「Araştırma Makalesi / Research Article」

Kastamonu Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi
Kastamonu University Journal of Faculty of Economics and Administrative Sciences

Haziran 2025 Cilt: 27 Sayı:1
iibfdergi@kastamonu.edu.tr

Başvuru Tarihi / Received: 05.10.2024
Kabul Tarihi / Accepted: 14.03.2025
DOI: 10.21180/iibfdkastamonu.1561910

Comparison of Some Performance Metrics Used in Multiple Classification Problems[1]

Ali Vasfi AĞLARCI[2], Cengiz BAL[3]

Abstract

The purpose of this research is to compare the performance metrics used in multiple classification problems in machine learning. For this purpose, simulation study was carried out under different scenarios by using 4 different classification methods and the performance metrics obtained were compared in this direction. While comparing the performance metrics in the study, the data to be used for classification purposes were derived under different scenarios, taking into account the effect of 4 factors. 90 different scenarios were created by considering the number of 3 different categories of the response variable, 5 different sample sizes, 3 different correlation structures, and the balanced and unbalanced distribution of the response variable. Accuracy, Kappa and CramerV metrics used in multiple classification problems were used as performance measures. Changes in performance metrics in the determined scenarios are summarized in tables and compared. As a result of the comparisons made with the simulation study, it has been seen that Kappa performance measure is a more accurate performance metric than the other two metrics in multi-class classification problems, and the method gives more reliable information about the classification success.

Keywords: *Classification Success, Classification Performance, Machine Learning, Simulation, Performance Metrics*

Jel Codes: *C15, C38, C88*

Çoklu Sınıflandırma Problemlerinde Kullanılan Bazı Performans Ölçütlerinin Karşılaştırılması

Öz

Bu araştırmanın amacı, makine öğrenmesinde çoklu sınıflandırma problemlerinde kullanılan performans metriklerini karşılaştırmaktır. Bu amaçla 4 farklı sınıflandırma yöntemi kullanılarak farklı senaryolar altında simülasyon çalışması yapılmış ve elde edilen performans metrikleri bu doğrultuda karşılaştırılmıştır. Çalışmada performans metrikleri karşılaştırılırken, sınıflandırma amacıyla kullanılacak veriler 4 faktörün etkisi dikkate alınarak farklı senaryolar altında türetilmiştir. Yanıt değişkeninin 3 farklı kategori sayısı, 5 farklı örneklem büyüklüğü, 3 farklı korelasyon yapısı ve yanıt değişkeninin dengeli ve dengesiz dağılımı dikkate alınarak 90 farklı senaryo oluşturulmuştur. Çoklu sınıflandırma problemlerinde kullanılan Accuracy, Kappa ve CramerV metrikleri performans ölçüsü olarak kullanılmıştır. Belirlenen senaryolardaki performans metriklerindeki değişimler tablolar halinde özetlenmiş ve karşılaştırılmıştır. Simülasyon çalışması ile yapılan karşılaştırmalar sonucunda, Kappa performans ölçütünün çok sınıflı sınıflandırma problemlerinde diğer iki metriğe göre daha doğru bir performans metriği olduğu ve yöntemin sınıflandırma başarısı hakkında daha güvenilir bilgi verdiği görülmüştür.

Anahtar Kelimeler: *Sınıflandırma Başarısı, Sınıflandırma Performansı, Makine Öğrenimi, Simülasyon, Performans Ölçümleri*

Jel Kodu: *C15, C38, C88*

[1] Bu çalışma, Eskişehir Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü bünyesinde Doç. Dr. Cengiz Bal danışmanlığında hazırlanan ‘‘Kronik Hepatit C Hastalığı Risk Belirlenmesinde Sıralı Lojistik Regresyon ve Makine Öğrenme Algoritmalarının Sınıflama Performansının Karşılaştırılması” başlıklı doktora tezinden üretilmiştir.

[1] **Sorumlu Yazar/Corresponding Author:** Dr. Öğr. Üyesi, Kastamonu Üniversitesi, Tıp Fakültesi, Temel Tıp Bilimleri Bölümü, Kastamonu, Türkiye. **E-posta:** avaglarci@kastamonu.edu.tr **Orcid no:** 0000-0002-9010-4537

[3] Prof. Dr., Eskişehir Osmangazi Üniversitesi, Tıp Fakültesi, Temel Tıp Bilimleri Bölümü, Eskişehir, Türkiye. **E-posta:** cengiz@ogu.edu.tr **Orcid no:** 0000-0002-1553-2902

Atıf/Citation: Ağlarcı, A. V., Bal, C. (2025), Comparison of Some Performance Metrics Used in Multiple Classification Problems, Kastamonu Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 27/1, s. 22-39.

INTRODUCTION

The general goal in machine learning is to predict an outcome using available data. If the result to be predicted is categorical, the related problem is called a classification problem, and if it is continuous, it is called a regression problem (Grandini et al., 2020).

Machine learning classification problems involving more than two classes are referred to as "multi-class classification." Evaluation of the success of the classification process is made through performance metrics. Different measurement metrics are available to test the performance of a multiclass classifier. Through these metrics, the performance of multiple classifiers (machine learning techniques) can also be compared (Grandini et al., 2020).

The most well-known classification performance metric is Accuracy (Gösgens et al., 2021). In most classification problems, the success of the classification made according to the Accuracy value is interpreted (Chen et al., 2020; Huang et al., 2009; Jeong et al.,2020). However, it has been reported that evaluating performance based on Accuracy value alone may lead to misinterpretations. It has been stated in studies in the literature that different metrics give different results (Ferri et al., 2009; Hossin & Sulaiman, 2015; Luque et al., 2019; Powers, 2011). In these studies, different metrics were compared in binary and multiple classification problems using some real data sets and presented to the literature (Rácz et al., 2019; Pereira et al., 2017; Patel & Markey, 2005; Ballabio et al., 2018). In some studies, it has been observed that new metrics have been developed in addition to known metrics (De Diego et al., 2022; Mingxing, 2021). One of the important elements in classification problems is balanced and unbalanced data sets. When the studies in the literature were examined, performance metrics were also compared in unbalanced data sets (Fatourechi et al., 2008; Jeni et al., 2013; Gu et al., 2009; Luque et al., 2019).

The aim of this research, unlike other studies, is not to compare classification and metrics using several data sets, but to conduct a comprehensive simulation study in which various features of the data set are taken into account. It is aimed to compare 3 performance metrics used for multi-class problems, taking into account the different characteristics of the data set. These are the Accuracy, Kappa, and CramerV metrics. It is seen that mostly F1 score, recall, precision and ROC metrics are used for binary classification (Dhasaradhan & Jaichandran, 2022; Folorunso et al., 2022; Kumar et al., 2019). However, since these metrics have to be calculated for each category in multiple classification problems, calculation and interpretation difficulties arise. Since multiple classifications were made in our study, these metrics were not used.

