Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 40, S. 29-33, Eylül 2022
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 40, pp. 29-33, September 2022
Copyright © 2022 EJOSAT
Research Article

Investigation of Covid-19 Infection with Clinical Data Using Decision Trees

Fırat Orhanbulucu[1,2], Fatma Latifoğlu[2]*

[1] Inonu University, Department of Biomedical Engineering, Malatya, Turkey (ORCID: 0000-0003-4558-9667), firat.orhanbulucu@inonu.edu.tr

*[2] Erciyes University, Department of Biomedical Engineering, Kayseri, Turkey (ORCID: 0000-0003-2018-9616), flatifoglu@erciyes.edu.tr

(1st International Conference on Innovative Academic Studies ICIAS 2022, September 10-13, 2022)

(**DOI:** 10.31590/ejosat.1171818)

ATIF/REFERENCE: Orhanbulucu, F. & Latifoğlu, F. (2022). Investigation of Covid-19 Infection with Clinical Data Using Decision Trees. *European Journal of Science and Technology*, (40), 29-33.

Abstract

The coronavirus disease, namely Covid-19 infection, which was declared a worldwide epidemic by the World Health Organization (WHO) in 2020, was first seen in Wuhan, China in the last months of 2019 and has affected the whole world. Early diagnosis of this rapidly spreading epidemic is important to prevent the disease. For this reason, methods such as image processing, deep learning, and machine learning have become important to detect the epidemic early. In this study, it has been tried to classify individuals who test positive and negative for Covid-19 based on some laboratory test results with several Decision Tree methods. Since the original form of the data set has an uneven distribution, the data set has been balanced by applying the oversampling and undersampling methods used for such data sets as a pre-processing study. Balanced dataset and original dataset using 5-Fold Cross Validation (CV), 10-Fold Cross Validation and Leave-One-Out (LOO)-CV, Random Forest (RF), Random Tree (RT), J48, ıt was analyzed with alternating decision tree (ADTree) and Function Trees (FT) classifiers. As a result of the examination, the most successful result was shown by the RF classifier with 87.5% success rates using CV-5 in the original data set, 93.3% using CV-10 and LOO-CV in the oversampling method, and 79% using CV-5 in the undersampling method. In addition to success rates, sensitivity-specificity metrics, which are important for patient and healthy diagnosis, were examined in terms of each classification algorithm and CV value.

Keywords: Covid-19; Decision Tree; Random Forest; Oversampling.

Karar Ağaçları Kullanılarak Klinik Verilerle Covid-19 Enfeksiyonunun İncelenmesi

Öz

2020 yılında Dünya Sağlık Örgütü (WHO) tarafından dünya çapında salgın ilan edilen koronavirüs hastalığı yani Covid-19 enfeksiyonu, ilk olarak 2019 yılının son aylarında Çin'in Wuhan kentinde görülmüş ve tüm dünyayı etkisi altına almıştır. Hızla yayılan bu salgının erken teşhisi, hastalıktan korunmak için önemlidir. Bu nedenle görüntü işleme, derin öğrenme, makine öğrenmesi gibi yöntemler salgını erken tespit etmek için önemli hale geldi. Bu çalışmada çeşitli Karar Ağacı yöntemleri ile bazı laboratuvar test sonuçlarına göre Covid-19 testi pozitif ve negatif çıkan bireyler sınıflandırılmaya çalışılmıştır. Veri setinin orijinal formu eşit olmayan bir dağılıma sahip olduğundan, bu tür veri setleri için kullanılan aşırı örnekleme ve eksik örnekleme yöntemleri bir ön işleme çalışması olarak uygulanarak veri seti dengelenmiştir. Dengeli hale getirilen veri seti ve orjinal veri seti 5-Fold Cross Validation (CV) , 10-Fold Cross Validation ve Leave-One-Out (LOO)-CV kullanılarak Random Forest (RF), Random Tree (RT), J48, Alternating decision tree (ADTree) ve Function Trees (FT) sınıflandırıcıları ile incelenmiştir. İnceleme sonucunda en başarılı sonuç orijinal veri setinde CV-5 kullanılarak %87,5, aşırı örnekleme yönteminde CV-10 ve LOO-CV kullanılarak %93,3 ve eksik örnekleme yönteminde CV-5 kullanılarak %79 ile RF sınıflandırıcısı göstermiştir. Başarı oranlarının yanı sıra hasta ve sağlıklı teşhisi için önemli olan duyarlılık-özgüllük metrik değerleri her bir sınıflandırma algoritması ve CV değeri bakımından incelenmiştir.

