Avrupa Bilim ve Teknoloji Dergisi
Sayı 52, S. 220-233, Kasım 2023
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 52, pp. 220-233, November 2023
Copyright © 2023 EJOSAT
Research Article

Diyabet Tahmininde Geleneksel Yöntemlerin Analizi ve Değerlendirilmesi

Hayriye Tanyıldız[1*], Canan Batur Şahin[2], Özlem Batur Dinler[3]

[1*] Malatya Turgut Özal University, Faculty of Engineering and Natural Sciences, Departmant of Computer Engineering, Malatya, Turkey, (ORCID: 0000-0002-6300-9016), hayriye.tanyildiz@tedas.gov.tr

[2] Malatya Turgut Özal University, Faculty of Engineering and Natural Sciences, Departmant of Computer Engineering, Malatya, Turkey, (ORCID: 0000-0002-2131-6368), canan.batur@ozal.edu.tr

[3] Siirt University, Faculty of Engineering, Departmant of Computer Engineering, Siirt, Turkey, (ORCID: 0000-0002-2955-6761), o.b.dinler@siirt.edu.tr

(İlk Geliş Tarihi 10 Temmuz 2023 ve Kabul Tarihi 30 Kasım 2023)

(**DOI:** 10.5281/zenodo.10439913)

ATIF/REFERENCE: Tanyıldız, H., Batur Şahin, C., Batur Dinler, Ö. (2023). Alysis and Evaluation of Conventional Methods for Diabetes Prediction. *Avrupa Bilim ve Teknoloji Dergisi*, (52), 220-233.

Öz

Dünya çapında milyonlarca insanı etkileyen kronik bir hastalık olan diyabet, vücudun kan şekeri düzeylerini etkili bir şekilde yönetememesiyle karakterize edilir. Kontrol edilmezse veya uygun şekilde yönetilmezse, bu durum kalp hastalığı, felç, böbrek yetmezliği ve hatta körlük gibi ciddi sonuçlara yol açabilir. Genetik ve yaşam tarzı faktörlerinin karşılıklı etkileşimi nedeniyle, diyabet insidansı artmakta ve diyabet acil müdahale gerektiren önemli bir küresel sağlık sorunu olarak konumlanmaktadır. Dünya Sağlık Örgütü (WHO), diyabetin küresel prevalansının 1980'den bu yana neredeyse iki katına çıktığını ve yetişkin nüfusta %4,7'den %8,5'e yükseldiğini bildirmektedir. Bu artış, hastalığın erken teşhisine ve etkin yönetimine yönelik stratejilerin aciliyetini ve önemini vurgulamaktadır. Böyle bir halk sağlığı sorunu karşısında sağlık hizmetleri bu salgınla mücadele için teknolojik gelişmelerden yardım istemektedir. Sağlık hizmetlerinde en umut verici teknolojik sınırlar arasında, çok büyük miktarda veriyi analiz edebilen, kalıpları tanımlayabilen ve sonuçları tahmin edebilen yapay zekanın (AI) bir alt kümesi olan Makine Öğrenimi (ML) yer alıyor. Makine öğrenimi, hasta sağlığına ilişkin değerli içgörüler sağlayarak, tedavi kararlarını bildirerek ve hatta bir kişinin gelecekte hastalığa yakalanma riskini tahmin ederek diyabet yönetiminde devrim yaratma potansiyeline sahiptir. Bu teknoloji, doğru kullanılırsa diyabetle mücadelede oyunu değiştirebilir. Bu bağlamda, diyabet riskini tahmin etmek için geleneksel sınıflandırıcı yöntemlerin kullanılması uygulanabilir ve etkili bir yaklaşım gibi görünmektedir. Bu yöntemler gelişmeye devam ettikçe, bu kronik hastalığın erken teşhisi ve etkili tedavisinde önemli bir rol oynamakta ve diyabet risk tahmininin doğruluğunu ve kesinliğini artırma sözü vermektedir.
Bu yazıda, diyabeti tahmin etmek için geleneksel sınıflandırıcı yöntemlerin nasıl kullanıldığını, bu teknolojinin hastalık teşhisindeki etkilerini ve gelişen bu alanın gelecekteki potansiyelini inceleyeceğiz.

Anahtar Kelimeler: Diyabet, Yapay zeka, Sınıflandırıcılar, Makine Öğrenmesi, Tahmin.

Analysis and Evaluation of Conventional Methods for Diabetes Prediction

Abstract

Diabetes, a chronic disease that affects millions of people worldwide, is characterized by the body's inability to manage blood sugar levels effectively. If left unchecked or not managed properly, this condition can lead to serious consequences such as heart disease, stroke, kidney failure, and even blindness. Due to the interplay of genetic and lifestyle factors, the incidence of diabetes is increasing, positioning it as a significant global health problem requiring urgent attention.
The World Health Organization (WHO) reports that the global prevalence of diabetes has nearly doubled since 1980, rising from 4.7% to 8.5% in the adult population. This increase highlights the urgency and importance of strategies aimed at early diagnosis and effective management of the disease. In the face of such a public health problem, health services seek help from technological developments to combat this epidemic. Among the most promising technological frontiers in healthcare is Machine Learning (ML), a subset of artificial intelligence (AI) that can analyze vast amounts of data, identify patterns and predict outcomes. Machine learning

[*] Sorumlu Yazar:hayriye.tanyildiz@tedas.gov.tr

has the potential to revolutionize diabetes management by providing valuable insights into patient health, informing treatment decisions, and even predicting a person's risk of developing the disease in the future. This technology, if used properly, could change the game in the fight against diabetes. In this context, the use of traditional classifier methods to estimate diabetes risk seems to be a viable and efficient approach. As these methods continue to evolve, they play an important role in the early detection and effective treatment of this chronic disease, promising to increase the accuracy and precision of diabetes risk estimation.
In this article, we will examine how traditional classifier methods are used to predict diabetes, the implications of this technology for disease diagnosis, and the future potential of this evolving field

Keywords: Diabetes, Artificial Intelligence, Classifiers, Machine Learning, Prediction.

