International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 9-No.1 (2023) pp. 36-41
http://dergipark.org.tr/en/pub/ijcesen

ISSN: 2149-9144

Research Article

Diabetes Prediction Using Colab Notebook Based Machine Learning Methods

Önder YAKUT*

Kocaeli University, Faculty of Technology, Department of Information Systems Engineering 41001, Kocaeli-Turkey
* **Corresponding Author**: **Email:** onder.yakut@kocaeli.edu.tr - **ORCID:** 0000-0003-0265-7252

Article Info:

DOI: 10.22399/ijcesen.1185474
Received : 07 October 2022
Accepted : 16 March 2023

Keywords :

Cloud Computing
Diabetes Prediction
Google Colaboratory
Machine Learning

Abstract:

Diabetes is getting more and more common around the world. People suffer from diabetes or live at risk associated with this disease. It is necessary to prevent health problems caused by diabetes, to reduce the risk of diabetes and to reduce the load of diabetes on the health system. Therefore, it is important to diagnose and treat diabetic patients early. In this study, Pima Indian Diabetes (PID) database was used to predict diabetes. The PID database was divided into 2/3 for the training dataset and 1/3 for the test dataset. Then, the test and training datasets were fed into the machine learning models using five-fold cross-validation. Random Forest Classifier, Extra Tree Classifier and Gaussian Process Classifier machine learning methods were used to predict whether individuals have diabetes or not. In this study, the proposed method with the highest prediction accuracy was determined as the Random Forest Classifier. The proposed method's accuracy was 81.71%, precision was 88.79%, recall was 84.83%, F-score was 86.76% and ROC AUC was 88.03%. The proposed method was developed to assist clinicians in predicting the diagnosis of diabetic patients. The machine learning methods developed in this study were applied using Colab Notebook a Google Cloud Computing service.

1. Introduction

Insulin is a hormone that regulates blood sugar in the human body. Diabetes is a chronic disease that occurs when the pancreas cannot produce the necessary insulin, or the body cannot use the insulin produced effectively. Over time, diabetes causes serious damage to the cardiovascular system, eyes, kidneys, and nerves [1]. Diabetes does not only affect the individual who is sick. It is also a disease that affects the family of the sick individual and the whole of society. Care and treatment costs due to diabetes and the complications it causes increase rapidly and puts a load on the health system. In addition, the patient's quality of life decreases and this situation negatively affects the patient's family. Diabetes has become a global problem. Approximately 422 million people have diabetes, according to the World Health Organization. Most of these people live in low and middle-income countries. 1.6 million people die each year due to diabetes [2].

Predicting people with diabetes using machine learning methods will make the job of clinicians easier. Clinicians will ensure that people with diabetes are diagnosed and treated at an early stage. Thus, the load on the health system will be reduced and healthcare expenditures will be reduced. With the predetermination of diabetes, the disease will affect the lives of individuals less and increase the quality of life of individuals. Some studies in the literature that predict diabetes are given below. In these studies, the Pima Indians Diabetes database was used to predict diabetes. Febrian et al. used k-Nearest Neighbor and Naive Bayes algorithms comparatively to predict diabetes. In the proposed method, the Naive Bayes algorithm produced the best result for the PID data set [3]. Kibria et al. proposed a weighted voting classifier model to successfully predict diabetes risk. In the proposed method, ensemble learning was developed by combining Random Forest and XGBoost machine learning methods [4]. Chang et al. proposed an e-diagnosis system to detect and classify diabetes as an application of the Internet of Medical Things. They used Naive Bayes, Random Forest, and Decision Tree machine learning algorithms for classification in the proposed system [5]. Krishnamoorthi et al. Logistic Regression, k-Nearest Neighbor, Support Vector Machine and Random Forest machine learning algorithms were used for diabetes prediction. In the study, they proposed a smart

