AITA
ARTIFICIAL INTELLIGENCE
THEORY AND APPLICATIONS

■ RESEARCH ARTICLE

Diabetes Risk Prediction with Machine Learning Models

Gözde Özsezer [a,b†], **Gülengül Mermer** [b]

[a] Institute of Health Sciences, Ege University, İzmir, Turkey
[b] Department of Public Health Nursing, Nursing Faculty, Ege University, İzmir, Turkey
† gozde.ozsezer@ege.edu.tr, corresponding author

RECEIVED JUNE 8, 2022
ACCEPTED AUGUST 24, 2022

CITATION Özsezer, G., & Mermer, G. (2022). Diabetes Risk Prediction with Machine Learning Models. *Artificial Intelligence Theory and Applications, 2*(2), 1-9.

Abstract

Diabetes mellitus (DM) is one of the most common chronic diseases worldwide, which is a major public health problem. The aim of this study is to predict DM risk with machine learning (ML) models using available data. In the analytical study, the “Diabetes Health Indicators Dataset” consisting of 253680 data and 21 variables collected annually by the CDC was used. KNN’s accuracy was 0.74, precision 0.31, recall 0.55, F1 score 0.39; Logistic regression’s accuracy was 0.72; precision 0.33, recall 0.74, F1 score 0.46; Decision tree’s was accuracy 0.84, precision 0.54 recall 0.15, F1 score 0.24; Random forest’s accuracy was 0.84, precision 0.56, recall 0.16, F1 score 0.25; Naive bayes's accuracy was 0.84, precision 0.52, recall 0.19, F1 score 0.28. In this study, ML algorithms were used for DM risk estimation. According to the experimental results, when the data set is divided into random training (80%) and testing (20%), the accuracy values of random forest and decision tree algorithms are very close to each other (RF: 0.848, DT: 0.847). Therefore, it can be said that the two best algorithms for diabetes risk estimation are random forest and decision tree.

Keywords: diabetes, risk, prediction, machine learning, artificial intelligence

1. Introduction

Diabetes Mellitus (DM) is a broad-spectrum chronic metabolic disease that occurs as a result of the effect of insulin metabolism, requires continuous medical care, and affects quality of life. [1-3]. It is clinically examined in 4 groups: Type 1 DM, Type 2 DM, Gestational DM and Secondary DM [4,5]. Type 1 DM is an insulin-dependent diabetes mellitus that occurs with absolute insulin deficiency as a result of the destruction of beta cells, which are responsible for insulin production in the pancreas. It is usually seen under the age of 35. The most common age group is 10-15 [4]. Type 2 DM is the most common and predominant type of diabetes in the world. It constitutes approximately 90% of all diabetes patients. [3-5]. Type 2 DM occurs as a result of insufficient insulin produced by the pancreas or the use of insufficient insulin. It is usually seen in individuals aged 40 and over [4-9]. Stress, sedentary life, irregular and unbalanced diet, genetics, being overweight and not exercising can be counted among the factors that accelerate the formation of type 2 DM [8-9].

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN : 978-605-69730-2-4 © 2022 University of Bakırçay

DM has become one of the biggest global epidemics of the 21st century. It is one of the most common non-communicable diseases worldwide [6]. The prevalence of DM is increasing rapidly in both developed and developing societies due to population growth, aging, urbanization, rapid changes in lifestyle, obesity and increase in physical inactivity. It is estimated that by 2045, 700 million people worldwide will have diabetes [7]. According to the data of the International Diabetes Federation (IDF), the prevalence of DM in 2021 was found to be 10.5% worldwide and 14.5% in Turkey [8]. According to the course of hyperglycemia, all patients with DM are at risk of developing acute and chronic complications affecting the whole body, especially cardiovascular, nervous system (neuropathy), eye (retinopathy) and kidney (nephropathy). Chronic complications are insidious and gradual complications that develop over time when hyperglycemia is not controlled [9].

DM is a major public health problem associated with increased mortality and morbidity. It has been declared a global epidemic by the World Health Organization [10]. It constitutes a large proportion of the resources allocated to health services worldwide [11]. DM development can be prevented or delayed with lifestyle changes studies, drug intervention studies, and early interventions [3,11]. Therefore, determining the population at high risk for DM will facilitate access to preventive health services.