1. Giriş

The increasing global prevalence of diabetes indicates an urgent need for advanced diagnostic and predictive tools. Currently, more than 537 million adults worldwide are living with diabetes, and this figure is predicted to increase to 784 million by 2045 [1]. Among the reasons for the increase in the number of diabetes are physical inactivity, unhealthy diet, and excessive stress factors due to urbanization.

The effects of diabetes are enormous and if left untreated, it can lead to serious complications such as kidney failure and blindness. The most important thing about diabetes is that it is often not diagnosed until complications arise. This delay in diagnosis is due to the insidious nature of the disease. Given these circumstances, the ability to predict diabetes risk and facilitate early detection is crucial.

In the last few decades, advances in technology have opened up new ways to predict and diagnose diabetes. Machine Learning (ML), a subset of artificial intelligence, has emerged as a powerful tool in healthcare due to its ability to process large datasets and identify patterns. Thanks to its capacity to include a wide variety of risk factors and to discern the complex relationships between them, it holds great promise in predicting disease risk, including diabetes. This article aims to explore how various classifier techniques can be used for diabetes risk estimation. The goal is to provide insight into how these techniques could potentially save millions of lives and significantly reduce the healthcare burden by enabling the early detection of diabetes.

In this research study conducted on the Pima Indian Diabetes (PID) dataset collection [13], a prediction accuracy of 82% was achieved using the Hidden Naïve Bayes classifier.

In study [14], 67% accuracy rate was obtained by using Random Forest algorithm on Pima dataset.

In this study for diabetes diagnosis [15], they presented an automated diagnostic system for diabetes on Linear Discriminant Analysis (LDA) and Morlet Wavelet Support Vector Machine Classifier (LDA–MWSVM).

In [16], obtained 44.12% accuracy with the Cosine KNN algorithm in their study to find the presence of diabetes using the Intermediate K-NN (K-Nearest Neighbor) and Cosine. The results show that accuracy success grows proportionally as the amount of sampling and the proportions of the training dataset increase.

In this study [14] detailed the investigations of CNN, CNN-LSTM, ConvLSTM and deep 1D convolutional neural network (DCNN) techniques for early diagnosis of diabetes and proposed a SMOTE-based deep LSTM method for diabetes prediction.

In [18], they used various Machine Learning techniques such as SVM, DT, KNN, Random Forest, Logistic Regression and Gradient Boosting and obtained 77 percent accuracy using the RF algorithm.

In the study [20], a framework was proposed for diabetes prediction that outperforms the different Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) methods. The proposed ensembling classifier outperforms the best results of the state-of-the-art results.

In a study [27], A and his colleagues used artificial neural networks to predict whether a person is diabetic or not. They predicted it with 87.3% accuracy.

In [28], In their study with the model they created with three different algorithms: Logistic Regression, Support Vector Machines and Random Forest, a and her colleagues achieved the highest accuracy value of 84% in the Random forest algorithm.

Machine learning techniques were used for the purpose of detecting diabetes in the current research. Thus, different machine learning-based classification algorithms, such as decision tree, SVM, AdaBoost, random forest, gradient boosting machines, KNN, XGBoost, CatBoost, light gradient boosting machine, linear discriminant analysis, Naïve Bayes, stochastic gradient descent, and quadratic discriminant analysis techniques, were employed. Afterward, the performance of the above-mentioned classifiers was assessed concerning precision, sensitivity, specificity, FPR, FDR, FNR, and F1 measures.

The novelty of this work is to apply an automated diabetes prediction to a dataset collected by the National Institute of Diabetes and Digestive and Kidney Diseases using machine learning techniques. In this study, we analyze large-scale machine learning techniques for the first time with approaches and measurement metrics not available in any other recent study.

The remaining part of the current work has the following organization. An explanation of the Material and methods is contained in Part 2. Part 3 describes the Application Results. Part 4 summarizes the conclusion and future research.

2. Materyal ve Metot

In this section, the machine learning methods used in the study to predict diabetes are examined. In this study, 13 different classification methods were used to predict diabetes. The software was developed using the Python programming language and the Colab editor

2.1. Decision Tree

The decision tree represents a popular machine computer, which is capable of analyzing a set of decision management configuration datasets. It is a tree-structured classifier, in which internal nodes refer to a dataset's features, branches refer to the decision rules, and every leaf node refers to the outcome. In decision trees, the start is at the tree's root for predicting a record's class label. The values of the root attribute are compared to the record's attribute. The purpose is to establish a model predicting a target variable's value as a result of learning simple decision rules obtained from data features. The said rules are organized in a tree-like model in which every feature forms a decision node.

