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

Vol. 10-No.2 (2024) pp. 134-140
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Predicting Diabetes Disease Using Machine Learning Classification Algorithms and Comparison of Algorithm Performances

Seher ARSLANKAYA [1]*, Gülsüm Saltan YAŞLI[2]

[1]Sakarya University, Engineering Faculty, Industrial Engineering Department, Sakarya-Turkey
* **Corresponding Author** : **Email:** aseher@sakarya.edu.tr - **ORCID:** 0000-0001-6023-2901

[2]Sakarya University, Engineering Faculty, Industrial Engineering Department, Sakarya-Turkey
Email: gulsum.saltan@ogr.sakarya.edu.tr - **ORCID:** 0000-0001-6412-5951

Article Info:

DOI: 10.22399/ijcesen.233
Received : 09 November 2024
Accepted : 30 April 2024

Keywords

Data mining
Classification
Naïve Bayes
Logistic regression
MATLAB

Abstract:

Data is always produced everywhere in the universe. In our globalizing world, where the internet is rapidly spreading, countless data set examples can be given. We produce data when we shop at a bookstore, make a transfer from our bank account, travel with an airline, or get tested in a hospital. Test results of hospital patients, production reports of a manufacturing factory, and exam results of students are examples of data sets. This data pile grows daily, and the data has no meaning. Data gains meaning by being processed by various tools and transformed into information. Today, information has become the most incredible power. In this study, a prediction study of diabetic patients was made with ten different algorithms from classification methods, one of the machine learning methods. At the end of the prediction study, the algorithms' performances were compared. The unique aspect of this study is that the performance of the algorithms is compared by making predictions with ten different classification methods instead of just one or a few classification methods. In this study, the performances of the algorithms were compared by testing the trained models with test data. In line with these results, the highest accuracy rate, lowest error rate and highest sensitivity value were obtained with SVM, Bagged decision tree and K-NN algorithms. When performance data is evaluated in terms of test time, the shortest test time was obtained with the K-NN algorithm. In summary, it was tested with the K-NN algorithm in the shortest time possible and the test results were predicted with 100% accuracy.

1. Introduction

Machine learning includes methods that produce meaningful information from big data and make predictions used in critical decision processes [1]. Machine learning is the general name of algorithms that model a problem according to the data set of that problem. There are multiple analyses used to obtain meaningful information within the data stack. Some multiple methods and algorithms perform these analyses. In summary, machine learning aims to reveal the information hidden under the data.

Machine learning is one of the most important solutions for using data effectively. As it is used in every field, it is essential to improve the quality of service by interpreting data from data sets in healthcare and revealing valuable information.

At the same time, machine learning can be used in many areas, such as in the field of healthcare. In this study, data in the field of health were used. Machine learning classification algorithms were used in the study. Data sets contain many data about patients and diseases. As a result of literature research, it has been seen that classification algorithms are frequently used with health data sets, and efficient results are obtained. For this reason, a health data set was studied with machine learning classification algorithms in this study.

The following classification algorithms were used in the study. First of all, the algorithms were trained with the data set. Algorithms trained with the test data set, which was not included in the training set, were tested. According to the test results obtained, the studied algorithms were compared with some performance indicators. The unique aspect of this study is that the performance of the algorithms is compared by making predictions with ten classification methods instead of just one or a few

classification methods. The main headings of the classification algorithms examined in the study are as follows.

- Decision tree algorithm (Fine Tree)
- Logistic regression algorithm
- Naive Bayes algorithm (Kernel naive Bayes)
- Support vector machine algorithm (Fine Gaussian SVM)
- K-Nearest neighbour algorithm (Fine KNN)
- Kernel approximation algorithm (SVM kernel)
- Batch classifier algorithm (Bagged trees – RUS Boosted trees)

Alan Mathison Turing, who first introduced the concept of artificial intelligence, questioned whether machines could think. Artificial intelligence was born in the early 1940s with the devices produced to decipher messages during World War II.

Tracking health parameters with wearable technologies can also be an example of artificial intelligence studies. For example, smartwatches measure heart rate and sleep data. Smartphones and intelligent assistant applications can also be given as examples in the same way. The "Neyim Var" artificial intelligence application, accessed through the e-pulse program launched in Turkey in 2021, can also be an example of artificial intelligence studies. People who enter the application first state their complaints about their health problems. Afterward, the program asks questions about the complaints submitted. The data obtained from the patient's answers, previous diseases, and tests are processed by artificial intelligence algorithms and turned into meaningful information. As a result, a possible diagnosis is determined for the patient. The patient is advised to visit the outpatient clinic at the last stage.