In the simulation study, 3 performance metrics were compared under different scenarios such as the correlation structure of the data set, sample size, number of response variable categories, balanced and unbalanced distribution. More than one classifier was used when comparing the three metrics. The methods used are K-nearest neighbor, random forest, ordinal logistic regression, decision tree (CART) methods.

The confusion matrix is used to calculate many metrics. In the field of machine learning, and especially in the statistical classification problem, the confusion matrix, also known as the error matrix, is a special tabular layout that allows the performance of an algorithm to be visualized. Each row of the matrix represents instances in a real class, while each column represents instances in a predicted class and vice versa (Stehman, 1997; Powers, 2011). The name of the confusion matrix comes from the fact that it makes it easy for the system to tell if it isconfusing classes (i.e. often mislabeling one with the other). In our study, the confusion matrix was used in the calculation of performance metrics.

1. MATERIAL AND METHOD

In the study, classification was made using four methods and performance metrics were calculated. These are K-nearest neighbor (KNN), random forest (RF), ordinal logistic regression (OLR), decision tree (CART) methods.

The KNN method, which is one of the machine learning classification methods, is a classification method that determines the class in which the observations will take place and the nearest neighbor according to the k-value. KNN makes classification with the help of distance or proximity calculation. The purpose of the method is to assign individuals or objects to predetermined classes or groups in the most accurate way, by making use of the properties of these objects. With the help of the learning data set, the observation to be classified is classified in the same data set with the k closest observations and the most similar ones. More detailed information about the theoretical infrastructure of the KNN method can be obtained from the relevant source (Bridge, 2013; Bishop, 2007).

The RF method is an ensemble learning algorithm. In Ensemble Learning, the results produced by more than one classifier are combined to produce a single result on behalf of the ensemble. In these methods, the predictions produced by more than one classifier are voted and the class with the most votes is given as the class prediction of the community. Decision trees are the basis of the Random Forest method proposed by Breiman (2001). More detailed information about the theoretical infrastructure of the RF method can be obtained from the relevant source (Breiman, 2001).

The OLR method is a logistic analysis method. Logistic regression analysis examines the relationship between the dependent variable and one or more independent variables. In logistic regression analysis, the independent variables can be categorical or continuous, while the dependent variable is categorical. In logistic regression analysis, binary logistic regression analysis is applied when the dependent variable has two categories. Multiple logistic regression analysis is used when the dependent variable has more than two categories. In this technique, there is no restriction on the number of categories for the dependent variable and there is no natural ordering. In ordinal logistic regression analysis, on the other hand, the dependent variable consists of more than two categories and the categories are ordered in a hierarchical order. There is no strict rule about whether the independent variables are categorical or continuous. More detailed information about the theoretical background of the OLR method can be obtained from the book “Applied Logistic Regression” by Hosmer et al. (2013).

The CART method is an algorithm used to create a decision tree. It is used for both classification and regression purposes. The CART algorithm is based on entropy and uses Twoing and Gini techniques to calculate the branching criterion. The CART algorithm allows the relevant group to be divided into two more homogeneous subgroups. In other words, each branch is divided into subgroups and the decision tree grows. In the separation process, Gini or Twoing is used if the dependent variable is categorical, and least squares deviation is used if it is continuous. Here, variability in independent variable impurity and variation measures (Gini, Twoing, least squares deviation) is used to produce homogeneous groups of the dependent variable. There is no limitation on the data types of dependent and independent variables in the CART algorithm. Both dependent and independent variables can be categorical (ordinal/nominal) or continuous data type. In the CART algorithm, if the dependent variable is categorical, a classification tree is formed, and if it is continuous, a regression tree is formed. The main aimfor this method is to separate the units in such a way that homogeneous classes are formed at the decision points. More detailed information about the theoretical infrastructure of the RF method can be obtained from the relevant source (Breiman et al., 1993).

While comparing performance metrics in the study, the data to be used for classification purposes were derived through simulation under different scenarios, taking into account the effect of 4 factors. The scenarios are summarized in Table 1. 3 different categories of response variable (3, 4 and 5 categories), 5 different sample sizes (100, 250, 500, 1000, 2500), 3 different correlation structures (low, medium, high), cases where the response variable is distributed balanced and unbalanced 90 different scenarios were created by taking this into consideration. The correlation between response variable and independent variables was determined as low, medium and high levels. The correlation between independent variables was determined at a low level. Using the "BinOrdNonNor" package in the R program, 10 variables (1 ordinal, 3 binary, 6 continuous) were derived. The response variable is the ordinal variable. In each scenario, 75% of the derived data was used as training data and 25% as test data. The classification steps are shown in Figure 1. Each scenario was repeated 1000 times and the average of the performance values was taken. All calculations were performed using the R program.

Table 1: Simulation Scenarios

Correlation Structure	Sample Size	Response Variable Number of Categories	Response Variable Category Distribution
*Low [0-0.3] *Medium [0.4-0.6] *High [0.7-0.9]	*100 *250 *500 *1000 *2500	*3	Equal (0.33:0.33:0.33) Not Equal (0.6:0.3:0.1)
		*4	Equal (0.25:0.25:0.25:0.25) Not Equal (0.5:0.25:0.15:0.1)
		*5	Equal (0.2:0.2:0.2:0.2:0.2) Not Equal (0.45:0.2:0.15:0.1:0.1)

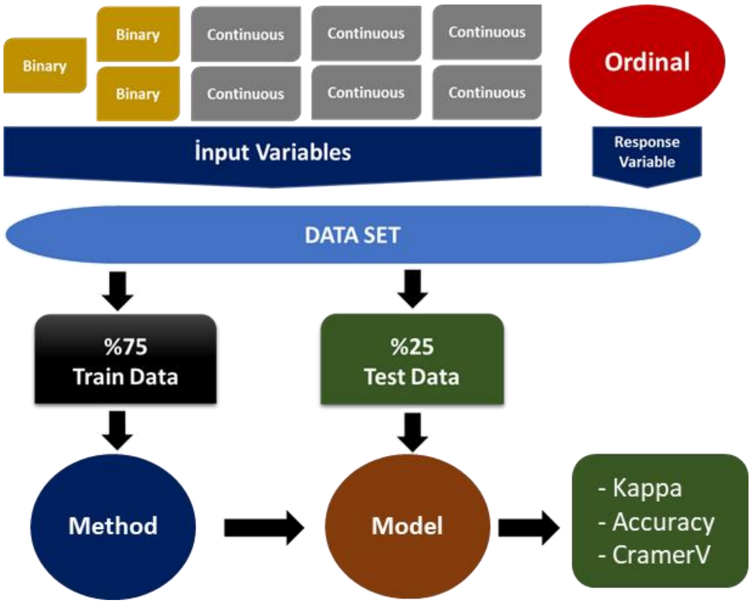

Figure 1: Classification Steps

1.1. Performance Measures

Performance metrics used are Kappa, Accuracy and CramerV coefficient. In addition to this, the number of failure to estimate any class in the classification process in 1000 repetitions was also recorded and expressed with Nan.