2.2. Support Vector Machine (SVM)

The purpose of this algorithm is to provide the most effective division by creating numerous vectors to separate data belonging to two different classes in a linear or non-linear way. This method, which is especially preferred in large data sets, makes it possible to get fast results. In addition, the ability to separate the data in linear or non-linear forms and the ability to find the best option among the infinite decomposition possibilities available has provided high accuracy results [7].

2.3. AdaBoost

The ensemble, which is formed by the combination of individual students and naturally their decisions, is called collective learning. In general, classification success in collective learning applications is higher than in single learning. AdaBoost is among the most used boosting algorithms and was first proposed by Freund and Schapire [19].

2.4. Random Forest

Collective learning is the combination of individual students and naturally their decisions. In general, classification success in collective learning applications is higher than in single learning. Random forest represents a classifier, which includes a number of decision trees in different subsets of the particular dataset and averages the said dataset with the objective of enhancing the prediction accuracy. The concept of ensemble learning, which represents the process of combining multiple classifiers for the solution of a complex problem and enhancing the model's performance, constitutes its basis [2].

2.5. Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) is a limitation by training a group of decision-making tree classifiers iteratively and aiming to optimize over a long period of time and reveal a powerful classifier [11]. Hiding GBM has the potential to provide high accuracy at the limits of generalization. Gradient boosting represents a powerful boosting algorithm combining a number of weak learners into strong ones, in which every novel model is trained using gradient descent with the objective of minimizing the loss function, e.g., the mean square error or cross-entropy of the previous model. At every iteration, the algorithm computes the loss function's gradient on the basis of the present group's estimates, following which it trains a novel weak model for the purpose of minimizing the gradient in question. Afterward, the novel model's predictions are added to the community, and the process is repeated until meeting a stopping criterion.

2.6. The K-Nearest Neighbor (KNN)

The Nearest Neighbor (kNN) algorithm was first proposed in the early 1950s. KNN algorithm draws attention, especially with its low computational cost and complexity. Therefore, it did not gain popularity until computing power became available. One of the supervised learning methods, the k Nearest Neighbor algorithm is a versatile algorithm that can be used both in classification and regression. To define it in its simplest form, the data of an unknown class is compared with other data in the training set and a distance measurement is made. According to the calculated distance, the most optimal class is found for the data that has not yet been assigned to a class [10].

2.7. XGBoost

XGBoost is a high-performance and effective gradient boosting library. Gradient boosting is a machine learning technique that usually combines a set of predictions of simple models (weak learners) such as decision trees. The new model attempts to correct the errors of the previous model, so that it creates a series of models and then combines them to form a result [3].

2.8. CatBoost

CatBoost is a Gradient Boosted Decision Tree (GBDT) algorithm that quickly processes categorical features. Unlike deep learning models, it can achieve effective results without the need for large datasets. This is a high-performance, easy-to-use algorithm that automatically processes categorical data. While traditional GBDT algorithms process categorical features in the preprocessing stage, CatBoost handles these features throughout the training process. Although there are different methods for using categorical features in gradient boosting, these methods may lead to deviations in estimates [4].

2.9. Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) is a type of gradient boosting method, and the term light refers to the lightweight version of this method, which is claimed to make the gradient boosting framework using tree-based learning methods faster, distributed, high-performance and efficient. It has the advantage of being able to process large-scale datasets and offer faster training times [12].

2.10. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a method that allows dividing the *p* features in the *X* data set into two or more real groups, and it ensures that the newly observed units are correctly assigned to the determined classes through the determined differential functions [5].

2.11. Naive Bayes

Naive Bayesian classifier is a simple probability classification that calculates a set of probabilities by quantifying the given dataset's frequency and combination of values [6]. The advantage of the Naive Bayes Classifier is that it can work quickly when applied to large and diverse data.

2.12. Stochastic Gradient Descent (SGD)

The Stochastic Gradient Descent (SGD) algorithm only considers a randomly selected sample, instead of using all the training data, while changing the weight values when classifying. This algorithm, in which a single point is examined, makes it possible to obtain faster results. In alternative terms, the SGD algorithm only processes a randomly selected sample instead of going through the entire training set to adjust the weight values during the classification process. This one-point focus approach allows the algorithm to produce results faster [8].

2.13. Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) is better known for providing classification and size reduction. As the name suggests, QDA is often used as a dimensionality reduction technique and a classifier. It is a variant of linear discriminant analysis (LDA), whereas QDA can only serve as a classifier [9].

3. Aplication Results

3.1. Dataset

This dataset, originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases, has been published online with the objective of diagnostically predicting whether a patient has diabetes, based on certain diagnostic measurements included in the dataset[26]. The dataset is comprised of two classes, labeled as 1 for "Diabetes" and 0 for "Non-Diabetes".

There are 154 records in the dataset. Out of these records, 99 are not diagnosed with diabetes, and 55 are diagnosed with diabetes.

In the partitioning of the dataset, considering the principle that models' discovery and adaptation abilities will increase due to the expansion of the search space with fundamental understanding and maximum resource utilization[29], the dataset has been divided into 80% training data and 20% test data.