Anahtar Kelimeler: Kovid19; Karar ağacı; Rastgele Orman; Aşırı Örnekleme.

*[2]Corresponding Author**:** flatifoglu@erciyes.edu.tr, +90 352 207 6666 (32977)

1. Introduction

Covid-19 infection, known as coronavirus disease, was first seen in Wuhan, China in the last months of 2019 and affected the whole world in 2020 [1]. This disease spread rapidly among people, causing a new epidemic in many countries. Coronavirus disease, which can be transmitted by small respiratory droplets, coughing, or sneezing when closely interacted with infected people, shows symptoms such as shortness of breath and cough. Although Covid-19 infection may show serious complications in some patients, some patients can overcome this disease asymptomatically [1, 2].

It has been stated that machine learning methods or image processing methods that can be applied to chest or lung images can play an important role in defining Covid-19 disease [3]. In the literature, several studies have been conducted using machine learning and image processing methods for the detection or analysis of Covid-19 disease. In the study, the effect of coronavirus disease on the region, spreading rate, and weather conditions were examined using the Support Vector Regression (SVR) method [4]. In a study, Convolutional Neural Network (CNN) model has been proposed to automatically detect Covid-19 patients or healthy individuals from chest X-ray images. With the proposed model, the success rate was found to be 96.78% [5]. De Moraes Batista, et al. collected data from 235 patients in emergency care and tried to predict Covid-19 patients using five machine learning methods [6]. Uneven data distributions are seen especially in rare epidemic diseases such as Covid-19. Oversampling and undersampling methods are used to eliminate such unbalanced data distributions. In a study, it was tried to predict Covid-19 patients based on laboratory test results commonly collected from suspected Covid-19 case applications using the Synthetic Minority Sampling Method (SMOTE) and Artificial Neural Networks (ANN) algorithm. As a result of the study, the success rate was found to be 86% in the original data set, while it was observed that the success rate increased to 90% when the unbalanced data distribution was eliminated by using the SMOTE-based method [7]. It has also been observed that studies have been conducted to estimate the number of cases using machine learning methods [8].

In this study, individuals with negative (healthy) or positive (patient) Covid-19 test results, according to widely measured laboratory values, who were taken from people who applied to the hospital with the suspicion of Covid-19 were examined using decision trees methods. In the examination, decision trees such as Random Forest (RF), Random Tree (RT), J48, and Function Trees (FT), Alternating decision tree (ADTree) classification methods, which are frequently used in the patient-healthy distinction in Biomedical studies, were used. Since the data set examined in the study showed an unbalanced distribution, the unbalanced distribution was eliminated by applying oversampling and undersampling methods, and the balanced data set was examined with machine learning methods and compared with the results of the data set with the unbalanced distribution. Cross-Validation values, which are important in classification studies, were selected as 5, 10, and Leave One Out in the study, and a comparison was made in terms of accuracy, sensitivity, and specificity metric values.

