*Avrupa Bilim ve Teknoloji Dergisi*
*Özel Sayı, S. 268-272, Eylül 2020*
*© Telif hakkı EJOSAT'a aittir*
***Araştırma Makalesi***

www.ejosat.com ISSN:2148-2683

*European Journal of Science and Technology*
*Special Issue, pp. 268-272, September 2020*
*Copyright © 2020 EJOSAT*
***Research Article***

# Prediction of Diabetes Mellitus by using Gradient Boosting Classification

Fatema Nusrat [1 *], Betül Uzbaş [2], and Ömer Kaan Baykan [3]

[1] Konya Technical University, Faculty of Engineering and Natural Science, Department of Computer Engineering, Konya, Turkey (ORCID: 0000-0001-8495-4925)
[2] Konya Technical University, Faculty of Engineering and Natural Science, Department of Computer Engineering, Konya, Turkey (ORCID: 0000-0002-0255-5988)
[3] Konya Technical University, Faculty of Engineering and Natural Science, Department of Computer Engineering, Konya, Turkey (ORCID: 0000-0001-5890-510X)

(1st International Conference on Computer, Electrical and Electronic Sciences ICCEES 2020 – 8-10 October 2020)

(**DOI:** 10.31590/ejosat.803504)

**ATIF/REFERENCE:** Nusrat, F., Uzbaş, B. & Baykan, Ö. K. (2020). Prediction of Diabetes Mellitus by using Gradient Boosting Classification, *European Journal of Science and Technology*, (Special Issue), 268-272.

**Abstract**

Diabetes has become a pervasive and endemic health problem worldwide. It is a chronic disease and also life-threatening. It can cause health problems in many organs such as the heart, kidneys, eyes, nerves, and blood vessels. To reduce the fatality rate from diabetes, early prevention techniques are needed. Nowadays, machine learning techniques are used to predict or detect different life-threatening diseases like cancer, diabetes, heart diseases, thyroid, etc. In this study, a prediction model of diabetes mellitus was presented using the Pima Indian dataset. Three different machine learning techniques that Decision Tree (DT), Random Forest (RF) and, Gradient Boosting (GB) algorithm were used to predict diabetes mellitus and the performance analysis was performed. Confusion matrix, accuracy, F1 score, precision, recall, Cohen's kappa were evaluated and also a ROC curve was plotted. Out of the three techniques, the best results have been achieved with GB.

**Keywords:** Diabetes, Gradient Boosting, Machine Learning

# Gradient Boosting Classification kullanarak Diabetes Mellitus Tahmini

**Öz**

Diyabet, dünya çapında yaygın ve endemik bir sağlık sorunu haline gelmiştir. Bu hastalık, kronik ve ayrıca yaşamı tehdit eden bir hastalıktır. Kalp, böbrekler, gözler, sinirler ve kan damarları gibi birçok organda sağlık sorununa yol açabilir. Diyabet kaynaklı ölüm oranını azaltmak için erken önleme tekniklerine ihtiyaç duyulmaktadır. Günümüzde makine öğrenmesi teknikleri kanser, diyabet, kalp hastalıkları, tiroid vb. gibi hayatı tehdit eden farklı hastalıkları tahmin etmek veya tespit etmek için kullanılmaktadır. Bu çalışmada Pima Indian veri setini kullanarak bir şeker hastalığı tahmin modeli sunulmuştur. Çalışmada şeker hastalığını tahmin etmek için Karar Ağacı (KA), Rastgele Orman (RO) ve Gradyan Artırma (GA) algoritmaları olmak üzere üç farklı makine öğrenmesi tekniği uygulanmış ve performans analizi yapılmıştır. Karmaşıklık matrisi, doğruluk, F1 skoru, kesinlik, geri çağırma, Cohen'in kappa'sı değerlendirilmiş ve ayrıca ROC eğrisi çizdirilmiştir. Üç teknikten, GA ile en iyi sonuçlar elde edilmiştir.

**Anahtar Kelimeler:** Diyabet, Gradyan Arttırma, Makina Öğrenmesi

## 1. Introduction

Diabetes mellitus is a chronic disease which threatens human being's health life. It is increasing rapidly worldwide. Long-lasting disease as diabetes mellitus specified by hyperglycemia. In the blood, a high level of sugar or glucose indicates hyperglycemia. Nowadays diabetes has become a prevalent health problem worldwide. It is slowly damaging different parts of our body and creates serious complications. There are many types of diabetes such as type 1, type 2, auto-immune mediated diabetes, gestational diabetes [1]. Type 1 diabetes is also called Immune-Mediated Diabetes. In the world, 5-10% of people have type 1 diabetes. An absolute lack of insulin in the body and obstruction of pancreatic secretion is the main cause of type 1 diabetes mellitus. On the other hand, type 2 is also known as non-insulin-dependent diabetes. About 90-95% of people have type 2 diabetes which is more prevalent [2]. Gestational diabetes occurs in pregnant women who have no diabetes in previous [1]. In the world, adult diabetic people has increased from 108 million people in 1980 to 422 million people in 2014. Also, a study is observed in 2014, in East Asia and South Asia the number of diabetic patients 106 million and 86 million respectively [3]. For diabetes mellitus, Asia is the highest risk zone in the whole world [4].

