**Advances in Artificial Intelligence Research (AAIR)**
**ISSN 2757-7422**
**www.dergipark.com/aair/**

**Vol. 2 (No. 1), pp. 7-14, 2022**
**doi: 10.54569/aair.1018444**
**Published online: Feb 16, 2022**

# Classification of Iris Flower by Random Forest Algorithm

Hilmi Cenk Bayrakçı [1,*], Abdullah Burak Keşkekçi [2], Recep Arslan [3]

[1] Isparta University of Applied Sciences, Dept. of Mechatronics Engineering, Isparta, Turkey;
[2] Isparta University of Applied Sciences, Dept. of Mechatronics Engineering, Isparta
[3] Isparta University of Applied Sciences, Dept. of Mechatronics Engineering, Isparta, Turkey;

**Abstract**

With the introduction of artificial intelligence into our lives, artificial intelligence researches and applications in different fields such as agriculture, health, military and engineering applications have become very popular iris flower was classified using the popular Random Forest, support vector machine and Artificial neural network machine learning classifiers with high accuracy rates. As a result of the classification, the performance of the trained models was evaluated according to the confusion matrix, sensitivity, specificity, accuracy, F1 score, ROC curve and AUC evaluation criteria. The random forest algorithm was the most successful among the trained algorithms with an accuracy rate of 97%.

***Keywords:*** *Random Forest; iris dataset; classification*.

## 1. Introduction

Today, with the rapid development of technology, artificial intelligence methods are used in many areas. Artificial intelligence is a method that can model human thought and decision-making ability [1]. Artificial intelligence is a set of software and hardware systems that imitate human intelligence, can reason, and can make decisions in line with the experiences gained [2-4]. One of the frequently used branches of artificial intelligence is machine learning. Machine learning is fed from different sciences such as computer sciences, statistics and engineering. machine-learning; It performs different tasks such as classification, clustering, regression, pattern recognition, density estimation, outlier detection and information extraction [5]. Machine learning tasks are divided into two main headings, supervised learning and unsupervised learning [6]. While input and output sets are given in supervised learning, it is a learning method without output sets in unsupervised learning. For example, classification and regression operations are examples of supervised learning, while clustering and density are examples of unsupervised learning [7].

In this study, classification of the iris dataset was performed by using Random Forest (RF) machine learning classifier, which is fast, frequently used and has high accuracy [8]. At the end of the training, the performance of the model was evaluated according to the confusion matrix, sensitivity, originality, accuracy and F1 score evaluation criteria.

## 2. Material and Method

The iris data set was used in the study. For this data set, the data set was modelled using the RF algorithm with a software prepared in the Python programming language.

### 2.1. Material

Materials such as iris data set, RF algorithm and performance evaluation criteria used in the material part of the study will be given as sub-titles based on both in-formation and literature.

### 2.1.1. Iris Dataset

The iris data set used in the study consists of three different classes, setose, versicolor and virginica, and 150 rows and 4 columns. There are 50 different examples for each class. In the study, it was tried to model with RF algorithm using these data. In Table 1, the characteristics and statistical information of the iris data set are given.

*Corresponding author
*E-mail address:* cenkbayrakci@isparta.edu.tr

Received: 04/Nov/2021; Received in revised form: 04/Jan/2022; Accepted: 17/04/2022.

**Table 1.** *Characteristics and statistical information of the Data Set*

| Feature | Meaning | Range | Average |
|---|---|---|---|
| sepal length | Lower leaf length | 4.3-7.9 | 5.843 |
| sepal width | Lower leaf width | 2.0-4.40 | 3.057 |
| petal length | Upper leaf length | 1.0-6.9 | 3.758 |
| petal width | Upper leaf width | 0.1-2.5 | 1.199 |

### 2.1.2. RF Algorithm

RF is a community machine learning algorithm developed by Leo Breiman [9]. Ensemble classification algorithms contain multiple classifiers instead of a single classifier [10]. The structure of the RF algorithm includes more than one decision tree, it gives results by taking the average of the decision trees [11]. A trained k-number trees are collected in an RF model defined in Equation 1 [12].

$$H(X,\theta_j) = \sum_{i=0}^{k} h_i(x,\theta_j) \quad ,(j = 1,2,3,\dots,m) \tag{1}$$

In the equation, H (X, θ_j) is a meta decision tree classifier. x represents the input feature vector of the training dataset and θ_j is an independent and uniformly distributed random vector that determines the growth process of the tree [1].