Today, many multidisciplinary studies are carried out to support the prevention and treatment of DM. These studies have become a research priority globally [10,12]. Therefore, multidisciplinary follow-up, prevention and control of DM becomes more and more important [10,12]. Since each discipline has its own potential and added value, the combination of these is thought to be a solution [12].

Many recent risk scorings are used to estimate DM risk [13-17]. Some models are based on non-laboratory clinical variables (non-invasive), while some models include biological variables (invasive). Although invasive risk scoring is more successful, non-invasive risk scoring also predicts DM risk with high success and cost-effectiveness [18-24]. Data sets are created for early diagnosis of DM and for taking necessary precautions, evaluating and processing the symptoms of DM (polyphagia, polydipsia, polyuria, sudden weight loss, obesity, etc.) in digital environments. It is very important to develop systems that can help health professionals for the early diagnosis of diabetes from symptom, body mass index, age, etc. data. It is expected that artificial intelligence will lead to quite radical changes in health sciences. Artificial intelligence and machine learning (ML) give very successful results in diagnosis and diagnosis based on existing data. ML is artificial intelligence applications that can learn and perceive through data [24,26].

ML has been used successfully in many situations where it is difficult or not possible to use traditional algorithms to accomplish any task. The most important advantage of ML is its ability to make consistent and high-performance predictions using complex and non-linear relationships between features. [26]. Therefore, they can describe complex relationships that are not directly visible to humans.

The aim of this study is to predict diabetes risk with machine learning models.

2. Material and Method

2.1. Dataset Information

In the study, the Diabetes Health Indicators Dataset, which is collected annually by the Centers for Disease Control and Prevention (CDC) in the form of a telephone

questionnaire, was used. The survey collects responses from more than 400,000 individuals each year on health-related risk behaviors, chronic health conditions, and use of preventative services. There are 3 individual files in the dataset. Each file has 21 feature variables. These characteristics are either questions asked directly to the participants or variables calculated based on individual participant responses. The features definitions of the variables are as in Table 1:

Table 1. Features definitions

Features	Label	Values
Diabetes_012	Presence of diabetes	0.no diabetes, 1.prediabetes, 2.diabetes
HighBP	High Blood Pressure	0.no high BP, 1.high BP
HighChol	High Cholesterol	0.no high cholesterol, 1.high cholesterol
CholCheck	Cholesterol Check	0.no cholesterol check in 5 years, 1.yes cholesterol check in 5 years
BMI	Body Mass Index	
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	0.no, 1.yes
Stroke	(Ever told) you had a stroke.	0.no, 1.yes
HeartDiseaseorAttack	Coronary heart disease (CHD) or myocardial infarction (MI)	0.no, 1.yes
PhysActivity	Physical activity in past 30 days - not including job	0.no, 1.yes
Fruits	Consume Fruit 1 or more times per day	0.no, 1.yes
Veggies	Consume Vegetables 1 or more times per day	0.no, 1.yes
HvyAlcoholConsum	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)	0.no, 1.yes
AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc	0.no, 1.yes
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?	0.no, 1.yes
GenHlth	Would you say that in general your health is: scale 1-5	1.excellent, 2.very good, 3.good, 4.fair, 5.poor
MentHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how	
PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30	
DiffWalk	Do you have serious difficulty walking or climbing stairs?	0.no, 1.yes
Sex		0.female, 1.male
Age	13-level age category	1.18-24 9.60-64, 13.80 or older
Education	Education level scale 1-6	1.Never attended school or only kindergarten, 2.Grades 1 through 8 (Elementary), 3.Grades 9 through 11 (Some high school), 4.Grade 12 or GED (High school graduate), 5.College 1 year to 3 years (Some college or technical school), 6.College 4 years or more (College graduate)
Income	Income scale scale 1-8	1.less than $10,000, 5.less than $35,000, 8.$75,000 or more

The descriptions of the files are as follows:

- ***Diabetes 012 health indicators:*** The clean dataset consists of 253680 survey responses. There is a class imbalance in this dataset.
- ***Diabetes binary 5050split health indicators:*** The clean dataset consists of 70692 survey responses. This dataset is balanced.
- ***Diabetes binary health indicators:*** The clean dataset consists of 253680 survey responses. This dataset is not balanced.