In the year of 2015, South East Asia (SEA) has approximately 78.3 million (8.5%) populations are suffering from type 2 diabetes mellitus who are adults. Nowadays adults' diabetes rates are so high than previous [5]. During pregnancy, the SEA region also observed 24.2% of women affected by gestational diabetes which is threatening for the child. Another statistic also found that the prevalence of diabetes will reach 4.4% in 2030 for all age-groups worldwide and the total number of diabetes patients is expected to increase from 171 million in 2000 to 366 million in 2030. [6].

Nowadays machine learning techniques are used to predict or detect different life-threatening diseases like cancer, diabetes, heart diseases, thyroid, etc. So, this research is to design a diabetes risk prediction model using Pima Indian diabetes dataset.DT, RF, and GB algorithms were used to predict diabetes mellitus and the performance was analyzed. Confusion matrix, accuracy, F1 score, precision, recall, Cohen's kappa were evaluated and also a ROC curve was plotted.

### 1.1 Related Work

Many researchers have worked a lot of research on medical data of diabetes mellitus. Different machine learning algorithms have been used to predict or detect diabetes. In this section, we have explained some previous works which have been done by different machine learning techniques.

Dewangan, Amit et al. [7] constructed the C4.5 model by using Pima Indian Dataset. On that 75-25% training-testing partitions accuracy was 77.08% and, 76.22% in the case of 85-15% training-testing partitions and 75.32% in the case of 90-10% training-testing partitions. Karthikeyani et al. [8] constructed a model that used the partial least squares method to extract features and Linear Discriminate Analysis (LDA) method for predicting diabetes mellitus. The accuracy of that model was 74.40%. Parashar et al. [9] proposed a classification technique which was the LDA method and then combined Support Vector Machine (SVM) with Feed Forward Neural Networks (FFNN). The accuracy of the SVM model was 75.65%. Al Helal, Mustakim, et al. [10] constructed three classification models are the KNN, Naïve Bayes, and RF then their final accuracy was according to 66.19%, 72.66%, 73.72%. They were used in the Weka tool.

## 2. Material and Method

The flow chart of the overall proposed model is described in Figure 1. In this dataset, there has no missing value. Then we used K-fold cross-validation where k=5 which means it divided the dataset into 5 data subsets. Then three classification techniques such as DT, RF, and GB algorithms are used for prediction. Then the last step is to calculate accuracy, F1-score, precision, recall, and Cohen's Kappa, and also a ROC curve was plotted.

### 2.1. Dataset Description

By using the Pima Indian Diabetes Dataset we have performed this study. This dataset is open and available from the University of California, Irvine UCI machine learning respiratory [11]. This dataset has 768 records with 9 attributes including the outcome attribute. In the outcome total, 768 records there are 268 cases are "tested positive" which means the patient has diabetes and 500 cases are "tested negative" which indicates the patient has no diabetes. This is also a two-class problem with numerical values. Table 1 has described the detailed attribute information of the dataset.

*Table 1. Attribute information in the dataset*

| Number | **Attributes** | **Description** |
|---|---|---|
| ***1*** | *Pregnancies* | *Number of times pregnant* |
| ***2*** | *Insulin* | *2-Hour serum insulin (µU/ml)* |
| ***3*** | *BMI* | *Body mass index (weight in kg/(height in m)2)* |
| ***4*** | *Age* | *Age(years)* |
| ***5*** | *Glucose* | *Plasma glucose concentration a 2 hours in an oral glucose tolerance test* |
| ***6*** | *Blood Pressure* | *Diastolic blood pressure (mm Hg)* |
| ***7*** | *Diabetes PedigreeFunction* | *Diabetes pedigree function* |
| ***8*** | *Skin Thickness* | *Triceps skinfold thickness (mm)* |
| ***9*** | *Outcome* | *range of value: 0 and 1(0 means no 1 means yes)* |

