction. Preprocessing involves the handling of missing values and data normalization. First, the input dataset is converted into a preprocessed dataset. In the GBM method, the missing values in the

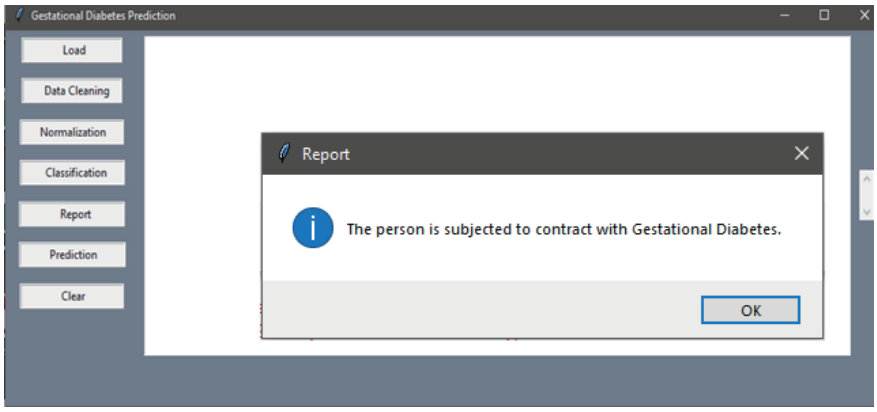

Fig. 18. (Color online) Prediction report (positive).

dataset are replaced with the group mean of the corresponding columns. The min-max method scales the data of the given column. After the data is preprocessed, it is classified by a supervised machine learning algorithm. Firstly, a classifier is built using a set of rules based on which the future class or data is classified. Classification is an important task in machine learning and data mining. In the current study, the RF algorithm is used with estimator selection for classification. Finally, this model is applied to predict GDM using patient health data. The proposed model compared with the other classifiers yields high accuracy rate.

6. Conclusion

The aim of this work is to develop a novel approach to predicting GDM using machine learning classifiers and data mining methods. The significant classifiers are taken into account, namely, LR, SVM, NB, KNN, and RF. In preprocessing, the GBM method and min-max normalization techniques are used to improve the data quality in the dataset. To evaluate the classifiers in terms of their accuracy rate, the confusion matrix and ROC curve are used. The results showed that the RF classifier with tuned parameters achieved higher accuracy than did the other classifiers. The generic nature of the GDM dataset contains correlated attribute values, which require an internally combined approach to obtain better results. The RF algorithm uses attribute values with regression effectively while using the GDM dataset. The performance evaluation results have also proven that RF is a suitable approach to predicting GDM in earlier stages. For the prediction of GDM with similar real-time datasets, the proposed model can also be enhanced by using combined techniques such as ensemble methods.

References

1 M. Hod, A. Kapur, and H. D. McIntyre: Am. J. Obstet. Gynecol. **221** (2019) 109. https://doi.org/10.1016/j.ajog.2019.01.206
2 L. McCloskey, E. Quinn, O. Ameli, T. Heeren, M. Craig, A. Lee-Parritz, R. Iverson, B. Jack, and A. Bernstein: Women Health Iss. **29** (2019) 480. https://doi.org/10.1016/j.whi.2019.08.003
3 P. S. Muller and M. Nirmala: Proc. 2016 Online Int. Conf. Green Engineering and Technologies (IC-GET) (IEEE, 2016) 1–4. https://doi.org/10.1109/GET.2016.7916859

4 E. Cambau, P. Saunderson, M. Matsuoka, S. T. Cole, M. Kai, P. Suffys, P. S. Rosa, D. Williams, U. D.Gupta, M. Lavania, N. Cardona-Castro, Y. Miyamoto, D. Hagge, A. Srikantam, W. Hongseng, A. Indropo, V. Vissa, R. C. Johnson, B. Cauchoix, V. K. Pannikar, E. A. W. D. Cooreman, V. R. R. Pemmaraju, and L. Gillini: Clin. Microbiol. Infect. **24** (2018) 1305. https://doi.org/10.1016/j.cmi.2018.02.022

5 J. B. Pernabas, S. F. Fidele, and K. K. Vaithinathan: Egypt. Inform. J. **20** (2019) 117. https://doi.org/10.1016/j.eij.2019.01.001

6 D. A. J. M. Schoenaker, Y. Vergouwe, S. S. Soedamah-Muthu, L. K. Callaway, and G. D. Mishra: Diabetes. Res. Clin. Pract. **146** (2018) 48. https://doi.org/10.1016/j.diabres.2018.09.021

7 A. Iyer, S. Jeyalatha, and R. Sumbaly: Int. J. Data. Min. Knowl. Manage. Process. **5** (2015) 1. https://doi.org/10.5121/ijdkp

8 R. Milewski, A. Kuczyńska, B. Stankiewicz, and W. Kuczyński: Adv. Med. Sci. **62** (2017) 202. https://doi.org/10.1016/j.advms.2017.02.001.