The open access dataset was retrieved from Kaggle on March 5, 2022. Data analysis was done with Phyton 3.0 programming language using numpy, pandas, matplotlib, seaborn, sciktlearn, imblearn libraries. With data pre-processing, outliers and missing data were removed. Clean data were classified with ML models. 80% of the data were randomly separated as training data and 20% as test data.

2.2. Machine Learning Classification Methods

In the study, ML models were applied to predict diabetes in the early stages. These models are K-Nearest Neighbor (KNN), Logistic regression, Decision tree, Random forest ve Naive Bayes. The prediction rate of the models was evaluated with accuracy, precision, recall and F1 Score. The data set was randomly split into training (80%) and testing (20%). The methodology of the research is shown in Figure 1.

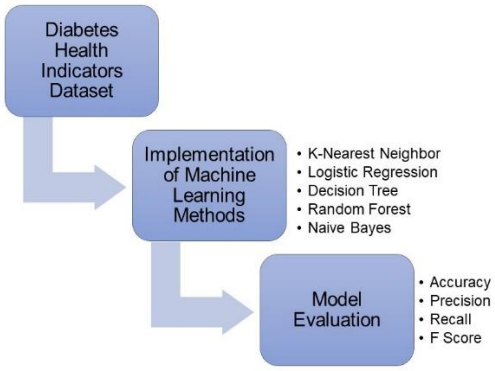

Figure 1. The Flowchart of the Study

2.2.1. K-Nearest Neighbor

K nearest neighbor is a simple yet effective machine learning algorithm. Training data is represented in a graph and an assumption that examples of the same classes will be closely positioned. When predicting the label of an instance, position of that instance at graph is determined by using its features and the k neighbors that are closest to that point is found. Labels of these neighbors are considered and prediction of the model is returned as the most seen label among neighbors [27].

2.2.2. Logistic Regression

Logistic regression is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry [28].

2.2.3. Decision Tree

Decision tree is a machine learning method that visualizes how the created model predicts data. It builds a tree in which nodes of the tree represent features, branches represent which direction must be taken after each node and leaves represent predictions. Classes of given data can be predicted by traversing from root to leaves by

choosing regarding branches. Decision tree also shows the importance of features, the most important and elective feature takes place at the root node. In the presented study, models have been created by using Gini Information Gain with two-level pruning settings [29].

2.2.4. Random Forest

When a part of the decision tree is built incorrectly, it causes the model to make false predictions. Random forest is a ML technique that aims to solve this overfitting issue. In this approach, predictions of several randomly created decision trees are combined and the most voted label is returned as a label of given data [30].

2.2.5. Naïve Bayes

Naive Bayes is a machine learning algorithm that is based on Bayes theorem. It makes an assumption that all attributes are independent so it does not produce good results when the dataset size is large and it has a lot of features [31].

2.3. Model Performance Comparison Metrics

Many different criteria are used in the performance comparison of ML models. In this study, different performance comparisons such as accuracy, precision, recall and F1 score were calculated.

2.3.1. Accuracy

The accuracy of a test is its power to accurately distinguish between sick and healthy individuals [32]. When calculating the accuracy of the diagnostic test, the rate of true positive and true negative is calculated for all patients and healthy individuals. The accuracy value takes a value between 0 and 1. In the formula given with (1), it is shortened as TP: True positive, TN: True negative, FP: False positive, FN: False negative.

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

2.3.2. Precision

The precision value is calculated when the positive predictive value of the diagnostic test or the positive predictions are actually positive (2). In other words, it is defined as the probability that an individual with a positive diagnostic test result will become ill [33, 34].

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

2.3.3. Recall

Recall estimates how many True Positives the model has captured (3). By the same logic, when False Negative has a huge cost, Recall is the model metric used to select the best model [34].

