Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 44-48, 11, 2023

Araştırma Makalesi

IJANSER
https://alls-academy.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 44-48, 11, 2023

Research Article

Buğday Tohumu Sınıflandırmasının Karar Ağacı Algoritmasıyla Gerçekleştirilmesi ve Değişken Eğitim Verisine Göre Başarı Kıyaslaması

Ahmet Çelik[1*]

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

[*]*ahmet.celik@dpu.edu.tr*

(Geliş Tarihi: 03 Aralık 2023, Kabul Tarihi: 11 Aralık 2023)

(2nd International Conference on Frontiers in Academic Research ICFAR 2023, December 4-5, 2023)

ATIF/REFERENCE: Çelik, A. (2023). Buğday Tohumu Sınıflandırmasının Karar Ağacı Algoritmasıyla Gerçekleştirilmesi ve Değişken Eğitim Verisine Göre Başarı Kıyaslaması. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(11), 44-48.

Özet – Buğday, tüm dünyada gıda alanında en önemli bir tahıl türüdür. İnsanlık için ihtiyaç olan birçok gıdanın temel yapı taşı buğdaya ve buğday alt ürünlerine bağlıdır. Buğday türlerini tanımlamak ve sınıflandırmak kalite kontrol sürecinde çok önemli bir aşamayı göstermektedir. Kalite sınıflandırmasının yapılması verimliği arttıran en önemli aşamadır. Çünkü buğday türleri, farklı biçimlerde işlenerek farklı alt ürünler oluşturulmaktadır. Buğday sınıflandırması çoğu ülkede hala çalışanlar tarafından görsel inceleme sonucu elle sınıflandırılmaktadır. Bu durum hem zaman alıcıdır hem de güvenilirlik düzeyi düşüktür. Makine öğrenme algoritmaları sınıflandırma işlemlerinde sıklıkla kullanılmaktadır. Makine öğrenme algoritmalarıyla, önceden hazırlanmış veri seti bilgilerine karar verilip sınıflandırma yapılabilmektedir. Bu çalışmada, UCI (University of California, Irvine) veri depolama alanında açık kaynak olarak yayınlanmış olan buğday tohumu (Wheat Seed) veriş eti kullanılmıştır. Bu veri seti içinde Kama, Rosa ve Kanada buğday sınıflarına ait 210 kayıt vardır. Her bir kayıt için 7 adet öznitelik bilgisi yer almaktadır. Çalışmada, buğday tohumu seti üzerinde makine öğrenme algoritmalarından biri olan C4.5 karar ağacı algoritması kullanılmıştır. Sınıflandırma başarı metriği için F1 score kullanılmıştır. F1 skor başarı metriği kullanılarak güvenilir ve tutarlı başarı sonuçları elde edilmektedir. Elde edilen sonuçlara göre en yüksek F1 skor başarı metriğine göre 0.903 bulunmuştur. Çalışmada, ek olarak farklı eğitim ve test verisi boyutlarına göre F1 skor başarı metriği sonuçları elde edilerek, kıyaslama yapılmıştır.

Anahtar Kelimeler – Makine öğrenmesi, Karar Ağacı, F1 Skor, Sınıflandırma, Öznitelik, Buğday Tohumu

I. GİRİŞ

Buğday, tüm dünyada en yaygın kullanılan bir temel tahıl grubu olduğundan, üretiminin devamlı ve yeterli düzeyde olması gerekmektedir. Dünya insan nüfusunun artışına bağlı olarak, tahıl ihtiyacı da giderek artmaktadır. Ancak son yıllarda yaşanan küresel iklim değişiklikleri verimin düşmesine sebep olmaktadır. Bununla birlikte şehirleşme, tarımsal alanların azalması ve çiftçilik yapan kişilerin tarımsal üretimden uzaklaşmaları da tahıl üretiminin azalmasına sebep olmaktadır.