The Kappa coefficient gives information about the classification performance of a classifier. The Kappa coefficient ranges from -1 to +1. It is stated that there is no fit when it is less than zero, and the performance of the classifier increases as it gets closer to 1 (Landis & Koch, 1977; McHugh, 2012).

The Accuracy value is a performance measure that is calculated by dividing the number of correctly classified numbers by the total (Metz, 1978).

Cramer's V coefficient measures the strength of the relationship between two variables in an IxJ-dimensional confusion matrix, independent of the number of rows and columns, and takes values between 0 and 1. 0 indicates no relationship and 1 indicates full relationship in square type tables. When considered as a performance measure for classification purposes, one of the two variables refers to the actual class variable and the other to the predicted class variable. In the CramerV coefficient formula, n represents the number of observations and k represents the number of response variable categories (Fávero et al., 2023).

How the Kappa, Accuracy and CramerV values were calculated for the three categories is shown with the help of the confusion matrix given in Table 2 (Equations 1, 2 and 3).

Table 2: Confusion Matrix for Three Classes

		Actual Class			
		A	**B**	**C**	**Total**
Predicted Class	**A**	Da	Yab	Yac	Ta=Da+Yab+Yac
	B	Yba	Db	Ybc	Tb=Yba+Db+Ybc
	C	Yca	Ycb	Dc	Tc=Yca+Ycb+Dc
Total		Ga=Da+Yba+Yca	Gb=Yab+Db+Ycb	Gc=Yac+Ybc+Dc	GT=Ta+Tb+Tc GT=Ga+Gb+Gc

$$Kappa = \frac{P_0 - P_e}{1 - P_e} \quad P_0 = \frac{D_a + D_b + D_c}{GT} \quad P_e = (\frac{G_a}{GT} x \frac{T_a}{GT}) + (\frac{G_b}{GT} x \frac{T_b}{GT}) + (\frac{G_c}{GT} x \frac{T_c}{GT}) \tag{1}$$

$$Accuracy = (D_a + D_b + D_c) / GT \tag{2}$$

$$CramerV = \sqrt{\frac{\chi^2}{nx(k-1)}} \tag{3}$$

2. RESULTS

Performance metric values obtained using four different classification methods are given in Tables 3, 4, 5 and 6. Accuracy, Kappa, CramerV and Nan values are given in the tables according to the level of correlation, the balanced and unbalanced distribution of the response variable, the change in the category of the response variable (3, 4, 5), and the change in the sample size (100, 250, 500, 1000, 2500). Performance metrics shown in the tables are the average of 1000 repetitions. The Nan values in the tables represent the number of times that the classifier could not predict any class at all in 1000 iterations. For example, in the case of a balanced distribution in the low correlation data structure, in the scenario where the response variable has 5 categories and the sample size is 100, it is seen that no category of the response variable can be predicted at all in 109 of the 1000 repetitive classifications made with the KNN method (Table 3).

Performance metrics as a result of classification with the KNN method:

When we compare the performance metrics (Table 3) obtained as a result of the classification made with the KNN method, it is seen that the Accuracy (accuracy) and CramerV values are higher than the Kappa values, although it is seen that the 3 performance measures increase as the sample size increases in the scenarios where the response variable is balanced distributed in the low correlation data structure. As the number of response variable categories increased, Accuracy and Kappa values decreased, while CramerV values increased at 100 sample sizes and decreased at other sample sizes.

In scenarios where the response variable was unbalanced distributed in the low correlation data structure, it was observed that Accuracy and Kappa values increased as the sample size increased, while CramerV values first decreased and then increased. When the Nan value, which is the number of unpredictability, was examined, it was seen that the Nan values decreased as the sample size increased. It was observed that CramerV value decreased as Nan value decreased in small sample sizes (100, 250). As the number of response variable categories increased, Accuracy and Kappa values decreased, while CramerV values increased (except for 2500 sample size). When the metrics were compared among themselves according to the balanced and unbalanced distribution in the low correlation data structure, Kappa and CramerV values were higher in the balanced distribution than the unbalanced distribution, while the Accuracy values in the unbalanced distribution scenarios were higher in small sample sizes (100, 250, 500). Considering the Nan values, it was observed that it was higher in the case of unbalanced distribution than in the case of balanced distribution.

In the scenarios where the response variable is balanced distributed in the medium correlation data structure, it was observed that the metric values increased as the sample size increased. It is observed that Accuracy values are higher than CramerV values and CramerV values are higher than Kappa values. It was observed that 3 performance metrics decreased as the number of response variable categories increased. In the case of unbalanced distribution in the medium correlation data structure, it is seen that Accuracy and Kappa values increase as the sample size increases. This interpretation cannot be made for the CramerV value in all categories. When the metric values are compared in the case of balanced and unbalanced distribution in the medium correlation data structure, the accuracy values are higher in the case of unbalanced distribution, while the Kappa and CramerV values are higher in the case of balanced distribution.

Similar comments can be made in the high correlation data structure. Accuracy values were found to be higher in the case of unbalanced distribution than in the case of balanced distribution. Kappa and CramerV values are higher in cases of balanced distribution. The metrics are listed in descending order of Accuracy, CramerV, and Kappa. The difference between Accuracy Kappa in the low correlation data structure is greater than the difference in the high correlation data structure.

Table 3: The Metric Values Obtained as a Result of the Classification Made with the KNN Method

Low Correlation - Balanced Distribution

Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,697	0,780	0,836	0,883	0,928	0,655	0,730	0,793	0,849	0,906	0,612	0,696	0,765	0,823	0,888
Kappa	0,538	0,668	0,754	0,824	0,893	0,532	0,637	0,723	0,798	0,875	0,507	0,617	0,705	0,778	0,859
CramerV	0,628	0,703	0,770	0,828	0,895	0,640	0,693	0,749	0,808	0,878	0,667	0,685	0,732	0,793	0,865
Nan	2	0	0	0	0	27	0	0	0	0	109	1	0	0	0

Low Correlation - Unbalanced Distribution

Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,751	0,792	0,830	0,867	0,909	0,704	0,753	0,800	0,846	0,898	0,691	0,745	0,783	0,831	0,886
Kappa	0,469	0,578	0,663	0,739	0,825	0,509	0,603	0,683	0,758	0,841	0,533	0,626	0,687	0,758	0,838
CramerV	0,623	0,621	0,665	0,727	0,807	0,675	0,667	0,711	0,765	0,838	0,709	0,684	0,713	0,763	0,833
Nan	518	113	5	0	0	383	21	0	0	0	503	24	0	0	0

