Avrupa Bilim ve Teknoloji Dergisi
Sayı 54, S. 20-31, Aralık 2024

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 54, pp. 20-31, December 2024

Research Article

Toprak Horizonların Sınıflandırılması için Farklı Makine Öğrenmesi Yöntemlerinin Karşılaştırılması

Zülküf Güman[1*], Hakan Tekin[2], Berhan Aksakal[3], Yasemin Gültepe[4]

[1*] Atatürk Üniversitesi, Mühendislilk Fakültesi, Bilgisayar Mühendisliği Bölümü, Erzurum, Türkiye, (ORCID: 0000-0001-5777-4267), zulkuf.guman13@ogr.atauni.edu.tr
[2] Atatürk Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Erzurum, Türkiye (ORCID: 0000-0001-8340-5094), hakan.tekin21@ogr.atauni.edu.tr
[3] Atatürk Üniversitesi, Mühendislik Fakültesi, Zirrat Mühendisliği Bölümü, Erzurum, Türkiye (ORCID: 0000-0002-9385-0789), berhan.aksakal07@ogr.atauni.edu.tr
[4] Atatürk Üniversitesi, Mühendislik Fakültesi, Yazılım Mühendisliği Bölümü, Erzurum, Türkiye (ORCID: 0000-0002-8684-9907), yasemingultepe@atauni.edu.tr

(İlk Geliş Tarihi 26 Mayıs 2023 ve Kabul Tarihi 6 Şubat 2024)

(**DOI:** 10.5281/zenodo.14175847)

ATIF/REFERENCE: Güman Z, Tekin H, Aksakal B, Gültepe Y (2024). Toprak Horizonların Sınıflandırılması için Farklı Makine Öğrenmesi Yöntemlerinin Karşılaştırılması. *Avrupa Bilim ve Teknoloji Dergisi*, (54), 20-31.

Öz

Toprak derinliği, eko-hidrolojik modellemesi karbon depolama hesaplama ve arazi değerlendirme için kritik öneme sahiptir. Sınırlı sayıda seyrek örnekle, geniş bir karmaşık arazi alanında toprak derinliğinin nasıl tahmin edileceği hala bir sorundur. Bu çalışmanın temel amacı, toprak özellikleri arasındaki ilişkiye dayalı olarak toprak derinliğini tahmin etmek için makine öğrenmesi algoritmaları kapsamlı bir şekilde karşılaştırmaktır. Bu amaçla, deneylerde doğruluk, duyarlılık ve özgüllük gibi iyi bilenen performans ölçütleri kullanılarak çeşitli modeller oluşturulmuş ve karşılaştırılmıştır. Bu çalışma toprak derinliği dağılım haritası, özellikle toprak derinliği verisi olmayan yerler için gelecekteki uygulamalar için yararlı olabileceği düşünülmektedir. Toprak Horizonlarının sınıflandırılması, bu problem için bu öznitelikleri kullanarak sınıflandırmayı başarıyla gerçekleştirmiştir.

Anahtar Kelimeler: CMRB dataset, Makine öğrenme, Toprak derinliği tahmini.

Comparison of Different Machine Learning Methods for Classification of Soil Horizons

Abstract

XSoil depth is critical for eco-hydrological modeling, carbon storage calculation and land evaluation. How to estimate soil depth over a large area of complex terrain with a limited number of sparse samples remains a challenge. The main aim of this study is to comprehensively compare machine learning algorithms to predict soil depth based on the relationship between soil properties. For this purpose, various models were created and compared in experiments using well-known performance measures such as accuracy, sensitivity and specificity. This study soil depth distribution map is thought to be useful for future applications, especially for places where soil depth data is not available. Classification of soil horizons has successfully performed the classification using these features for this problem.

Keywords: CMRB dataset, Machine Learnin, Soil depth estimation.

[*] Sorumlu Yazar: yasemingultepe@atauni.edu.tr

1. Giriş

Toprak özellikleri, toprak verimliliği ve çevresel analiz yöntemlerinde çok önemlidir. Toprak, esas olarak yaşamı destekleyen organik mineraller ve kaya parçacıklarından oluşan yer kabuğunun en üst tabakasıdır (Goldhaber ve Banwart, 2015). Toprak profili, yüzeye paralel uzanan katmanlardan oluşan toprağın dikey bir enine kesitidir. Bu katmanlar, toprak horizonları olarak bilinir (Hartemink ark., 2020). Toprak sınıflandırılması, toprak ayırt edici özelliklerine ve kullanım seçimlerini belirleyen kriterlere dayalı olarak sistematik olarak sınıflandırılmasıyla ilgilenmektedir (Ramesh ve Ramar, 2011; Raunak, 2018).

