Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **23** (2023) 055102 (1142-1149) AKU J. Sci. Eng. **23** (2023) 055102 (1142-1149)

DOI: 10.35414/akufemubid.1263900

Araştırma Makalesi / Research Article

Determination of the Classification Success of KNN Algorithm Distance Metric Methods on Wheat Seeds Dataset

Ahmet ÇELİK[1]

[1] *Kütahya Dumlupınar Üniversitesi, Tavşanlı Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Kütahya.*

e-mail: ahmet.celik@dpu.edu.tr ORCID ID: http://orcid.org/0000-0002-6288-3182

Geliş Tarihi: 12.03.2023 Kabul Tarihi: 25.09.2023

Keywords
Machine learning; Classification; Seeds dataset; KNN algorithm; Distance metric methods; Random sampling.

Abstract

Machine learning algorithms are widely used in product sorting processes in the food industry. The attributes of the products are used in the classification process. Attributes vary for each product. In this study, using the k nearest neighbor (KNN) algorithm, the classification of the wheat groups of Kama, Rosa and Canada was performed. The Seeds dataset provided in UCI (University of California, Irvine) machine learning open source data storage was used. There are 70 examples of each wheat class in the data set. In addition, the classification estimation success of distance metrics and the number of training data was measured. Each of the wheat samples was randomly selected and a soft X-ray technique was used to visualize the inner core structure of the wheat in the experimental environment with high quality. According to the training rates ranging from 50% to 90% of the data set, the classification success of the KNN algorithm was tested. In the KNN algorithm, the neighborhood values 1, 3 and 5 were selected to affect the classification success. The successes of the Euclidean, Chebyshev, Manhattan and Mahalanobis distance metric methods of the KNN algorithm were tested according to each k neighborhood value. According to the results obtained, with the Mahalanobis metric method, a classification success rate of 0.9924 accuracy was obtained according to the AUC (Area Under the Curve) success metric by using the neighborhood value of k = 3. In the literature, there is no study comparing the KNN algorithm, neighborhood values and distance vectors together on food data sets using varying training and test data. Therefore, it is thought that the study will make an important contribution to the literature.

KNN Algoritması Uzaklık Metrik Yöntemlerinin Buğday Tohumları Veri Seti Üzerinde Sınıflandırma Başarısının Tespit Edilmesi

Anahtar kelimeler
Makine öğrenmesi; Sınıflandırma; Seeds veri seti; KNN algoritması; Uzaklık metrik yöntemleri Rastgele örnekleme.

Öz

Makine öğrenmesi algoritmaları, gıda sektöründe ürün sınıflandırma işlemlerinde yaygın olarak kullanılmaktadır. Sınıflandırma işleminde ürünlerin öznitelikleri kullanılmaktadır. Öznitelikler her ürüne göre değişiklik göstermektedir. Bu çalışmada, k en yakın komşu (KNN) algoritması kullanılarak, Kama, Rosa ve Kanada buğday gruplarının sınıflandırması gerçekleştirilmiştir. UCI (University of California, Irvine) makine öğrenme açık kaynak veri depolama alanında temin edilen Seeds veri seti kullanılmıştır. Veri setinde her buğday sınıfına ait 70 örnek mevcuttur. Ayrıca uzaklık metriklerinin ve eğitim veri sayısının sınıflandırma tahmin başarısı ölçülmüştür. Her bir buğday örneği rastgele seçilerek, deney ortamında buğdayların iç çekirdek yapısının yüksek kalitede görselleştirilmesi için yumuşak bir X-ışını tekniği kullanılmıştır. Veri setinin %50 ile %90 arasında değişen eğitim oranlarına göre KNN algoritmasının sınıflandırma başarısı test edilmiştir. KNN algoritmasında sınıflandırma başarısını etkilen k komşuluk değeri 1, 3 ve 5 seçilmiştir. Her k komşuluk değerine göre KNN algoritmasının Euclidean, Chebyshev, Manhattan ve Mahalanobis uzaklık metrik yöntemlerinin başarıları test edilmiştir. Elde edilen sonuçlara göre Mahalanobis metrik yöntemiyle, k=3 komşuluk değeri kullanılarak, AUC(Area Under the Curve: Eğri Altındaki Alan) başarı metriğine göre, 0.992 doğrulukta sınıflandırma başarısı elde edilmiştir. Literatürde, değişen eğitim ve test verileri kullanılarak gıda veri setleri üzerinde, KNN algoritmasının, komşuluk değerlerinin ve uzaklık vektörlerinin birlikte kıyaslandığı bir çalışmaya rastlanmamıştır. Bundan dolayı yapılan çalışmanın, literatüre önemli katkı sağlayacağı düşünülmektedir.