Medium Correlation - Balanced Distribution

Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,835	0,893	0,923	0,946	0,964	0,770	0,845	0,889	0,922	0,951	0,725	0,807	0,858	0,901	0,940
Kappa	0,748	0,838	0,885	0,918	0,947	0,687	0,791	0,851	0,896	0,935	0,649	0,757	0,822	0,876	0,925
CramerV	0,792	0,852	0,892	0,920	0,947	0,752	0,813	0,860	0,899	0,937	0,741	0,787	0,836	0,881	0,926
Nan	2	0	0	0	0	16	0	0	0	0	78	0	0	0	0

Medium Correlation - Unbalanced Distribution

Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,865	0,902	0,929	0,949	0,966	0,806	0,859	0,893	0,924	0,953	0,769	0,824	0,866	0,904	0,941
Kappa	0,730	0,811	0,865	0,903	0,937	0,687	0,779	0,834	0,883	0,928	0,658	0,746	0,810	0,864	0,916
CramerV	0,770	0,798	0,845	0,888	0,927	0,756	0,778	0,828	0,873	0,920	0,764	0,760	0,804	0,853	0,907
Nan	327	14	0	0	0	251	9	0	0	0	414	21	0	0	0

High Correlation - Balanced Distribution

Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,884	0,936	0,960	0,974	0,985	0,837	0,889	0,929	0,955	0,977	0,787	0,851	0,898	0,934	0,965
Kappa	0,820	0,902	0,938	0,960	0,978	0,774	0,849	0,903	0,939	0,969	0,724	0,809	0,869	0,916	0,956
CramerV	0,840	0,900	0,935	0,958	0,977	0,799	0,856	0,900	0,934	0,965	0,782	0,819	0,867	0,912	0,952
Nan	2	0	0	0	0	48	0	0	0	0	180	0	0	0	0

High Correlation - Unbalanced Distribution

Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,895	0,937	0,961	0,975	0,986	0,850	0,892	0,930	0,955	0,976	0,814	0,856	0,899	0,931	0,963
Kappa	0,806	0,886	0,931	0,956	0,976	0,763	0,834	0,893	0,931	0,964	0,721	0,792	0,856	0,903	0,947
CramerV	0,821	0,884	0,924	0,952	0,973	0,791	0,834	0,879	0,922	0,957	0,789	0,792	0,837	0,884	0,935
Nan	39	0	0	0	0	133	1	0	0	0	391	20	0	0	0

Performance metrics as a result of classification with OLR method:

The performance metrics calculated as a result of the classification made with the OLR method are given in Table 4. It was observed that Kappa values increased, while CramerV and Nan values tended to decrease as the sample size increased in low correlation data structure and balanced distribution scenarios. While CramerV values were higher than Kappa values, Accuracy values were higher than Kappa and CramerV values.

In the low correlation data structure and unbalanced distribution scenarios, Kappa and CramerV values tended to decrease as the sample size increased. While the Nan values decreased depending on the sample size in the 3 and 4 category cases, they increased in the 5 category cases. Accuracy values were higher than CramerV values and CramerV values were higher than Kappa values.

In the low correlation data structure, Accuracy values in unbalanced distribution scenarios are higher than Accuracy values in balanced distribution scenarios. Kappa values are higher in balanced distribution scenarios. To interpret for CramerV values, unbalanced distribution values were higher in small sample sizes, while balanced distribution values were mostly higher in large sample sizes.

In the balanced distribution scenarios in the medium correlation data structure, Accuracy and Kappa values increased as the sample size increased, while CramerV values decreased (the case with 4 and 5 categories). It is seen that the Accuracy values are higher than the other two metrics, and the Kappa values are lower.

In the medium correlation data structure and unbalanced distribution scenarios, Accuracy and kapa values increase as sample size increases, while Nan and CramerV values decrease. It was observed that CramerV values were positively correlated with Nan values as the sample size increased. Accuracy values were higher than Kappa and CramerV values.

In the medium correlation data structure, the Accuracy and Kappa values in the unbalanced distribution scenarios are higher than the Accuracy and Kappa values in the balanced distribution scenarios. While unbalanced distribution values are higher in 100 sample sizes for CramerV values, balanced distribution values are higher in other sample sizes.

In the high correlation data structure and balanced distribution scenarios, Accuracy and Kappa values increased as the sample size increased, while CramerV values increased in the 3-category case and decreased in the 5-category case. Accuracy values were higher than Kappa and CramerV values.

In the high correlation data structure and unbalanced distribution scenarios, Accuracy and Kappa values increased as the sample size increased, while CramerV values increased in the 3-category case and decreased in the 5-category case. While CramerV values were higher than Kappa values in small sample sizes, Kappa values were higher than CramerV values in large sample sizes. In all cases, the Accuracy values are higher than the Kappa and CramerV values.

Accuracy values in unbalanced distribution scenarios in high correlation data structure are higher than in balanced distribution scenarios. Kappa values were found to be higher in balanced distribution scenarios. CramerV values are higher in unbalanced distribution scenarios in 3-category case than in balanced distribution scenarios in 4- and 5-category cases.

Table 4: The Metric Values Obtained as a Result of the Classification Made with the OLR Method

Low Correlation - Balanced Distribution

Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,435	0,449	0,453	0,458	0,459	0,349	0,352	0,360	0,361	0,360	0,294	0,298	0,299	0,298	0,298
Kappa	0,127	0,161	0,172	0,182	0,187	0,103	0,121	0,137	0,143	0,145	0,084	0,105	0,114	0,116	0,120
CramerV	0,326	0,268	0,243	0,235	0,232	0,381	0,275	0,240	0,223	0,212	0,434	0,299	0,245	0,215	0,197
Nan	166	102	58	15	0	536	491	369	218	55	862	790	714	601	366

Low Correlation - Unbalanced Distribution

Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,589	0,604	0,608	0,610	0,611	0,488	0,504	0,505	0,507	0,507	0,450	0,457	0,460	0,458	0,460
Kappa	0,116	0,129	0,129	0,130	0,130	0,093	0,091	0,086	0,084	0,080	0,082	0,077	0,071	0,065	0,064
CramerV	0,355	0,256	0,229	0,206	0,189	0,422	0,291	0,230	0,194	0,167	0,464	0,299	0,258	0,188	NaN
Nan	611	575	494	365	169	857	757	612	424	249	994	993	998	996	1000