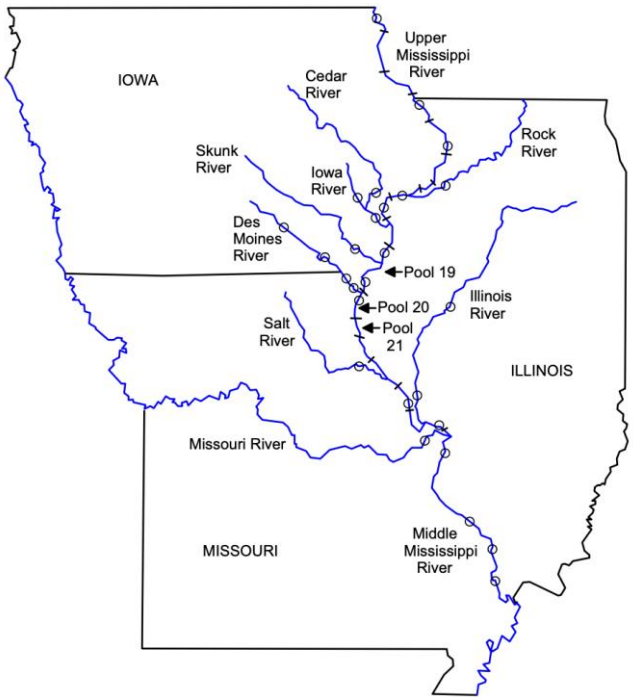
Bu karmaşık toprak özelliklerinin ölçülmesi, karmaşık ve zaman alıcı laboratuvar prosedürleri gerektirir. Toprağı analiz etmek için taşınabilir X-ışını flüoresansı (pXRF) veya görünür ve yakın kızılötesi (Vis-NIR) spektroskopisi gibi proksimal sensörlerin kullanılması, elemental konsantrasyonları ölçmek için hızlı bir yol veya topraktan veri toplamak için diğer alternatifler sundukları için popülerlik kazanmaktadır (Pham ve ark., 2021).

Veri toplama süresi zordur, zaman almaktadır ve pahalıdır. Ayrıca toplanan verilerin kullanımı (yani analiz veya üst düzey özellik tahminleri) zaman alan ve hataya açık işlemdir. Toprağın doğru analizi, sağlıklı (veya kaliteli) ve sürdürülebilir bir toprak yönetimi için büyük önem arz etmektedir. Topraktan elde edilen devasa veriler açık kaynak kodlu algoritmalar ve yapay zekanın bir alt kategorisi olan makine öğrenmesi (ML) prosedürlerinin hızlı bir şekilde kullanımını gerektirmiştir. Son araştırmalar gösteriyor ki bu şekilde büyük karmaşık yapılı veri setleri üzerinde yapay zeka (YZ) ve makine öğrenme yöntemleri diğer klasik tahmin yöntemlerine göre doğruluğu daha yüksek sonuçlar elde edilmektedir (Côté ve ark., 2022).

Bu çalışmada diğer çalışmalardan farklı olarak toprak horizonlarının ilk defa çeşitli makine öğrenme yöntemleri ile sınıflandırılması başarı ile sağlanmıştır. Bu bağlamda Orta Mississippi Nehir Havzası LTAR sahasına ilişkin temel veriler kullanılmıştır. Tükenmekte olan Mississippi Nehri Vadisi Alüvyal Akifer (MRVAA) bağlamında, büyük veriler, sulama kararlarının, kullanımlar ve zaman içinde en faydalı ve değerli olan yeraltı suyu stoklarının tahsisi olacak şekilde verilmesini sağlamada önemli bir role sahiptir (Nelson ve ark., 2022). Şekil 1'de Mississippi nehri havzasının orta kısmının haritası gösterilmiştir (Whitledge ve ark., 2018).

Orta Mississippi Nehir Havzası LTAR sahasındaki bozulmamış toprak örneklerinde 2016-2018 yılları arasında Nitrojensiz Hava Devridaim Metodu (N-FARM) aracılığıyla dinitrojen (N_2) ve nitröz oksit (N_2O) üretimi doğrudan ölçülerek yerinde denitrifikasyon oranları ölçülmüştür. 10 günlük laboratuvar inkübasyonları, mikrobiyal solunum ve potansiyel net azot mineralizasyonu ve nitrat azotu konsantrasyonları dahil olmak üzere, ilişkili çeşitli toprak örneklerinden veriler sunulmuştur (Weitzman ve ark., 2018).

Bu çalışmada makine öğrenmesi algoritmaları kullanılarak toprak horizonlarını tahmin etmek amaçlanmıştır. Çalışma kapsamında Yapay Sinir Ağları (YSA), Destek Vektör Makineleri (DVM), k-En Yakın Komşuluk (k-NN), Karar Ağaçları (KA), Lojistik Regresyon (LR), Rastgele Orman (RO), Naive Bayes (NB), AdaBoost algoritmaları uygulanarak sınıflandırma modelleri oluşturulmuştur ve ardından veriler için sınıf tahmini yapılarak, modellerin tahmin sonuçlarının performans ölçüm değerleri elde edilmiştir. Çalışmanın ikinci bölümünde çalışmada kullanılan materyal ve metotlar belirtilmiştir. Üçüncü bölümde, deneysel sonuçlar ve tartışmaya yer verilerek veri kümesi üzerinden yapılan deneysel sonuçlar ayrıntılı olarak verilmiştir. Son olarak yapılan çalışmaya dair elde edilen bilimsel bulgulara ve sonuçlara dayanan tartışmaya yer verilerek, toprak horizonlarını sınıflandırma probleminin geliştirilmesi konusunda çeşitli öneriler dördüncü bölümde sunulmuştur.



Şekil 1. Mississippi nehri havzasının orta kısmının haritası (Figure 1. Map of the middle part of the Mississippi river basin)

2. Materyal ve Metot

Bu bölümde, deneysel çalışmalarda toprak seviyesi tespiti için kullanılan yöntemler ve oluşturulan modellerin başarısını değerlendirmek için kullanılan veri seti hakkında açıklamalar özetlenmiştir.

2.1. Veri Seti

Çalışmada Uzun Vadeli Tarımsal Araştırma ağının Orta Mississippi Nehri Havzası sahasındaki bozulmamış toprak veri seti (Weitzman ve ark., 2018) kullanılmıştır. Sınıflandırma modeli elde edilecek veri setinde toprak horizonları ile ilgili üst toprak (surface) ve alt toprak (confining) olmak üzere iki sınıf bulunmaktadır. Veri setinde uzmanlar tarafından sınıflandırılmış, 110 adet üst toprak, 110 adet alt toprak olmak üzere toplamda adet toprak verisi bulunmaktadır.