1. Introduction

The classification of food products according to their quality or qualities is of great importance in both production and consumption stages. This classification needs to be done very quickly and accurately.

Bread, which is the basic building block in the food sector, is a very important product in human nutrition all over the world. The most important raw material of bread is wheat. Wheat is also a very important grain that is used in many food sectors. Wheat and the products obtained by grinding wheat are the raw materials of many foodstuffs in the food industry. In the wheat-based food industry, raw material quality is the factor that affects the final product characteristics the most. Wheat quality is very important for the farmer who grows the wheat, the flour mill that grinds the wheat, and the producers who process the end products. In order for the wheat to be used to be of high quality and homogeneous, different species must be separated quickly (Bilgiçli and Soylu 2017).

In 2022, during the grain crisis caused by the Russian-Ukrainian war, grain shipments from Ukraine could not be made, which led to an increase in food prices in many countries and the threat of hunger in underdeveloped countries. Thanks to Turkey's mediation, grain shipments have started and the danger of hunger has decreased in food prices have been eliminated (Int. Rfn 1).

Machine learning algorithms, which are a subset of artificial intelligence techniques, are widely used in many areas (Song *et al*. 2021). Using the KNN algorithm, a machine learning algorithm, disease classification by Deivasikamani *et al*. (2022), image classification by Çelik (2022), and fault classification by Cheng and Yuan (2013) were successfully performed.

After obtaining wheat images with a camera system, it will be of great benefit to attribute these images and classify them quickly and accurately with machine learning methods. Making the right classification has a direct impact on the increase in the quality of both production and consumer products.

In the literature, there are studies comparing the Deep Neural Network Application and Classification models for the classification of wheat seeds taken from the UCI Machine Learning Repository (Eldem 2020, Yasar *et al*. 2016, Kayabasi *et al*. 2018). Margapuri *et al*. (2021) proposed an application for seed classification. They obtained 94.6% classification success rate. Özkan *et al*. (2021) proposed a smart machine learning system for classification of wheat seeds. In the study, AlexNet and VGG1 models used for classification. Çınar and Koklu (2022) performed the classification of rice species with machine learning algorithms using morphological features, shape features, color features. Thirunavukkarasu *et al*. (2018) used k-Nearest Neighbors (KNN) to classify with different tools. Sabancı and Akkaya (2016) used the WEKA program to classify the wheat seed data obtained from the UCI machine learning data repository. The classification success rate of the KNN algorithm was calculated for different number of neighboring values. When k=4 neighbor value is used, the highest success rate is 95.71%. However, they did not perform the comparison of distance metrics. Hussain and Ajaz (2015) used the Weka classification tool to classify the seed dataset with other machine learning algorithms. The classifiers used from these methods are Multilayer Perceptron, Logistics, SMO, NaiveBayes Updateable, Naïve Bayes, Bayes Net, MultiClass Classifier. Mladenova and Valova (2021) used the KNN algorithm to classify fake news and click bait headlines on Bulgarian Facebook Pages. In the study, the success of the Euclid, Manhattan, Minkowski and Chebyshev distance metrics of the KNN algorithm was tested. A fixed number was used for training and testing data. In addition, no comparison of the Mahalanobis distance metric direction was made. Dilki and Başar (2020) determined and compared the success rates of the Euclidean, Manhattan, Chebysev and Minkowski distance measures of the k-nearest neighbor algorithm in the bankruptcy estimation of enterprises.

In this study, classification of seeds dataset (Int. Rfn 2) data was performed by using KNN algorithm. The factors affecting the success of the KNN algorithm are the k neighborhood value and the distance metric methods used in the KNN algorithm.

In the study, 4 commonly used distance metric methods of the KNN algorithm were used and the training and test data were used for the test at different rates. The KNN algorithm selected the k neighborhood value 1, 3 and 5 and tested the successes of the Euclidean, Chebyshev, Manhattan and Mahalanobis distance metric methods. The test was repeated 10 times and their average success was recorded. According to the results obtained, the highest successful classification results were obtained by using the Mahalanobis distance metric method.

2. Material and Method

In this study, a data set of wheat seeds containing 210 pieces of data was successfully classified using KNN machine learning algorithm. Within the data set, there are records of the Kama, Roza and Canadian wheat classes.