### 2.1.3. Support Vector Machine Algorithm

Today, Support Vector machines (SVM) are a widely used method in many fields such as engineering, healthcare and agriculture [13]. SVM is divided into two main classes, Support Vector Classification (SVC) and Support Vector Regression (SVR) [14]. SVR is generally used in measurable learning and nonlinear regression Using Equation 2, the most accurate function is found with SVR in the hypothesis function set. In the equation, x represents the training data set, w represents the weight vector and b represents the threshold value [14].

$$\{f \mid f(x) = w^T x + b,\ w \in \mathbb{R}^d,\ b \in \mathbb{R}\} \tag{2}$$

### 2.1.4. Artificial Neural Networks (ANN)

Artificial neural networks (ANN), one of the artificial intelligence methods, have been used quite frequently in recent years [15]. ANN expresses a mathematical model that is structurally similar to biological neurons [16]. ANN is a method of estimating an output sequence from input variables entered as a sequence as a mathematical model [17]. ANNR is a model that shows a nonlinear match between an input vector and an output vector, as seen in Figure 1. The $i_1, i_2$ and $i_3$ in the figure represent the input vectors, and the $O_1$ and $O_2$ values represent the output vectors [18].

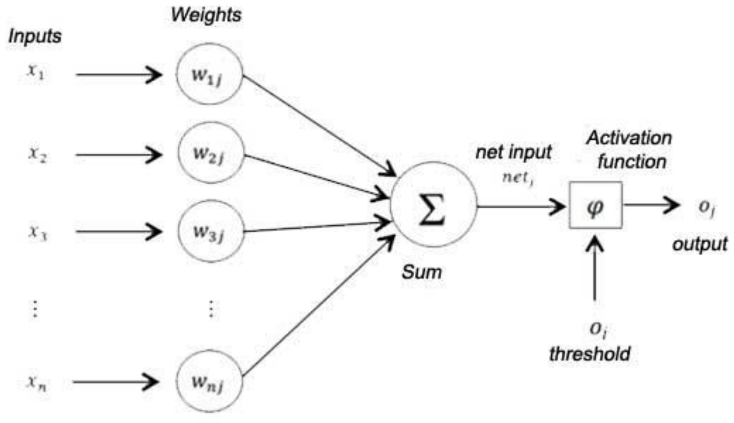

**Figure 1.** *Structure of ANN [18]*

### 2.1.5. Performance Evaluation Criteria

All Different evaluation criteria such as Receiver Operating Area Under the ROC Curve (AUC), Characteristic Curve (ROC), sensitivity, authenticity, accuracy and F1 score are frequently used in classification processes. The ROC curve is a probability curve for different classes. In the ROC curve, False

Positive Rate (FPR) values on the horizontal axis and True Positive Rate (TPR) values on the vertical axis are used. The AUC value is the area under the ROC curve. Equations for sensitivity, Specificity, accuracy and F1 score are given in 2-5 [18-23].

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

$$Specificity \ = \frac{TN}{TN+FP} \tag{3}$$

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \tag{4}$$

$$F1\ Score = \frac{2*TP}{2*TP+FN+FP} \tag{5}$$

### 2.2. Method

In the study, the workflow diagram of the classification of iris flower with random forest is given in figure 2. First, the open access iris dataset was taken. In the second stage, the data set was divided into two as training 80% and test 20%. During the training phase, three different artificial intelligence algorithms, namely RF, SVM and ANN, were trained. The number of estimators was set to 80 and maximum depth 10 without final training in the RF algorithm. In the SVM algorithm, the thickness of tube (ε) was chosen as 1.4, the penalty factor (C) was 25 and the kernel coefficient values (γ) was 0.000125. The ANN model created in the study includes a single hidden layer. 20 neurons and relu activation functions were used in the hidden layer. An ANN model with four inputs and three outputs, sepal length, sepal width, petal length and petal width, is designed. In the training process of the ANN model, the value of the batch size is 10, the number of epoch is 50 and the optimization method is man. The final models to be obtained after the training process was completed were tested on the test data set. In the last stage, the trained model-in test data set was evaluated according to the confusion matrix, sensitivity, specificity, accuracy and F1 score, ROC and AUC evaluation criteria.