Table 1. Dataset Description

Number	Attribute	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mm U/ml)
6	BMI	Body mass index
7	DiabetesPedigreeFunction	Diabetes pedigree function
8	Age	Ager (years)
9	Outcome	Class variable (0 or 1)

. To classify the dataset, we employed a variety of classifiers including SVM, LightGBM, CatBoost, Random Forest, Decision Tree, XGBOOST, Gradient Boosting Machines, Stochastic Gradient Descent, Naïve Bayes, Quadratic Discriminant Analysis, KNN, and Linear Discriminant Analysis. We allocated 80% of the maternal health risk data set for training purposes, reserving the remaining 20% for testing. The data, classified via six distinct classifiers, was analysed to determine the confusion matrix and accuracy ratio.

The first model used in the study to predict maternal risk health is the Decision Tree. The confusion matrix of the Decision Tree method is given in Figure 1.

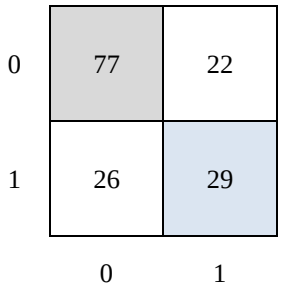

Figure 1. Confusion Matrix of Decision Tree

Meta Parameters used in the Decision Tree classifier is shown in Table 2

Table 2. Meta Paramaters of Decision Tree

criterion	gini
splitter	best
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	None
random_state	1

In Figure 1, the Decision Tree model's confusion matrix was analyzed, yielding an accuracy score of 68.83% when assessing the test data. Out of the 154 data points set aside for testing, the Decision Tree classifier correctly predicted 103 and incorrectly identified 22. When looking at the 99 non-diabetic samples, 77 were accurately categorized, but 22 were misinterpreted as being diabetic. Of the 55 diabetic samples, 26 were incorrectly labeled, while 29 were accurately identified. The performance measurement metrics of the Decision Tree method are presented in Table 15.

The confusion matrix of the SWM Classifier is shown in Figure 2.

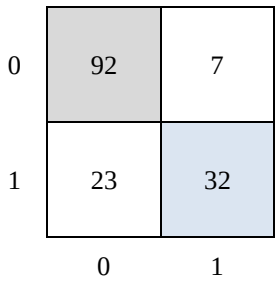

Figure 2. Confusion Matrix of SVM

Meta Parameters used in the SVM classifier is shown in Table 3

Table 3. Meta Paramaters of SVM

C	1.0
kernel	rbf
gama	scale
probability	False
tol	Ie-3
cache_size	200
verbose	False
Max_iter	-1

In Figure 2, the SVM model's confusion matrix was analyzed, yielding an accuracy score of 80.52% when assessing the test data. Out of the 154 data points set aside for testing, the SWM classifier correctly predicted 124 and incorrectly identified 30. When looking at the 99 non-diabetic samples, 92 were accurately categorized, but 22 were misinterpreted as being diabetic. Of the 55 diabetic samples, 23 were incorrectly labeled, while 32 were accurately identified. The performance measurement metrics obtained in the SWM are presented in Table 15.

The confusion matrix of the AdaBoost Classifier is shown in Figure 3.

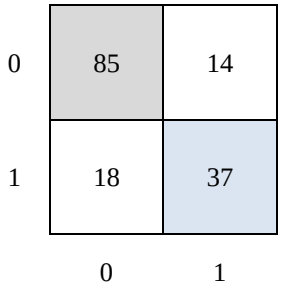

Figure 3. Confusion Matrix of AdaBoost

Meta Parameters used in the AdaBoost classifier is shown in Table 4

Table 4. Meta Paramaters of AdaBoost

C base_estimator	DecisionTreeClassifier(max_depth=1)
n_estimators	50
learning_rate	1.0
algorithm	SAMME.R
random_state	1

In Figure 3, the AdaBoost model's confusion matrix was analyzed, yielding an accuracy score of 79.22% when assessing the test data. Out of the 154 data points set aside for testing, the AdaBoost classifier correctly predicted 122 and incorrectly identified 32. When looking at the 99 non-diabetic samples, 85 were accurately categorized, but 14 were misinterpreted as being diabetic. Of the 55 diabetic samples, 18 were incorrectly labeled, while 37 were accurately identified. The performance measurement metrics obtained in the AdaBoost are presented in Table 15.

The confusion matrix obtained in the Random Forest Classifier is shown in Figure 4.

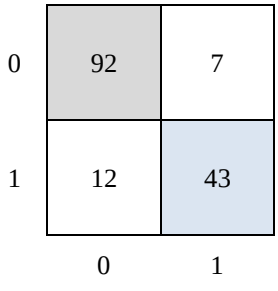

Figure 4. Confusion Matrix of Random Forest

Meta Parameters used in the Random Forest classifier is shown in Table 5

Table 5. Meta Paramaters of Random Forest

n_estimators	100
random_state	1

In Figure 4, the Random Forest model's confusion matrix was analyzed, yielding an accuracy score of 87.66% when assessing the test data. Out of the 154 data points set aside for testing, the Random Forest classifier correctly predicted 136 and incorrectly identified 19. When looking at the 99 non-diabetic samples, 92 were accurately categorized, but 7 were misinterpreted as being diabetic. Of the 55 diabetic samples, 12 were incorrectly labeled, while 43 were accurately identified. The performance evaluation indicators garnered from the Random Forest model are outlined in Table15.

The confusion matrix obtained in the Gradient Boosting Machines classifier is shown in Figure 5.