Figure 1 shows a graphic of the designed model. In the study, training and test data of different sizes were selected from the Seeds dataset and classification was performed with the KNN algorithm depending on the parameters of 1, 3 and neighborhood value. The achievements of the KNN algorithm, Euclidean, Chebyshev, Manhattan and Mahalanobis distance metric methods were also compared. The selection of training and test data was randomly selected using the Random Sampling Method. In order to prove the accuracy of the study results, the classification of the training and test data selected by the Random Sampling method was applied 10 times on the separately designed model and the results were recorded.

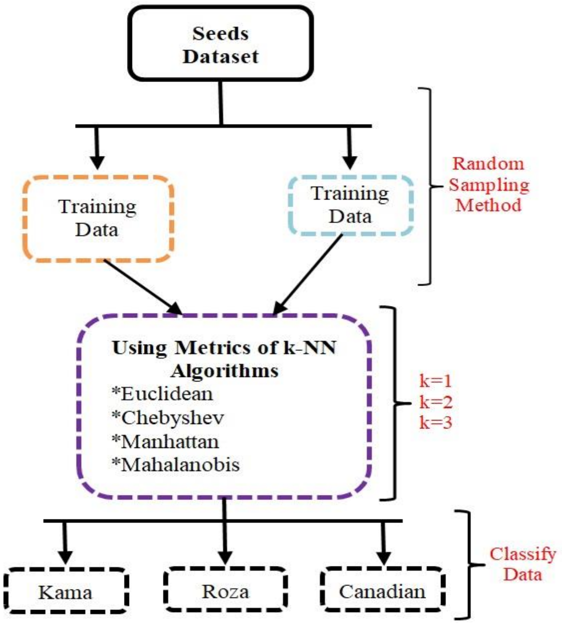

Figure 1: The designed model diagram

2.1 k-Nearest Neighbor Network

This algorithm is a classification algorithm proposed by Cover and Hart in 1967 (Cover and Hart 1967). In KNN, data is divided into subgroups. New unclassified data are classified according to their similarity to previously classified records (Taunk *et al*. 2019). This classification is classified by looking at the near neighbor value of the number K (Donuk and Hanbay 2021). The KNN algorithm commonly calculates the proximity rates of the data using Euclidean, Chebyshev, Manhattan and Mahalanobis distance metrics.

a) Euclidean distance metric

The Euclidean distance metric is shown on equation 1.

$$\mathrm{d_{Euclidean}}(\mathrm{X_i}, \mathrm{Y_i}) = \sqrt{\sum_{i=1}^{n}(\mathrm{X_i} - \mathrm{Y_i})^2} \qquad (1)$$

$\mathrm{X_i}$ is the i'th sample value, $\mathrm{Y_i}$ is the sample in the data set. n is the number of attributes. $\mathrm{d_{Euclidean}}(\mathrm{X_i}, \mathrm{Y_i})$, is the distance result $\mathrm{X_i}$ and $\mathrm{Y_i}$ (Silahtaroğlu 2016, Akbaş and Berber 2020, Durak 2011).

b) Chebishev distance metric

The maximum difference over any of the values is calculated. It is defined in equation 2 (Berber 2020, Durak 2011).

$$\mathrm{d_{L\infty}}(\mathrm{X_i}, \mathrm{Y_i}) = \max_{i=1,2,\ldots m}|\mathrm{X_i} - \mathrm{Y_i}| \qquad (2)$$

c) Manhattan distance metric

The metric is also known as the L1 norm or linear distance. This is also a commonly used distance measure. It got its name from the rectangular grid models of the streets in downtown Manhattan. It is defined in equation 3 (Durak 2011).

$$\mathrm{d}_{\mathrm{manhattan}}(\mathrm{X_i}, \mathrm{Y_i}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{X_i} - \mathrm{Y_i}| \qquad (3)$$

d) Mahalanobis distance metric

Mahalanobis distance between two samples (x, y) of a random variable is defined in equation 4 (Durak 2011).

$$\mathrm{D}_{Mahalanobis}(\mathrm{X_I}, \mathrm{Y_I}) = \sqrt{(\mathrm{X_I} - \mathrm{Y_I})^{\mathrm{T}} \sum^{-1}(\mathrm{X_I} - \mathrm{Y_I})} \qquad (4)$$

$\sum^{-1}$ is the inverse of covariance matrix (Durak 2011).