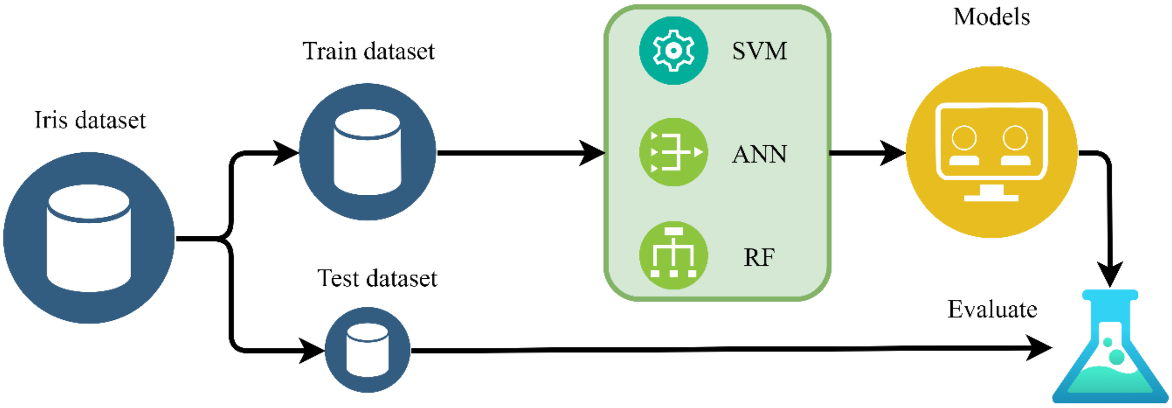

**Figure 2.** *Workflow diagram*

### 3. Research Findings and Discussion

In the study, classification process was carried out on the iris dataset consisting of 150 different samples and containing iris flower information with the random forest, support vector machine and artificial neural networks machine learning algorithm using the Python programming language. As a result of training random forest important measurements were made on 3 attributes. The graph of the importance of the attributes is shown in Figure 3.

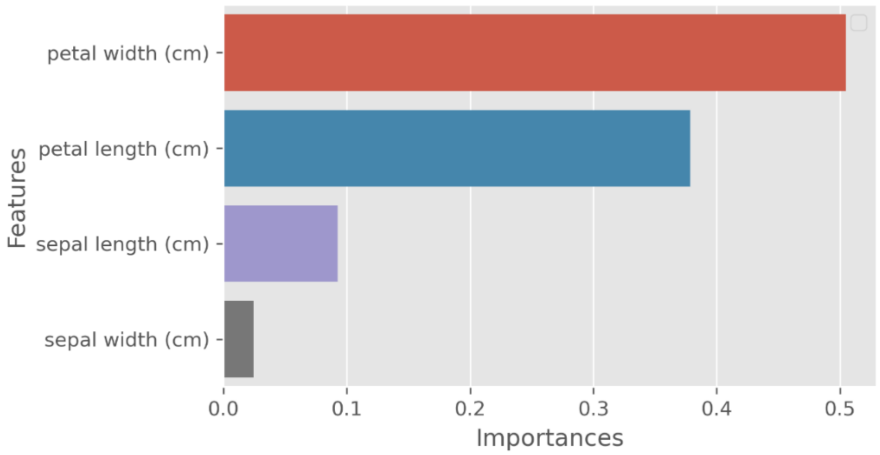

**Figure 3.** *Significance graph of attributes*

When Figure 3 is examined, it is seen that the "petal length" attribute has the most effect on the trained model and is the most important parameter. The "sepal width" attribute is the least important attribute. The trained model of RF was tested with 30 data. The confusion matrix of RF formed after the test and estimation process is given in figure 4.