Medium Correlation - Balanced Distribution

Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,739	0,758	0,763	0,768	0,769	0,661	0,670	0,673	0,679	0,680	0,584	0,598	0,600	0,602	0,605
Kappa	0,604	0,635	0,643	0,651	0,653	0,542	0,558	0,563	0,571	0,573	0,473	0,494	0,499	0,502	0,506
CramerV	0,667	0,671	0,674	0,675	0,677	0,652	0,631	0,628	0,625	0,626	0,645	0,605	0,592	0,589	0,585
Nan	0	0	0	0	0	6	0	0	0	0	83	0	0	0	0

Medium Correlation - Unbalanced Distribution

Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,797	0,812	0,817	0,819	0,820	0,705	0,718	0,722	0,723	0,726	0,644	0,649	0,654	0,654	0,655
Kappa	0,609	0,645	0,654	0,661	0,662	0,537	0,562	0,569	0,573	0,577	0,483	0,499	0,509	0,511	0,512
CramerV	0,684	0,665	0,668	0,668	0,667	0,670	0,623	0,612	0,608	0,609	0,671	0,591	0,570	0,561	0,555
Nan	104	3	0	0	0	176	7	0	0	0	422	55	3	0	0

High Correlation - Balanced Distribution

Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,852	0,872	0,880	0,883	0,884	0,801	0,825	0,828	0,829	0,832	0,753	0,771	0,778	0,779	0,781
Kappa	0,774	0,805	0,817	0,822	0,824	0,729	0,763	0,766	0,769	0,773	0,683	0,709	0,718	0,719	0,723
CramerV	0,793	0,807	0,815	0,817	0,820	0,770	0,769	0,766	0,770	0,768	0,757	0,734	0,728	0,726	0,725
Nan	0	0	0	0	0	7	0	0	0	0	91	1	0	0	0

High Correlation - Unbalanced Distribution

Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,871	0,887	0,893	0,897	0,899	0,818	0,834	0,840	0,843	0,844	0,775	0,787	0,791	0,793	0,795
Kappa	0,767	0,800	0,812	0,820	0,823	0,723	0,748	0,758	0,763	0,765	0,677	0,698	0,705	0,708	0,712
CramerV	0,794	0,809	0,816	0,818	0,819	0,765	0,754	0,752	0,754	0,754	0,752	0,710	0,700	0,695	0,692
Nan	10	0	0	0	0	67	0	0	0	0	290	6	0	0	0

Performance metrics as a result of classification with the CART method:

The metric values calculated as a result of the classification made by the CART method are given in Table 5. In the low correlation data structure and balanced distribution scenarios, Accuracy and Kappa values increased, while CramerV values decreased as the sample size increased. It was observed that Accuracy and CramerV values were higher than Kappa values. It was observed that Nan values first decreased and then increased as the sample size increased. It is seen that the number of unpredictability of one of the response variable categories in the sample size of 2500 is very high. As the number of categories increases, Accuracy and Kappa values decrease, while CramerV and Nan values increase.

In the low correlation data structure and unbalanced distribution scenarios, Accuracy values increased as the sample size increased, Kappa values first increased and then tended to decrease, CramerV values decreased, and Nan values first decreased and then increased. Accuracy and CramerV values were found to be higher than Kappa values. As the number of categories increased, Accuracy and Kappa values decreased, while CramerV and Nan values increased.

In the low correlation data structure, the Accuracy values in the unbalanced distribution scenarios were observed to be higher than the values in the balanced distribution scenarios. While Kappa and CramerV values in unbalanced distribution scenarios are higher in small sample sizes, Kappa and CramerV values in balanced distribution scenarios are higher in large sample sizes.

In the medium correlation data structure and balanced distribution scenarios, Accuracy and Kappa values first increased, then tended to decrease as the sample size increased, and CramerV values decreased. Nan values first decreased and then tended to increase in case of 4 and 5 categories as the sample size increased. Accuracy and CramerV values are higher than Kappa values. As the number of categories increased, Accuracy and Kappa values decreased, Nan values increased, CramerV values increased in 100 sample sizes and decreased in other sample sizes.

In the medium correlation data structure and unbalanced distribution scenarios, it was observed that Accuracy values generally increased as the sample size increased, Kappa values first increased and then tended to decrease, while CramerV values decreased. It was observed that Nan values first decreased and then increased as the sample size increased. Nan values show an inverse relationship with Kappa values. As the number of categories increased, the Accuracy and Kappa values decreased, while the CramerV values increased in the small sample size and started to decrease as the sample size increased.

In the medium correlation data structure, the Accuracy and Kappa values in the unbalanced distribution scenarios were observed to be higher than the values in the balanced distribution scenarios. For CramerV values, while the values in the unbalanced distribution scenarios were higher in small sample sizes, the values in the balanced distribution scenarios were higher in large sample sizes.

It has been observed that Accuracy and Kappa values tend to increase and then decrease as the sample size increases in the high correlation data structure and balanced distribution scenarios, CramerV values first increase and then decrease in the 3-category case, and decrease in the 4- and 5-category case. As the sample size increased, it was observed that the Nan values first decreased and then started to increase, especially in cases with 4 and 5 categories. Accuracy values were higher than Kappa and CramerV values. It is seen that as the number of categories increases, the acccuracy and Kappa values start to decrease and the Nan values increase. It is seen that CramerV values increase in 100 sample sizes as the number of categories increases, but begin to decrease in other sample sizes.

In the high correlation data structure and unbalanced distribution scenarios, the change findings obtained for performance metrics are similar to the findings obtained in balanced distribution.

In the high correlation data structure, the Accuracy values in unbalanced distribution scenarios were higher than the values in the balanced distribution scenarios. Considering the Kappa values, while the values in the balanced distribution scenarios are higher in the 3-category case, the values in the unbalanced distribution scenarios in the 4 and 5-category cases are higher. When the CramerV values are examined, the values in the balanced distribution scenarios were observed to be higher in the other sample sizes except 100 sample sizes.