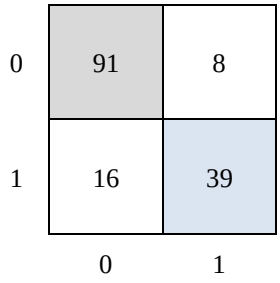

Figure 5. Confusion Matrix of Gradient Boosting Machines

Meta Parameters used in the Gradient Boosting Machines classifier is shown in Table 6

Table 6. Meta Paramaters of SVM

loss	deviance
Learning_rate	0.1
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0

In Figure 6, the Gradient Boosting Machines model's confusion matrix was analyzed, yielding an accuracy score of 84.42% when assessing the test data. Out of the 154 data points set aside for testing, the Gradient Boosting Machines classifier correctly predicted 130 and incorrectly identified 24. When looking at the 99 non-diabetic samples, 91 were accurately categorized, but 8 were misinterpreted as being diabetic. Of the 55 diabetic samples, 16 were incorrectly labeled, while 39 were accurately identified. The performance evaluation indicators garnered from the Gradient Boosting Machines model are outlined in Table 15.

The confusion matrix obtained in the KNN classifier is shown in Figure 6.

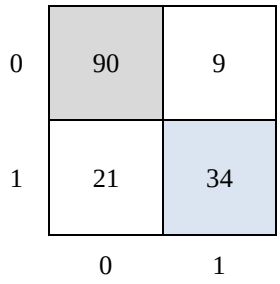

Figure 6. Confusion Matrix of KNN

Meta Parameters used in the KNN classifier is shown in Table 7

Table 7. Meta Paramaters of KNN

n_neighbors	5

In Figure 6, the KNN model's confusion matrix was analyzed, yielding an accuracy score of 79.97% when assessing the test data. Out of the 154 data points set aside for testing, the KNN classifier correctly predicted 124 and incorrectly identified 30. When looking at the 99 non-diabetic samples,90 were accurately categorized, but 9 were misinterpreted as being diabetic. Of the 55 diabetic samples, 21 were incorrectly labeled, while 34 were accurately identified. The performance evaluation indicators garnered from the Gradient Boosting Machines model are outlined in Table 15.

The confusion matrix obtained in the XGBOOST classifier is shown in Figure 7.

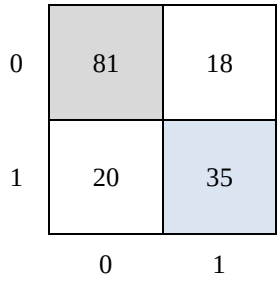

Figure 7. Confusion Matrix of XGBOOST

Meta Parameters used in the XGBOOST classifier is shown in Table 8

Table 8. Meta Paramaters of XGBOOST

max_depth	6
n_estimators	100
learning_rate	0.3
objective	binary:logistic
booster	1 gbtree

In Figure 7, the Xgbost model's confusion matrix was analyzed, yielding an accuracy score of 75.32% when assessing the test data. Out of the 154 data points set aside for testing, the Xgbost classifier correctly predicted 126 and incorrectly identified 30. When looking at the 99 non-diabetic samples, 81 were accurately categorized, but 18 were misinterpreted as being diabetic. Of the 55 diabetic samples, 20 were incorrectly labeled, while 35 were accurately identified. The performance evaluation indicators garnered from the Xgboost model are outlined in Table 15.

The confusion matrix obtained in the Catboost classifier is shown in Figure 8.

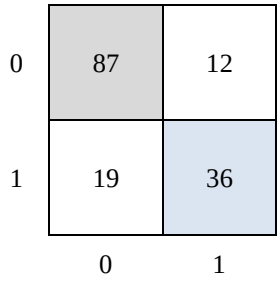

Figure 8 Confusion Matrix of CATBOOST

Meta Parameters used in the CATBOOST classifier is shown in Table 9

Table 9. Meta Paramaters of CATBOOST

iterations	100
depth	6
learning_rate	0.03
l2_leaf_reg	3
booster	254
verbose	500
Od_type	IncToDec

In Figure 8, the Catboost model's confusion matrix was analyzed, yielding an accuracy score of 79.87% when assessing the test data. Out of the 154 data points set aside for testing, the Catboost classifier correctly predicted 133 and incorrectly identified 31. When looking at the 99 non-diabetic samples, 87 were accurately categorized, but 12 were misinterpreted as being diabetic. Of the 55 diabetic samples, 19 were incorrectly labeled, while 36 were accurately identified. The performance evaluation indicators garnered from the Catboost model are outlined in Table 15.

The confusion matrix of the LightGBM Classifier is shown in Figure 10.

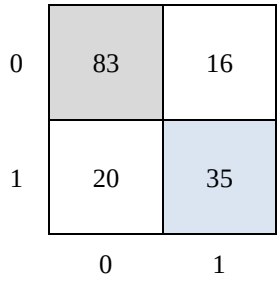

Figure 9 Confusion Matrix of LightGBM

Meta Parameters used in the LightGBM classifier is shown in Table 10

Table 10. Meta Paramaters of LightGBM

boosting_type	gbdt
num_leaves	31
learning_rate	0.1
n_estimators	100
subsample	1.0

In Figure 9, the LightGBM model's confusion matrix was analyzed, yielding an accuracy score of 76.62% when assessing the test data. Out of the 154 data points set aside for testing, the LightGBM classifier correctly predicted 138 and incorrectly identified 36. When looking at the 99 non-diabetic samples, 83 were accurately categorized, but 16 were misinterpreted as being diabetic. Of the 55 diabetic samples, 20were incorrectly labeled, while 35 were accurately identified. The performance evaluation indicators garnered from the LightGBM model are outlined in Table 15.