Khatri vd. [1] yaptıkları çalışmada, k-NN (k en yakın komşu), Karar ağacı ve Naive algoritmasıyla Buğday Tohumu veri seti üzerinde, sınıflandırma yaparak, AUC(Doğruluk) metriğinde göre başarı kıyaslama yapmışlardır.

Çelik [2], k-NN (k en Yakın Komşu) algoritması kullanarak sınıflandırma uygulamasını yapmıştır.

Sınıflandırma işleminde %60 ile %90 arasında değişen eğitim veri oranları kullanılmıştır. Çalışmada, bu algoritmanın uzaklık metriklerinin başarıları kıyaslanmış ve Mahalanobis metrik yönteminin daha başarılı olduğu tespit edilmiştir.

Eldem [3] Derin Sinir Ağı (DNN) modeli kullanarak, buğday veri seti üzerinde sınıflandırma uygulaması yapmıştır. Yapılan çalışmada, %70 eğitim verisi ve %30 test kullanılmıştır.

II. MATERYAL VE YÖNTEM

Bu çalışmada, karar ağacı algoritmalarından biri olan C4.5 makine öğrenme algoritması kullanılmıştır. Performans değerlendirmesi için F1 skor metriği kullanılmıştır. Tasarlanan model Orange platformu üzerinde tasarlanmıştır.

A. *C4.5 Karar Ağacı Algoritması*

Sınıflandırma algoritmalarından biri olan karar ağacı, anlaşılması kolay ve hızlı çalışan bir algoritmadır[4]. Bu algoritmada, karar ağaç dalı kriterlerinin belirlenmesi, en önemli aşamadır. Bu aşamada ağaç yapısının hangi özellik yapılacağı belirlenmektedir[5]. C4.5 makine öğrenme algoritması, J. Ross Quinlan isim araştırmacı tarafından, 1993 yılında geliştirilmiştir. Makine öğrenme algoritmaları içinde en yaygın kullanılan, başarılı bir sınıflandırma algoritmasıdır[6]. C4.5 algoritması, hem kategorik hem de sayısal veri setleri kullanılabilmektedir[7].

Bu çalışmada, rastgele seçilen örnekler üzerinde tahmin işleminin başarı ölçümü gerçekleştirilmiştir. Şekil 1 üzerinde karar ağacı algoritmasıyla rastgele seçilen örnekler üzerinde yapılan tahmin başarıları görülmektedir.

Tree	error	Seed Class	İd
1.00 : 0.00 : 0.00 → Canadian	0.000	Canadian	168
1.00 : 0.00 : 0.00 → Canadian	1.000	Kama	60
0.00 : 1.00 : 0.00 → Kama	0.000	Kama	36
1.00 : 0.00 : 0.00 → Canadian	0.000	Canadian	208
0.00 : 1.00 : 0.00 → Kama	0.000	Kama	13
0.00 : 1.00 : 0.00 → Kama	0.000	Kama	33
0.02 : 0.00 : 0.98 → Rosa	0.021	Rosa	135
0.00 : 1.00 : 0.00 → Kama	0.000	Kama	69
0.02 : 0.00 : 0.98 → Rosa	0.021	Rosa	86
0.00 : 1.00 : 0.00 → Kama	0.000	Kama	15
0.02 : 0.00 : 0.98 → Rosa	0.021	Rosa	121
1.00 : 0.00 : 0.00 → Canadian	0.000	Canadian	153
1.00 : 0.00 : 0.00 → Canadian	0.000	Canadian	148

Şekil 1. Örneklerin tahmin başarıları

B. *Buğday Tohumu (Seeds) Veri seti*

UCI (California Üniversitesi, Irvine) veri deposu birçok bilimsel araştırmada, araştırmacılar tarafından kullanılmıştır. UCI veri deposu içindeki veri setleri, açık kaynaklı olarak paylaşılmaktadır. Bu açık kaynak kütüphaneden içinden sınıflandırma, kümeleme ve tahmin işlemleri için birçok veri setine kolayca ulaşılabilmektedir[8, 9].