Machine learning is a subset of the concept of artificial intelligence. Although the concepts of artificial intelligence and machine learning are confused, they do not mean the same thing. Machine learning can be defined as algorithms gaining the ability to learn by training them with data.

Pala M. et al. (2019) studied the performance of K-NN and decision tree algorithms in diagnosing breast cancer. As a result, the best result was achieved with the K-NN algorithm at 97.30% [2].

Taşcı M. et al. (2020) worked with WEKA and 10 data mining classification algorithms in their heart disease diagnosis study. The purpose of the study is to predict the presence of heart disease. As a result, the K-NN algorithm was found to have the highest performance [3].

Bilge H. et al. (2015) worked on the ovarian cancer data set with data mining methods. The study mentioned the current state of data mining studies in medicine, and relevant studies on cancer data sets were examined. Ovarian cancer data were modelled with various classification methods, and the accuracy rates of the algorithms were compared [4].

Çataloluk H. (2012), health data were studied with KNN and K-means algorithms. Using these algorithms, a decision-making system that can be used by doctors working in the relevant field and makes predictions for making diagnoses has been designed [5].

Gültepe Y. (2019) made air pollution prediction and evaluations with machine learning algorithms. As a result of the study showed that it gave the most successful prediction results with an accuracy rate of 87% in the artificial neural network model, 99% in the random forest algorithm, and 99% in the decision tree. The linear regression method has been observed to perform poorly, with an accuracy rate of 30% [6].

Kemalbay G. et al. (2020) predicted the direction of stock market index movement with multiple logistic regression and k-nearest neighbour algorithms. According to the results, logistic regression analysis achieved better prediction performance with an accuracy rate of 81% against the k-nearest neighbour algorithm on BIST100 data within the given period [7]

2. Material and Methods

Machine learning is a set of methods that can automatically detect relationships in data and use these detected relationships to predict the future or make decisions under uncertainty. Classification techniques are the most used techniques of machine learning. Classification in machine learning is dividing the available data into classes according to specific algorithm rules and predicting which class the new data will belong to based on the results [8].

Machine learning stands as a crucial component within the realm of artificial intelligence, empowering machines to acquire knowledge and execute designated functions [9]. Indeed, machine learning comprises a collection of methods and algorithms capable of forecasting forthcoming occurrences or categorizing data through the acquisition of discernible patterns from existing datasets [10].

In summary, it can be decided which class newly encountered data will belong to by classification methods [11].

2.1 Decision tree

The decision tree method is one of the classification methods. A decision tree is a directional tool consisting of a root node with no input and internal nodes receiving one input [12]. The basic structure of the decision tree consists of the following 3 sections: Nodes, branches, leaves.

Decision trees are essential in transforming data into information because they are simple and easily understandable [13]. The classification model is easy to understand and performs faster than other classification methods. There are many decision tree algorithms: ID3 algorithm, C4.5 algorithm, CART algorithm, SPRINT algorithm, and SLIQ algorithm [14].

CART Tree/Fine Tree is an algorithm used to create a decision tree. In other words, it is also called the Gini algorithm. This algorithm divides the node into two branches at each decision node. The point at which the nodes will be divided, and the division value are calculated by looking at the Gini index value [15]. The index value is the proportion of assets in the data set. The distributions of assets with the same index value are also interpreted as the same. In order to find which, attribute the division will start on, the smallest one is selected among the index values calculated for each attribute. The abovementioned steps are repeated on the remaining data set, and other divisions are calculated [16].

2.2 Logistic regression

This method aims to establish a model that can appropriately describe the relationship between dependent and independent variables in a way that best fits by using the least number of variables. Recently, it has been used at a very high rate in biology, medicine, and economics [17]. Logistic regression is used to determine the cause-effect relationship with independent variables in cases where the dependent class variable is more than two categorical variables. Logistic regression is a method used to determine the cause-and-effect relationship with explanatory variables in cases where the response variable is observed in categorical, binary, triple, and multiple categories. The expected values of the outputs according to the inputs are obtained as probabilities [18].