Table 5: The Metric Values Obtained as a Result of the Classification Made with the CART Method

Low Correlation - Balanced Distribution															
Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,372	0,379	0,393	0,406	0,407	0,287	0,286	0,302	0,311	0,311	0,235	0,235	0,245	0,253	0,252
Kappa	0,048	0,063	0,085	0,106	0,107	0,037	0,043	0,066	0,077	0,081	0,032	0,039	0,051	0,061	0,064
CramerV	0,297	0,190	0,169	0,161	0,152	0,366	0,230	0,186	0,160	0,115	0,429	0,266	0,204	0,175	NaN
Nan	10	0	7	340	758	55	0	44	848	997	170	2	69	982	1000
Low Correlation - Unbalanced Distribution															
Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,508	0,511	0,546	0,589	0,599	0,397	0,401	0,442	0,491	0,500	0,336	0,342	0,390	0,446	0,451
Kappa	0,051	0,065	0,081	0,054	0,016	0,047	0,054	0,060	0,027	0,004	0,033	0,048	0,055	0,020	0,001
CramerV	0,298	0,194	0,155	0,137	NaN	0,376	0,237	0,183	0,146	NaN	0,438	0,274	0,205	NaN	NaN
Nan	312	57	52	775	1000	333	57	101	973	1000	584	90	241	1000	1000
Medium Correlation - Balanced Distribution															
Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,550	0,581	0,593	0,590	0,580	0,446	0,467	0,479	0,473	0,462	0,364	0,388	0,399	0,394	0,386
Kappa	0,316	0,368	0,389	0,383	0,370	0,252	0,286	0,303	0,296	0,282	0,195	0,231	0,246	0,240	0,232
CramerV	0,461	0,454	0,456	0,448	0,431	0,470	0,425	0,417	0,397	0,377	0,497	0,410	0,388	0,363	0,345
Nan	4	0	0	0	0	36	1	0	2	33	150	2	0	31	207
Medium Correlation - Unbalanced Distribution															
Category	3					4					5				
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,648	0,668	0,682	0,684	0,682	0,532	0,548	0,564	0,565	0,563	0,466	0,480	0,495	0,502	0,502
Kappa	0,325	0,375	0,400	0,398	0,386	0,260	0,300	0,321	0,312	0,299	0,231	0,260	0,277	0,271	0,256
CramerV	0,473	0,447	0,447	0,437	0,421	0,509	0,426	0,409	0,388	0,368	0,538	0,420	0,386	0,350	0,325
Nan	210	24	0	3	34	273	21	2	16	228	848	53	22	224	785
High Correlation - Balanced Distribution															

Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,595	0,632	0,647	0,647	0,641	0,488	0,526	0,542	0,542	0,538	0,426	0,454	0,471	0,472	0,469
Kappa	0,381	0,441	0,464	0,462	0,449	0,302	0,358	0,378	0,375	0,361	0,265	0,306	0,326	0,322	0,307
CramerV	0,488	0,501	0,512	0,501	0,481	0,493	0,466	0,459	0,436	0,409	0,518	0,446	0,426	0,396	0,366
Nan	7	0	0	0	1	66	0	3	6	115	200	4	2	53	557
High Correlation - Unbalanced Distribution															
Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,650	0,682	0,694	0,694	0,687	0,551	0,578	0,594	0,598	0,595	0,504	0,525	0,545	0,554	0,556
Kappa	0,372	0,438	0,462	0,458	0,436	0,309	0,361	0,384	0,379	0,358	0,281	0,325	0,350	0,345	0,324
CramerV	0,503	0,495	0,510	0,498	0,479	0,513	0,461	0,452	0,430	0,400	0,538	0,445	0,419	0,386	0,350
Nan	53	1	0	0	0	187	3	1	13	248	443	49	20	193	866

The metric values calculated as a result of the classification made by the RF method are given in Table 6. In the low correlation data structure and balanced distribution scenarios, Accuracy and Kappa values tended to increase as sample size increased, while CramerV values decreased. As the number of categories increased, the Accuracy and Kappa values decreased, while the CramerV values increased for 100, 250 and 500 sample sizes and decreased for 1000 and 2500 sample sizes. Nan values increased as CramerV values increased with the increase in the number of categories in 100 sample sizes. Kappa values were lower than the other two performance metrics.

In the low correlation data structure and unbalanced distribution scenarios, Accuracy and Kappa values tended to increase as the sample size increased, while CramerV values showed a decrease. Kappa values were found to be lower than other performance measures. As the number of categories increased, the Accuracy and Kappa values decreased, while the CramerV values increase, except for the 2500 sample size.

In the low correlation data structure, the Accuracy values in the unbalanced distribution scenarios were found to be higher than the values in the balanced distribution scenarios. Kappa values in balanced distribution scenarios are higher than the values in unbalanced distribution scenarios. CramerV values, on the other hand, are higher in unbalanced distribution scenarios in small sample sizes, while values in balanced distribution scenarios are higher in large sample sizes.

In the medium correlation data structure and balanced distribution scenarios, Accuracy and Kappa values increased as the sample size increased, CramerV values increased in the 3-category case, first decreased and then increased in the 4- and 5-category cases. It was observed that as the number of categories increased, all three performance measures decreased. Kappa values were lower than the other metrics.

In the medium correlation data structure and unbalanced distribution scenarios, Accuracy and Kappa values increased as the sample size increased, while CramerV values first decreased and then tended to increase. Kappa values were observed to be lower than the other metrics. It was observed that all three performance measures decreased as the number of categories increased (excluding 100 sample sizes for CramerV values).

In the medium correlation data structure, Accuracy values in unbalanced distribution scenarios are higher than the values in balanced distribution scenarios. When we look at the Kappa values, while the values in the balanced distribution scenarios are mostly higher in the 3 and 4 category cases, the values in the unbalanced distribution scenarios in the 5-category case are higher. For CramerV values, while the values in the unbalanced distribution scenarios were higher at 100 sample sizes, the values in the balanced distribution scenarios were found to be higher in other sample sizes.

In the high correlation data structure and balanced distribution scenarios, performance metric values increased as the sample size increased and decreased as the number of categories increased (except for CramerV values in 100 samples). Accuracy values were higher than CramerV values and CramerV values were higher than Kappa values.

Accuracy and Kappa values increased as the sample size increased in the high correlation data structure and unbalanced distribution scenarios, while CramerV values increased in the 3-category case and first decreased and then increased in the 5-category case. The increase in the number of categories mostly decreased the performance metrics.

In the high correlation data structure, the Accuracy values in the unbalanced distribution scenarios were found to be higher than the values in the balanced distribution scenarios. When Kappa values are examined, Kappa values in balanced distribution scenarios are observed to be higher than the values in unbalanced distribution scenarios. Similar interpretations cannot be made for CramerV values for all scenarios.