Veri seti içindeki 210 adet buğday tohumunun Kama, Rosa ve Canadian buğday sınıf bilgileri vardır. Veri seti içinde her bir sınıfa ait 70 kayıt vardır[2].

Tablo 1 üzerinde, buğday tohumu veri seti içindeki kayıtlar hakkında bilgi verilmektedir. Veri seti içindeki her kayıt gerçek buğday örneklerinden elde edilmiştir[10]. Veri seti içindeki her bir kaydın 7 geometrik öznitelik verisi vardır. Bunlar; Alan (A), Çevre (P), Kompaktlık C=4*pi/P^2, Çekirdek Uzunluğu, Çekirdek Genişliği, Asimetri Katsayısı ve Çekirdek Oluk Uzunluğu kullanılmıştır[11,8].

Tablo 1. Veri seti kayıtlarının özellikleri [2].

Öz nitelikler	Sınıflar	Kayıt Özelliği	Kayıt Sayısı
Alan (A)	Rosa Kama Canadian	Gerçek	210
Çevre (P)			
Kompaktlık C=4*pi/P^2			
Çekirdek Uzunluğu			
Çekirdek Genişliği			
Asimetri Katsayısı			
Çekirdek Oluk Uzunluğu			

C. F1 Skor Başarı metriği

F1 Skor, ikili sınıflandırma ve ayrıca çok etiketli sınıflandırma bağlamında doğruluk ve tutarlılık açısından yüksek doğruluk oranı göstermektedir[12].

F1 skor, Kesinlik ve Duyarlılık değerlerinin harmonik ortalaması alınarak hesaplanmaktadır. Kesinlik hesabı denklem 1'de, Duyarlılık hesabı, denklem 2 de gösterilmektedir[13-15].

$$Kesinlik = \frac{(Doğru\ Pozitif)}{(Doğru\ Pozitif)+(Yanlış\ Pozitif)} \quad (1)$$

$$Duyarlılık = \frac{(Doğru\ Pozitif)}{(Doğru\ Pozitif)+(Yanlış\ Negatif)} \quad (2)$$

Tahmin işleminde, (Doğru Pozitif), doğru tahmin edilen Pozitif veri sayısını göstermektedir. $(Yanlış\ Negatif)$, ise yanlış tahmin edilen Negatif veri sayısını göstermektedir. F skor hesaplaması ise denklem 3 de gösterilmektedir[13].

$$F\ skor = 2 * \frac{Kesinlik*Duyarlılık}{Kesinlik+Duyarlılık} \quad (3)$$

III. BULGULAR

Bu çalışmada değişen eğitim verilerine göre elde edilen başarılar ölçülmüş ve her sınıf için ayrı ayrı ROC eğri grafikleri oluşturulmuştur. En yüksek başarı %70 eğitim, %30 test veri miktarı seçildiğinde elde edilmiştir. Şekil 2 üzerinde Rosa, Kama ve Canadian sınıflandırmasının en yüksek tahmin (yüzde) başarılarının karmaşıklık matrisi gösterilmiştir. En başarılı sınıflandırma %91.3 başarıyla Canadian ve Roca buğday sınıflandırmasında elde edilmiştir. En düşük sınıflandırma %88.2 başarıyla Kama buğday sınıflandırmasında elde edilmiştir.