2.2 Seeds dataset

UCI (University of California, Irvine) is widely used by researchers. UCI library is open source data source. Many data sets can be accessed for classification and prediction in the library (Dua and Graff 2019). In the seeds data set, there are records of 210 wheat products belonging to 3 classes. The dataset includes classes Kama, Roza, and Canadian, each with 70 records (Charytanowicz *et al*. 2010, Int. Source 2). Table 1 shows the data set properties. The attributes of the wheat belonging to the classes consist of real data. 7 real geometric attribute data are used for each wheat product as attribute; Area (A), Perimeter(P), Compactness C=4*pi/P^2, Length of Kernel, Width of Kernel, Asymmetry Coefficient and Length of Kernel Groove were used (Kayabasi *et al*. 2018, Dua and Graff 2019).

Table 1. Attributes of Seeds data sets

Classes	Attributes	Feature	Number of Samples
Kama	Area (A)		
	Perimeter(P)		
Roza	Compactness C=4*pi/P^2		
	Length of Kernel	Real	210
	Width of Kernel		
Canadian	Asymmetry Coefficient		
	Length of Kernel Groove		

3. Result and Discussion

In this study, Seeds, data set containing 210 records was used. Repeated tests (by Random Sampling method) are shown on Table 2, Table 3 and Table 4 by selecting random data 10 times on different sizes, training and test data, k neighbor values and 4 KNN algorithm distance data in the data set.

On the tables, in the Education Percentage section, the rates of the data set ranging between 60%, 70%, 80% and 90% were used for the training data.

In the Training Data and Testing Data sections, 147 training, 63 test data were used when 60% of the Seeds dataset was selected for training, 126 training, 84 test data, 70% for training, 147 training, 63 test data, 80% for training, 168 training, 42 test data and 90% for training.

In the sections shown by the numbers [1-10] in the tables, it is shown how many tests of randomly selected training and test data are applied. Then, the classification achievements were recorded separately and the average of the results was calculated in the Mean section. Thus, the accuracy of the tests has been proven.

On Table 2, the classification successes of the tests repeated 10 times by using the k=1 neighborhood value of the KNN algorithm with the 4 distance metric method are shown. The highest success rate was obtained by Mahalanobis and the lowest success rate was obtained by the Chebyshev distance metric method.

On Table 3, the classification successes of 10 repeated tests using the KNN k=3 neighborhood value are shown with 4 distance metric methods. The highest success rate was obtained by Mahalanobis and the lowest success rate was obtained by the Chebyshev distance metric method.

On Table 4, the classification successes of the test, which was repeated 10 times by using the KNN k=5 neighborhood value, with the 4 distance metric method are shown. The highest success rate was again achieved by the Mahalanobis distance metric method. However, the lowest success rate was achieved by the Euclidean method at the rate of 3 educations and the Manhattan distance metric method at the rate of 1 education.

Table 2. Classification achievements tested 10 times for k=1 neighbors of distance metrics based on different training and test dimensions

Education Percentage	Training Data	Testing Data	Distance Metrics	1	2	3	4	5	6	7	8	9	10	Mean	k neighbor value
60	126	84	Euclidean	0.865	0.915	0.948	0.875	0.882	0.882	0.922	0.917	0.931	0.919	0.905	k=1
60	126	84	Chebyshev	0.891	0.88	0.922	0.866	0.883	0.891	0.921	0.917	0.913	0.892	0.897	
60	126	84	Manhattan	0.865	0.923	0.921	0.884	0.9	0.909	0.905	0.909	0.931	0.919	0.906	
60	126	84	Mahalanobis	0.894	0.93	0.928	0.902	0.927	0.893	0.909	0.9	0.904	0.92	0.910	
70	147	63	Euclidean	0.914	0.892	0.911	0.911	0.905	0.951	0.939	0.89	0.907	0.89	0.911	
70	147	63	Chebyshev	0.902	0.904	0.9	0.888	0.892	0.928	0.939	0.879	0.894	0.904	0.903	
70	147	63	Manhattan	0.914	0.917	0.9	0.923	0.917	0.94	0.927	0.867	0.918	0.879	0.910	
70	147	63	Mahalanobis	0.965	0.917	0.886	0.872	0.867	0.939	0.964	0.891	0.931	0.925	0.915	
80	168	42	Euclidean	0.931	0.947	0.929	0.981	0.878	0.93	0.929	0.931	0.93	0.889	0.927	
80	168	42	Chebyshev	0.931	0.929	0.929	0.963	0.859	0.948	0.929	0.931	0.913	0.889	0.922	
80	168	42	Manhattan	0.931	0.947	0.929	0.981	0.861	0.93	0.948	0.933	0.93	0.868	0.925	
80	168	42	Mahalanobis	0.928	1	0.931	0.906	0.914	0.96	0.894	0.909	0.931	0.929	0.930	
90	189	121	Euclidean	0.908	0.9	0.956	0.853	0.969	0.925	0.866	0.932	0.97	0.875	0.915	
90	189	121	Chebyshev	0.908	0.9	0.904	0.853	0.969	0.925	0.866	0.932	0.97	0.844	0.907	
90	189	121	Manhattan	0.908	0.859	0.913	0.882	0.969	0.925	0.927	0.932	0.97	0.917	0.920	
90	189	121	Mahalanobis	0.913	0.964	1	0.83	1	0.962	0.97	0.797	1	0.896	0.933	