Table 6: The Metric Values Obtained as a Result of the Classification Made by the RF Method

Low Correlation - Balanced Distribution															
Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,415	0,411	0,420	0,424	0,429	0,313	0,319	0,321	0,323	0,329	0,259	0,256	0,259	0,262	0,269
Kappa	0,103	0,108	0,126	0,132	0,143	0,063	0,083	0,091	0,096	0,105	0,053	0,060	0,069	0,076	0,085
CramerV	0,311	0,222	0,198	0,188	0,187	0,376	0,250	0,204	0,181	0,169	0,438	0,278	0,213	0,178	0,159
Nan	10	0	0	0	0	78	1	0	0	0	198	3	0	0	0
Low Correlation - Unbalanced Distribution															
Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,591	0,584	0,588	0,592	0,595	0,475	0,477	0,482	0,483	0,487	0,422	0,425	0,429	0,434	0,437
Kappa	0,086	0,087	0,093	0,098	0,102	0,060	0,071	0,074	0,075	0,079	0,053	0,063	0,067	0,068	0,070
CramerV	0,331	0,230	0,186	0,158	0,148	0,404	0,267	0,201	0,167	0,142	0,452	0,296	0,222	0,173	0,138
Nan	696	509	315	132	12	758	368	119	18	0	860	463	175	32	0
Medium Correlation - Balanced Distribution															
Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,667	0,690	0,706	0,722	0,735	0,550	0,579	0,592	0,610	0,628	0,468	0,491	0,511	0,525	0,542
Kappa	0,491	0,532	0,558	0,582	0,602	0,388	0,435	0,454	0,479	0,503	0,322	0,358	0,386	0,405	0,427
CramerV	0,573	0,589	0,601	0,619	0,633	0,557	0,538	0,543	0,555	0,570	0,561	0,510	0,502	0,511	0,524
Nan	2	0	0	0	0	34	0	0	0	0	131	3	0	0	0
Medium Correlation - Unbalanced Distribution															
Category			3					4					5		

N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,740	0,763	0,771	0,784	0,794	0,632	0,647	0,658	0,668	0,683	0,570	0,579	0,586	0,594	0,607
Kappa	0,462	0,527	0,551	0,581	0,604	0,383	0,425	0,451	0,471	0,499	0,336	0,370	0,388	0,406	0,429
CramerV	0,601	0,563	0,577	0,593	0,613	0,603	0,529	0,526	0,533	0,545	0,614	0,506	0,485	0,481	0,487
Nan	430	101	5	0	0	450	56	0	0	0	622	112	6	0	0
High Correlation - Balanced Distribution															
Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,734	0,770	0,789	0,809	0,829	0,642	0,673	0,694	0,716	0,740	0,566	0,602	0,619	0,639	0,665
Kappa	0,586	0,646	0,677	0,708	0,739	0,503	0,551	0,581	0,612	0,646	0,435	0,486	0,510	0,537	0,572
CramerV	0,648	0,664	0,687	0,706	0,737	0,618	0,604	0,615	0,633	0,658	0,616	0,567	0,567	0,580	0,604
Nan	17	0	0	0	0	118	1	0	0	0	290	16	0	0	0
High Correlation - Unbalanced Distribution															
Category			3					4					5		
N	100	250	500	1000	2500	100	250	500	1000	2500	100	250	500	1000	2500
Accuracy	0,768	0,795	0,817	0,833	0,852	0,686	0,705	0,720	0,737	0,759	0,644	0,650	0,664	0,679	0,695
Kappa	0,556	0,620	0,666	0,698	0,735	0,480	0,528	0,557	0,588	0,627	0,445	0,471	0,498	0,525	0,554
CramerV	0,643	0,651	0,682	0,707	0,740	0,627	0,591	0,600	0,614	0,637	0,638	0,556	0,542	0,546	0,562
Nan	101	5	0	0	0	293	25	2	0	0	665	153	18	0	0

CONCLUSION

In this study, a comparison of the metrics that provide information about the performance of classifications made with machine learning methods was made. Three performance metrics were compared by performing a simulation study for multiple classification problems. Four different classification methods were used in the study, in which sample size, correlation structure, number of categories and category distribution changes were taken into account.

F1 Score, Precision, Sensitivity (Recall), etc. used in binary classification problems in the research metrics were not evaluated. Since these metrics are calculated separately for each category, computational difficulty (response variable becomes more complex as the number of categories increases) and evaluation complexity occur for multiple classification problems.

According to the simulation results, the findings of the metric values obtained as a result of classification with different machine learning methods are as follows:

- Looking at the tables, it was seen that all three metrics were affected by the sample size, the number and distribution of the response variable categories, and the correlation structure.
- When all scenarios of the simulation study are examined, the fact that Accuracy values are higher in unbalanced distribution scenarios and Nan values are higher in unbalanced distribution scenarios shows that this metric can be misleading in the measurement of classification success. It is seen in the literature that the accuracy value is used as a measure of classification performance in most classification studies (Chen et al., 2020; Huang et al., 2009; Jeong et al.,

2020). As a result of the findings, it is recommended that the Accuracy value should not be used alone as a performance measure, especially in unbalanced distributed data sets.

• Looking at the results for Kappa values in the classification process in which four methods were used, it was observed that the values in the balanced distribution scenarios were higher than the values in the unbalanced distribution scenarios, although not in all scenarios. When the CramerV values are examined, it is observed that while the unbalanced distribution values are higher in small sample sizes, the values in the balanced distribution scenarios are higher in large sample sizes. The fact that Nan values are higher in unbalanced distribution scenarios, especially in small sample sizes, than in balanced distribution scenarios shows that the Kappa metric is more reliable than the Accuracy and CramerV metrics in measuring classification success.

• Looking at the scenarios, it was seen that Accuracy values were higher than CramerV values and CramerV values were higher than Kappa values in all methods. The Accuracy metric tends to be higher than the other two metrics. The difference between metric values at low correlation levels is higher than the difference between metric values at the high correlation levels. Increasing the correlation level from low to high decreased the difference between the metric values. Nan values are higher in scenarios with low correlation data structure. As the correlation level increases, the Nan values decrease and the difference between the Kappa metric and the Accuracy metric decreases, suggesting that the Kappa metric is a more sensitive measure in interpreting the classification performance of the classifier.

• According to the classification results using the four methods, Kappa and Accuracy values increased with the increase in sample size, while CramerV values decreased in some scenarios. However, Nan values decreased as the sample size increased. It is expected that the classification performance will increase with the decrease of the nan value, that is, the number of unpredictability. However, the decrease in CramerV values along with Nan values in some scenarios (especially in unbalanced distribution scenarios) suggests that this metric may be misleading in interpreting the classification performance.

• When the tables are examined, it is seen that Nan values increase with the increase in the number of categories. While Accuracy and Kappa values decreased as the number of categories increased, it was observed that CramerV values increased with Nan values in small sample sizes (100 and 250) in some scenarios, which could be misleading as a classification performance metric.

• When we look at the relationship between Kappa and Nan values, it is seen that there is a negative relationship. It can be said that the response of Kappa values to the change in Nan values is more reliable when compared to the other two metrics. For example, when we look at the scenarios in the classification made by the CART method, mostly Nan values first decrease and then increase according to the sample size, while the opposite is observed in Kappa values. While the Nan values are high in unbalanced distributions, the fact that Kappa values are mostly observed high in balanced distributions is an indicator for this.

• In addition, considering the Nan values (number of unpredictability), it can be seen in line with the findings that the Kappa performance metric is a more accurate measure for classification performance measurement than the other two metrics. The fact that Nan values are higher in unbalanced distribution scenarios, Accuracy values are high in the unbalanced distribution data structure, and the change in CramerV values is positive with the change of Nan values in some scenarios supports this conclusion about the Kappa metric.