Tablo 1'de 15 bağımsız ve 1 bağımlı değişken olmak üzere veri setinde kullanılan değişkenler gösterilmiştir. 15 bağımsız değişken toprak horizonlarının sınıflandırılması için önemli faktörlerdir ve bu faktörler makine öğrenmesi yöntemleri için öznitelik olarak ifade edilmektedir. Tablo 1'de görüldüğü gibi "Derinlik sınıfı" bağımlı değişkendir ve toprağın horizonunu ifade etmektedir. Bu bilgiler doğrultusunda bir toprak derinliği 0-10 arası ise üst toprak (pozitif); 51-90 arası ise alt toprak (negatif) olarak kodlanmıştır.

Tablo 1. Çalışmada Kullanılan Veri Setinin Özellikleri

Özellik	**Ayrıntılar**
Eyalet	LTAR konumunun ABD'deki eyaleti
Tarih	Tarlada toprak numunesinin alındığı tarih
Saat	Örnekleme zaman sırası
Konum	Site içinde alan tanımlama
Plot	Örnek arsa konumu
Derinlik	Toprak örnekleme derinliği aralığı
Derinlik Sınıfı	Toprak örnekleme derinliğinin kalitatif tanımlaması
GWC	Gravimetrik su içeriği
CO_2	Karbondioksit akışı
N_2O	Nitröz oksit akışı
N_2	Dinitrojen akışı
RESPC	Toprak solunumu
NO_3	Toprak nitrat
NH_4	Toprak amonyum
MIN	Potansiyel net N mineralizasyonu (laboratuar inkübasyonu)
NIT	Potansiyel net nitrifikasyon (laboratuar inkübasyonu)

2.2. Toprak Örnekleri ve Analizi

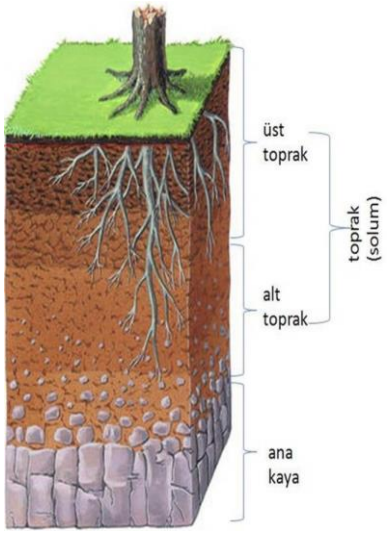
Makine öğrenmesi, insanların öğrenme şeklini taklit etmek için verilerin ve algoritmaların kullanımına odaklanan ve örneklerden ve gözlemlerden anlamlı ilişkileri ve kalıpları otomatik olarak öğrenmeye çalışan bir YZ ve bilgisayar bilimi dalıdır (Sarker, 2022). Makine öğrenmesi gibi büyük veri analitik araçlarıyla birlikte büyük çevresel veri setlerinde son gelişmeler toprak fosforunun (P) gibi karmaşık çevresel değişkenlerdeki uzamsal kalıpları değerlendirmek ve tahmin etmek için fırsatlar yaratmıştır. (Dolp ve ark., 2022) çalışmasında ABD'de Yukarı Mississippi Nehri Havzası (Upper Mississippi River Basin, UMRB) için 100 m'lik bir ızgara ölçeğinde topraktaki toplam P değerini tahmin etmek için toprak türü, arazi örtüsü, arazi kullanımı, topografya, besin girdileri ve iklimin özelliklerini özetleyen yaklaşık 300 jeo-uzamsal özellik ile birlikte veri kümesi kullanılmıştır. Sonuç olarak bölgedeki su kalitesini iyileştirmeye yönelik koruma planlaması ve modelleme çabalarını iyileştirmek için kullanılabilir.

Uzun vadeli aylık akış tahminleri, bir nehir havzası sisteminde karar vermek için gereklidir. Literatürde YSA ile yapılan birçok mevsimsel tahmin çalışması mevcuttur. YSA modellerindeki girdi değişkenleri dikkate alınarak seçilirse, eğitim modellerinin sayısı daha az olsa bile aylık tahminlerin önemli ölçüde iyileştirilebileceği görülmektedir (Chandrasekaran ve ark., 2013).

Su miktarının ve kalitesinin etkin yönetimi için, mevcut yüzey suyunun kirlilik seviyesinin tahmin edilmesi gereklidir. (Khoi ve ark., 2022) çalışmasında 5 adet artırma tabanlı algoritma (uyarlamalı artırma, gradyan artırma, histogram tabanlı gradyan artırma, hafif gradyan artırma ve aşırı gradyan artırma), üç karar ağacı dahil olmak üzere on iki makine öğrenmesi modellerinin performansını değerlendirmeyi amaçlamaktadır. Sonuçlar, on iki makine öğrenimi modelinin hepsinin su kalitesi indeksi (WQI)'yi tahmin etmede iyi performansa sahip olduğunu, ancak aşırı gradyan artırmanın (XGBoost) en yüksek doğrulukla (R^2 = 0,989 ve RMSE = 0,107) en iyi performansa sahip olduğunu göstermiştir. (Rizal ve ark., 2022) çalışmasında Malezya'daki Langat Nehrinin kalitesini izlemek ve

bozulmaya karşı korumak için nehir suyu kalitesini modellemek için makine öğrenmesi yöntemlerinden YSA, Gauss Süreç Regreysonu ve DVM kullanılmıştır. Model performans metriklerine bakıldığında, YSA modeli tüm modellerden daha iyi performans gösterirken, GPR ve DVM modelleri aşırı uyum özelliği sergilemiştir.