2. Materials and Methods

2.1. Dataset

The data set used in the study was taken from the Kaggle platform, which is used as an open data set source for machine learning studies. Information on the data set was obtained from patients who came to Israelita Albert Einstein Hospital in São Paulo, Brazil with suspicion of Covid-19 [9]. The data set consists of 5644 people in total. Of these 5644 people, 5086 have negative test results, 558 of them are positive people with positive test results. “Hematocrit, Hemoglobin, Platelets, Red Blood Cells, Mean Platelet Volume, Lymphocytes, Leukocytes, Mean Corpuscular Hemoglobin, Mean Corpuscular Hemoglobin Concentration, Basophils, Eosinophils, Monocytes, Red Blood cell Distribution Width laboratory measurement values” and patient ID quality, A total of 111 attribute values were entered into the data set, such as which service they were referred to or whether the patient was admitted to the normal room. As a result of the examination, it was found that 5050 people contained blank data or missing values and non-quantitative values that did not affect the analysis. The data set was downloaded to 594 people after discarding empty data and patient ID information without any effect, values such as which service the patient was referred to, whether the patient was admitted to the normal room, and empty values. Hematocrit, Hemoglobin, Platelets, Red Blood Cells, Mean Platelet Volume, Lymphocytes, Leukocytes, Mean Corpuscular Hemoglobin, Mean Corpuscular Hemoglobin Concentration, Basophils, Eosinophils, Monocytes, Red Blood cell Distribution Width laboratory measurement values for 14 studies in total were quantitative. It has been examined. Of the 594 people, 513 were negative and 81 were positive (coronavirus patients). Figure 1 shows the distribution of the data to be examined in the study. Since the data showed an uneven distribution oversampling and undersampling processes were applied and the data were analyzed by balancing both upwards and downwards to make our classification result healthier.

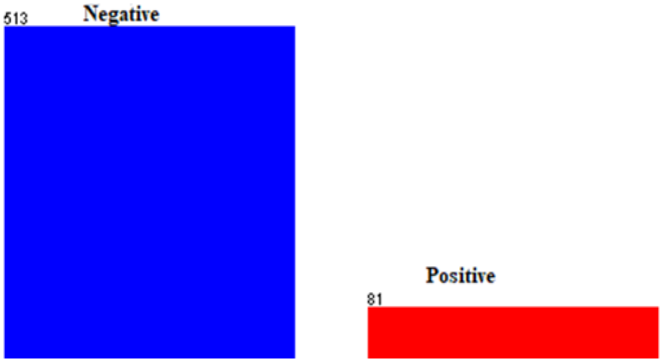


Figure 1. Distribution of the data to be examined.

2.2. Imbalance Process for Data

In order to prevent the imbalance in the data set, the oversampling and undersampling methods are applied to the data set as a preprocess. As Oversampling, Synthetic minority over-sampling technique (SMOTE) method, which is frequently applied in studies showing such unbalanced data distributions, and Spread Subsampling (SS) method as undersampling process were used.

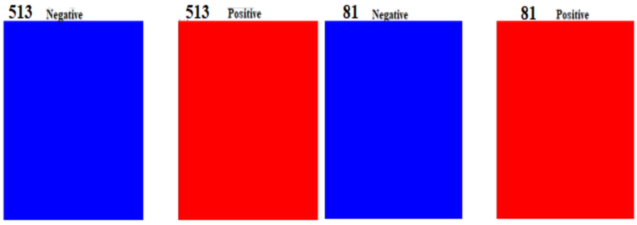
2.2.1. Oversampling

The Oversampling method shows high speed sampling. This method makes samples in the minority class closer to the number of samples in the majority class by randomly generating samples synthetically according to their closest neighbors. The imbalance in terms of the number of samples produced and the number of samples between the majority class is eliminated. The advantage of this method is that there is no data loss, unlike the undersampling method [10]. Among the oversampling methods, the most frequently used method is the SMOTE method developed by Chawla, et.al in 2002 [11]. In the study, this method was also preferred and the data set was balanced so that 513 negative-513 positives.

2.2.2. Undersampling

The undersampling sampling method removes the imbalance between classes by randomly removing the samples in the majority class to eliminate the imbalance in the data set. The disadvantage of this method is the possibility of discarding important data in the majority class [10, 12]. The SS undersampling method data set preferred in the study was set to be 81 negative and 81 positive.

In Figure 2, data distributions formed after SMOTE method (a) and SS method (b) are applied to the data set used in the study are shown.



(a) (b)

Figure 2. Data distribution analysis (a) SMOTE (b) Spread Subsampling

2.3. Classification

Classification is the process of distributing data to classes in a controlled manner after passing the training and testing phase. Primarily, the data to be trained is created and the distribution of the training data generated to the classification algorithm method is taught. Then, the classification process is performed most accurately by testing the data whose class is not known. The classification methods preferred from the biomedical studies and in the distinction between sick and healthy were used in the study. In the study, the classification processes were made using the Waikato Environment for Knowledge Analysis (WEKA) Version 3.8.3 program [13]. Some information about the classification methods used is given below.