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

2.3.4. F1- Score

Depending on the problem being attempted to solve, in most cases a higher priority can be assigned to maximize precision or recall. However, there is a simpler statistic that takes into account both precision and recall, and attempts are made to maximize this number to improve the model. The F1 score is essentially a statistic that is the harmonic mean of precision and recall [35]. The formula for the F1 score is entirely dependent on precision and recall (4).

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \qquad 4$$

Confusion matrix is an ***nxn*** matrix that n denotes the number of labels of a given dataset [36]. Each row represents actual labels and each column represents predicted labels. Confusion matrix shows the performance of the model as illustrated in Figure 2.

	Actually Positive (1)	**Actually Negative (0)**
Predictied Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 2. Standard Form of Confusion Matrix

3. Experimental Results

13.9% of the participants have DM. The visualization of the features is shown in Figure 3 and the correlation matrix in Figure 4.

Figure 3. Distribution of attributes

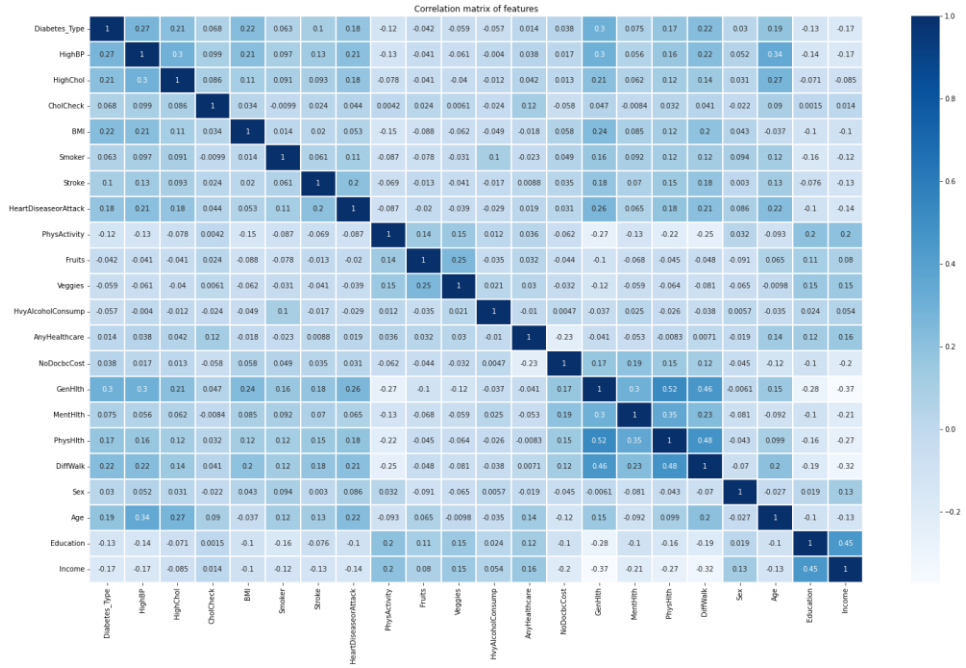

Figure 4. Correlation matrix

Accuracy, precision, recall and F1 score performance values of ML models were compared (Table 2).

Table 2. Comparative performance of ML models

Model	Accuracy	Precision	Recall	F1 score
KNN	0.742663	0.313909	0.551049	0.399971
Logistic regression	0.724925	0.332380	0.749332	0.460498
Decision tree	**0.847791**	**0.549535**	**0.157886**	**0.245297**
Random forest	**0.848875**	**0.560582**	**0.163705**	**0.253408**
Naïve bayes	0.790189	0.369035	0.477905	0.416473

When diabetes classification performances of ML models are compared, it is seen that the most successful methods are random forest and decision tree. The confusion matrix in classification of random forest and decision tree models, which are successful methods in classifying diabetes, is given in Figure 5.