diabetes mellitus prediction framework [6]. Bhoi et al. proposed a model to predict diabetes in females in the PID dataset. In the proposed model, Classification Tree, Support Vector Machine, k-Nearest Neighbor, Naive Bayes, Random Forest, Neural Network, AdaBoost and Logistic Regression machine learning methods were used [7]. Guldogan et al. compared the prediction of the Multilayer Perceptron and Radial Based Function models to classify Type 2 Diabetes Mellitus [8]. Maulidah et al. proposed a method for diagnosing diabetes using the Naive Bayes algorithm and Particle Swarm Optimization technique utilization of the PID database [9]. Tigga et al. proposed the Random Forest Classifier method to predict Type 2 diabetes risk using the PID database [10]. Jakka et al. compared various machine learning methods to predict diabetic patients with high accuracy using the PID database [11]. Sisodia et al. used the Decision Tree, Support Vector Machine, and Naive Bayes machine learning classification algorithms to detect diabetes at an early stage [12]. Feng et al. proposed a variable-coded hierarchical fuzzy model for use in classification problems. The proposed method was used to diagnose diabetes. [13]. In this study, a machine learning method was proposed to be used in decision support systems for the early diagnosis of diabetes, which is becoming increasingly common. The proposed method has been thought to help patients start treatment early by identifying patients at risk for diabetes. The proposed method was developed to predict whether an individual had diabetes based on diagnostic criteria. For this purpose, training and test datasets were created using the patient data in the PID database. The PID database was divided into 2/3 for the training dataset and 1/3 for the test dataset. Thus, the developed machine learning models were fed with this training and test data set using five-fold cross-validation. The machine learning models developed in the study gained robust characteristics through rigorous training and testing. When the performance criteria obtained as a result of the experimental studies were evaluated, it is suggested that the proposed method has been used in computer-aided diagnosis systems to assist clinicians in decision-making.

The paper is structured as follows: material and methods are included in section 2. The experimental study is included in section 3. The experimental results are included in section 4. The discussion is included in section 5. The conclusions are included in section 6.

2. Material and Methods

In this section, the PID database used in the study is presented. Random Forest Classifier, Extra Tree Classifier and Gaussian Process Classifier machine learning methods used in the study are explained. In addition, it defined how the metrics that calculate the performance of the machine learning methods in the study are calculated.

2.1 Data Set

The data set used was obtained from the Pima Indians Diabetes (PID) Database of the National Institute of Diabetes and Digestive and Kidney Diseases. The PID database contains 768 records each with 9 attributes.

Table 1. *List of PID database's features [14].*

No	Name	Descriptions
1	Pregnancies	The number of pregnancies
2	Glucose	PGC 2 hours in an OGTT
3	Blood Pressure	Diastolic blood pressure
4	Skin Thickness	Triceps skinfold thickness
5	Insulin	2 Hours serum insulin
6	BMI	Body mass index
7	Diabetes Pedigree Function	Diabetes pedigree function
8	Age	Age
9	Outcome	Class label

The data set consists of two classes. The number of non-diabetic patients is 500 and it is labelled as 0. The number of patients with diabetes is 268 and is labelled 1. All records in the data set consist of women aged between 21 and 81 [14].

Table 1 shows the name and description of the PID database features in the data set. In this study, the PID database is divided into two a training and test data set. The training data set consists of 512 records (2/3). The training set was used to train the machine learning models. The test data set consists of 256 records. The test data set was used to test the machine learning models that have been trained.

2.2 Machine Learning Methods

In this study, Random Forest Classifier, Extra Tree Classifier and Gaussian Process Classifier machine learning methods used to predict whether individuals have diabetes or not are explained below.

2.2.1 Random Forest Classifier

Random Forest is a method that aims to improve the classification result by using more than one decision tree. The number of decision trees used in the Random Forest is parametric. Decision trees formed within the scope of this parameter are formed from subsets chosen randomly from the data set. Training takes place on randomly selected subsets and prediction is made on each decision tree. As a result

of these estimates, the decision tree and estimate with the highest success rate are selected as the result [15, 16].

2.2.2. Extra Trees Classifier

Extra Trees is a tree-based community learning algorithm proposed by Geurts et al in 2006. Extra Trees is an algorithm that continues its process through this ensemble by combining predictions from multiple decision trees to derive the classification result. While Random Forest calculates the best variable selection during the best split Extra Trees chooses a random variable and a random cut point instead of the calculation. Thus, the diversity between trees increases and the number of splits decreases. Since the calculation cost for the split process decreases, the training time of the model is also shortened [16, 17].

2.2.3. Gaussian Process Classifier

The Gaussian Process defines various kernel functions on the input data and creates an output with the weighted sums of these functions. Radial-based functions are used as kernel functions. These functions produce a Gaussian output according to the distance of the input data from a point [18].

