EUROPEAN MECHANICAL SCIENCE

Research Article

e-ISSN: 2587-1110

Performance comparison machine learning algorithms in diabetes disease prediction

Aslı Göde[1*], **Adnan Kalkan**[1]

[1]Department of Management Information Systems/Burdur Mehmet Akif Ersoy University, Türkiye

Orcid: A. Göde (0000-0001-7785-6200), A. Kalkan (0000-0002-2270-4100)

Abstract: Machine learning has been widely used in the field of medicine with the developing technology in recent years. Machine learning is a field that is also used in the diagnosis of diabetes and helps experts make decisions. Diabetes is a lifelong disease that is common worldwide and in our country. The main purpose of this study is to diagnose diabetes early using different machine learning classification algorithms. Another purpose of the study is to compare the success of the machine learning models used. Early diagnosis of diabetes allows to lead a healthy and normal life. In this context, it has been tried to diagnose diabetes early by using the machine learning techniques Decision Tree, Random Forests, K-Nearest Neighbor and Support Vector Machines classifiers on the Pima Indians Diabetes dataset. The dataset includes 9 features and 768 samples. Success evaluation of classifiers was made using Accuracy, Precision, Recall, F1-Score and AUC metrics. Random Forests gave the best results with 80 percent accuracy. This paper is to examine the association of different machine learning techniques usage, diabetes data diagnostic capabilities, diagnosis of diabetes in women diabetes patients and comparison of performances for machine learning techniques. Implications for theory and practice have been discussed. In this study, comparisons were made using different algorithms from the classification algorithms used in the literature and contributed to the literature in this field.

Keywords: Diabetes Disease; Machine Learning; Classification; Random Forests

1. Introduction

In today's world, diabetes mellitus, also known as diabetes among people, is accepted as one of the most important health problems. Diabetes is a disease that is frequently seen throughout the world and in our country, and the number of patients is increasing day by day [1]. It is known that if the necessary precautions are not taken, diabetes will increase day by day and it will bring other diseases. Diabetes is a lifelong disease that occurs when the body does not use the insulin hormone produced by the pancreas or cannot produce enough insulin hormone. Diabetes is the presence of sugar in the urine, which should not contain sugar, and the level of glucose (sugar) in the blood rises above the required value. Diabetes risk is a big problem in our country. According to the results of the research conducted in Turkey in 2019, 3 million 600 thousand people in our country have diabetes. However, 1 million 200 thousand of these patients have not been diagnosed yet. People do not consult a doctor, thinking that the symptoms of weakness, debility, thirst and hunger are caused by fatigue or stress. This poses a great risk of diabetes. Early diagnosis of diabetes ensures a normal and healthy life [2].

Machine Learning (ML) is a sub-field of artificial intelligence that can make decisions and use it in solving another problem by creating a model from the knowledge and experience that a computer has obtained from past information [3]. Machine learning is a computer science that aims to learn computers through experience [4]. Machine learning combines elements from statistics, understanding relationships from data, with elements from computer science, developing algorithms to manage data [5]. The more data is used in machine learning, the better machine learning works [6]. Machine learning algorithms are grouped as supervised learning, unsupervised learning, semi-supervised learning and reinforced learning [7]. Machine learning is applied in wide variety of fields namely: virtual personal assistants (like Apple-Siri), pattern recognition, computer games, natural language processing, data mining, traffic prediction, robotics, online transportation network (Ola Cabs), online fraud prediction, product recommendation, share market prediction, medical diagnosis (e.g. diabetes, cancer, tumor), crime prediction through video surveillance system, agriculture advisory, search engine result refining (e.g. Google search engine), social media services, BoTs, E-mail spam filtering [8].

In this study; It is aimed to diagnose diabetes disease

* Corresponding author.
Email: agode@mehmetakif.edu.tr

European Mechanical Science (2023), 7(3): 178-183
https://doi.org/10.26701/ems.1335503
Received: July 31, 2023 — Accepted: August 23, 2023

early, which is of great importance for doctors and diabetes patients. Early diagnosis of diabetes gives people the chance for a normal and healthy life. People can lead a normal life with treatment methods such as physical activities, nutrition plan and medication. Diabetes can lead to serious diseases if not treated early. These diseases are nervous disease, kidney disease and cardiovascular diseases. For this reason, a study was conducted for the diagnosis of diabetes. Decision Tree, Random Forest, K-Nearest Neighbour and Support Vector Machines from machine learning methods were used as the working model. The aim of the study was to help doctors and patients diagnose diabetes early and reduce the number of patients. At the same time, the performances of machine learning algorithms were compared. Thus, it is aimed to contribute to the literature.