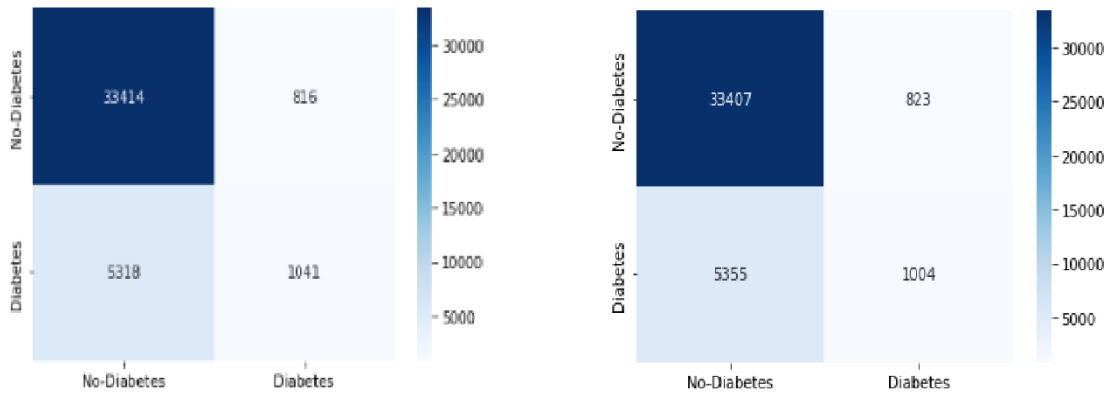

Figure 5. Confusion matrix for random forest and decision tree classification

4. Conclusion

ML methods have been used a lot lately for early diagnosis and planning in the field of health. ML models are very useful, especially in costly chronic diseases such as diabetes. While diabetes is a major public health problem, early detection is important. Individuals from different age groups are at risk of DM. In this study, it is emphasized that early detection of diabetes and clues for early diagnosis are very important. In this study, the predictive values of ML models for early prediction of diabetes risk were compared. According to the experimental results, when the data set is divided into random training (80%) and testing (20%), the accuracy values of random forest and decision tree algorithms are very close to each other (RF: 0.848, DT: 0.847). Therefore, it can be said that the two best algorithms for diabetes risk estimation are random forest and decision tree.

References

[1] World Health Organisation (WHO). 2016. "Global report on diabetes" https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf Accessed: 17 December 2021.

[2] Türkiye Endokrinoloji ve Metabolizma Derneği (TEMD). Diabetes Mellitus Çalışma ve Eğitim Grubu. "Diabetes Mellitus ve Komplikasyonlarının Tanı, Tedavi ve İzlem Klavuzu 2019". https://temd.org.tr/admin/uploads/tbl_kilavuz/20190819095854-2019tbl_kilavuzb48da47363.pdf Accessed: 27 December 2021.

[3] Tekir, O., Çevik, C., Kaymak, G. Ö., & Kaya, A. (2021). The Effect of Diabetes Symptoms on Quality of Life in Individuals with Type 2 Diabetes. Acta Endocrinologica (Bucharest), 17(2), 186.

[4] TÜRKDİAB (2019). Diyabet Tanı ve Tedavi Rehberi. Güncellenmiş 9. Baskı. Armoni Nüans Baskı Sanatları A.Ş. İstanbul, s. 16.

[5] World Health Organization (2019). Classification of Diabetes Mellitus 2019. ISBN: 9789241515702.

[6] Guo, Y., Zhao, J., Wang, H., Liu, S., Huang, T., & Chang, G. (2020). Metabolic disorder-related hypertension. In Secondary hypertension (pp. 507-545). Springer, Singapore.

[7] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... & IDF Diabetes Atlas Committee. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice, 157, 107843.

[8] International Diabetes Federation (IDF). "IDF Diabetes Atlas 10th Edition 2021". https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf Son erişim tarihi: 27 Aralık 2021.

[9] Ulusal Diyabet Konsensus Grubu. "TÜRKDİAB Diyabet Tanı ve Tedavi Rehberi 2019". https://www.turkdiab.org/admin/PICS/files/Diyabet_Tani_ve_Tedavi_Rehberi_2019.pdf Accessed: 27 Aralık 2021.

[10] World Health Organization (2016). Global Report on Diabetes, https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=1C047E5A6F657E8A51DB41D7512B089E?sequence=1 Accessed 24 May 2022.

[11] Diabetes Prevention Program Research Group. (2009). 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. The Lancet, 374(9702), 1677-1686.