2.3 Performance Metrics

These are the metrics used to evaluate the prediction result of the developed machine learning model. These metrics are calculated using the parameters True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). To calculate the prediction result of the developed model, the formulas for performance metrics are given below [19, 20].

$$Accuracy = \frac{(TP+TN)}{(TP+FN+FP+TN)} \quad (1)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

$$F1-Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

3. Experimental Study

In this study, a method is proposed to predict whether an individual is diabetic or not. The block diagram of the proposed method is shown in Fig. 1. The proposed method was developed using cloud computing-based Google Colab [21] and Google Drive service. Users must have a Google account to use these services. By using this account, the data set is uploaded to Google Drive and becomes accessible via cloud storage. Through Google

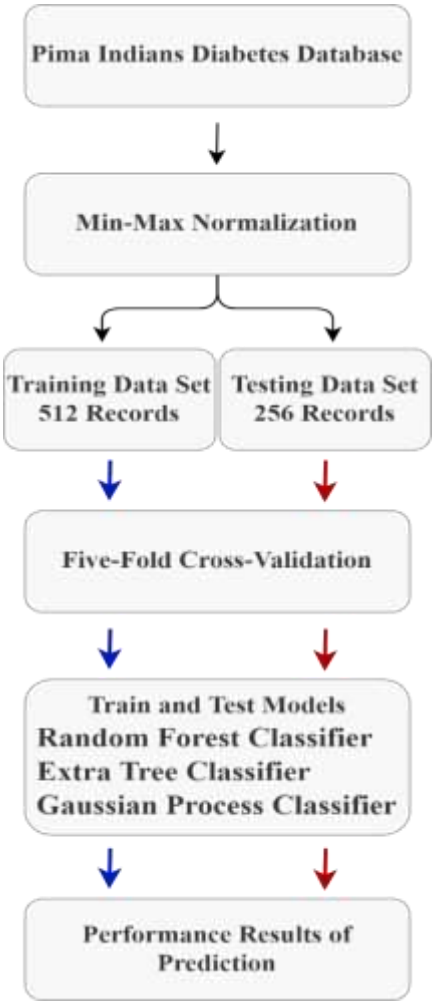

Figure 1. *Block diagram of the proposed method for diabetes prediction.*

Colab, pre-processing can be done on the data set in Python and the prediction results can be evaluated by developing machine learning models. Thus users can meet their software and hardware needs through the Google Cloud Computing Platform saving time and money.

In the proposed method, the PID database was loaded on Google Drive and made accessible via Google Colab. PID database features were normalized according to min-max normalization using Python programming language via Google Colab. Then, the PID database was divided into a training data set (512 records) and a test data set (256 records). With these data sets, machine learning models were fed by using the cross-validation technique. In this study, Random Forest Classifier, Extra Tree Classifier and Gaussian Process Classifier machine learning methods were used. Models were created using machine learning methods. Thus, the created models were trained, and training performance results were obtained. Then, the final prediction results were obtained by testing the trained models.

The proposed method was developed using classification-based machine learning methods that predict diabetes. The prediction results of the models belonging to these methods were obtained by using the PID database. The prediction results of the models were compared with each other, and the most successful model was proposed for diabetes prediction.

4. Experimental Results

In this study, the machine learning methods used the PID database to predict whether individuals have diabetes or not. The prediction results of the developed models are shown in Table 2. The models in Table 2 are listed from high performance to low performance. When Fig. 2 is examined, it is seen that the training results of the machine learning models in the study

Table 2. *Prediction results of developed models trained and tested using PID database.*

Methods		Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Random Forest Classifier	Training	82.08	84.64	77.09	93.81
	Testing	81.17	86.76	88.79	84.82
Extra Trees Classifier	Training	81.43	84.17	77.09	92.66
	Testing	78.57	84.93	86.97	83.04
Gaussian Process Classifier	Training	80.82	85.55	88.92	82.43
	Testing	77.49	84.15	87.89	80.70

In Fig. 2, the ROC AUC (Receiver Operating Characteristic Area Under Curve) values obtained from the training and test results of each model are shown. The diabetes prediction performance of the developed machine learning models in Fig. 2 has been compared.

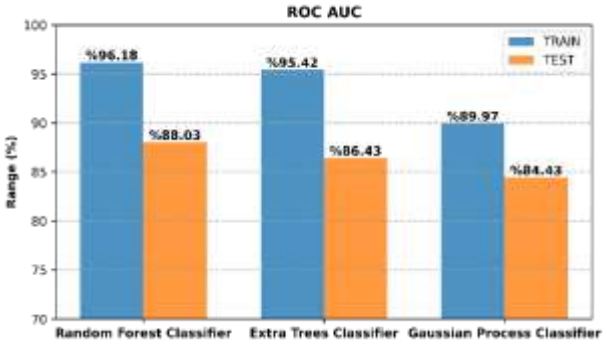

Figure 2. *ROC AUC of the developed machine learning models.*

are higher than the test results. When the test results of the machine learning models in the study were analyzed, the Random Forest Classifier performance result was 88.18%. The Extra Trees Classifier achievement result was 86.43%. The Gaussian Process Classifier performance result was 84.43%. When the performance values of the machine learning models in the study are investigated, the classifier with the highest value is Random Forest Classifier. Extra Trees Classifier and Gaussian Process Classifier are classifiers with second and third performance values respectively. The results of the second and third classifiers are comparatively lower than Random Forest Classifier.