2. Literature Review

Looking at the literature, different machine learning techniques have been used to diagnose diabetes. Faruque et al. [9] compared Support Vector Machines, Naive Bayes, Decision Tree and K-nearest neighbor algorithms to diagnose diabetes in adults and revealed that the decision tree model gave the best results. Haq et al. [10] used the clinical dataset to diagnose diabetes with Decision Tree, Adaboost and Random Forest algorithms. They concluded that the best classification model in the model comparison is the decision tree. Dritsas and Trigka [11] used algorithms such as Naive Bayes, Support Vector Machines, Logistic Regression, Artificial Neural Networks, Adaboost, K-Nearest Neighbor and Random Forest to diagnose type 2 diabetes. As a result of the study, they revealed that the best classification model is the K-Nearest Neighbor and Random Forest. Khanam and Foo [12] used machine learning and Artificial Neural Network methods to diagnose diabetes. They found that Logistic Regression and Support Vector Machines from Machine Learning models achieved the best results. In the Artificial Neural Network method, they reached an accuracy of 88.6%. Ayon and Islam [13] used deep learning methods to diagnose diabetes. As a result of the study; 5-fold cross-validation has achieved an accuracy rate of 98.35% and 10-fold cross-validation has achieved an accuracy rate of 97.11%. Baser et al. [14], using Machine Learning Techniques (Decision Tree, K-Nearest Neighbor, Logistic Regression, Naive Bayes, and Random Forest) classified diabetes patient data from 130 hospitals in the USA. As a result of the study, they revealed that the Random Forest model made the best classification. Er and Isık [15] used convolutional Neural Network and long-short-term memory methods as a hybrid to diagnose diabetes, and obtained an accuracy of 86.45%.

3. Materials and Methods

3.1. Research design

The main purpose of this study is to diagnose diabetes early using different machine learning classification algorithms. Another purpose of the study is to compare the success of the machine learning models used. In the study, comparisons were made using different algorithms from the classification algorithms used in the literature and contributed to the literature in this field. This work was carried out on hardware with 8 GB RAM, Intel Core i5-7200U processor, NVIDIA GeForce 940MX graphics card. It is coded using the Python programming language on the Anaconda platform. In this section, the methodology of the study is explained. The flow chart of the study is shown in Figure 1.

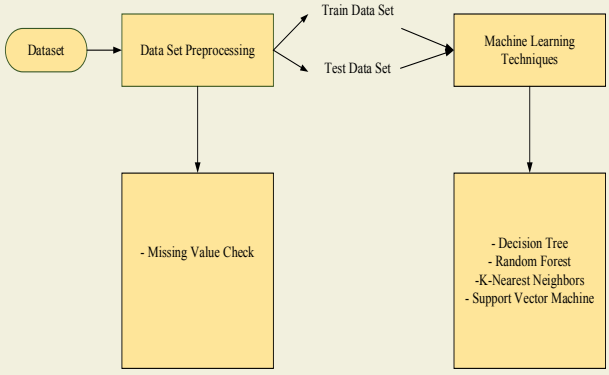

Figure 1. Flow chart of the study.

3.2. Dataset used in experiments

The Pima Indians Diabetes Dataset available on Kaggle was used in the study [16]. The dataset includes diabetes findings of women 21 years of age and older, obtained from the National Institute of Diabetes and Digestive and Kidney. There are 9 qualifications (8 attributes and 1 class variable) and 768 samples in the dataset. The dataset is divided into 70% training and 30% testing. The characteristics of the dataset are presented in Table 1.

Table 1. Properties of the Dataset

Features	Data Type
Pregnancies	int64
Glucose	int64
Blood Pressure	int64
Skin Thickness	int64
Insulin	int64
BMI	float64
Diabetes Pedigree Function	float64
Age	int64
Outcome	int64

3.3. Dataset preprocessing

For machine learning algorithms to give better results, preprocessing applications are made on the data. The preprocessing stage is one of the important steps to increase the classification performance and achieve the most accurate result. Missing values were checked for the dataset and no missing values were found. However, the min values of blood pressure, glucose, skin thickness, insulin and body mass index attributes are seen as

0 in the dataset. The Skewness method was used to fill in these data. Table 2 shows the dataset values obtained as a result of preprocessing.