Mississippi nehri havzası üzerine yapılan literatür çalışmaları incelendiğinde su tahminleme işlemlerinin yapıldığı çalışmalar göze çarpmaktadır. Toprak horizonlarının farklılıkların toprak öznitelikleri üzerindeki etkili olduğu bilinmekte olup, bu açıdan ülkelerin kendi içinde değerlendirilmeleri önemlidir. Bu sebeple gerçekleştirilen bu çalışma, literatürde Orta Mississippi Nehri Havzasının toprak horizonları üzerine olan literatürü güçlendirecektir. Ayrıca makine öğrenmesi tabanlı algoritmaların toprak horizonlarının modellenmesinde literatürdeki çalışmaların eksikliği görülmekte olup farklı makine öğrenmesi yöntemlerinin başarılı sonuçlar verdiği de yine yapılan çalışma ile gösterilmiştir. Bu çalışmada kullanılan veri setinde toprak horizonlarından sadece 2 çeşit (üst toprak ve alt toprak) örneklemeler mevcuttur. Şekil 2'de farklı toprak katmanları üst toprak, alt toprak ve ana kaya olmak üzere üçe ayrılmıştır (Esra, 2020).



Şekil 2. Toprak profili (Figure 2. Soil profile)

2.3. Makine Öğrenmesi Sınıflandırma Algoritmaları

Makine öğrenmesi (ML), sistemlere özel olarak programlanmadan deneyimlerden otomatik olarak öğrenebilen, kalıpları tanımlayabileceği ve minimum insan müdahalesi ile karar verebileceği fikrine dayanan bir tür yapay zekadır (Gültepe, 2019; Wong, 2021). ML, veri sorunlarını çözmek için farklı algoritmalara dayanır. Kullanılan algoritmanın türü, çözmek istediğiniz sorunun türüne, değişken sayısına, ona en uygun modelin türüne vb. bağlıdır. Bu çalışmada en başarılı sınıflandırıcıyı tespit edebilmek amacıyla sıklıkla tercih edilen sınıflandırıcılardan YSA, DVM, k-NN, KA, LR, RO, NB ve AdaBoost kullanılmıştır. Bu sekiz yöntemin sonuçlarına göre aralarından toprak horizonlarının sınıflandırılmasında kullanılabilecek en iyi sonucu veren yöntemin bulunması amaçlanmıştır. Aşağıda popüler olarak kullanılan makine öğrenme yöntemleri kısaca tanıtılmıştır.

Bu yöntemler, Python 3.7 programlama dili yardımıyla Jupyter Notebook (Anaconda3) editöründe uygulanmıştır. Bu çalışmada Python 3.7 programlama dilinin kullanılmasının nedeni, Python kütüphaneleri ile yapay zeka algoritmalarının uygulanmasında kolaylık sağlanmasıdır (Raschka ve ark., 2020).

2.3.1 Destek Vektör Makineleri (DVM)

DVM, sınıflandırma ve regresyon problemlerinde kullanılan denetimli bir öğrenme sistemidir. DVM, daha az hesaplama gücü ile önemli doğruluk değerleri elde edildiği için birçok çalışmada tercih edilmektedir (Ayhan ve Erdoğmuş, 2014; Guo ve ark., 2005; Wang, 2005). En az sayıda eğitim verisi örneği aracılığıyla farklı veri sınıfları arasında en uygun hiper düzlemi bulmaya çalışır. Bu hiper düzlem, sınıflar arasındaki sınırları en üst düzeye çıkarır. Eğitim veri setinde "destek vektörleri", hiper düzleme en yakın olan veri noktaları olarak kabul edilir (Boser ve ark., 1992; Foody ve ark., 2007; Gültepe, 2022; Sarmadian ve ark., 2014).

2.3.2 Yapay Sinir Ağları (YSA)

YSA, genellikle insan beyninin yapısını oluşturan biyolojik sinir ağlarına dayalı hesaplamalı bir ağdır. İnsan beyninde birbirine bağlı nöronlar olduğu gibi, yapay sinir ağlarında da ağların çeşitli katmanlarından birbirine bağlı nöronlar bulunur (Yang ve Wang, 2020). Bir YSA girdi katmanı, gizli katman ve çıktı katmanı olmak üzere üç katmandan oluşmaktadır. Her katmandaki nöronlar belirli bir ağırlıkla birbirine bağlıdır. Bu ağırlıklar, hedef değerlere yeterince yakın olana kadar yinelemeli olarak güncellenir. Ağırlıklar ayarlandığında sistem eğitilmiş olarak ifade edilebilir. Bu aşamadan sonra test işlemi yapılabilir (Zakaria ve ark., 2014).

2.3.3 k-En Yakın Komşuluk

k-NN, basitliği ve etkinliği nedeniyle yaygın olarak kullanılan bir örüntü sınıflandırma tekniğidir (Pandith ve ark., 2020). k-NN algoritması, eğitim veri noktalarını içeren bir özellik alanı gerektirir. Bu 'özellik benzerliği', özellik uzayındaki mevcut veri noktalarına benzerliğine dayalı olarak yeni veri noktasını tahmin etmek için kullanılır. Algoritma, bilinmeyen bir veri noktası ile en yakın 'k' eğitim

veri noktası arasındaki mesafeleri belirler ve yeni noktayı o sınıfa ait olarak sınıflandırır. 'k' değeri, eğitim veri setinden seçilen veri noktalarının sayısına bağlıdır. Algoritma öncelikle mesafeyi hesaplamak için bir metrik seçerek test veri noktasını en iyi şekilde çizecek sayısal bir özellik aramaya başlar. Yeni veri noktası ile 'k' noktası arasındaki mesafeler, bir mesafe ölçüm metriği kullanılarak hesaplanır (Brownlee, 2016; Harison 2018; Gültepe, 2022).