5. Discussions

In Table 2, diabetes prediction made with the PID database was ranked from high performance to low performance. Random Forest Classifier method results were obtained Accuracy of 81.17%, F1-Score of 86.76%, Precision of 88.79%, and Recall of 84.82%. The performance results of the Extra Tree Classifier method were found Accuracy of 78.57%, F1-Score 84.93%, Precision of 86.97%, and Recall of 83.04%. The results of the Gaussian Process Classifier method achieved an Accuracy of 77.49%, F1- Score of 84.15%, Precision of 87.89%, and Recall of 80.70%. When the performances of the developed models were compared it was seen that the results were in the same range and close to each other. The machine learning methods in this study were compared with each other. So, it was concluded that the method with the highest performance was the Random Forest Classifier method.

In this study, ROC AUC results were obtained to show how well the machine learning methods predicted. The larger the area covered by the ROC AUC, the more effective the model was in distinguishing classes within the data set [22]. The ROC AUC values of the machine learning models in this study are shown in Fig. 2. When the ROC AUC values of the models were analyzed performances of models were listed as Random Forest Classifier 88.03%, Extra Tree Classifier 86.43% and Gaussian Process Classifier 84.43%. These results were evaluated according to the gradation of the AUC table [23]. Thus, it was concluded that the results obtained were quite close to each other and at the same time satisfactorily good.

When the results of the models in Table 2 and Figure 2 are examined, the method with the highest performance was the Random Forest Classifier. Within the scope of this study, the Random Forest Classifier method was proposed to estimate whether individuals had diabetes or not. The results of the Random Forest Classifier model are compared with

Table 3. *Comparison of the proposed method with studies in the literature.*

Model	Accuracy (%)
Proposed Method (Random Forest Classifier)	81.71
Naive Bayes [3]	78.57
Voting Classifier (XGB + RF) [4]	90.00
Random Forest [5]	79.57
Logistic Regression [6]	86.00
Logistic Regression [7]	76.80
Multilayer Perceptron [8]	78.10
Naive Bayes + Particle Swarm Optimization [9]	77.34
Random Forest [10]	75.00
Logistic Regression [11]	77.60
Naive Bayes [12]	76.30
Hierarchical Fuzzy Rule-based Evolutionary [13]	79.17

other studies in the literature in Table 3. In Table 3, the prediction accuracy of studies that predict diabetes using the PID database is compared with the prediction accuracy of the proposed method. As a result of the comparison, it is observed that the prediction degree of the proposed method is in the same range and close to each other with the models in the literature. High-performance models in the literature have produced better results by using the ensemble approach and editing the dataset. The data set used in our study is divided into 66% (2/3) for training and 34% (1/3) for testing. In such a case, the trained models are trained with less data. It is also being tested with more data. As a result, the model produces more durable, valid and reliable results. Also, the proposed method makes predictions using a single machine learning model. This situation affects the performance result of the proposed method. The Logistic Regression [6] model used the data set as 80% for training and 10% for testing. In this case, the Logistic Regression [6] model was trained with more data and tested with less data. Thus, the performance of the model is increased. The Voting Classifier (XGB + RF) [3] model used the data set as 70% for training and 30% for testing. They reduced the possibility of making mistakes in machine learning methods by reducing the number of features by making feature selections. In addition, they used two classifiers based on the ensemble learning approach in the prediction. As a result, they achieved a high level of accuracy.

When the ROC AUC result of the Random Forest Classifier shown in Figure 2 is examined, it is determined that the performance of this method is satisfactorily high. Also, the accuracy of the Random Forest Classifier is given in Table 2 as 81.71%. Therefore, the Random Forest Classifier method was proposed in this study, considering it useful and sufficient to predict diabetes. It is concluded that the proposed method will help clinicians' decision-making processes in computer-aided diagnosis systems.

6. Conclusions

In this study, a method that predicts whether individuals have diabetes is proposed by using Random Forest Classifier, Extra Tree Classifier and Gaussian Process Classifier methods. When the prediction results are analyzed, the Random Forest Classifier method, which has a satisfactory degree of success has been proposed within the scope of this study. The proposed method has been developed to assist clinicians in predicting the diagnosis of diabetic patients. It is thought that the proposed method can be useful in the decision support process by using computer-aided diagnosis systems. In future studies, a diabetes data set can be created for larger audiences and includes more records. The new data set can be analyzed using more advanced machine learning methods.