2.3 Naïve bayes

Naive Bayes algorithm is one of the classification algorithms of data mining. It is widely used in classifying text documents. It seems to be preferred due to its high applicability and performance rate. Some assumptions are made in the application of this algorithm. Accepting the independence of attributes from each other can be an example. If the qualities affect each other, it is not easy to calculate the probability here. It is accepted that all qualities are equally important [19].

The basis of the algorithm is based on Bayes' theorem. It is a lazy learning algorithm but can also work on imbalanced datasets [20]. The algorithm works by calculating the probability of each state for an element and classifying it according to the one with the highest probability value. High accuracy rates can be achieved with a small amount of training data [21]. The kernel naive Bayes approach was used in this study.

2.4 Support vector machines

Support Vector Machines (SVM) are one of the most well-known methods for binary classification. This method can be viewed as a perceptron trying to find a hyperplane that separates the data. It does this in a way that ensures that the hyperplane separates the data in the best way [22]. It is based on creating the maximum margin between the decision surfaces of two linearly separable classes, defined as support vectors, and determining the class boundaries [23].

Radial SVM is localized along the entire x-axis and gives finite responses. Kernel functions return the child product between two points in an appropriate property space. Thus, it provides fast classification performance by defining the concept of similarity with minimal computational cost, even in very high-dimensional areas. The core takes the data and converts it into the required form. The kernel method increases classification performance when used with nonlinear data [24].

For this reason, kernels are selected according to the data type used - examples of commonly used kernels: linear kernel, polynomial kernel, gaussian kernel, Bayesian kernel. The Gaussian kernel is also known as the radial basis function RBF kernel. It can be expanded to infinite dimensions with the help of the Taylor Series expansion [25].

When the support vector algorithm tries to classify the data linearly, the classification does not occur in some cases. To fix this situation, kernel cheating is used. This kernel trick is called the kernel trick. The algorithm in which this method is used is called the kernel approach [26]. This method is used when the data cannot be classified linearly [27].

2.5 K-Nearest neighbour

The K-NN algorithm is one of the most well-known classification algorithms. The learning process is carried out through training data. A new data set is compared with the samples in the training set and classified according to their similarity. The examples in the training set are called n-dimensional numerical attributes. All training samples are kept in an n-dimensional sample space, each representing a point in the n-dimensional space. When a new example is encountered, the k closest to the relevant example is determined from the training set. The class label of the new example is assigned according to the

majority vote of the class labels of its k nearest neighbours. The new observation is divided into the required classes with a simple voting method [28].
K-neighbour count: Represents the amount of nearest neighbours of a point of unknown class in the new data set. The k quantity of the algorithm is selected to classify. The accuracy rate of the algorithm depends on the K parameter; the accuracy rate may vary depending on the k parameter value [29]. Distance types: Various distance types can be used in classification studies with the K-nearest neighbour algorithm. The distance type is decided according to the characteristics of the data in the data set. The types of distances used can be listed as Euclidean distance, Minkowski distance, Manhattan distance, Chebyschev distance, Dilca distance, and Hamming distance. Euclidean distance was used in this study [30].

2.6 Ensemble classifiers

The bootstrap method was developed by Breiman. An ensemble is created by applying estimators to bootstrap samples from the original data set. Here, bootstrapping is used to select returns and create subsamples randomly. Bagging can also increase the predictive validity of an inconsistent predictor variable. It makes them more convenient by using variables that have a low amount of bias but high variance. According to experimental results, it has been shown that this method provides more efficient results than single trees [31].
In the boosting method, inferences are made from the collection of trees obtained by giving different weights to the data set. First, all observations are weighted equally. As the tree community begins to expand, weightings are adjusted based on problem information. While the weight of misclassified observations is increased, the weight of rarely misclassified observations is reduced [32].

2.7 Model performance indicators

The matrix used to evaluate the performance of the classification model is called the confusion matrix. It is used to show the prediction results of a classifier. The confusion matrix used in the study is shown in Table 1 [33]. True negative (TN): Indicates the number of results in the model where negative results were predicted as unfavourable. False negative (FN): The number of times positive results in the model were predicted as unfavourable.
False positives (FP): The number of times negative results in the model were predicted as positive.
True positive (TP): The number of times positive results in the model are predicted to be positive.
The most commonly used, simple, and decisive criterion in measuring model performance is the accuracy value of the model.