	Canadian	Kama	Rosa	∑
Canadian	91.3 %	11.8 %	0.0 %	23
Kama	8.7 %	88.2 %	8.7 %	19
Rosa	0.0 %	0.0 %	91.3 %	21
∑	23	17	23	63

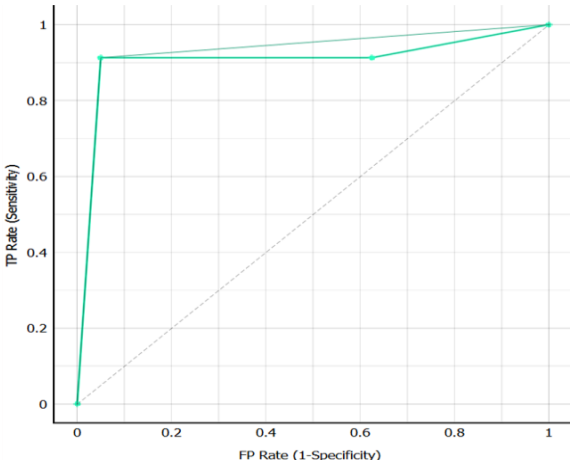
Şekil 2 Sınıflandırma başarısının Karmaşıklık Matrisi (Yüzde Oranlarına Göre)

Şekil 3 üzerinde ise Rosa, Kama ve Canadian sınıflandırmasının en yüksek tahmin başarılarının veri sayına göre karmaşıklık matrisi gösterilmiştir. Canadian sınıflandırmasında, 23 örnekten 21 tanesi doğru sınıflandırılmıştır. Kama sınıflandırmasında, 17 örnekten 15 tanesi doğru sınıflandırılmıştır. Rosa sınıflandırmasında da 23 örnekten 21 tanesi doğru sınıflandırılmıştır.

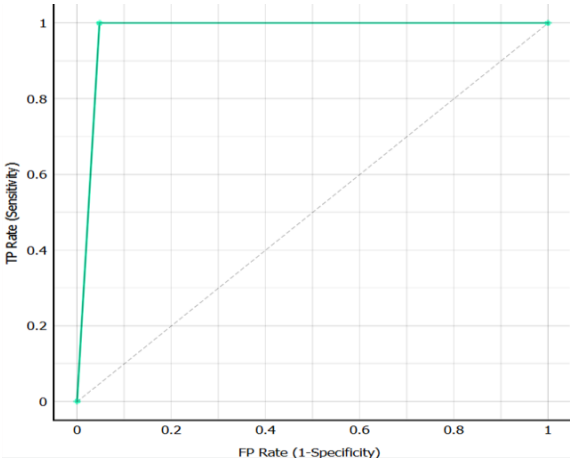
	Canadian	Kama	Rosa	∑
Canadian	21	2	0	23
Kama	2	15	2	19
Rosa	0	0	21	21
∑	23	17	23	63

Şekil 3 Sınıflandırma başarısının Karmaşıklık Matrisi (Sayısal Veri Miktarına Göre)

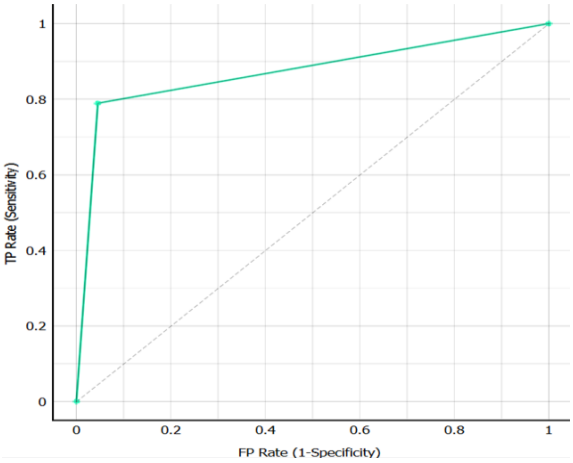
Çalışmada, Canadian buğday sınıflandırması başarısını gösteren ROC eğri grafiği şekil 4 üzerinde gösterilmiştir. Rosa buğday sınıflandırması başarısını gösteren ROC eğri grafiği şekil 5 üzerinde gösterilmiştir. Kama buğday sınıflandırması başarısını gösteren ROC eğri grafiği ise şekil 6 üzerinde gösterilmiştir. Şekillere göre en yüksek başarılı sınıflandırmanın Rosa buğday sınıflandırmasında yapıldığı, en düşük sınıflandırmanın ise Kama buğday sınıflandırmasında yapıldığı görülmüştür.