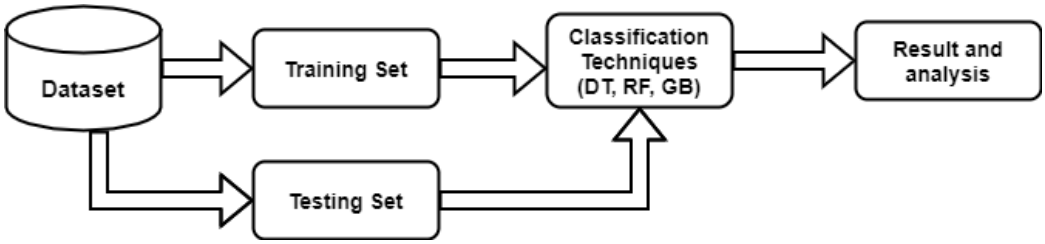


*Figure 1: Proposed Model of this research*

### 2.2. Decision Tree Method

Decision Tree is a supervised classification technique. It is a tree structure flow chart. It has a root node, internal nodes, and leaf nodes. When the decision tree has so many nodes then it prunes some node is called the pruning method [12]. In our study, we have done pruning where max_depth=3.

### 2.3. Random Forest Method

Random forest is a supervised machine learning algorithm. It is an ensemble learning method based on Bagging. It uses for regression and classification problems. It selects samples randomly from the dataset then builds a decision tree for each sample. A prediction result is measured from each decision tree. Then vote the prediction result after that the most votes consider the final prediction model [13].

### 2.4. Gradient Boosting Method

Gradient boosting is a machine learning technique that converts weak learners into strong learners. It is an ensemble learning method which also uses for regression and classification problems. The idea of gradient boosting was originated by Leo Breiman. There are three elements in gradient boosting. They are loss function, weak learner, and additive model [14].

### 2.5. Accuracy Measure

By using the confusion matrix we are calculated accuracy, F1-score, recall, precision. F1-score is the harmonic mean of precision and recall [15]. Also, a ROC (Receiver Operating Characteristic) curve was plotted and measured AUC value to analyze the performance of classification techniques. ROC curve is plotted by two parameters: TPR (True Positive Rate) and FPR (False Positive Rate). Also, Cohen's kappa was calculated which is a coefficient of statistics. It is a quantitative measure that measures the agreement between two raters. The range of kappa's value between 0 -1 where 0 means there is random agreement among raters and 1 means that there is a complete agreement among the raters [16].

## 3. Results and Discussion

We are used Python3 to implement the model. Several additional Python libraries are imported to solve the algorithm much efficiently. We have imported the necessary libraries like pandas, NumPy, scikit-learn, matplotlib, seaborn, and also imported our dataset into the Jupyter notebook. We have created three different classification models like DT, RF, and GB by using the Scikit-learn library. We are imported three classifier DecisionTreeClassifier (), RandomForestClassifier (), GradientBoostingClassifier (). Then we have measured the Confusion matrix, Accuracy, Precision, Recall, F1-score, AUC, and Cohen's Kappa. The confusion matrix of the DT, RF, and GB as shown in Table 2 is obtained in the analysis of diabetes mellitus. The comparison of Accuracy, Recall, Precision, F1-score, AUC, and Cohen's kappa of classification techniques is shown in Table 3.

### 3.1. Discussion

According to Table 3, the accuracy of DT has obtained as 0.7369, which is less than RF. While RF has 0.7450 accuracy, finally as the top best classifier, GB has 0.7630 accuracy. So GB has the highest performance that predicts diabetes mellitus. Precisions were obtained 0.6854, 0.7737, 0.6854 and recalls were obtained 0.4552, 0.3955, and 0.5932 for DT, RF, and GB respectively. And, DT, RF, and GB have 0.5470, 0.5234, 0.630 for F1-score values, respectively. Cohen's Kappa Statistic has also calculated for the classifiers and for DT, RF, and GB, 0.3722, 0.3761, and 0.4616 Cohen's Kappa have obtained. As the last comparison criteria; DT, RF, and GB have obtained 0.6842, 0.7996, and 0.8280 AUC value. According to the performance comparisons, among those three classifiers, the GB is the best classifier for the prediction. Also, figure 2 is shown a ROC curve of the GB model.

We have compared our results with the other researchers' works. They also used the same dataset. Table 4 shows a comparison between previous works and our study. GB classifier is also good for predictive accuracy other than a single predictive model like linear regression, naïve Bayes, support vector machines.