Table 1. *Confusion matrix*

	No disease	**There is illness**
No disease	TP	FN
There is illness	FP	TN

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

The error rate is calculated by dividing the number of incorrectly classified samples by the total number of samples [34].

$$Error\ rate = \frac{FP + FN}{TP + FP + TN + FN} \quad (2)$$

Precision measures the number of positive class predictions that belong to the positive class [34].

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

3. Results and Discussions

A ready data set was obtained from the Kaggle platform for the study to be carried out in the field of health. The resulting data set was arranged to be studied with machine learning classification algorithms.
The data set is divided into two parts. The first piece is the training data set, and the second data set is the test data set. The training data set was used to obtain a learned model by training classification algorithms. After the learned model is obtained, the learned models are tested using the data in the test data set with known results. In this way, the study is completed to compare the performance of the algorithms.
The obtained results and performances were compared using the confusion matrix.
When the data set is examined, the inputs and categorical output information are summarized in the table 2. The data set consists of continuous and categorical data.
In the table 2, the Conclusion heading represents the categorical output, and the other headings represent the inputs. The data in question is divided into two. The first data set was used to train the algorithms for prediction. The second data set was used to make predictions with the trained model. It was observed that the data such as glucose, skin thickness, insulin, and blood pressure were written as 0.

Table 2. *Data structure*

Input name	Input definition	Limits
Number of pregnancies	Number of times an individual becomes pregnant	0-17
Glucose	2nd hour glucose concentration	56-198
Blood pressure	Diastolic blood pressure (mm Hg)	24-110
skin thickness	Triceps skinfold thickness (mm)	7-63
Insulin	2-hour serum insulin (mu U/ml)	14-846
BMI	Body mass index	18.2-67.1
Pedigree function:	Diabetes pedigree function	0.085-2.42
Age	Individual's age	21-81
Conclusion	Categorical variable 0 - 1 0: Not a diabetic 1: Diabetic patient	0-1

The studies were carried out by removing the rows containing these data from the data set. All algorithms were worked on MATLAB's "Classification Learner" interface. Calculations regarding the performance

Table 3. *Algorithm training results*

Classification	Algorithm	Training accuracy	Training durations (sec)
Decision tree	Fine tree	75,5%	5,2557
Logistic regression	Logistic regression	78,60%	4,3848
Naïve bayes	Kernel naïve bayes	76,00%	3,7379
Support vector machine	Fine Gaussian SVM	66,30%	2,4842
Classification	**Algorithm**	**Training accuracy**	**Training durations (sec)**
K-Nearest neighbour	Fine KNN	73,50%	2,3749
Kernel approach	SVM kernel	69,40%	4,6108
Batch classifier	Bagged trees	77,30%	7,4513
Batch classifier	RUS Boosted trees	76,80%	6,7348

of the algorithms were made on MS Excel. The accuracy rates and training times obtained after training the training data are as in the Table 3. The test data were tested with the models obtained by the algorithms trained with the training data and confusion matrices were obtained.

The confusion matrices obtained as a result of the test study are given in table 4-11 on an algorithm basis.

Table 4. *Decision tree*

	No disease	There is illness
No disease	31	1
There is illness	2	15

Table 5. *Logistic regression*

	No disease	There is illness
No disease	30	2
There is illness	9	8

Table 6. *Kernel naive bayes*

	No disease	There is illness
No disease	30	2
There is illness	5	12

Table 7. *Fine Gaussian SVM*

	No disease	There is illness
No disease	32	0
There is illness	0	17

Table 8. *K-Nearest neighbour*

	No disease	There is illness
No disease	32	0
There is illness	0	17

Table 9. *SVM Kernel*

	No disease	There is illness
No disease	32	0
There is illness	7	10

Table 10. *Bootstrap aggregating*

	No disease	There is illness
No disease	32	0
There is illness	0	17

Table 11. *RUS Boosted trees*

	No disease	There is illness
No disease	30	2
There is illness	1	16

For this purpose, the accuracy rate, error rate and precision values of the models were calculated on an algorithm basis. While making these calculations, confusion matrix performance criteria and formulas were taken as basis. Accuracy rate is calculated by dividing the number of all correct predictions by the total number of predictions. The error rate is calculated by subtracting the accuracy rate from 100. The precision value is obtained by dividing the number of correctly predicted positive observations to the total number of positively predicted observations. The performance results obtained are as in Table 12.