3.4. Machine learning techniques

Machine learning, instead of trying to learn by coding information into a computer; is an artificial intelligence subfield that learns by extracting meaningful relationships and patterns from examples and observations [17]. Machine learning techniques and data processing tasks are shown in Figure 2.

In our study, classification was made with machine learning methods using a diabetes dataset. Decision Tree, Random Forest, K-Nearest Neighbor and Support Vector Machines from machine learning classifiers were used. Classification algorithms suitable for the data types in the used data set were selected. Selected algorithms are often used in classification problems. In addition, the selected algorithms are algorithms that prevent overfitting. Thus, it is aimed to achieve high accuracy rates.

a) Decision tree (DT)

The main structure of a decision tree consists of units called nodes, leaves and branches. The uppermost part of the tree is called the root and the lowermost part is called the leaf. The part between the root and the leaves is expressed as a branch [19]. A widely used machine learning is a classification algorithm. Decision trees continuously divide the dataset into sub-branches according to the splitting criterion. It decides on the information it collects from these sections [20]. Figure 3 shows the structure of the decision tree.

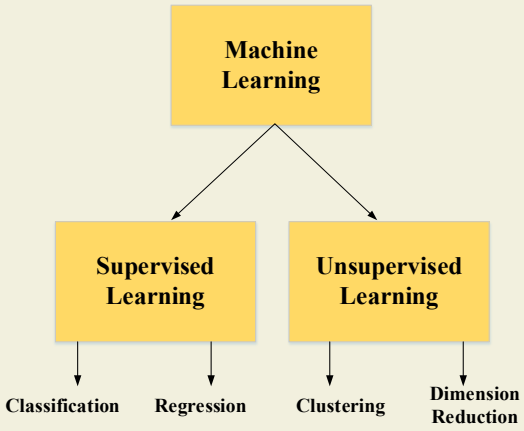

Figure 2. Machine learning techniques and data processing tasks [18].

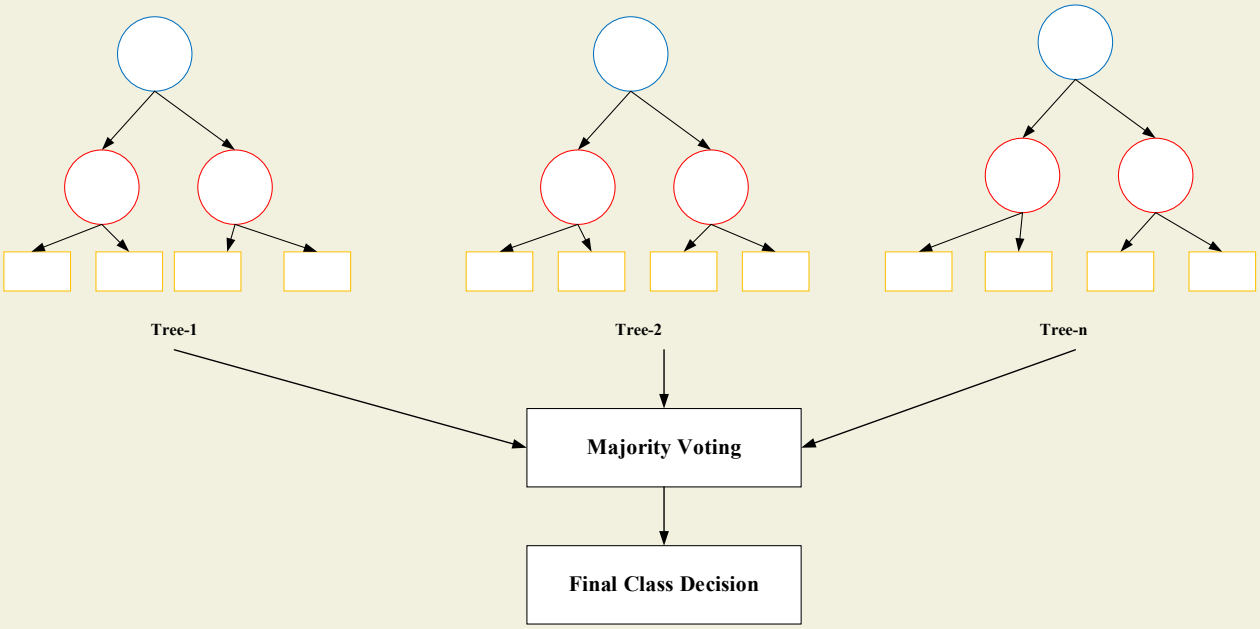

Figure 4. Random forest structure [23].