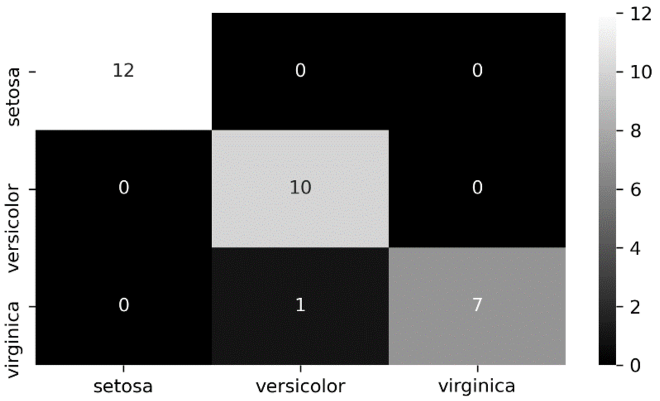

**Figure 4.** *Confusion Matrix of RF*

When the confusion matrix given in Figure 4 is examined, the number of correctly and incorrectly estimated by RF samples belonging to three different classes is seen. As shown in the confusion matrix, 29 out of 30 data were predicted correctly. It is seen that the model obtained in this way correctly detects the classes with an accuracy of approximately 97%. Since the data set used in the modelling includes more than two classes, the macro calculation method, which is one of the multi-class evaluation techniques, was preferred. Evaluation results are given in Table 2. When Table 2 is examined, it is seen that the model obtained is quite successful. The accuracy of the trained model was determined as 97%.

**Table 2.** *evaluation results of RF*

| Evaluation Criteria | Value |
|---|---|
| Sensitivity | 0.96 |
| Specificity | 0.97 |
| F1 score | 0.97 |
| AUC | 0.97 |
| Accuracy | 0.97 |

As seen in Table 2, the AUC value was obtained as 0.97. It has been observed that the AUC value is almost close to 1. The ROC curve is given in figure 4 to examine the accuracy of the model graphically. When examined in Figure 4, the ROC curve of each class is given and the AUC values are shown. When the curve is examined, it is seen that it gives a result close to an ideal ROC curve in the detection of iris flower.

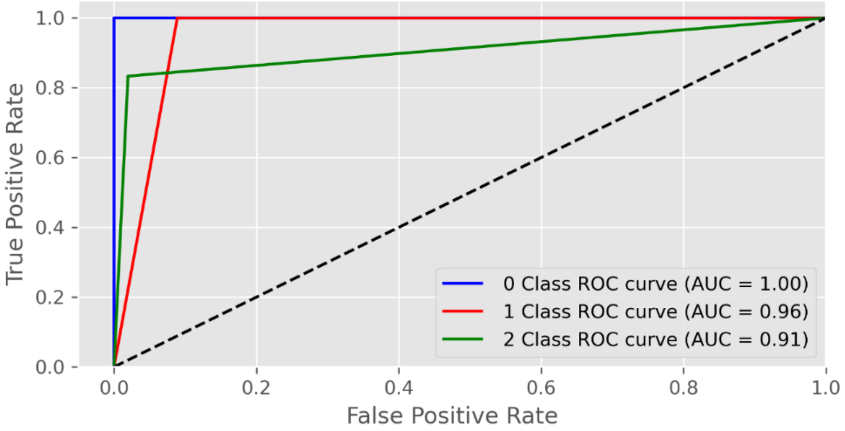

**Figure 4.** *ROC curve and AUC values of RF*

After training RF model to compare it with other machine learning model we tried SVM and ANN models. The trained SVM model was tested with 30 data. The confusion matrix of SVM formed after the test and estimation process is given in figure 5.