Table 12. *Algorithm test results*

Algorithm	Test accuracy	Test error %	Test precision
Fine Tree	93,9%	6,1%	93,9%
Logistic regression	77,6%	22,4%	76,9%
Kernel naive bayes	85,7%	14,3%	85,7%
Fine Gaussian SVM	100%	0,0%	100%
Fine KNN	100%	0,0%	100%
SVM kernel	85,7%	14,3%	82,1%
Bagged trees	100%	0,0%	100%
RUS Boosted trees	93,9%	6,1%	96,8%

Based on the confusion matrix, the accuracy and precision values of the test data were calculated in Table 13. Looking at the performance results; the highest accuracy rate, lowest error rate and highest precision value were obtained with SVM, Bagged decision tree and K-NN algorithms.

4. Conclusions

In this study, the performance of the algorithms is compared by making predictions with ten classification methods instead of just one or a few. In this case, it shows the original aspect of this study. The confusion matrices obtained with the test data and the accuracy rate, error rate and precision value and the performances of the algorithms are compared in Table 12.

Finally, the performances were compared in terms of the accuracy rate obtained when testing the trained models and the test data and the testing time as a different parameter.

The test accuracy rates, and algorithm test times obtained as a result of prediction studies with classification algorithms are given in Table 13. Among the algorithms studied, five have a test accuracy rate of 100%. Among these, the algorithm with the shortest training time is the "K-Nearest Neighbour" algorithm. Looking at the support vector machine algorithm, the test accuracy rate is 100%, and the training time is very close to the k-nearest neighbour algorithm. The K-Nearest Neighbour algorithm was tested with diabetes data in 2.3749 seconds, and 100% accuracy and precision were achieved on the test data. In summary, the most reliable prediction results on the studied data set were achieved with the "K-Nearest Neighbour algorithm." The algorithm that gives the closest results to the algorithm in question is the

Table 13. *Comparison of algorithms*

Classification	Algorithm	Test accuracy	Durations (sec)
Support vector machine	Fine Gaussian SVM	100%	2,4842
K-Nearest neighbour	**Fine KNN**	**100%**	**2,3749**
Batch classifier	Bagged trees	100%	7,4513
Decision tree	Fine Tree	93,9%	5,2557
Batch classifier	RUS Boosted trees	93,9%	6,7348
Naive Bayes	Kernel naive bayes	85,7%	3,7379
Kernel approach	SVM kernel	85,7%	4,6108
Logistic regression	Logistic regression	77,6%	4,3848

"Support vector machine" algorithm. In conclusion, considering the results in this study, both approaches gave better results than the other approaches studied. As in the methods studied in the literature, the "K-Nearest Neighbour Algorithm" showed the highest performance in the prediction study, with ten different classification algorithms.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Bramer M. (2016), *Principles of data mining*, Springer London.

[2] Pala M., Çimen M., Boyraz Ö., Yıldız M., Boz A. (2019), Comparative performance analysis of decision tree and K-NN algorithms in diagnosing breast cancer, *Published in 7th International Symposium on Innovative Technologies in Engineering and Science 22-24 November 2019 (ISITES2019 Şanlıurfa-Turkey)*

[3] Taşçı M., Şamlı R. (2020), Heart disease diagnosis with data mining, *European Journal of Science and Technology Special Issue, pp. 88-95, April 2020.*

[4] Bilge H., Demircioğlu H. (2015), Analysis of gene expressions in the ovarian cancer dataset with data mining, *Marmara Journal of Science and Technology 4: 125-134, DOI: 10.7240/mufbed.89154*

[5] Çataloluk H. (2012), *Disease diagnosis using data mining methods on real medical data*, Bilecik University, Department of Computer Engineering, master's thesis.

[6] Gültepe Y. (2019), A comparative evaluation on air pollution prediction with machine learning algorithms, *European Journal of Science and Technology* 16;8-15.

[7] Kemalbay G., Alkış B. (2021), Prediction of stock market index movement direction with multiple logistic regression and k-nearest neighbor algorithm, *Pamukkale University Journal of Engineering Sciences*, 27(4), 556 569.