Şekil 4. Canadian ROC eğrisi



Şekil 5. Rosa ROC eğrisi



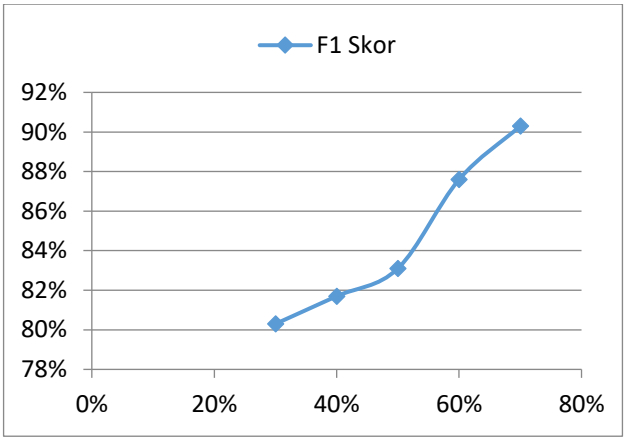
Şekil 6. Kama ROC eğrisi

Bu çalışmada, değişen eğitim oranlarına göre C4.5 karar ağacı algoritmasıyla elde edilen tahmin başarı oranları (yüzde) tablo 2 üzerinde gösterilmektedir. Değişken eğitim oranları, veri seti içinden %70 ile %30 arasında seçilmiştir. Bu eğitim oranlarına karşılık, test oranları veri seti içinden %30 ile %70 arasında rastgele seçilerek tahmin başarıları tespit edilmiştir.

Tablo 2. Değişen Eğitim Oranına Göre Başarı

Eğitim Oranı	Test Oranı	F1 Skor
%70	%30	%90.3
%60	%40	%87.6
%50	%50	%83.1
%40	%60	%81.7
%30	%70	%80.3

Tablodan elde edilen eğitim ve test verilere göre elde edilen, F1 skor başarı değerleri Şekil 7 üzerinde gösterilmektedir. En yüksek %90.3 sınıflandırma başarısı %70 eğitim, %30 test verisi seçildiğinde elde edilmiştir.



Şekil 7. Değişen Eğitim Oranlarına göre F1 Skor Metriğine Sınıflandırma Başarı Grafiği

IV. TARTIŞMA

Khatri vd. (2022) yaptıkları çalışmada, Buğday Tohumu Sınıflandırmasını, veri seti içinden, %70 eğitim ve %30 test verileriyle, Karar ağaçları kullanarak yaptıklarında, sadece AUC başarı metriğine gore %94 başarı elde etmişlerdir. Çalışmada değişken veri oranlarına göre ve başarı değişimi gösterilmemiştir.

Çelik (2023) yaptığı çalışmada, k-NN makine öğrenme algoritmasının Mahalanobis metrik yöntemini ve k=3 komşuluk değerini kullanılarak AUC (Area Under the Curve) başarı metriğine göre %99,24 oranında sınıflandırma başarısı elde edilmiştir.
Eldem (2020) yaptığı çalışmada, veri seti üzerinde, %70 eğitim ve %30 test verileriyle, Derin Sinir Ağı modeli kullanarak sınıflandırma yaptığında, %100 oranı elde etmiştir.

Bu çalışmada %30 ile %70 arasında değişen eğitim oranlarına göre C4.5 karar ağacı makine öğrenme algoritmasının sınıflandırma başarısı ölçülmüştür. En yüksek başarı %90.3 elde edilmiştir. Başarı metriği için F1 Skor kullanılmıştır.