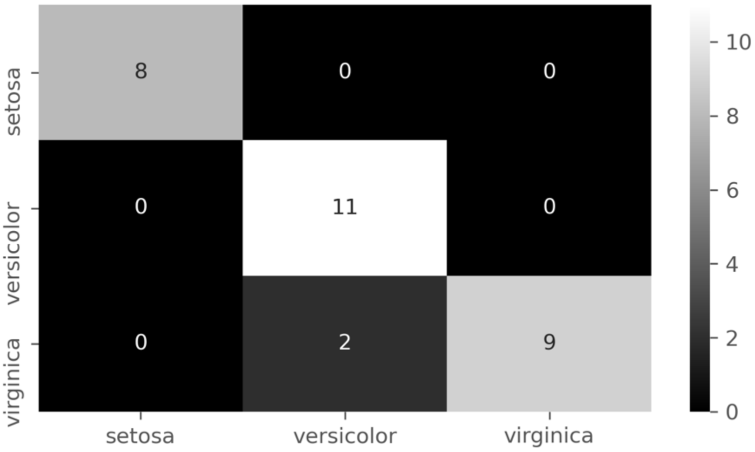

**Figure 5.** *Confusion Matrix of SVM*

When the confusion matrix of SVM given in Figure 5 is examined, the number of correctly and incorrectly estimated samples belonging to three different classes is seen. As shown in the confusion matrix, 29 out of 30 data were predicted correctly. It is seen that the model obtained in this way correctly detects the classes with an accuracy of approximately 93%. Since the data set used in the modelling includes more than two classes, the macro calculation method, which is one of the multi-class evaluation techniques, was preferred. Evaluation results are given in Table 3. When Table 3 is examined, it is seen that the model obtained is quite successful. The accuracy of the trained model was determined as 93%.

**Table 3.** *evaluation results of SVM*

| Evaluation Criteria | Value |
|---|---|
| Sensitivity | 0.94 |
| Specificity | 0.95 |
| F1 score | 0.93 |
| AUC | 0.95 |
| Accuracy | 0.93 |

As seen in Table 3, the AUC value was obtained as 0.95. It has been observed that the AUC value is almost close to 1. The ROC curve is given in figure 6 to examine the accuracy of the model graphically. When examined in Figure 6, the ROC curve of each class is given and the AUC values are shown. When the curve is examined, it is seen that it gives a result close to an ideal ROC curve in the detection of iris flower.

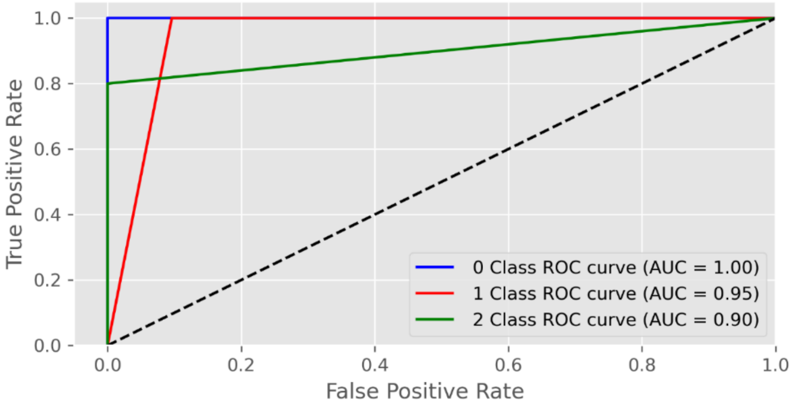

**Figure 6.** *ROC curve and AUC values of SVM*

After training RF model to compare it with other machine learning model we tried SVM and ANN models. The trained ANN model was tested with 30 data. The confusion matrix of ANN formed after the test and estimation process is given in figure 7.