As a result, in this research, three performance metrics used in multi-class classification problems were compared by performing a simulation study. As a result of the findings listed above, it is thought that the Kappa metric is a more accurate performance metric than the other two metrics, and it gives more reliable information about the classification success of the method. It is recommended to use the Kappa metric to measure the performance of the classifier for multiple classification problems.

ETİK BEYAN VE AÇIKLAMALAR

Etik Kurul Onay Bilgileri Beyanı

Çalışma, etik kurul izni gerektirmeyen bir çalışmadır.

Yazar Katkı Oranı Beyanı

Yazarlar tüm çalışmaları birlikte yürütmüştür.

Çıkar Çatışması Beyanı

Çalışmada potansiyel bir çıkar çatışması bulunmamaktadır.

REFERENCES

Ballabio, D., Grisoni, F. & Todeschini, R. (2018). Multivariate Comparison of Classification Performance Measures. *Chemometrics and Intelligent Laboratory Systems*, *174,* 33-44. https://doi.org/10.1016/j.chemolab.2017.12.004.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. New York: Springer. ISBN: 0-387- 31073-8.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. F., Jerome, O. A., Richard, S. J. & Stone, C. (1993). *Classification and Regression Trees*. New York: Chapman & Hall.

Bridge, D. (2013). *Classification: K-nearest Neighbours*. Online Courses. Retrieved from www.cs.ucc.ie/~dgb/courses/tai/notes/handout4.pdf, Accessed time: 12.08.2024.

Chen, P., Lien, C., Wu, W., Lee, L. & Shaw, J. (2020). Gait-Based Machine Learning for Classifying Patients with Different Types of Mild Cognitive Impairment. *Journal of Medical Systems, 44*(6),107-120.

De Diego, I. M., Redondo, A. R., Fernández, R. R., Navarro, J. & Moguerza, J. M. (2022). General Performance Score for Classification Problems. *Applied Intelligence*, *52*(10), 12049-12063.

Dhasaradhan, K. & Jaichandran, R. (2022). Performance Analysis of Machine Learning Algorithms in Heart Disease Prediction. *Concurrent Engineering, 30*(4), 335-343.

Fatourechi, M., Ward, R. K., Mason, S. G., Huggins, J., Schlögl A. & Birch, G. E. (2008). Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets. *Seventh International Conference on Machine Learning and Applications 2008*, 777-782.

Fávero, L.P., Belfiore, P. & Souza, R.F. (2023). *Bivariate Descriptive Statistics.* In: L. P. Fávero, P. Belfiore & R. F. Souza (Eds.), Data Science, Analytics and Machine Learning with R (pp. 63-71). Academic Press. https://doi.org/10.1016/B978-0-12-824271-1.00003-2.

Ferri, C., Hernández-Orallo, J. & Modroiu, R. (2009). An Experimental Comparison of Performance Measures for Classification. *Pattern Recognition Letters*, *30*(1), 27–38.

Folorunso, S. O., Awotunde, J. B., Adeniyi, E. A., Abiodun, K. M. & Ayo, F. E. (2022). *Heart Disease Classification Using Machine Learning Models*. In: S. Misra, J. Oluranti, R. Damaševičius & R. Maskeliunas (Eds.), Communications in Computer and Information Science (pp. 35-49). Springer, Cham. https://doi.org/10.1007/978-3-030-95630-1_3

Gösgens, M., Zhiyanov, A., Tikhonov, A. & Prokhorenkova, L. (2021). Good Classification Measures and How to Find Them. *35th Conference on Neural Information Processing Systems 2021*, 1-12.

Grandini, M., Bagli, E. & Visani, G. (2020) Metrics for Multi-Class Classification: An Overview. arXiv 2020, (1-17). https://doi.org/10.48550/arXiv.2008.05756.

Gu, Q., Zhu, L. & Cai, Z. (2009). *Evaluation Measures of the Classification Performance of Imbalanced Data Sets*. In: Z. Cai, Z. Li, Z. Kang & Y. Liu (Eds.), Communications in Computer and Information Science (pp. 461-471). Berlin: Springer. https://doi.org/10.1007/978-3-642-04962-0_53

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied Logistic Regression*. New York: John Wiley and Sons.

Hossin, M. & Sulaiman, M. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1-11.

Huang, C., Yang, Y., Yang, D. & Chen, Y. (2009). Frog Classification Using Machine Learning Techniques. *Expert Systems with Applications*, *36*(2), 3737-3743.

Jeni, L. A., Cohn, J. F. & Torre, F. D. (2013). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. *Humaine Association Affective Computing and Intelligent Interaction Conference 2013*, 245-251.

Jeong, B., Cho, H., Kim, J., Kwon, S., Hong, S., Lee, C. & Heo, T. (2020). Comparison between Statistical Models and Machine Learning Methods on Classification for Highly Imbalanced Multiclass Kidney Data. *Diagnostics*, *10*(6), 415.

Kumar, A., Sushil, R. & Tiwari, A. K. (2019). Significance of Accuracy Levels in Cancer Prediction Using Machine

Learning Techniques. *Bioscience Biotechnology Research Communications*, *12*(3), 741-747.

Landis, J. R. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159-174.

Luque, A., Carrasco, A., Martín, A. & Heras, A. (2019). The Impact of Class Imbalance in Classification Performance Metrics Based on the Binary Confusion Matrix. *Pattern Recognition*, *91*, 216–231.

McHugh, M. L. (2012). Interrater Reliability: The Kappa Statistic. *Biochemia Medica*, *22*(3), 276–282.

Metz, C. E. (1978). Basic Principles of ROC Analysis (PDF). *Seminars in Nuclear Medicine*, *8*(4), 283–298. doi:10.1016/s0001-2998(78)80014-2.

Mingxing, G. (2021). A Novel Performance Measure for Machine Learning Classification. *International Journal of Managing Information Technology*, *13*(1), 1-19.

Patel, A. C. & Markey, M. K. (2005). Comparison of Three-Class Classification Performance Metrics: A Case Study in Breast Cancer CAD. Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment 2005. https://doi.org/10.1117/12.595763

Pereira, L. & Nunes, N. (2017). A Comparison of Performance Metrics for Event Classification in Non-Intrusive Load Monitoring. *IEEE International Conference on Smart Grid Communications 2017*, 159-164.

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Rácz, A., Bajusz, D. & Héberger, K. (2019). Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. *Molecules*, *24*(15), 1-18.

Stehman, S. V. (1997). Selecting and Interpreting Measures of Thematic Classification Accuracy. *Remote Sensing of Environment*, *62*(1), 77–89.