Table 3. Classification achievements tested 10 times for k=3 neighbors of distance metrics based on different training and test dimensions

Education Percentage	Training Data	Testing Data	Distance Metrics	1	2	3	4	5	6	7	8	9	10	Mean	k neighbor value
60	126	84	Euclidean	0.964	0.967	0.939	0.932	0.96	0.952	0.97	0.971	0.966	0.967	0.958	k=3
60	126	84	Chebyshev	0.972	0.975	0.94	0.931	0.948	0.955	0.979	0.959	0.96	0.955	0.957	
60	126	84	Manhattan	0.947	0.968	0.947	0.96	0.976	0.946	0.957	0.965	0.963	0.967	0.959	
60	126	84	Mahalanobis	0.979	0.969	0.949	0.97	0.966	0.975	0.969	0.973	0.971	0.964	0.968	
70	147	63	Euclidean	0.979	0.973	0.964	0.98	0.956	0.982	0.968	0.979	0.971	0.996	0.974	
70	147	63	Chebyshev	0.979	0.973	0.946	0.969	0.953	0.979	0.954	0.973	0.959	0.984	0.966	
70	147	63	Manhattan	0.973	0.961	0.962	0.967	0.97	0.98	0.969	0.981	0.977	0.995	0.973	
70	147	63	Mahalanobis	0.971	0.964	0.968	0.984	0.977	0.998	0.984	0.972	0.973	0.998	0.978	
80	168	42	Euclidean	0.966	0.957	0.975	0.958	0.976	0.994	0.976	0.972	0.959	0.966	0.969	
80	168	42	Chebyshev	0.93	0.939	0.938	0.926	0.955	0.99	0.96	0.956	0.962	0.967	0.952	
80	168	42	Manhattan	0.973	0.959	0.949	0.96	0.976	0.993	0.976	0.969	0.994	0.952	0.970	
80	168	42	Mahalanobis	0.983	0.958	0.994	0.999	0.98	0.997	0.996	0.976	0.967	0.986	0.983	
90	189	121	Euclidean	0.97	0.955	0.967	0.989	0.985	0.988	0.991	0.995	0.943	0.959	0.974	
90	189	121	Chebyshev	0.941	0.957	0.967	0.984	0.977	0.989	0.991	0.991	0.909	0.959	0.966	
90	189	121	Manhattan	0.97	0.914	0.967	0.989	0.99	0.988	0.998	0.998	0.923	0.959	0.969	
90	189	121	Mahalanobis	0.977	0.986	0.977	1	0.992	0.996	0.998	1	1	0.998	0.992	

Table 4. Classification achievements tested 10 times for k=5 neighbors of distance metrics based on different training and test dimensions

Education Percentage	Training Data	Testing Data	Distance Metrics	1	2	3	4	5	6	7	8	9	10	Mean	k neighbor value
60	126	84	Euclidean	0.957	0.983	0.971	0.974	0.979	0.972	0.986	0.989	0.978	0.959	0.974	k=5
60	126	84	Chebyshev	0.964	0.98	0.976	0.979	0.975	0.98	0.978	0.985	0.979	0.958	0.975	
60	126	84	Manhattan	0.964	0.985	0.964	0.974	0.989	0.973	0.987	0.979	0.977	0.954	0.974	
60	126	84	Mahalanobis	0.968	0.985	0.985	0.971	0.988	0.98	0.984	0.989	0.965	0.975	0.979	
70	147	63	Euclidean	0.986	0.977	0.977	0.981	0.975	0.984	0.971	0.952	0.975	0.962	0.974	
70	147	63	Chebyshev	0.986	0.977	0.971	0.978	0.964	0.986	0.976	0.961	0.972	0.957	0.972	
70	147	63	Manhattan	0.987	0.978	0.991	0.977	0.979	0.985	0.972	0.937	0.967	0.971	0.974	
70	147	63	Mahalanobis	0.992	0.982	0.991	0.996	0.982	0.995	0.976	0.957	0.979	0.975	0.982	
80	168	42	Euclidean	0.96	0.975	0.985	0.914	0.979	0.974	0.988	0.951	0.97	0.953	0.964	
80	168	42	Chebyshev	0.954	0.977	0.987	0.911	0.982	0.978	0.992	0.967	0.969	0.967	0.968	
80	168	42	Manhattan	0.963	0.978	0.988	0.922	0.983	0.974	0.992	0.958	0.974	0.957	0.968	
80	168	42	Mahalanobis	0.997	0.992	0.987	0.96	0.972	0.994	0.997	0.997	0.973	0.968	0.983	
90	189	121	Euclidean	0.986	0.965	0.973	0.99	0.978	0.993	0.971	0.992	0.954	0.962	0.976	
90	189	121	Chebyshev	0.988	0.966	0.989	0.966	0.957	0.991	0.989	0.994	0.954	0.962	0.975	
90	189	121	Manhattan	0.995	0.961	0.984	0.991	0.974	0.997	0.968	0.994	0.954	0.962	0.978	
90	189	121	Mahalanobis	1	0.974	0.993	0.993	1	1	0.996	1	0.963	0.976	0.989	