9 R. Coppi, and P. D'Urso: Comput. Stat. Data. Anal. **43** (2003) 149. https://doi.org/10.1016/S0167-9473(02)00226-8

10 P. S. Kumar and V. Umatejaswi: Int J Sci Re. **7** (2017) 705. http://www.ijsrp.org/research-paper-0617/ijsrp-p6689.pdf

11 I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda: Comput. Struct. Biotechnol. J. **15** (2017) 104. https://doi.org/10.1016/j.csbj.2016.12.005

12 S. Nagarajan, R. M. Chandrasekaran, and P. Ramasubramanian: Int. J. Curr. Res. Acad. Rev. **2** (2014) 91. http://www.ijcrar.com/vol-2-10/Srideivanai%20Nagarajan,%20et%20al.pdf

13 Z. Omiotek, A. Burda, and W. Wójcik: Expert Syst. Appl. **40** (2013) 6684. https://doi.org/10.1016/j.eswa.2013.03.022

14 R. Asaoka, K. Hirasawa, A. Iwase, Y. Fujino, H. Murata, N. Shoji, and M. Araie: Am. J. Ophthalmol. **174** (2017) 95. https://doi.org/10.1016/j.ajo.2016.11.001

15 L. Dora, S. Agrawal, R. Panda, and A. Abraham: Expert Syst. Appl. **114** (2018) 313. https://doi.org/10.1016/j.eswa.2018.07.039

16 Z. Guo, Y. Wan, and H. Ye: Neurocomputing **360** (2019) 185. https://doi.org/10.1016/j.neucom.2019.06.007

17 E. Tutubalina, Z. Miftahutdinov, S. Nikolenko, and V. Malykh: J. Biomed. Inf. **84** (2018) 93. https://doi.org/10.1016/j.jbi.2018.06.006

18 M. F. Wahid, R. Tafreshi, M. Al-Sowaidi, and R. Langari: J. Comput. Sci. **27** (2018) 69. https://doi.org/10.1016/j.jocs.2018.04.019

19 B. F. Sallis, L. Erkert, S. Moñino-Romero, U. Acar, R. Wu, L. Konnikova, W. S. Lexmond, M. J. Hamilton, W. A. Dunn, Z. Szepfalusi, J. A. Vanderhoof, S. B. Snapper, J. R. Turner, J. D. Goldsmith, L. A. Spencer, S. Nurko, and E. Fiebiger: J. Allergy Clin. Immunol. **141** (2018) 1354. https://doi.org/10.1016/j.jaci.2017.11.027

20 T. Shu, B. Zhang, and Y. Y. Tang: Inf. Sci. **467** (2018) 477. https://doi.org/10.1016/j.ins.2018.08.011

21 S. Nami and M. Shajari: Expert Syst. Appl. **110** (2018) 381. https://doi.org/10.1016/j.eswa.2018.06.011.

22 A. Statnikov, L. Wang, and C. F. Aliferis: BMC Bioinform. **9** (2008) 319. https://doi.org/10.1186/1471-2105-9-319

23 M. Z. Alam, M. S. Rahman, and M. S. Rahman: Inf. Med. Unlocked **15** (2019) 1. https://doi.org/10.1016/j.imu.2019.100180

24 T. K. Ho: Proc. 3rd Int. Conf. Document Analysis and Recognition (IEEE, 1995) 278. https://doi.org/10.1109/ICDAR.1995.598994

25 E. M. Kleinberg: IEEE Trans. Pattern Anal. Mach. Intell. **22** (2000) 473. https://doi.org/10.1109/34.857004

26 R. Yang, G. Zhang, F. Liu, Y. Lu, F. Yang, F. Yang, M. Yang, Y. Zhao, and D. Li: Ecol. Indic. **60** (2016) 870 https://doi.org/10.1016/j.ecolind.2015.08.036

27 H. Wei, C. Hu, S. Chen, Y. Xue, and Q. Zhang: Inf. Sci. **477** (2019) 399. https://doi.org/10.1016/j.ins.2018.10.056

28 J. A. Cook: Pattern Recognit. Lett. **85** (2017) 35. https://doi.org/10.1016/j.patrec.2016.11.015

29 R. J. Simpson, T. A. Johnson, and I. A. Amara: Am. Heart J. **116** (1988) 1663. https://doi.org/10.1016/0002-8703(88)90791-0

30 K. H. Knuth: Digit Signal Process. **95** (2019) 102581. https://doi.org/10.1016/j.dsp.2019.102581

31 I. Dejanović, R. Vaderna, G. Milosavljević, and Ž. Vuković: Knowl-Based Syst. **115** (2017) 1. https://doi.org/10.1016/j.knosys.2016.10.023

32 M. H. Moghadam and S. M. Babamir: Proc. 2014 4th Int. Conf. Computer and Knowledge Engineering (ICCKE, 2014) 775–780. https://doi.org/10.1109/ICCKE.2014.6993419

33 A. Rangel and J. A. Clithero: Curr. Opin. Neurobiol. **22** (2012) 970. https://doi.org/10.1016/j.conb.2012.07.011

34 V. Gajera, Shubham, R. Gupta, and P. K. Jana: Proc. 2016 2nd Int. Conf. Applied and Theoretical Computing and Communication Technology (iCATccT, 2016) 812–816. https://doi.org/10.1109/ICATCCT.2016.7912111.

Appendix

Comparison of various Machine Learning models (PIMA dataset).

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	71.05	50.00	71.00	59.00
K-Nearest Neighbor	76.97	77.00	77.00	77.00
Random Forest	77.63	79.00	78.00	78.00
Decision Tree	65.13	69.00	65.00	66.00
Naïve Bayes	76.97	78.00	77.00	77.00
Support Vector Machine	71.05	50.00	71.00	59.00
XG Boosting	75.65	77.00	76.00	76.00

Comparison of various Machine Learning models with SMOTE (PIMA dataset).

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	77.92	78.00	78.00	78.00
K-Nearest Neighbor	72.72	76.00	73.00	73.00
Random Forest	78.57	80.00	79.00	79.00
Decision Tree	66.88	66.00	67.00	67.00
Naïve Bayes	74.02	74.00	74.00	74.00
Support Vector Machine	74.67	75.00	75.00	75.00
XG Boosting	74.67	75.00	75.00	75.00