V. SONUÇLAR

Çalışmada, yaygın kullanılan buğday tohumu veri seti içindeki 3 buğday sınıfının, C4.5 karar ağaçları algoritmasıyla %90.3 başarıyla gerçekleştirilebileceği gösterilmiştir. Çalışma, doğruluğu kanıtlanmış F1 skor başarı metriğine göre başarı kıyaslaması yapılmıştır. Ayrıca değişken eğitim oranlarıyla sınıflandırma başarı

kıyaslaması yapılarak eğitim oranının sınıflandırma başarısına etkisi gösterilmiştir. Daha önceki çalışmalarda, daha yüksek başarı elde edilmesine rağmen yapılan bu çalışmada değişken eğitim oranı kullanılması ve sadece F1 skor metriğinin kullanılması farklılığını göstermektedir.

KAYNAKLAR

[1] A, Khatri, S. Agrawal, J. Chatterjee (2022). "Wheat Seed Classification: Utilizing Ensemble Machine Learning Approach", Scientific Programming, 2022: 1-9.

[2] A. Çelik, "Determination of the Classification Success of KNN Algorithm Distance Metric Methods on Wheat Seeds Dataset", Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi, 2023, 23: 1142–1149.

[3] A. Eldem. "An Application of Deep Neural Network for Classification of Wheat Seeds", Avrupa Bilim Ve Teknoloji Dergisi, 2020, 19: 213-220.

[4] H. Küçükönder, K.K, Vursavuş, F. Üçkardeş. "K-Star, Rastgele Orman ve Karar Ağacı (C4.5) Sınıflandırma Algoritmaları ile Domatesin Renk Olgunluğu Üzerinde Bazı Mekanik Özelliklerin Etkisinin Belirlenmesi", Türk Tarım–Gıda Bilim ve Teknoloji Dergisi, 2015, 3: 300-306.

[5] T. Kavzoğlu, İ. Çölkesen. "Karar Ağaçları İle Uydu Görüntülerinin Sınıflandırılması: Kocaeli Örneği", Harita Teknolojileri Elektronik Dergisi, 2010, 2: 36-45.

[6] G. Andrienko, N. Andrienko. "GIS Visualization Support To The C4. 5 Classification Algorithm of KDD", In Proceedings of the 19th International Cartographic Conference, pp: 1-7, 14-21 August 1999, Ottawa.

[7] E. Akbal, Ş. Doğan, N. Varol, "Karar Ağaçları ile Telefon Dolandırıcılığı Verilerinin Analizi", Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 2017, 29: 171–177.

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

[9] [9]: University of California, [Online]. UCI UCI Machine Learning Repository, Available: https://archive.ics.uci.edu/ml/datasets/seeds (accessed 15.11.2023)

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

[11] A. Kayabasi, A. Toktas, K. Sabanci, E. Yigit. "Automatic classification of agricultural grains: Comparison of neural networks", Neural New World, 2018, 28: 213-224.

[12] Lipton, Zachary & Elkan, Charles & Narayanaswamy, Balakrishnan. "F1-Optimal Thresholding in the Multi-Label Setting", 2014. arxiv: https://arxiv.org/abs/1402.1892

[13] S. Demirel, "Makine Öğrenme Algoritmalarıyla Akciğer X-Ray Görüntü Özniteliklerini Kullanarak Pnömoni Tespiti ve Sınıflandırılması", Yüksek Lisans Tezi, Kütahya Dumlupınar Üniversitesi, Kütahya, 2022.

[14] V. Yarğı, S. Postalcıoğlu, "EEG İşareti Kullanılarak Bağımlılığa Yatkınlığın Makine Öğrenmesi Teknikleri ile Analizi", ECJSE, 2021, 8: 142–154..

[15] W. Powers, A. Ailab, "Evaluation: from precision, recall and F-measure to ROC informedness, markedness and correlation", J. Mach. Learn. Technolgy, 2008, 2: 2229-3981.