Average classification success rates for k=1 on Table 5 are given with different training data dimensions.

Table 5. The average classification success for the k=1 neighbor value of the distance metrics

	Training data dimensions (%)			
Distance Metric Name	**60**	**70**	**80**	**90**
Euclidean	0.905	0.911	0.927	0.915
Chebyshev	0.897	0.903	0.922	0.907
Manhattan	0.906	0.910	0.925	0.920
Mahalanobis	0.910	0.915	0.930	0.933

When the training data was selected at 60% (test data at 40%), the highest success rate was found to be 0.9107 by the Mahalanobis distance metric method. The lowest success rate was found to be 0.8976 by Chebyshev distance metric method.

When the training data was selected as 70% (test data as 30%), the highest success rate was found to be 0.9157 by the Mahalanobis distance metric method. The lowest success rate was found to be 0.903 by the Chebyshev distance metric method.

When the training data was selected at 80% (test data at 20%), the highest success rate was found to be 0.9302 by the Mahalanobis distance metric method. The lowest success rate was found to be 0.9221 by Chebyshev distance metric method.

When training data was selected at 90% (test data at 10%), the highest success rate was found to be 0.9332 by the Mahalanobis distance metric method. The lowest success rate was found to be 0.9071 by Chebyshev distance metric method.

Figure 2 shows a graph of the data obtained from Table 5 for the neighbor value k=1.

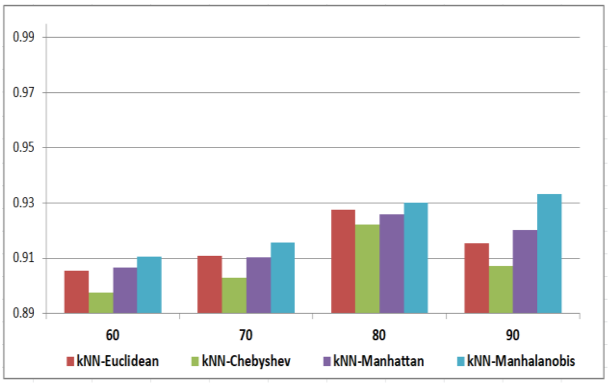

Figure 2: Classification success graph of KNN distance metric methods for training rates between 60% and 90% for k=1.

Average classification success rates for k=3 on Table 6 are given with different training data dimensions.

Table 6. The average classification success for the k=3 neighbor value of the distance metrics

	Training data dimensions (%)			
Distance Metric Name	**60**	**70**	**80**	**90**
Euclidean	0.958	0.974	0.969	0.974
Chebyshev	0.957	0.966	0.952	0.966
Manhattan	0.959	0.973	0.970	0.969
Mahalanobis	0.968	0.978	0.983	0.992

When the training data was selected as 60% (test data was 40%), the highest success rate was found to be 0.9685 by the Mahalanobis distance metric method. The lowest success rate was found to be 0.9574 by Chebyshev distance metric method.

When the training data was selected as 70% (test data as 30%), the highest success rate was found to be 0.9789 with the Mahalanobis distance metric method. The lowest success rate was found to be 0.9669 by Chebyshev distance metric method.

When the training data was selected as 80% (test data was 20%), the highest success rate was found to be 0.9836 with the Mahalanobis distance metric method. The lowest success rate was found to be 0.9523 by the Chebyshev distance metric method.