2.3.1. Alternating decision tree (ADTree)

The Adtree algorithm consists of decision states that specify the outcome of an action. Detailed information about ADTree, which consists of decision node and forecast node layers, is given in [14].

2.3.2. Random Forest (RF)

The RF classification method was developed by Breamin, and this method extracts the small information contained in the data set and enables differentiation [15]. It has been stated that the RF method, which is also defined as the collection of decision trees, requires very little pre-processing and is successful on unstable data sets in studies [14-16].

2.3.3. Random Tree (RT)

The trees created as a result of the RT classification algorithm are randomly selected from the possible tree cluster, and there is a chance to sample each tree equally. It has been stated that the trees show a similar distribution and that the models created by many random trees can generally have a high rate of accuracy [17].

2.3.4. J48

J48 is a C4.5 decision tree developed by J. Ross Quinalan for classification of nonlinear and small size data [18]. Decision tree selection is important in solving the classification problem. In this method, a tree is created to create the classification model and the classification process is performed over the remaining data, ignoring the missing data [19].

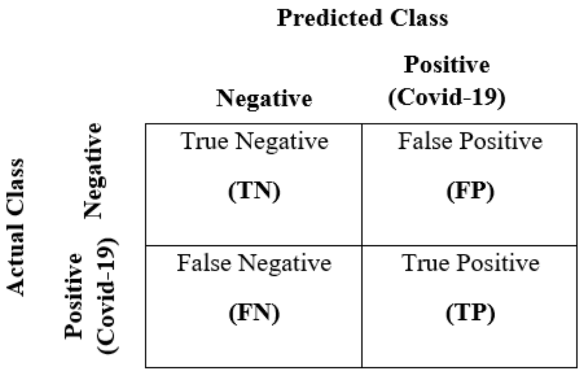
2.3.5. Function Trees (FT)

FT is an algorithm that can be implemented with four different models. The FT algorithm can be viewed as a generalization of multivariate trees and can fix data in a sample space by dividing it [20].

3. Research Results

In this study, healthy or sick individuals were classified using RF, RT, J48, ADTree, FT machine learning methods according to the results of the Covid-19 test in line with the laboratory values obtained by measuring from those who applied to the hospital with the suspicion of Covid-19. Since data belonging to positive (patient) and negative (healthy) classes were unevenly distributed at the beginning, SMOTE and SS procedures were applied as a pre-treatment. As a result of the SMOTE method, the minority class was multiplied and the unbalanced distribution between the two classes was equalized with 513 people in both classes. As a result of the SS method, the data in the majority class was reduced and the data was adjusted so that 81 people were in both classes. As a result of the classification process, data distribution at the beginning, resulting from SMOTE analysis and SS analysis were examined in terms of 5-10-fold Cross-Validation (CV) and Leave-One-Out (LOO) CV. In the 10-fold cross-validation method, the data is divided into 10 equal parts and 10% of the data is divided into parts as the test, 90% is the training data, and then the trained training parts are used to predict the tested part. The result of the classification is estimated by repeating this process 10 times and taking the average of the results. Similarly, in the 5-fold cross-validation process, the data is divided into 5 equal parts and the remaining 4 layers for a solid test are used as a training set and this process is repeated 5 times. In the LOO-CV process, which is preferred in cases where the number of samples is less in the literature, the data set as much as the sample number is divided into pieces, and each sample is used as both training and test data.

According to the Confusion Matrix is given in Figure 3, the accuracy, sensitivity, and precision metric values were calculated from equation (1) (2) (3) for the classification algorithm and CV values used in the study. Accuracy value refers to the rate of successfully classified sample, Sensitivity value refers to the number of positive samples correctly classified, ie disease diagnosis rate. Specificity metric value is used to measure the proportion of correctly classified negative models, that is, the determination of healthy individuals [21]. The results resulting from the calculations are shown in Table 1.