The confusion matrix of the Naïve Bayes Classifier is shown in Figure 10.

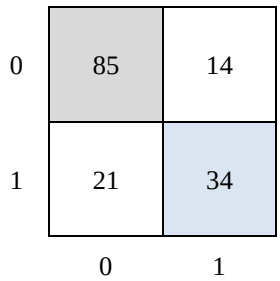

Figure 10. Confusion Matrix of Naïve Bayes

Meta Parameters used in the Naïve Bayes classifier is shown in Table 11

Table 11. Meta Paramaters of Naïve Bayes

priors	None
var_smoothing	1e-9

In Figure 10, the Naïve Bayes model's confusion matrix was analyzed, yielding an accuracy score of 77.27% when assessing the test data. Out of the 154 data points set aside for testing, the Naïve Bayes classifier correctly predicted 129 and incorrectly identified 35. When looking at the 99 non-diabetic samples, 85 were accurately categorized, but 14 were misinterpreted as being diabetic. Of the 55 diabetic samples, 21 were incorrectly labeled, while 34 were accurately identified. The performance evaluation indicators garnered from the Naïve Bayes model are outlined in Table 15.

The confusion matrix of the Linear Discriminant Analysis (LDA) Classifier is shown in Figure 11.

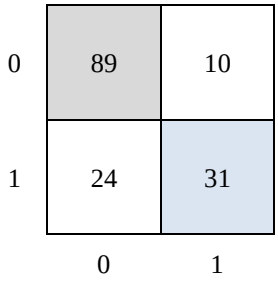

Figure 11. Confusion Matrix of LDA

Meta Parameters used in the LDA classifier is shown in Table 12

Table 12. Meta Paramaters of LDA

priors	None
n_components	min(n_classes - 1, n_features)
store_covariance	True
tol	0.0001
store_covariance	False

In Figure 11, the Linear Discriminant Analysis model's confusion matrix was analyzed, yielding an accuracy score of 77.92% when assessing the test data. Out of the 154 data points set aside for testing, the Linear Discriminant Analysis classifier correctly predicted 120 and incorrectly identified 34. When looking at the 99 non-diabetic samples, 89 were accurately categorized, but 10 were misinterpreted as being diabetic. Of the 55 diabetic samples, 24 were incorrectly labeled, while 31 were accurately identified. The performance evaluation indicators garnered from the Linear Discriminant Analysis model are outlined in Table 15.

The confusion matrix of the Stochastic Gradient Descent (SGD) Classifier is shown in Figure 12.

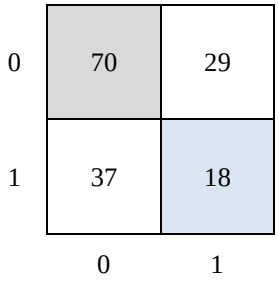

Figure 12. Confusion Matrix of SGD

Meta Parameters used in the SGD classifier is shown in Table 13

Table 13. Meta Paramaters of SGD

penalty	12
reg_param	0
shuffle	True
tol	1e-3

In Figure 12, the Stochastic Gradient Descent model's confusion matrix was analyzed, yielding an accuracy score of 57.14% when assessing the test data. Out of the 154 data points set aside for testing, the Stochastic Gradient Descent Analysis classifier correctly predicted 120 and incorrectly identified 34. When looking at the 99 non-diabetic samples, 70 were accurately categorized, but 29 were misinterpreted as being diabetic. Of the 55 diabetic samples, 37 were incorrectly labeled, while 18 were accurately identified. The performance evaluation indicators garnered from the Linear Discriminant Analysis model are outlined in Table 15.

The confusion matrix of the Quadratic Discriminant Analysis (QDA) Classifier is shown in Figure 13.

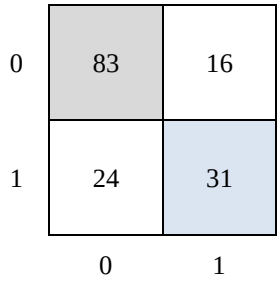

Figure 13. Confusion Matrix of QDA

Meta Parameters used in the ODA classifier is shown in Table 14

Table 14. Meta Paramaters of ODA

priors	None
reg_param	0
store_covariance	True
tol	1.0e-4

In Figure 13, the Quadratic Discriminant Analysis model's confusion matrix was analyzed, yielding an accuracy score of 74.03% when assessing the test data. Out of the 154 data points set aside for testing, the Quadratic Discriminant Analysis classifier correctly predicted 114 and incorrectly identified 40. When looking at the 99 non-diabetic samples, 83 were accurately categorized, but 16 were misinterpreted as being diabetic. Of the 55 diabetic samples, 24 were incorrectly labeled, while 31 were accurately identified. The performance evaluation indicators garnered from the Quadratic Discriminant Analysis model are outlined in Table 15.

Table 16 presents the accuracy results from the six classifiers used in the study.