Table 2. Dataset Values

	Pregnancy	Glucose	Blood Pressure	Skin Thickness	Insulin	Body Mass Index	Diabetes History	Age	Result
Mean	3.84	121.94	73.10	30.85	176.11	32.53	0.47	33.24	0.34
St. deviation	3.36	30.61	12.76	10.64	99.44	6.92	0.33	11.76	0.47
Min. value	0.00	44.00	24.00	7.00	14.00	18.20	0.07	21.00	0.00
%25	1.00	99.75	64.00	23.00	102.75	27.50	0.24	24.00	0.00
%50	3.00	117.00	72.00	30.00	165.00	32.40	0.37	29.00	0.00
%75	6.00	141.25	80.00	39.00	236.00	36.80	0.62	41.00	1.00
Max. value	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00	1.00

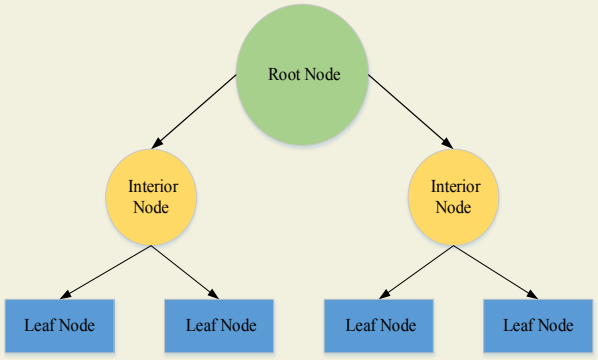

Figure 3. Decision tree structure [21].

b) Random forest (RF)

The random forest method creates a large number of decision trees and randomly selects the best feature from the child nodes [22]. The difference between a random forest and a decision tree is that the random forest method randomly performs the process of finding the root node and splitting it into nodes. Figure 4 shows the structure of the random forest method.

c) K-Nearest neighbor (K-NN)

The K-Nearest Neighbor classification was introduced by Cover and Hart in 1967. The K-NN classifier, thanks to its simplicity of implementation and effective efficiency, is among the top 10 classifiers among data mining classifiers [24]. The K-NN classifier finds the best class for the new sample by measuring the distance between the new sample to be classified and the training samples [25]. The K-NN structure is as presented in Figure 5.

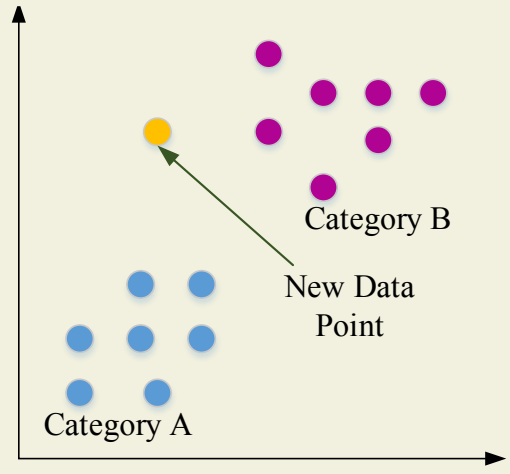

Figure 5. KNN structure [26].

d) Support Vector Machines (SVM)

Support Vector Machines draw a line to separate points on a plane. It aims to have the drawn line at the maximum distance for the points of both classes [27]. In short, SVM finds a decision boundary between the two classes that are furthest from any point.

There are two classes, red and blue, as shown in Figure 6.

The purpose of classification is to decide in which class the future new data will be. The area between the line is called the margin. SVM tries to find the most appropriate correct range [28].