Author Statements:

- The conducted research is not related to either human or animal use.
- The authors declare that they have equal right on this paper.
- The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- The authors declare that they have nobody or no company to acknowledge.

References

[1]Diabetes Overview, (2023). *https://www.who.int/news-room/fact-sheets/detail/diabetes*

[2]Diabetes, (2023). *https://www.who.int/health-topics/diabetes#tab=tab_1*

[3]Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216; 21-30. https://doi.org/10.1016/j.procs.2022.12.107

[4]Kibria, H. B., Nahiduzzaman, M., Goni, M. O. F., Ahsan, M., & Haider, J. (2022). An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors*, 22(19); 7268. https://doi.org/10.3390/s22197268

[5]Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, 1-17. https://doi.org/10.1007/s00521-022-07049-z

[6]Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*, https://doi.org/10.1155/2022/1684017

[7]Bhoi, S. K. (2021). Prediction of diabetes in females of pima Indian heritage: a complete supervised learning approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10); 3074-3084. https://doi.org/10.17762/turcomat.v12i10.4958

[8]Güldoğan, E., Zeynep, T. U. N. Ç., Ayça, A. C. E. T., & ÇOLAK, C. (2020). Performance evaluation of different artificial neural network models in the classification of type 2 diabetes mellitus. *The Journal of Cognitive Systems*, 5(1); 23-32.

[9]Maulidah, N., Abdilah, A., Nurlelah, E., Gata, W., & Hasan, F. N. (2020). Seleksi Fitur Klasifikasi Penyakit Diabetes Menggunakan Particle Swarm Optimization (PSO) Pada Algoritma Naive Bayes. *Elkom: Jurnal Elektronika dan Komputer*, 13(2); 40-48.

[10]Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*, 167, 706-716. https://doi.org/10.1016/j.procs.2020.03.336

[11]Jakka, A., & Vakula Rani, J. (2019). Performance evaluation of machine learning models for diabetes prediction. *Int. J. Innov. Technol. Explor. Eng.(IJITEE)*, 8(11). 10.35940/ijitee.K2155.0981119

[12]Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, 132;1578-1585. https://doi.org/10.1016/j.procs.2018.05.122

[13]Feng, T. C., Li, T. H. S., & Kuo, P. H. (2015). Variable coded hierarchical fuzzy classification model using DNA coding and evolutionary programming. *Applied Mathematical Modelling*, 39(23-24);7401-7419. https://doi.org/10.1016/j.apm.2015.03.004

[14]Pima Indians Diabetes Database, (2022). https://data.world/data-society/pima-indians-diabetes-database

[15]Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[16]Yakut Ö., Bolat E. D. (2020). Arrhythmia Diagnosis from ECG Signal Using Tree-based Machine Learning Methods. *International Journal of Mathematic Engineering and Natural Sciences*, 4(16);954-964. https://doi.org/10.38063/ejons.361

[17]Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1); 3-42. https://doi.org/10.1007/s10994-006-6226-1

[18]MacKay, D. J. (1998). Introduction to Gaussian processes. *NATO ASI series F computer and systems sciences*, 168; 133-166.

[19]Yakut, Ö. (2020, November 19-21). *Cloud Computing Based Voting Classifier Method Used For Survival Prediction of Heart Failure Patients*. International Conference on Engineering Technologies, Konya-Turkey. https://icente.selcuk.edu.tr/

[20]Yakut, Ö. (2020, November 28-29). *Comparison of Clustering Methods For Early Stage Diabetes Risk Prediction Using Cloud Computing*. International Black Sea Coastline Countries Symposium 5, Zonguldak-Turkey.

[21]Google Colab Notebbook, (2022). https://colab.research. google.com

[22]Yakut, Ö., Bolat, E. D. (2020). An Efficient Arrhythmic Heartbeat Classification Method Using ECG Morphology Based Features. *Euroasia Journal of Mathematics, Engineering, Natural & Medical Sciences*, 7(13);200-212. https://doi.org/10.38065/euroasiaorg.403

[23]Yakut, Ö., Timuş, O., Bolat, E. D. (2016). HRV Analysis Based Arrhythmic Beat Detection Us-ing kNN Classifier. *International Journal of Biomedical and Biological Engineering*, 10(2); 60-63. doi.org/10.5281/zenodo.1339067