Table 16 shows the accuracy rates of various classifiers for diabetes prediction. Accuracy metric was used to measure the performance of each classifier on the data set. Accuracy is defined as the ratio of the model's correct predictions to the total predictions.

Looking at Table 16, we see that the Random Forest classifier has the highest accuracy rate of 87.66%. This indicates that the Random Forest model outperforms other classifiers in this particular diabetes prediction task. However, the Gradient Boosting Machines, KNN, and SVM classifiers also perform quite well, with an accuracy rate of over 80%. Gradient Boosting Machines classifier has the second highest accuracy with 84.42, while KNN and SVM are third with 80.52%. On the other hand, the SGD classifier had the lowest accuracy rate of 57.14%, outperforming other classifiers in this task

Table 15. Performance metrics (%).

	Accuracy	Precission	Sensitivity	Specificity	FPR	FDR	FNR	F1
Decision Tree	68.83	77.78	74.76	56.86	43.14	22.22	25.24	76.24
SVM	80.52	92.93	80.00	82.05	17.95	7.07	20	85.98
AdaBoost	79.22	85.86	82.52	72.55	27.45	14.14	17.48	76.24
Random Forest	87.66	92.93	88.46	86.00	14.00	07.07	11.54	90.64
Gradient Boosting M.	84.42	91.92	85.05	82.98	17.02	08.08	14.95	88.35
KNN	80.52	88.89	81.48	76.09	23.91	11.11	18.52	85.02
XGBOOST	75.32	81.82	80.20	66.04	33.96	18.18	19.80	81.00
CATBOOST	79.87	87.88	82.08	75.00	25.00	12.12	17.92	84.88
LightGBM Classifier	76.62	83.84	80.58	68.63	31.37	16.16	19.92	82.18
Naïve Bayes Classifier	77.27	85.86	80.19	70.83	29.17	14.14	19.81	82.93
LDA Classifier	77.92	89.90	78.76	75.61	24.39	10.10	21.24	82.96
SGD Classifier	57.14	70.71	65.42	38.30	61.70	29.29	34.58	67.96
QDA Classifier	74.03	83.84	77.57	65.96	34.04	16.16	22.43	80.58

Table 16. Accuracy rates of classifiers (%)

Decision Tree	LightGBM	CatBoost	Random Forest	Gradient Boosting Machines	KNN
68.83	76.62	79.87	87.66	84.42	80.52
QDA	**SGD**	**Naïve Bayes**	**XGBOOST**	**AdaBoost**	**SVM**
74.03	57.14	77.27	75.32	79.22	80.52
LDA	-	-	-	-	-
77.92	-	-	-	-	-

4. Conclusions

Table 17. Comparison table of accuracy rates achieved by different models

Study Reference	Algorithm/Model Used	Accuracy Rate(%)
[13]	Naïve Bayes	82
[14]	Random Forest	67
[16]	CosineKNN-IntermediateKNN	44.12
[18]	Random Forest	77
[27]	Artificial Neural Networks	87.3
[28]	Random Forest	85
Our Work	**Random Forest**	**87.66**

The performance of models varies widely, from as low as 44.12% accuracy achieved with the Cosine KNN method in study [16] to as high as 87.3% in study [27] using artificial neural networks.

Random Forest seems to be a recurrent algorithm in multiple studies ([14], [18], [28]). It has showcased accuracy rates ranging between 67% to 84%, highlighting its robustness and reliability for this specific dataset.

The study [16], where the accuracy was found to be around 44.12% with the Cosine KNN algorithm, indicates that not all models are suitable for every type of data.

In this study, a problem that will help in the field of health care is discussed with different machine learning approaches. It is aimed to predict diabetes by analyzing using computer-based classifiers. Among the studied models, 87,66% accuracy value was obtained in Random Forest classifier. Considering the results of the models studied, it has been observed as a result of the results obtained that the users will be very helpful in detecting the diabetes risk at a very early stage.

In future studies, it is aimed to evaluate and analyze vital diseases such as diabetes with such approaches. With richer datasets, similar diseases will be evaluated and compared with deep learning approaches as well as machine learning approaches.

5. Acknowledge

The present paper does not include any research with human participants conducted by any of the authors.