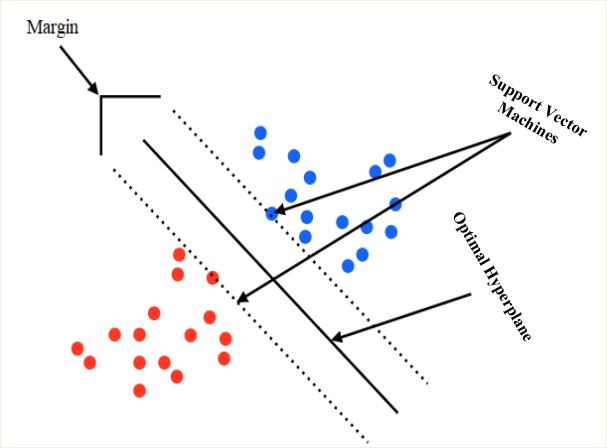

Figure 6. SVM structure [29].

3.5. Performance evaluation

The dataset, consisting of a total of 768 samples, was divided into 70% training and 30% testing. After the training process, classification success was checked. True negative (True Negative – TN) when the normally negative sample is correctly classified as negative, false positive (False Positive – FP) when the normally negative sample is misclassified as positive, and true positive (True Positive – TP) when the positive sample is correctly classified as positive. When a normally positive sample is misclassified as negative, it is defined as False Negative (FN). The matrix in which all these situations are shown is called the Confusion Matrix. Table 3 shows the confusion matrix.

Table 3. Confusion matrix

Estimated values	Actual values	
	Negative	Positive
Negative	True Negative (TN)	False Negative (FN)
Positive	False Positive (FP)	True Positive (TP)

When the test classes are compared with the classes produced by the system, how much of it is predicted correctly shows the overall classification accuracy of the model. Equation 1 shows the accuracy formula.

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

The proportion of positive samples that can be correctly detected by the classifier is called Sensitivity. Equation 2 shows the sensitivity formula.

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

The accuracy of positive predictions is found by the

Precision equation. Equation 3 shows the Precision formula.

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

Precision and sensitivity, in particular, a so-called F1-score is used to simply compare two classifiers. This criterion is the harmonic mean of precision and sensitivity. If the precision and sensitivity value are high, the F1-score will be high. Equation 4 shows the formula for the F1-score criterion.

$$F_1 = \frac{TP}{TP + \frac{FN+FP}{2}} \quad (4)$$

The ROC (Receiver Operating Characteristic) curve is a commonly used tool to predict classifier performance. The ROC curve is plotted according to the True Positive Rate (TPR) to the False Positive Rate (FPR). The area under the ROC curve in performance evaluation is called AUC (Area Under Curve). ROC-AUC value will have been 1 for a perfect classifier. The ROC-AUC value approaching 1 indicates the successful separation of positives from negatives. The ROC-AUC curve is shown in Figure 7.

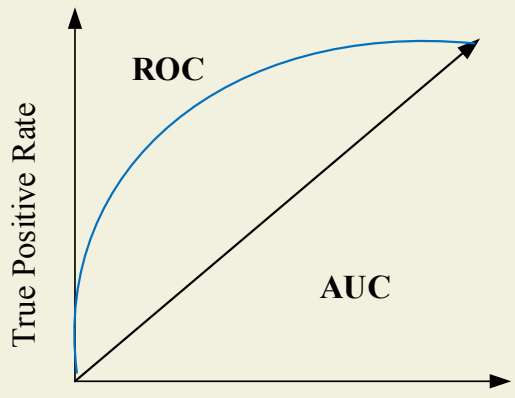

Figure 7. ROC-AUC curve [30].

4. Experimental Results

The dataset containing the diabetes findings was classified separately by Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (K-NN) and Support Vector Machines (SVM) models. Accuracy, Precision, Sensitivity, F1-score and Area Under Curve (AUC) values were calculated for each model and those are shown in Table 4.

Table 4. Success performance of models

Models	Accuracy	Precision	Sensitivity	F1-Score	AUC
DT	0.70	0.77	0.78	0.78	0.65
RF	0.80	0.83	0.88	0.85	0.74
KNN	0.78	0.77	0.95	0.85	0.68
SVM	0.77	0.78	0.93	0.85	0.68

According to Table 4, the RF model gave the best results in the classification of diabetes with 80% accuracy and 74% AUC rates. As shown in Figure 8, the DT model showed 70% success, the RF model 80% success, the K-NN model 78% success, and the SVM model 77% success.