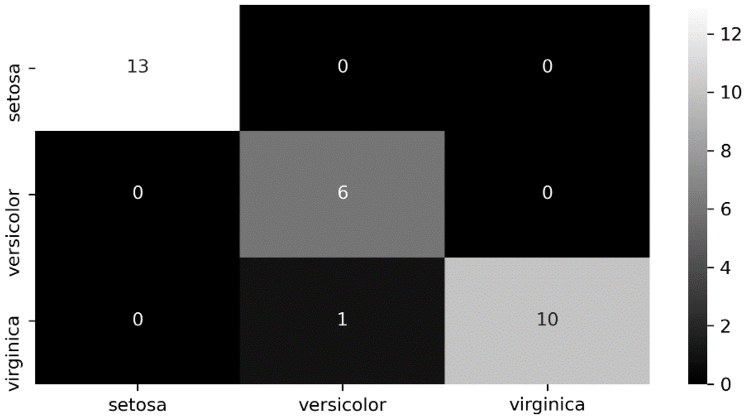

**Figure 7.** *Confusion Matrix of ANN*

When the confusion matrix of ANN given in Figure 7 is examined, the number of correctly and incorrectly estimated samples belonging to three different classes is seen. As shown in the confusion matrix, 29 out of 30 data were predicted correctly. It is seen that the model obtained in this way correctly detects the classes with an accuracy of approximately 97%. Since the data set used in the modelling includes more than two classes, the macro calculation method, which is one of the multi-class evaluation techniques, was preferred. Evaluation results are given in Table 4. When Table 4 is examined, it is seen that the model obtained is quite successful. The accuracy of the trained model was determined as 97%.

**Table 4.** *evaluation results of ANN*

| Evaluation Criteria | Value |
|---|---|
| Sensitivity | 0.97 |
| Specificity | 0.95 |
| F1 score | 0.96 |
| AUC | 0.95 |
| Accuracy | 0.97 |

As seen in Table 4, the AUC value was obtained as 0.97. It has been observed that the AUC value is almost close to 1. The ROC curve is given in figure 8 to examine the accuracy of the model graphically. When examined in Figure 8, the ROC curve of each class is given and the AUC values are shown. When the curve is examined, it is seen that it gives a result close to an ideal ROC curve in the detection of iris flower.

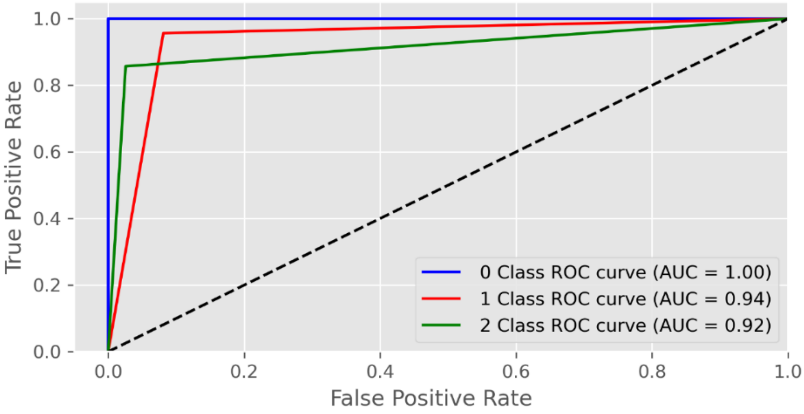

**Figure 8.** *ROC curve and AUC values of ANN*

## 3. Conclusion

AI methods are frequently used in agriculture, as in many different fields, like early detection of many plant diseases and plays an important role in determining whether the plants are suitable for harvesting. In the study, the type of iris flower was determined with a random forest, support vector machine and artificial neural networks algorithms and the classification accuracy was evaluated in terms of performance. Obtained performance results are given below.

- The confusion matrix is obtained for the random forest classifier. It was observed that only 1 out of 30 samples was detected incorrectly in the confusion matrix.
- The trained model by random forest was found to be successful with 97% accuracy, 97% specificity, 96% sensitivity, 97% F-measure and 97% AUC value.
- The confusion matrix is obtained for the support vector machine classifier. It was observed that only 2 out of 30 samples was detected incorrectly in the confusion matrix.
- The trained model by support vector machine was found to be successful with 93% accuracy, 95% specificity, 94% sensitivity, 93% F-measure and 95% AUC value.
- The confusion matrix is obtained for the artificial neural networks. It was observed that only 1 out of 30 samples was detected incorrectly in the confusion matrix.
- The trained model by random forest was found to be successful with 97% accuracy, 95% specificity, 97% sensitivity, 96% F-measure and 95% AUC value.

With the results obtained in the study, an artificial intelligence-based model has been proposed for the classification of iris flower. With this proposed model, it is aimed to contribute to the academic literature for the applications of artificial intelligence in the field of agriculture. In future academic studies, it is thought to increase the accuracy rate by using different artificial intelligence models.