References

[1] IDF Diabetes Atlas, “Diabetes around the world in 2021”, Accessed 13.09.2023, https://diabetesatlas.org/.
[2] Kalaycı, T. E. (2018). Kimlik hırsızı web sitelerinin sınıflandırılması için makine öğrenmesi yöntemlerinin karşılaştırılması. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24(5), 870-878.
[3] Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794). ACM. doi: 10.1145/2939672.2939785.
[4] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516.
[5] Karaıbrahımoglu, A. , Kara, Ü. , Kılıçoğlu, Ö. & Kara, Y. (2023). Prediction of absorption dose of radiation on Thorax CT imaging in geriatric patients with COVID-19 by classification algorithms . European Mechanical Science , 7 (2) , 89-98 . Retrieved from https://dergipark.org.tr/en/pub/ems/issue/76070/1262875.
[6] Saritas, M.M., Yasar, A. (2019) Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. International Journal of Intelligent Systems and Applications in Engineering 7(2), 88-91. (https://doi.org/10.18201// ijisae.2019252786).
[7] Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2010. A Practical Guide to Support Vector ClassificationDepartment of Computer Science, National Taiwan University, Taipei, Taiwan16.
[8] Chandrashekhar, A.M., Raghuveer, K. (2014). Amalgamation of K-means Clustering Algorithm with Standard MLP and SVM Based Neural Networks to Implement Network Intrusion Detection System. In: Kumar Kundu, M., Mohapatra, D., Konar, A., Chakraborty, A. (eds) Advanced Computing, Networking and Informatics- Volume 2. Smart Innovation, Systems and Technologies, vol 28. Springer, Cham. https://doi.org/10.1007/978-3-319-07350-7_31
[9] Chand, S., & Vishwakarma, V. P. (2022). Application of quadratic discriminant analysis algorithm for the classification of acute leukemia using microscopic image data. *Adv. Appl. Math. Sci.*, *21*, 2737-2750.
[10] Hu, L.Y., Huang, M.W., Ke, S.W., Tsai, C.F., (2016), "The Distance Function Effect on kNearest Neighbor Classification for Medical Datasets", Springer Plus, 5(1), 1-9.
[11]Friedman J., Greedy function approximation: A gradient boosting machine, Annals of Statistics, 29, 1189-1232, 2001.
[12]Sai, M.J., Chettri, P., Panigrahi, R. *et al.* An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. *Int J Comput Intell Syst* **16**, 14 (2023). https://doi.org/10.1007/s44196-023-00184-y
[13] Al-Hameli, B., Alsewari, A., Basurra, S., Bhogal, J. & Ali, M. (2023). Diabetes disease prediction system using HNB classifier based on discretization method. *Journal of Integrative Bioinformatics*, *20*(1), 20210037. https://doi.org/10.1515/jib-2021-0037.
[14]Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. Front Genet. 2018 Nov 6;9:515. doi: 10.3389/fgene.2018.00515. PMID: 30459809; PMCID: PMC6232260.
[15]Çalişir, D., & Doğantekin, E. (2011). An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. *Expert Systems with Applications*, *38*(7), 8311-8315.
[16] Nirupama S, & Jenila Rani D. (2022). Analysis And Comparison Of Diabetic Prediction Using Medium KNN Classifier And Cosine KNN Classifier. *Journal of Pharmaceutical Negative Results*, 386–394. https://doi.org/10.47750/pnr.2022.13.S04.043.
[17]Alex SA, Jhanjhi N, Humayun M, Ibrahim AO, Abulfaraj AW. Deep LSTM Model for Diabetes Prediction with Class Balancing by SMOTE. *Electronics*. 2022; 11(17):2737. https://doi.org/10.3390/electronics11172737
[18] Nahzat, S. & Yağanoğlu, M. (2021). Diabetes Prediction Using Machine Learning Classification Algorithms. European Journal of Science and Technology, (24), 53-59.

[19] Bulut, F. (2016). Determining Heart Attack Risk Ration Through AdaBoost/AdaBoost ile Kalp Krizi Risk Tespiti. *Celal Bayar University Journal of Science*, *12*(3), 459-472.

[20] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[21]Gayathri, S., Gopi, V.P. & Palanisamy, P. Diabetic retinopathy classification based on multipath CNN and machine learning classifiers. Phys Eng Sci Med 44, 639–653 (2021). https://doi.org/10.1007/s13246-021-01012-3

[22] Sharma, T., Shah, M. A comprehensive review of machine learning techniques on diabetes detection. Vis. Comput. Ind. Biomed. Art 4, 30 (2021). https://doi.org/10.1186/s42492-021-00097-7

[23] T. Gupta, M. R. A T, R. C and R. Kumar M, "Diabetes Prediction using different Machine Learning Classifiers," 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), Vellore, India, 2023, pp. 1-5, doi: 10.1109/ViTECoN58111.2023.10157531.

[24]Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(1-2), 1-10. https://doi.org/10.1049/htl2.12039

[25]Puneeth N. Thotad, Geeta R. Bharamagoudar, Basavaraj S. Anami, Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, Volume 17, Issue 1, 2023, 102690, ISSN 1871-4021, https://doi.org/10.1016/j.dsx.2022.102690.

[26]https://www.kaggle.com/datasets/mathchi/diabetes-data-set/code

[27]El-Jerjawi, N.S., & Abu-Naser, S.S. (2018). Diabetes Prediction Using Artificial Neural Network.

[28]D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, BC, Canada, 2018, pp. 924-928, doi: 10.1109/IEMCON.2018.8614871.

[29]Gökalp, O. (2021). Performance evaluation of Heuristic and Metaheuristic Algorithms for Independent and Static Task Scheduling in Cloud Computing. 2021 29th Signal Processing and Communications Applications Conference (SIU), 1-4.

[30]Tanyıldız, H. & Batur Şahin, C. (2023). Transfer Learning for Detection of Casting Defects Model In Scope of Industrial 4.0 . Türk Doğa ve Fen Dergisi , 12 (3) , 45-51 . DOI: 10.46810/tdfd.1236584.

[31]Şahin, C.B. (2023). Semantic-based vulnerability detection by functional connectivity of gated graph sequence neural networks. Soft Comput 27, 5703–5719 . https://doi.org/10.1007/s00500-022-07777-3.

[32]C. B. Şahin, (2021).DCW-RNN: Improving Class Level Metrics for Software Vulnerability Detection Using Artificial Immune System with Clock-Work Recurrent Neural Network," 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Kocaeli, Turkey, pp. 1-8, doi: 10.1109/INISTA52262.2021.9548609.