2.3.4 Karar Ağaçları

Aşırı miktarda veri içeren bir veri setini, kümeleme algoritması yardımı ile daha küçük kümelere/gruplara bölmek için kullanılan bir yöntemdir. En önemli denetimli öğrenme algoritmalarından biri olan karar ağacı hem regresyon hem de sınıflandırma için kullanılabilir. Bir karar ağacı; sınıfı belirlenmiş verilerden hareket ederek sınıfı belli olmayan verilerin hangi sınıfa ait olduğunu belirlemeye yarayan bir yöntemdir. En üstünde bulunan hücreye kök (root), kök hücre altında bulunan hücrelere düğüm (node), en altında bulunan hücrelere ise yaprak (leaf) denir. Bir karar ağacının karmaşıklığı düğüm sayısı ile doğru orantılıdır. Bölünme kullanıcının ihtiyacı karşılanana kadar devam eder (Gültepe ve Gültepe, 2020; Podgorelec ve Zorman, 2015).

2.3.5 Logistic Regresyon

Regresyon analizi, bir dizi bağımsız değişken ile bir bağımlı değişken arasındaki ilişkiyi matematiksel olarak tanımlar. Kullanabileceğiniz çok sayıda regresyon modeli vardır. Bu seçim genellikle bağımlı değişken için sahip olduğunuz veri türüne ve en iyi uyumu sağlayan model türüne bağlıdır. Regresyondan ziyade çoğunlukla ikili sınıflandırma için kullanılan doğrusal bir modeldir (Arabameri ve ark., 2018; Basu ve Pal, 2017; Stoltzfus, 2011).

2.3.6 Rastgele Orman

RF sınıflandırıcı, başlangıçta rastgele bir veri alt kümesi seçer ve çok sayıda karar ağacı oluşturulur ve her bir karar ağacının tahmini değer sonucu oluşur. Tahmin sonucu oluşan her değer için oylama gerçekleştirilir. Daha sonra algoritma son tahmin için en çok oylanan değeri seçerek sonuç oluşturur. Tek bir karar ağacının hataya yol açma olasılığı daha fazladır, ancak sınıflandırma işlemine birçok karar ağacı dahil edildiğinde hatanın azaldığını ve doğruluğun arttığını gözlemlenir. Bu algoritma, herhangi bir karar ağacından alınan her bir kararın etkisini değerlendirirken ağırlık kavramını kullanır. Hatası yüksek ağaca düşük ağırlık, düşük hatası olan ağaca yüksek ağırlık verilir (Anisha ve ark., 2021).

2.3.7 Naive Bayes

Hem sınıflandırma hem de tahmin için kullanılması oldukça basit bir metot olan Naive Bayes denetimli bir öğrenme yöntemidir. Her bir sınıf şartının birbirinden bağımsız olduğu varsayımına dayanarak çalışan bu algoritmada aynı zamanda en yüksek olabilirlik eğitimini bulmak için kapalı form ifadesi kullanıldığından hesaplama maliyeti düşüktür. Daha açıklayıcı bir ifade ile Naive Bayes "Bağımsız Özellik Modeli" ile çalışmaktadır. Bu yöntemin en önemli avantajlarından biri, sınıflandırma için tahmin yapabilmesi adına az miktarda eğitim verisi gerektirmesidir. Bayes teoremi olasılıkları önceki deneyimlere dayalı olarak tahmin eder (Bhargavi ve Jyothi, 2009; Priya ve ark., 2018).

2.3.8 AdaBoost

AdaBoost, en temel toplulukla öğrenme yöntemlerinden biridir. Adaboost algoritması her bir yinelemesinde yanlış sınıflandırılmış örneklere ilişkin ağırlık değerleri artırılırken, doğru sınıflandırılmış örneklere ilişkin ağırlık değerleri azaltılır. Böylelikle yanlış sınıflandırılan örneğin bir sonraki adımda doğru olarak sınıflandırılması amaçlanmıştır (Wang ve ark., 2022). Sonuçta bir önceki modelin yanlış sınıflandırılmış örneklerine duyarlı bir öğrenme modeli elde edilebilir (Coopersmith ve ark., 2014).

2.4. Performans Metrikleri

Makine öğrenme modelinin performansını değerlendirmek için performans değerlendirme metrikleri kullanılır. Performans metriklerinin seçimi, modelin performansının nasıl ölçüleceğini ve kıyaslama sonuçlarıyla nasıl karşılaştırılacağını etkilediği için çok önemlidir. Bu seçim modelin türüne ve uygulamasına bağlıdır. Veri setinde üst katman (pozitif) ve alt tabaka (negatif) olmak üzere iki sınıf bulunmaktadır. İki sınıflı sınıflandırmalarda elde edilen karışıklık matrisi, doğru pozitif (DP), doğru negatif (DN), yanlış pozitif (YP) ve yanlış negatif (YN) temel kriterleri içerir. Burada DP, doğru şekilde sınıflandırılan yüzey toprak sayısını, DN ise doğru şekilde sınıflandırılan alt tabaka toprak sayısını ifade eder. Ek olarak, YN yanlış sınıflandırılmış yüzey toprak sayısını verirken, YP yanlış sınıflandırılmış alt tabaka toprak sayısını verir.