### Declaration of interest

It was presented as a summary at the ICAIAME 2021 conference.

### References

[1] Özkan, İ. N. İ. K., & Ülker, E. (2017). Derin Öğrenme ve Görüntü Analizinde Kullanılan Derin Öğrenme Modelleri. Gaziosmanpaşa Bilimsel Araştırma Dergisi, 6(3), 85-104.

[2] Russell SJ, Norvig P. Artificial intelligence: a modern approach. 3rd ed. Pearson Education Inc:New Jersey;2016.

[3] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology. 2017;2(4):230-243.

[4] Aksoy, B, Halis, H, Salman, O. (2020). Elma Bitkisindeki Hastalıkların Yapay Zekâ Yöntemleri ile Tespiti ve Yapay Zekâ Yöntemlerinin Performanslarının Karşılaştırılması. International Journal of Engineering and Innovative Research, 2 (3), 194-210. DOI: 10.47933/ijeir.772514

[5] Mitchell T. Machine Learning. New York, USA, McGraw Hill, 1997.

[6] Kalaycı, T. E. (2018). Kimlik hırsızı web sitelerinin sınıflandırılması için makine öğrenmesi yöntemlerinin karşılaştırılması. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24(5), 870-878.

[7] Harrington P. Machine Learning in Action. New York, USA, Manning Publications, 2012.

[8] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.

[9] Breiman, L. Rastgele Ormanlar. Machine Learning 45, 5-32 (2001).

[10] Akar, Ö., & Güngör, O. (2012). Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. Jeodezi ve Jeoinformasyon Dergisi, ss, 139-146.

[11] Korkmaz, D., Çelik, H. E., & Kapar, M. Sınıflandırma ve Regresyon Ağaçları ile Rastgele Orman Algoritması Kullanarak Botnet Tespiti: Van Yüzüncü Yıl Üniversitesi Örneği. Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 23(3), 297-307.

[12] Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., & Li, K. (2016). A parallel random forest algorithm for big data in a spark cloud computing environment. IEEE Transactions on Parallel and Distributed Systems, 28(4), 919-933.

[13] Mohammadi, K., Shamshirband, S., Anisi, M. H., Alam, K. A., & Petković, D. (2015). Support vector regression based prediction of global solar radiation on a horizontal surface. Energy Conversion and Management, 91, 433-441.

[14] Zhang, L., Zhou, W. D., Chang, P. C., Yang, J. W., & Li, F. Z. (2013). Iterated time series prediction with multiple support vector regression models. Neurocomputing, 99, 411-422.

[15] Yang, Z., Cai, Y., Li, Q., Li, H., Jiang, Y., Lin, R., ... & Gao, X. (2021). Predicting particle collection performance of a wet electrostatic precipitator under varied conditions with artificial neural networks. Powder Technology, 377, 632-639.

[16] Taşar, B., Üneş, F., Demirci, M., & Kaya, Y. Z. (2018). Yapay sinir ağları yöntemi kullanılarak buharlaşma miktarı tahmini. DÜMF Mühendislik Dergisi, 9(1), 543-551.

[17] Taşar, B., Üneş, F., Demirci, M., & Kaya, Y. Z. (2018). Yapay sinir ağları yöntemi kullanılarak buharlaşma miktarı tahmini. DÜMF Mühendislik Dergisi, 9(1), 543-551.

[18] Budak, H., & Erpolat, S. (2012). Kredi Riski Tahmininde Yapay Sinir Ağları ve Lojistik Regresyon Analizi Karşılaştırılması. AJIT-e: Online Academic Journal of Information Technology, 3(9), 23-30.

[19] Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland, 19, 67.

[20] Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care & Pain, 8(6), 221-223.

[21] Eusebi, P. (2013). Diagnostic accuracy measures. Cerebrovascular Diseases, 36(4), 267-272.

[22] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), 6.

[23] Ekrem, Ö., Salman, O. K. M., Aksoy, B., & İnan, S. A. (2020). Yapay Zekâ Yöntemleri Kullanilarak Kalp Hastaliğinin Tespiti. Mühendislik Bilimleri ve Tasarım Dergisi, 8(5), 241-254.