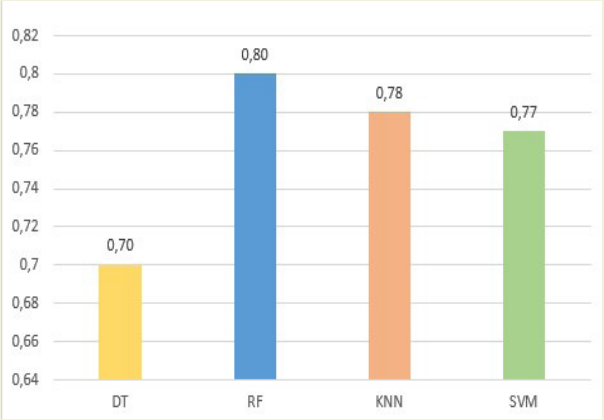

Figure 8. Success comparison of models

5. Conclusion

Diabetes is a type of disease that is at the forefront of today's diseases and is common in many parts of the world. Since diabetes causes the formation of other diseases, early diagnosis, taking precautions and starting treatment are of great importance. Early diagnosis provides the opportunity for a healthy and normal life. Therefore, early diagnosis of diabetes is an important issue.

Machine Learning is frequently used in the medical field as well as in many different fields. Thus, in this study, it was aimed to diagnose diabetes at an early stage. The dataset containing the diabetes findings firstly went through the preprocessing stage. In the preprocessing stage, missing value control was performed. No missing values were found. However, the min values of blood pressure, glucose, skin thickness, insulin and body mass index attributes are seen as 0 in the dataset. The Skewness method was used to fill in these data. Then, the data set consisting of 768 samples was trained and tested. For this, Decision Tree, Random Forests, K-Nearest Neighbor and Support Vector Machines from Machine Learning models were used. Accuracy, Precision, Sensitivity, F1-Score and AUC metrics were used to compare the performance of the models. All algorithms have been successfully trained and yield high accuracy results. When the classification accuracy of the models was compared, 70%, 80%, 78% and 77% results were obtained, respectively. It was revealed that the highest result was obtained from the Random Forests model with 80%. The Random Forests model gave the highest results not only in accuracy but also in precision, F1-score and AUC values.

Finally, the dataset in this study includes diabetes findings in women 21 years and older from the National Institute of Diabetes and Digestive and Kidney. There are 9

attributes (8 attributes and 1 class variable) and 768 samples in the data set. It has been trained and tested using machine learning methods Decision Tree, Random Forest, K-Nearest Neighbor and Support Vector Machines.

Machine learning techniques can give better and more successful results as the learning data increases. In this direction, more successful results can be obtained by using more machine learning models in future studies. The dataset can be enriched in terms of attributes.

References

[1] Çoşansu, G. (2015). Diyabet: küresel bir salgın hastalık. Okmeydanı Tıp Dergisi, 31:, 1-6. doi:10.5222/otd.2015.001.

[2] Türkiye Diyabet Vakfı, (accessed date: 01 January 2023). https://www.turkdiab.org/diyabet-hakkinda-hersey.asp?lang=TR&id=59

[3] Pulat, M., Kocakoç, I., D. (2021). Bibliometric analysis of published theses in the field of machine learning and decision trees in Turkey. Journal of Management and Economics, 28(2): 287-308. doi: 10.37990/medr.1077024.

[4] Bi, Q., Goodman, K., E., Kaminsky, J., Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. American Journal of Epidemiology, 188(12): 2222-2239. doi: 10.1093/aje/kwz189.

[5] Peng, G., C., Alber, M., Buganza Tepole, A., Cannon, W., E., De, S., Dura-Bernal, S., Kuhl, E. (2021). Multiscale modeling meets machine learning: What can we learn?. Archives of Computational Methods in Engineering, 28(3):1017-1037. doi:10.1007/s11831-020-09405-5.

[6] Benos, L., Tagarakis, A., C., Dolias, G., Berruto, R., Kateris, D., Bochtis, D. (2019). Machine learning in agriculture: A comprehensive updated review. Sensors, 21(11): 3758. doi: 10.3390/s21113758.

[7] Humelnicu, C., Ciortan, S., Amortila, V. (2019). Artificial neural network-based analysis of the tribological behavior of vegetable oil–diesel fuel mixtures. Lubricants, 7(4): 32. doi: 10.3390/lubricants7040032.

[8] Ray, S. (2019). A quick review of machine learning algorithms. COMITCon 2019 Conference Proceedings, p. 35-39. doi: 10.1109/COMITCon.2019.8862451.