When the training data was selected as 90% (test data 10%), the highest success rate was found to be 0.9924 with the Mahalanobis distance metric method. The lowest success rate was found to be 0.9665 by Chebyshev distance metric method.

Figure 3 shows a graph of the data obtained from Table 6 for the neighbor value k=3.

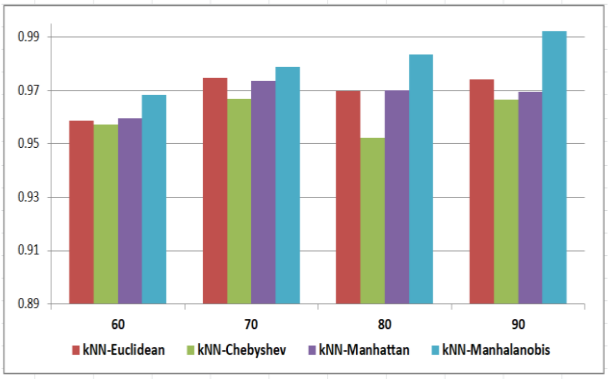

Figure 3: Classification success graph of KNN distance metric methods for training rates between 60% and 90% for k=3.

Average classification success rates for k=5 on Table 7 are given with different training data dimensions.

Table 7. The average classification success for the k=5 neighbor value of the distance metrics

	Training data dimensions (%)			
Distance Metric Name	60	70	80	90
Euclidean	0.974	0.974	0.964	0.976
Chebyshev	0.975	0.972	0.968	0.975
Manhattan	0.974	0.974	0.968	0.978
Mahalanobis	0.979	0.982	0.983	0.989

When the training data was selected as 60% (test data was 40%), the highest success rate was found to be 0.979 with the Mahalanobis distance metric method. The lowest success rate was found to be 0.974 by the Manhattan distance metric method.

When the training data was selected as 70% (test data 30%), the highest success rate was found to be 0.982 with the Mahalanobis distance metric method. The lowest success rate was found to be 0.974 by the Euclidean distance metric method.

When the training data was selected as 80% (test data as 20%), the highest success rate was found to be 0.983 with the Mahalanobis distance metric method. The lowest success rate was found to be 0.964 by the Euclidean distance metric method.

When the training data was selected as 90% (test data was 10%), the highest success rate was found to be 0.989 with the Mahalanobis distance metric method. The lowest success rate was found to be 0.976 by the Euclidean distance metric method. Figure 4 shows a graph of the data obtained from Table 7 for the neighbor value k=5.

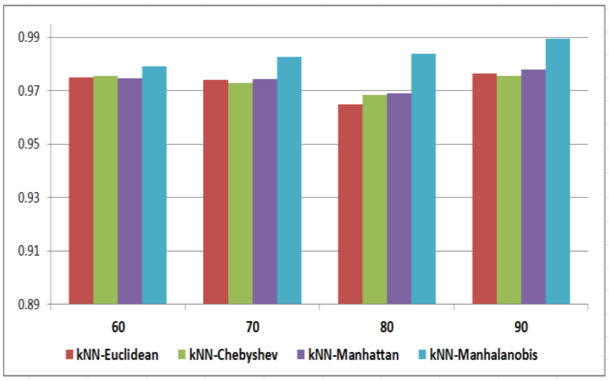

Figure 4: Classification success graph of KNN distance metric methods for training rates between 60% and 90% for k=5.

4. Conclusion

In this study, the classification success of the KNN machine learning algorithm on the wheat Seeds dataset was tested based on different sizes of training data. The Seeds dataset used in the study is shared as an open source from UCI storage. There are 210 data records in the data set, including 70 from the Kama, 70 from the Rosa and 70 from the Canadian classes.

In this study, 1, 3 and 5 were selected as the neighborhood (k) value of the KNN algorithm. The successes of the Euclidean, Chebyshev, Manhattan and Mahalanobis distance metric methods of the KNN algorithm were tested depending on each k neighborhood value.

According to the results obtained, with the Mahalanobis distance metric method, a classification success rate of 0.992 AUC was obtained when the neighborhood value of k = 3 was used. When the literature was examined, there was no study comparing the KNN algorithm both neighborhood values and Euclidean, Chebyshev, Manhattan and Mahalanobis distance metrics together on food data sets using varying education and test data. In this respect, the model developed in this study and its results will be able to serve as a source for future studies.

5. References

Akbaş, Y., Berber, T., 2020. Yanık Görüntülerinin Bulanık Kümelenmesinde Uzaklık Ölçülerinin Başarımlarının Değerlendirilmesi. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, **22**, 639-647.