*Table 2: Confusion Matrix Model*

| Classification Technique | TP | TN | FP | FN |
|---|---|---|---|---|
| ***Decision Tree*** | *24* | *89* | *11* | *29* |
| ***Random Forest*** | *20* | *94* | *6* | *33* |
| ***Gradient Boosting*** | *32* | *85* | *14* | *22* |

*Table 3: Comparison of the different Classification Technique*

| Classification Technique | Accuracy | Precision | Recall | F1 score | Cohens kappa | AUC |
|---|---|---|---|---|---|---|
| ***DT*** | *0.7369* | *0.6854* | *0.4552* | *0.5470* | *0.3722* | *0.6842* |
| ***RF*** | *0.7450* | *0.7737* | *0.3955* | *0.5234* | *0.3761* | *0.7996* |
| ***GB*** | *0.7630* | *0.6854* | *0.5932* | *0.6360* | *0.4616* | *0.8280* |

*Table 4. Comparison between previous works and our study*

| Method | Accuracy | Reference |
|---|---|---|
| ***C4.5 model*** | *75.32%* | *Dewangan, Amit et al. [7]* |
| ***LDA*** | *74.40%.* | *Karthikeyani et al. [8]* |
| ***SVM with FFNN*** | *75.65%.* | *Parashar et al. [9]* |
| ***KNN,*** | *66.19%* | *Al Helal, Mustakim, et al. [10]* |
| ***Naïve Bayes,*** | *72.66%* | |
| ***RF*** | *73.72%.* | |
| ***DT,*** | *73.69%* | *Our study* |
| ***RF,*** | *74.50%* | |
| ***GB*** | *76.30%* | |

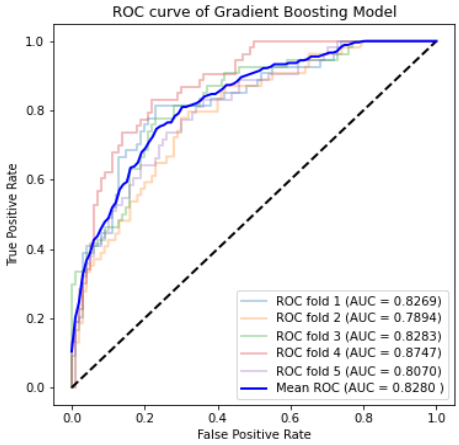


*Figure 2: ROC curve of the Gradient Boosting Model*

## 4. Conclusions

Detecting disease at an early stage is helpful for the medical center by using machine learning algorithms. The doctors can easily help the patients to identify their disease and also help them lead a better life. In this study, three classifier models, like DT, RF, and GB have experimented. On three classification techniques, the GB is the best classifier which can help doctors to diagnose or predict diabetes mellitus accurately.

## References

[1] Kerner, W., & Brückel, J. (2014). Definition, classification and diagnosis of diabetes mellitus. Experimental and clinical endocrinology & diabetes, 122(07), 384-386.

[2] Mellitus, D. (2005). Diagnosis and classification of diabetes mellitus. *Diabetes care*, *28*(S37), S5-S10.

[3] Priyadi, Akhmad, et al. (2019). An economic evaluation of diabetes mellitus management in South East Asia. Journal of Advanced Pharmacy Education & Research| Apr-Jun 9.2

[4] Chan, J. C., Malik, V., Jia, W., Kadowaki, T., Yajnik, C. S., Yoon, K. H., & Hu, F. B. (2009). Diabetes in Asia: epidemiology, risk factors, and pathophysiology. *Jama*, *301*(20), 2129-2140.

[5] Latif, Z. A., Ashrafuzzaman, S. M., Amin, M. F., Gadekar, A. V., Sobhan, M. J., & Haider, T. (2017). A Cross-sectional Study to Evaluate Diabetes Management, Control and Complications in Patients with type 2 Diabetes in Bangladesh. *BIRDEM Medical Journal*, *7*(1), 17-27.

[6] Wild, S., Roglic, G., Green, A., Sicree, R., & King, H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes care*, *27*(5), 1047-1053.
[7] kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *International Journal of Engineering and Applied Sciences*, *2*(5).
[8] Karthikeyani, V., & Begum, I. P. (2013). Comparison a performance of data mining algorithms (CPDMA) in prediction of diabetes disease. *International journal on computer science and engineering*, *5*(3), 205.
[9] Parashar, A., Burse, K., & Rawat, K. (2014). A Comparative approach for Pima Indians diabetes diagnosis using lda-support vector machine and feed forward neural network. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(11), 378-383.
[10] Al Helal, M., Chowdhury, A. I., Islam, A., Ahmed, E., Mahmud, M. S., & Hossain, S. (2019, February). An optimization approach to improve classification performance in cancer and diabetes prediction. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-5). IEEE.
[11] Dataset, P. I. D. UCI Machine Learning Repository, diambil dari http://archive. ics. uci. edu/ml/datasets. *Pima+ Indians+ Diabetes*.
[12] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.
[13] Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, *2*(1), 602-609.
[14] Breiman, L. (June 1997). Arcing The Edge (PDF). Technical Report 486. Statistics Department, University of California, Berkeley.
[15] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437.
[16] https://towardsdatascience.com/cohens-kappa-9786ceceab58