[9] Faruque, M., F., Sarker, I., H. (2019). Performance analysis of machine learning techniques to predict diabetes mellitus. ECCE 2019 Conference Proceedings, p. 1-4. doi: 10.1109/ECACE.2019.8679365.

[10] Haq, A., U., Li, J., P., Khan, J., Memon, M., H., Nazir, S., Ahmad, S., Ali, A. (2020). Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. Sensors, 20(9): 2649. doi: /10.3390/s20092649.

[11] Dritsas, E., Trigka, M. (2022). Data-driven machine-learning methods for diabetes risk prediction. Sensors, 22(14): 5304. doi: 10.3390/s22145304.

[12] Khanam, J., J., Foo, S., Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4): 432-439. doi: 10.1016/j.icte.2021.02.004.

[13] Ayon, S., I., Islam, M., M. (2019). Diabetes prediction: a deep learning approach. International Journal of Information Engineering and Electronic Business, 12(2): 21. doi: 10.5815/ijieeb.2019.02.03.

[14] Baser, B., O., Yangın, M., Sarıdas, E., S. (2021). Classification of diabetes with machine learning techniques. Journal of Suleyman Demirel University Science Institute, 25(1): 112-120. doi: 10.19113/sdufenbed.842460.

[15] Er, M., B., Isık, I. (2021). Prediction of Diabetes disease using LSTM-based deep networks. Journal of Turkish Nature & Science, 10(1): 68-74.

[16] Kaggle, (accessed date: 16 February 2023). https://www.kaggle.com/code/kwonnnyr/diabetes-prediction-using-random-forest/notebook.

[17] Janiesch, C., Zschech, P., Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3): 685-695. doi: 10.1007/s12525-021-00475-2.

[18] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2): 612-619.

[19] Kavzoglu, T., Colkesen, I. (2010). Classification of satellite images with decision trees. Electronic Journal of Map Technologies, 2(1):36-45.

[20] Suresh, A., Udendhran, R., Balamurgan, M. (2020). Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers. Soft Computing, 24(11): 7947-7953. doi:10.1007/s00500-019-04066-4.

[21] Banjongkan, A., Pongsena, W., Kerdprasop, N., Kerdprasop, K. (2021). A study of job failure prediction at job submit-state and job start-state in high-performance computing system: using decision tree algorithms. Journal of Advances in Information Technology, 12(2). doi: 10.12720/jait.12.2.84-92.

[22] Shah, K., Patel, H., Sanghvi, D., Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1): 1-16. doi: 10.1007/s41133-020-00032-0.

[23] Asha Kiranmai, S., Jaya Laxmi, A. (2018). Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. Protection and Control of Modern Power Systems, 3(1): 1-12. doi: 10.1186/s41601-018-0103-3.

[24] Gou, J., Ma, H., Ou, W., Zeng, S., Rao, Y., Yang, H. (2019). A generalized mean distance-based k-nearest neighbor classifier. Expert Systems with Applications, 115: 356-372. doi: 10.1016/j.eswa.2018.08.021.

[25] Kumbure, M., M., Luukka, P., Collan, M. (2020). A new fuzzy K-nearest neighbor classifier based on the Bonferroni mean. Pattern Recognition Letters, 140: 172-178. doi: 10.1016/j.patrec.2020.10.005.

[26] Asharf, J., Moustafa, N., Khurshid, N., Debie, E., Haider, W., Wahab, A. (2020). A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions. Electronics, 9(7): 1177. doi: 10.3390/electronics9071177.

[27] Dinh, T., V,. Nguyen, H., Tran, X., L., Hoang, N., D. (2021). Predicting rainfall-induced soil erosion based on a hybridization of adaptive differential evolution and support vector machine classification. Mathematical Problems in Engineering. 1-20. doi: 10.1155/2021/6647829.

[28] Guner, N., Comak, E. (2011). Predicting the success of engineering students in MathematicsI courses using support vector machines. Journal of Pamukkale University Engineering Science, 17(2): 87-96.

[29] Do, T., N. (2020). Automatic learning algorithms for local support vector machines. SN Computer Science, 1(1): 1-11. doi: 10.1007/s42979-019-0006-z.

[30] Carta, S., Ferreira, A., Reforgiato Recupero, D., Saia, R. (2021). Credit scoring by leveraging an ensemble stochastic criterion in a transformed feature space. Progress in Artificial Intelligence, 10(4): 417-432.