Modellerin performansları, doğruluk (Accuracy), kesinlik (Precision), duyarlılık (Sensitivity), yanlış negatif oranı (YNO) ve duyarlılık (Recall) gibi farklı metrikler kullanılarak ölçülür. Doğru pozitif oranı (DPO) olarak adlandırılan duyarlılık, doğru şekilde sınıflandırılan yüzey örneklerin tüm yüzey örneklere oranını ifade ederken özgünlük, doğru şekilde sınıflandırılan alt tabaka örneklerin tüm alt tabaka örneklere oranını ifade eder. YNO, yanlış sınıflandırılmış alt tabaka örneklerin tüm alt tabaka örneklere oranıdır.

Doğruluk, doğru sınıflandırılan örneklerin tüm örnekler içindeki oranını hesaplayarak genel sınıflandırma performansını gösterir. Kesinlik, doğru şekilde sınıflandırılan yüzey numunelerin, yüzey olarak tahmin edilen tüm örneklere oranını verir. Duyarlılık, yapılabilecek tüm pozitif tahminlerden yapılan doğru pozitif tahminlerin sayısını gösterir. F1 skor değeri, kesinlik ve duyarlılık değerlerinin harmonik ortalamasını göstermektedir. Bu metrikler sırasıyla denklem (1), (2), (3) ve (4)'da verilmiştir (Gültepe, 2021; Japkowicz, 2011).

Tablo 2. Karışıklık Matrisi

		Tahmin	
		Negatif	Pozitif
Gerçek	Negatif	DN	YP
	Pozitif	YN	DP

$$Doğruluk\ (Acc) = \frac{DP+DN}{DP+YP+YN+DN} \quad (1)$$

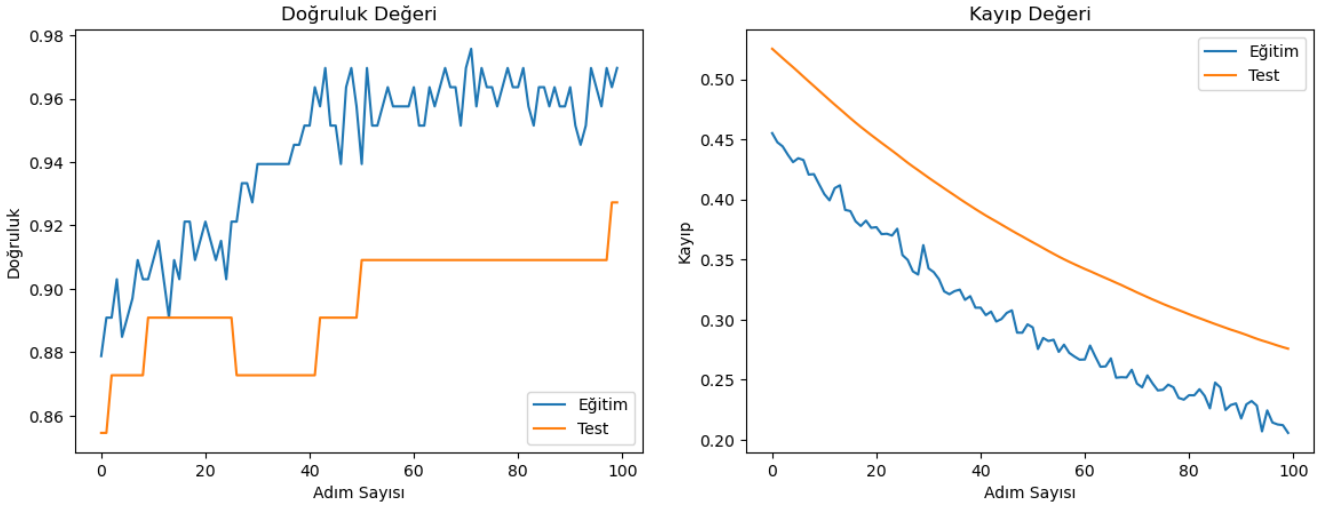
$$Kesinlik\ (Pre) = \frac{DP}{DP+YP} \quad (2)$$

$$Duyarlılık\ (Sen) = \frac{DP}{DP+YN} \quad (3)$$

$$F_1 = 2 * (\frac{\text{Kesinlik*Duyarlılık}}{\text{Kesinlik+Duyarlılık}}) \quad (4)$$