		Predicted Class	
		Negative	Positive (Covid-19)
Actual Class	Negative	True Negative (TN)	False Positive (FP)
	Positive (Covid-19)	False Negative (FN)	True Positive (TP)

Figure 3. Confusion Matrix

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

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

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

Table 1. Classification Results

		10-Fold CV			5-Fold CV			LOO-CV		
Pre-Processing Procedure	**Classification Method**	**Ac (%)**	**Sp (%)**	**S (%)**	**Ac (%)**	**Sp (%)**	**S (%)**	**Ac (%)**	**Sp (%)**	**S (%)**
Original Data Set	ADTree	86.2	93.0	43.2	85.2	93.2	34.6	86.2	91.8	50.6
	J48	86.7	94.5	38.3	86	93.4	39.5	84.3	92	35.8
	RF	**86.8**	96.5	24.7	**87.5**	96.9	28.4	**87.2**	96.5	28.5
	RT	83.6	90.4	40.7	83.5	89.7	44.4	84.3	91.4	39.5
	FT	86.2	95.0	29.6	87.3	93.2	50.6	86.5	93.0	45.7
Oversampling (SMOTE)	ADTree	84.4	79.9	88.9	84.7	81.5	87.9	84.2	80.5	87.9
	J48	85.3	82.7	88.1	84	81.1	86.9	85.2	81.3	89.3
	RF	**93.3**	92	94.5	**93**	92.2	94	**93.3**	91	95.7
	RT	86.1	85	87.3	84.5	82.7	86.4	85	83.2	86.9
	FT	84.9	83	86.9	83.9	80.5	87.3	85.8	82.8	88.7
Undersampling (SS)	ADTree	74.1	75.3	72.8	73.4	74.1	72.8	74.1	72.8	75.3
	J48	73.4	67.9	79	70.9	65.4	76.5	69.1	69.1	69.1
	RF	**76**	72.8	79	**79**	77.8	80.2	**77.8**	75.3	80.2
	RT	76.5	76.5	76.5	70.9	76.5	65.4	71.6	77.8	65.4
	FT	74.7	70.4	79.0	72.8	65.4	80.2	79	71.6	86.4

Ac: Accuracy, Sp: Specificity, S: Sensitivity, LOO-CV: Leave-One-Out Cross-Validation

4. Discussion and Conclusion

In this study, the data set created as a result of the laboratory test samples taken from people who applied to the hospital with the suspicion of Covid-19 was in its original form, accuracy with the ADTree, J48, RF, RT, and FT classification algorithms, which were balanced as a result of applying SS from Oversampling methods and Undersampling methods, Analyzed for sensitivity and specificity. In the analysis process, the number of CVs was determined as 5, 10, and LOO-CV, which is generally preferred.

As a result of the examination, the RF classification algorithm gave successful results both in its original form and in the balanced form of the data set. The success rate was 86.8% in the original data distribution, 93.3% with SMOTE analysis, and 76% in data distribution after SS analysis. It is stated in the literature that the RF algorithm gives successful results in unbalanced data distributions [16, 22]. In the study, a 93.3% success rate was obtained in the RF classifier, especially with SMOTE analysis, and a result that supports the literature has been presented. When Table 1 was examined, it was seen that the sensitivity rate in the original classification results was lower than the results obtained as a result of the application of SMOTE and SS methods. When the unbalanced distribution that occurred in the original state was eliminated, it was seen in the study that the success rate in the classification procedures increased with the sensitivity rate, which is especially important for the diagnosis of the patient. It has been observed that the downward balancing of the data set with the Undersampling method reduces the success rate in contrast to the oversampling method. When the values in Table 1 are examined in terms of CV, no significant differences were observed in success rates.

In the study using a similar data set in the literature, the data set was examined with the original state and SMOTE analysis with the ANN algorithm, and it was observed that SMOTE analysis increased the success rate [7]. Similarly in the literature, in another study conducted on data sets obtained as a result of blood count, RF, ANN, and glmnet classifiers were used to detect Covid-19 positive-negative carriers. As a result of the classification, success rates were found as 86% for RF, 80% for ANN, and 84% for glmnet [22]. It is stated in the literature that Covid-19 can be detected as a result of image processing methods, as well as Covid-19 can be detected with laboratory test results [5-8, 22].

In this study, it was stated that the oversampling-undersampling pretreatment methods are important to obtain healthier results for the analysis of data with the unbalanced distribution. As a result of this study, it was seen that the rapidly spreading coronavirus disease can be detected with some samples taken from the laboratory, especially due to the high success rate of the RF classification algorithm in the analysis made as a result of the SMOTE analysis.