[8] Aggarwal C. (2014), *Data classification: Algorithm and applications*, Taylor, and Francis.

[9] Dean J. (2014), *Big data, data mining, and machine learning*, Wiley.

[10] Shafaf N., Malek H. (2019), Applications of machine learning approaches in emergency medicine, a review article, *Archives of Academic Emergency Medicine.* 7 (1): e34.

[11] Poyraz O. (2012), *Data mining applications in medicine: Breast cancer data set analysis*, Trakya University, Institute of Science and Technology, master's thesis, Edirne.

[12] Dilki G., Başar Ö. (2020), Comparison of distance measures via k-nearest neighbor algorithm in predicting bankruptcy of businesses, *Istanbul Ticaret University Journal of Science*, 19(38);224-233.

[13] Moshkov M. (2020), *Comparative analysis of deterministic and nondeterministic decision trees*, Springer International Publishing.

[14] Kareem E. (2017), *An optimized decision tree classification algorithm for a data set*, Noor Publishing.

[15] Maimon O., Rokach L. (2014), *Data mining with decision trees: Theory and applications*, World Scientific Publishing Company.

[16] Adak M., Yurtay N. (2013), Development of a software that allows creating a decision tree using the Gini algorithm, *Journal of Information Technologies*, 6(3).

[17] Marvin L. (2017), *Statistics with MATLAB advanced regression*, CreateSpace Independent Publishing Platform.

[18] Schein A., Ungar L. (2007), Active learning for logistic regression: an evaluation, Mach Learn (2007) 68: 235–265, DOI 10.1007/s10994-007-5019-5.

[19] Bhambri P. (2020), *Design of paddy crop production technique: using k-means, naive bayes, KNN and SVM classifiers*, Lambert Academic Publishing.

[20] Sabry F. (2023), *Naive bayes classifier: Fundamentals and applications*, One Billion Knowledgeable.

[21] Hatipoğlu E. (13 June 2018), *Machine learning – classification- naive bayes- part 11*, Date of access: 31.10.2022, Access address: https://medium.com/@ekrem.hatipoglu/machine-learning-classification-naive-bayes-part-11-4a10cd3452b4

[22] Pala T. (2013), *Realization of medical decision support system with data mining methods*, Marmara University, Institute of Science and Technology, Department of Electronic Computer Education, master's thesis, Istanbul.

[23] Şekeroğlu S. (2010), *A data mining application in the service sector*, Istanbul Technical University, Institute of Natural and Applied Sciences, Department of Industrial Engineering, master's thesis, Istanbul.

[24] Suykens J., Signoretto M., Argyriou A. (2015), *Regularization, optimization, kernels, and support vector machines*, CRC Press Taylor and Francis Group.

[25] Campbell C., Ying Y. (2011), Learning with support vector machines, Morgan and Claypool Publishers.

[26] Kung S.Y. (2014), *Kernel methods and machine learning*, Cambridge University Press.

[27] Uğur B. (3 September 2020), *Support vector machine (SVM)*, Date of access: 21.10.2023, Access address: https://burakugur.medium.com/destek-vekt%C3%B6r-makinesi-svm-f07f57f0a641

[28] Akküçük U. (2011), *Data mining clustering and classification algorithms*, Yalın Publishing, Istanbul.

[29] Youqiang Z., Guo C., Bisheng W., Xuesong L. (2018), A novel ensemble method for k-nearest neighbor, *Pattern Recognition*, 85;13-25.

[30] Kılınç D., Borandağ E., Yücalar F., Tunalı V., Şimşek M., Özçift A. (2016), Scientific article classification using KNN algorithm and text mining with R language, *Marmara Journal of Science* 3: 89-94.

[31] Kozak J. (2018), *Decision tree and ensemble learning based on ant colony optimization*, Springer International Publishing.

[32] Ay Ş. (16 December 2019), *Ensemble learning — bagging and boosting*, Date of access: 20.10.2022, access address: https://medium.com/deep-learning-turkiye/ensemble-learning-bagging-ve-boosting-50643428b22b.

[33] Ranawana R., Palade V. (2006), *Optimized precision-a new measure for classifier performance evaluation*, 2006 IEEE Congress on Evolutionary Computation.

[34] Hossin M., Sulaiman M.N. (2015), A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining & Knowledge Management Process (IJDKP)* 5(2).