Bilgiçli, N., Soylu, S., 2017. Buğday ve Un Kalitesinin Sektörel Açıdan Değerlendirilmesi. *Bahri Dağdaş Bitkisel Araştırma Dergisi*, **5**, 58-67.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Lukasik, S., Zak, S. 2010. A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images. Information Technologies in Biomedicine, Springer-Verlag, Germany, 15-24.

Cheng Z., Yuan L., 2013. The application and research of fault detection based on PC-KNN in semiconductor batch process. *25th Chinese Control and Decision Conference (CCDC)*, 4209-4214

Cover, T.M., Hart, P.E., 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory,* **13**, 21-27.

Çelik, A., 2022. Improving Iris Dataset Classification Prediction Achievement by Using Optimum k Value of

KNN Algorithm. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, **3**, 23-30.

Çınar, İ., Koklu, M., 2022. Identification of Rice Varieties Using Machine Learning Algorithms. *Journal of Agricultural Sciences*, **28**, 307-325.

Deivasikamani, G., Akshay, C., Ananthakrishnan, T., Manoj R. C., 2022. Covid Cough Classification using KNN Classification Algorithm. *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 232-237.

Dilki, G., Başar, Ö.D, 2020. İşletmelerin İflas Tahmininde K-en yakın komşu Algoriması Üzerinden Uzaklık Ölçütlerinin Karşılaştırılması. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, **19**, 224-233.

Donuk, K., Hanbay, D., 2021. Sınıflandırma Algoritmalarına Dayalı VGG-11 ile Yüzde Duygu Tanıma. *Computer Science, 5th International Artificial Intelligence and Data Processing Symposium*, 359-365.

Dua, D., Graff, C., 2019. UCI Machine Learning Repository. Irvine, CA: *University of California, School of Information and Computer Science*.

Durak, B., 2011. A Classification Algorithm Using Mahalanobis Distance Clustering of Data with Applications on Biomedical Data Sets. Master of Science in Industrial Engineering Department. Middle East Technical University, Ankara, 104.

Eldem, A., 2020. An Application of Deep Neural Network for Classification of Wheat Seeds. *European Journal of Science and Technology*, **19**, 213-220.

Kayabasi, A., Toktas, A., Sabanci, K., Yigit, E., 2018. Automatic classification of agricultural grains: Comparison of neural networks. *Neural Netw World*. **28**, 213-224.

Lal, H., Raja, A., 2015. Seed Classification using Machine Learning Techniques. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, **2**, 1098-1102.

Margapuri, V., Penumajji, N., Neilsen, M., 2021. Seed Classification Using Synthetic Image Datasets Generated from Low-Altitude UAV Imagery. *20th IEEE International Conference on Machine Learning and Applications (ICMLA 2021)*, 116-121.

Mladenova, Valova, I., Analysis of the KNN Classifier Distance Metrics for Bulgarian Fake News Detection. *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 1-4.

Özkan, K., Seke, E., Işık, Ş., 2021. Wheat kernels classification using visible-near infrared camera based on deep learning. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, **27**, 618-626.

Sabancı, K., Akkaya, M., 2016. Classification of Different Wheat Varieties by Using Data Mining Algorithms. *International Journal of Intelligent Systems and Applications in Engineering*, **4**, 40-44.

Silahtaroğlu, G., 2016. Veri madenciliği (Kavram ve algoritmaları). 3. Basım, İstanbul, Türkiye: Papatya Yayıncılık Eğitim, 118-120.

Song, L., Deng, Y.Q., Zhu, Z.L., Hua, H.L., Tao, Z. Z., 2021. A Comprehensive Review on Radiomics and Deep Learning for Nasopharyngeal Carcinoma Imaging. *Diagnostics*, **11**, 1523.

Taunk, K, De, S, Verma, S, Swetapadma, A., 2019. A brief review of nearest neighbor algorithm for learning and classification. 2*019 International Conference on Intelligent Computing and Control Systems (ICCS 2019)*, 1255–1260.

Thirunavukkarasu, K., Singh, A. S., Rai, P., Gupta, S., 2018. Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 1-4.

Yasar, A., Kaya, E., Saritas, I., 2016. Classification of Wheat Types by Artificial Neural Network. *International Journal of Intelligent Systems and Applications in Engineering*, **4**, 12-15.

References of Internet

1-)https://www.bloomberght.com/tahil-anlasmasi-icin-tarihi-imzalar-atildi-2311295 (20.02.2023).

2-)https://archive.ics.uci.edu/ml/datasets/seeds (15.01.2023).