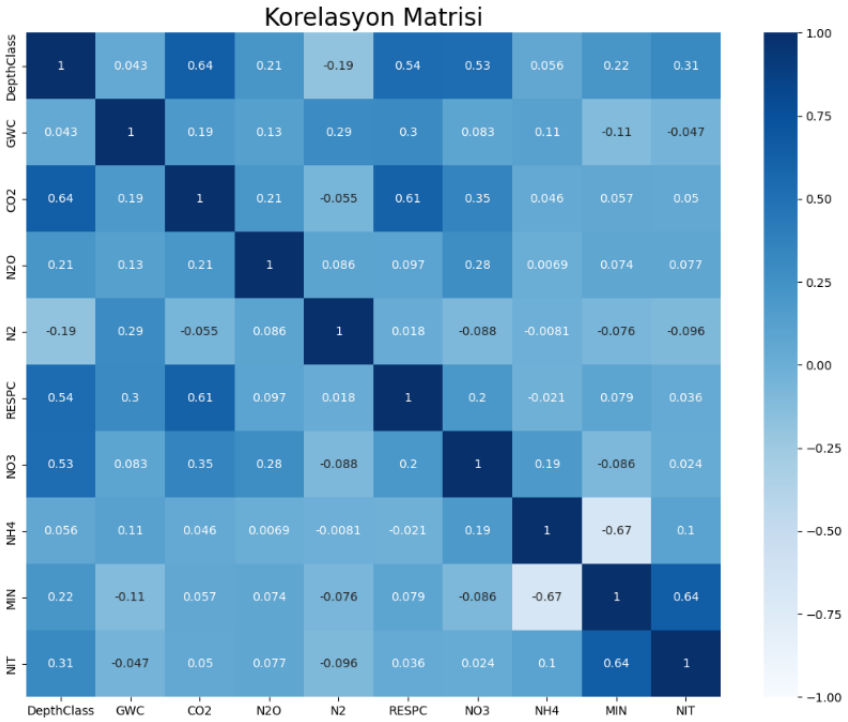
3. Araştırma Sonuçları ve Tartışma

Bu çalışmada Orta Mississippi Nehri Havzası sahasındaki bozulmamış toprak veri seti kullanılarak sekiz farklı makine öğrenmesi yöntemleri ile sınıflandırma yapılmıştır. YSA giriş, gizli ve çıktı katmanlarının sayıları sırasıyla 8-6-5-2 olarak belirlenmiş olup, geri yayılma öğrenimi ile 100 iterasyonlu olarak uygulanmıştır. YSA katmanlarındaki düğümlerde ReLU aktivasyon fonksiyonu, sınıflandırma katmanında ise Softmax aktivasyon fonksiyonu kullanılmıştır. YSA modeli kullanılarak toprak horizonlarının sınıflandırılmasından elde edilen eğitim sırasındaki doğruluk ve kayıp (loss) sonuç grafikleri, Şekil 3'de verilmiştir. Şekil 3'de, eğitim verileri ile test verileri arasında, değişimin orantılı bir şekilde olduğunu, doğruluk açısından yükselen eğilimde, kayıp değerleri açısında düşüş eğilimde olduğunu ve aralarında çok fark olmadığını gözlemlenmiştir.



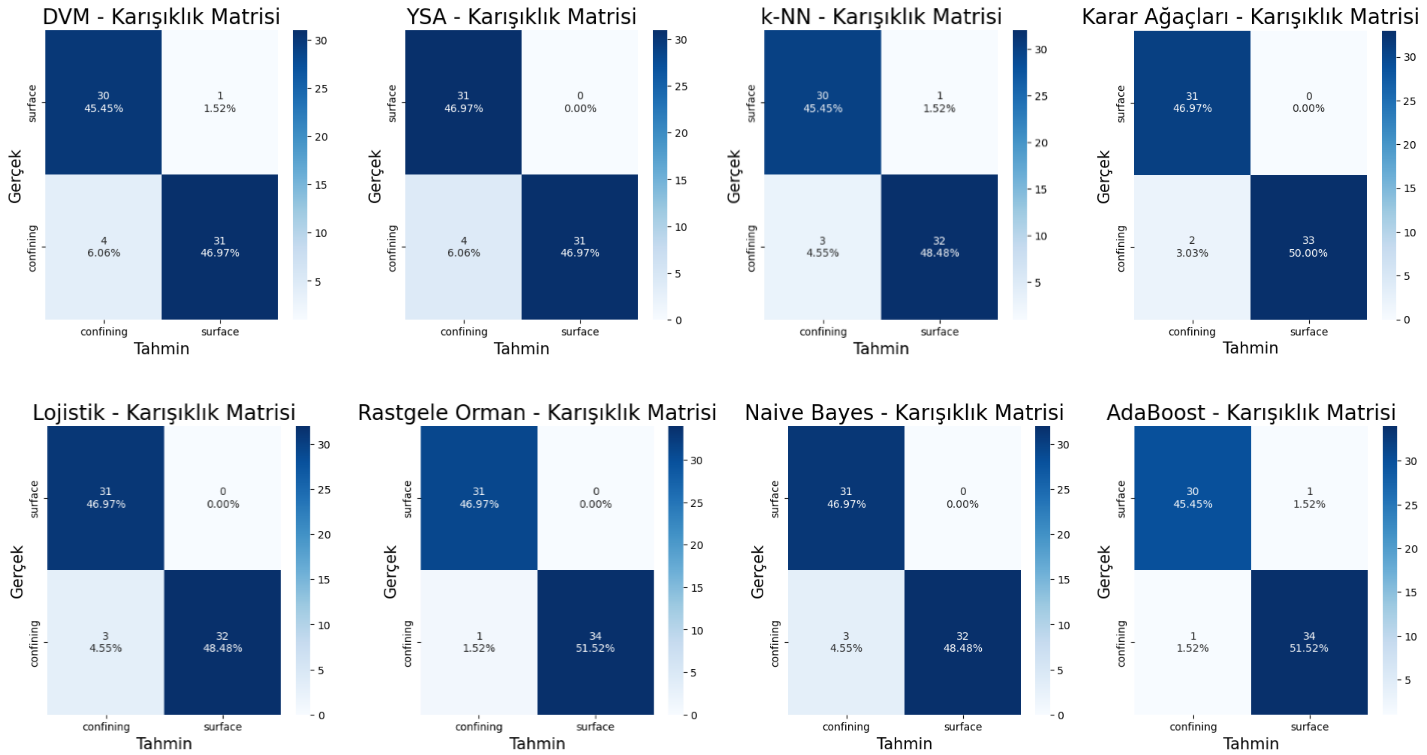
Şekil 3. YSA eğitim doğruluk ve kayıp değeri grafikleri (Figure 3. ANN training accuracy and loss graphs)

Çalışmada kullanılan veri setindeki 10 farklı öznitelikler arasındaki korelasyonu gösteren matris, Şekil 4'de görülmektedir. Korelasyon matrisine göre özelliklerin hiçbirinin birbiriyle yüksek oranda ilişkili olmadığı söylenebilir. Bu nedenle her özelliğin toprak horizon tahminine yönelik bireysel katkısı bulunmaktadır. Veri setinin %80'i eğitim, %20'i ise test verisi olarak ayrılmıştır.