References

[1] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Tan, W. (2020). A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine. (DOI: 10.1056/NEJMoa2001017)

[2] Hu, Z., Song, C., Xu, C., Jin, G., Chen, Y., Xu, X., ... & Shen, H. (2020). Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. Science China Life Sciences, 63(5), 706-711. (https://doi.org/10.1007/s11427-020-1661-4)

[3] Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., & Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of COVID-19. Plos one, 15(6), e0235187. (https://doi.org/10.1371/journal.pone.0235187)

[4] Yadav, M., Perumal, M., & Srinivas, M. (2020). Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos, Solitons & Fractals, 139, 110050. (https://doi.org/10.1016/j.chaos.2020.110050)

[5] Apostolopoulos, I. D., & Mpesiana, T. A. (2020). Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine, 1. (https://doi.org/10.1007/s13246-020-00865-4)

[6] de Moraes Batista, A. F., Miraglia, J. L., Donato, T. H. R., & Chiavegatto Filho, A. D. P. (2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv. (https://doi.org/10.1101/2020.04.04.20052092)

[7] Yavaş, M., Güran, A., & Uysal, M. Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması. Avrupa Bilim ve Teknoloji Dergisi, 258-264. (https://doi.org/10.31590/ejosat.779952)

[8] Ahmad, A., Garhwal, S., Ray, S. K., Kumar, G., Malebary, S. J., & Barukab, O. M. (2020). The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. Archives of Computational Methods in Engineering, 1-9. (https://doi.org/10.1007/s11831-020-09472-8)

[9] Einstein Data4u, E. Hospital Israelita Albert Einstein, Sao Paulo, Brazil. Diagnosis of Covid-19 and its clinical spectrum, URL: https://www.kaggle.com/einsteindata4u/datasets, (accessed 08/10/2020)

[10] Yıldırım, P. (2016). Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes. Procedia Computer Science, 83, 1013-1018. (https://doi.org/10.1016/j.procs.2016.04.216)

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357. (https://doi.org/10.1613/jair.953)

[12] Hernandez, J., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2013, November). An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. In Iberoamerican Congress on Pattern Recognition (pp. 262-269). Springer, Berlin, Heidelberg. (https://doi.org/10.1007/978-3-642-41822-8_33)

[13] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18. (https://doi.org/10.1145/1656274.1656278)

[14] Freund, Y., & Mason, L. (1999, June). The alternating decision tree learning algorithm. In icml (Vol. 99, pp. 124-133).

[15] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. (https://doi.org/10.1023/A:1010933404324)

[16] Buettner, R., Sauer, S., Maier, C., & Eckhardt, A. (2015, January). Towards ex ante prediction of user performance: a novel NeuroIS methodology based on real-time measurement of mental effort. In 2015 48th Hawaii International Conference on System Sciences (pp. 533-542). IEEE. (DOI: 10.1109/HICSS.2015.70)

[17] A. ONAN, “Comparative Performance Analysis of Decision Tree Algorithms in the Corporate Bankruptcy Prediction”, Bilişim Teknolojileri Dergisi, vol. 8, no. 1, 2015. (https://doi.org/10.17671/btd.36087)

[18] Quinlan, J. R. (1994). The minimum description length principle and categorical theories. In Machine Learning Proceedings 1994 (pp. 233-241). Morgan Kaufmann. (https://doi.org/10.1016/B978-1-55860-335-6.50036-2)

[19] Pradeep, K. R., & Naveen, N. C. (2016, December). Predictive analysis of diabetes using J48 algorithm of classification techniques. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I) (pp. 347-352). IEEE. (DOI: 10.1109/IC3I.2016.7917987)

[20] Gama, J. (2004). Functional trees. Machine learning, 55(3), 219-250.

[21] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1. (DOI: 10.5121/ijdkp.2015.5201).

[22] Banerjee, A., Ray, S., Vorselaars, B., Kitson, J., Mamalakis, M., Weeks, S., ... & Mackenzie, L. S. (2020). Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. International immunopharmacology, 86, 106705. (https://doi.org/10.1016/j.intimp.2020.106705).