[12] Özsezer Kaymak, G., & Tekir Ö. (2020). Diyabet Bakımında Yapay Zeka Kullanımı. Eds. (B. Tunçsiper, F. Taşpınar, Ö. Erkin Geyiktepe). Sağlık Bilimlerinde Multisipliner Yaklaşımlar 2. P.393-410.

[13] Lindstrom, J., & Tuomilehto, J. (2003). The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes care, 26(3), 725-731.

[14] Chen, L., Magliano, D. J., Balkau, B., Colagiuri, S., Zimmet, P. Z., Tonkin, A. M., ... & Shaw, J. E. (2010). AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Medical Journal of Australia, 192(4), 197-202.

[15] Balkau, B., Lange, C., Fezeu, L., Tichet, J., de Lauzon-Guillain, B., Czernichow, S., ... & Eschwège, E. (2008). Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). Diabetes care, 31(10), 2056-2061.

[16] Rosella, L. C., Manuel, D. G., Burchill, C., & Stukel, T. A. (2011). A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). Journal of Epidemiology & Community Health, 65(7), 613-620.

[17] Hippisley-Cox, J., Coupland, C., Robson, J., Sheikh, A., & Brindle, P. (2009). Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. Bmj, 338.

[18] Ergün, Ö. N., & Ilhan, H. O. (2021). Early Stage Diabetes Prediction Using Machine Learning Methods. Avrupa Bilim ve Teknoloji Dergisi, (29), 52-57.

[19] Bilgin, G. (2021). Makine öğrenmesi algoritmaları kullanarak erken dönemde diyabet hastalığı riskinin araştırılması. Journal of Intelligent Systems: Theory and Applications, 4(1), 55-64.

[20] Özkan, Y., Yürekli, B. S., & Suner, A. Diyabet tanısının tahminlenmesinde denetimli makine öğrenme algoritmalarının performans karşılaştırması. Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 12(1), 211-226.

[21] Cihan, P., & Coşkun, H. Diyabet Tahmini için Makine Öğrenmesi Modellerinin Performans Karşılaştırılması. 2021 29th Signal Processing and Communications Applications Conference (SIU).

[22] Akyol, K., & Karaci, A. Diyabet Hastalığının Erken Aşamada Tahmin Edilmesi İçin Makine Öğrenme Algoritmalarının Performanslarının Karşılaştırılması. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 9(6), 123-134.

[23] Harman, G. (2021). Destek Vektör Makineleri ve Naive Bayes Sınıflandırma Algoritmalarını Kullanarak Diabetes Mellitus Tahmini. Avrupa Bilim ve Teknoloji Dergisi, (32), 7-13.

[24] İlyas, Ö. (2020). Uzun Kısa Dönem Bellek Ağlarını Kullanarak Erken Aşama Diyabet Tahmini. Mühendislik Bilimleri ve Araştırmaları Dergisi, 2(2), 50-57.

[25] Sousa Lima, W., Souto, E., El-Khatib, K., Jalali, R., & Gama, J. (2019). Human activity recognition using inertial sensors in a smartphone: An overview. Sensors, 19(14), 3213.

[26] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

[27] Kramer, O. (2013). K-nearest neighbors. In Dimensionality reduction with unsupervised nearest neighbors (pp. 13-23). Springer, Berlin, Heidelberg.

[28] Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1), 12-18.

[29] Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 3(2), 334-337.

[30] Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

[31] Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15, 713-714.

[32] Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: simple definition and calculation of accuracy, sensitivity and specificity. Archives of Academic Emergency Medicine (Emergency), 3(2), 48-49.

[33] Koenig, I. R., Fuchs, O., Hansen, G., von Mutius, E., & Kopp, M. V. (2017). What is precision medicine?. European respiratory journal, 50(4).

[34] Torgo, L., & Ribeiro, R. (2009, October). Precision and recall for regression. In International Conference on Discovery Science (pp. 332-346). Springer, Berlin, Heidelberg.

[35] Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. arXiv preprint arXiv:1402.1892.

[36] Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. MAICS, 710, 120-127.

Acknowledgement

This article is an expanded and revised version of the oral presentation presented at the "III. International Congress of Artificial Intelligence in Health" held in Izmir on 11-13 May 2022. We thank the editors and referees for their contributions during the review and evaluation phase of the article.