Şekil 4. Veri setindeki toprak öznitelikleri için korelasyon matrisi (Correlation matrix for soil features in the dataset)

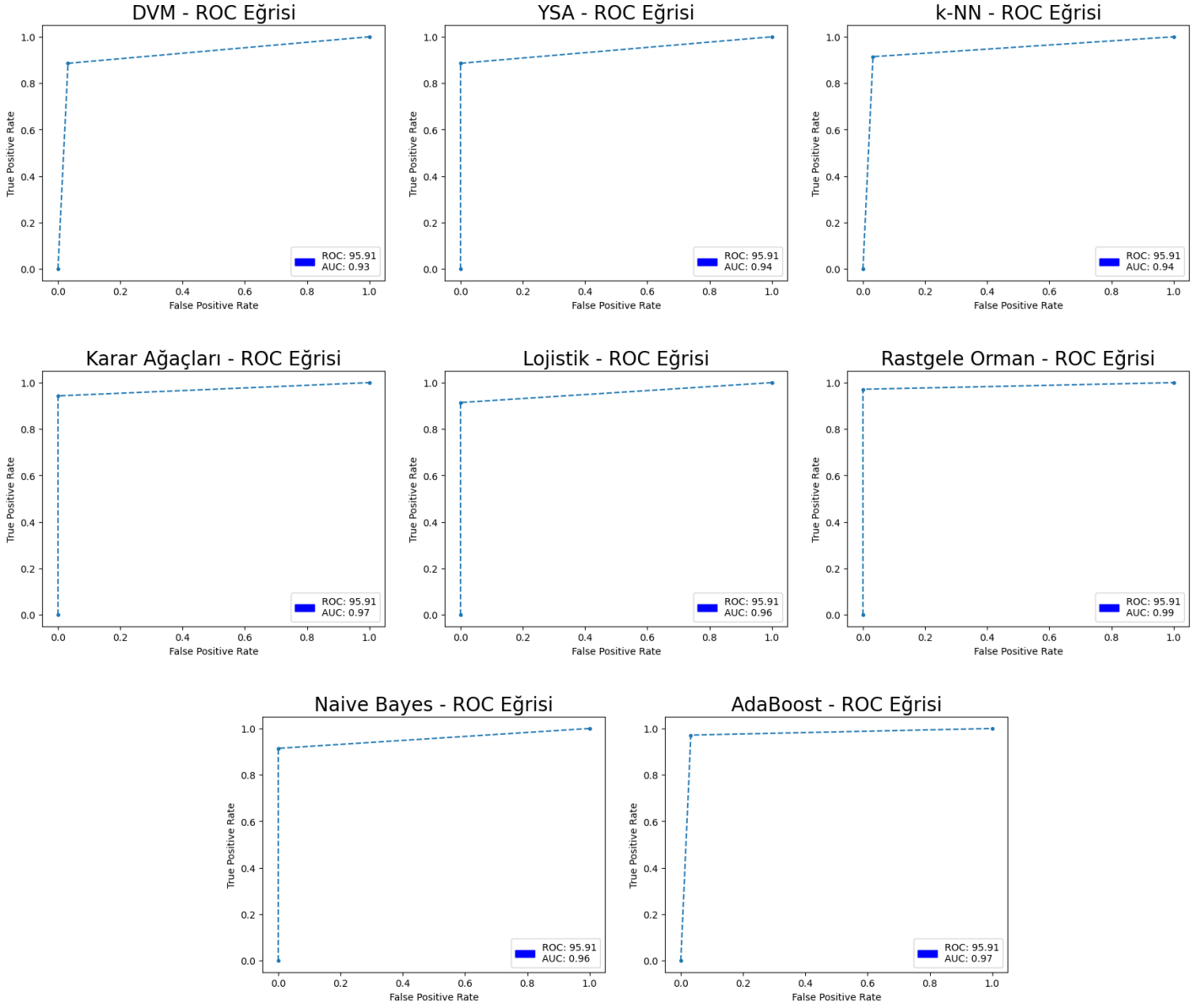
Karışıklık tablosu, makine öğrenmesi modelinin yapmış olduğu tahminlerin performansını değerlendirmek için kullanılır. Bu çalışmada kullanılan sekiz adet makine öğrenmesi modellerinin ikili sınıflandırmada yapmış oldukları tahminlerin doğru sınıfa ait olması konusunda başarısını gösteren karışıklık matrisleri sırasıyla Şekil 5'deki gibidir. Ortaya çıkan karışıklık matrislerinde her bir kategorideki sayılar ve kategorilerin veri setini yüzde kaçlık oranda temsil ettiği gösterilmektedir. Genel olarak, Rastgele Orman ve Adaboost modelleri en yüksek başarıya sahip olduğu görünmektedir.



Şekil 5. Makine öğrenmesi yöntemlerinin tahmin karışıklık matrisleri (Prediction confusion matrices of machine learning methods)

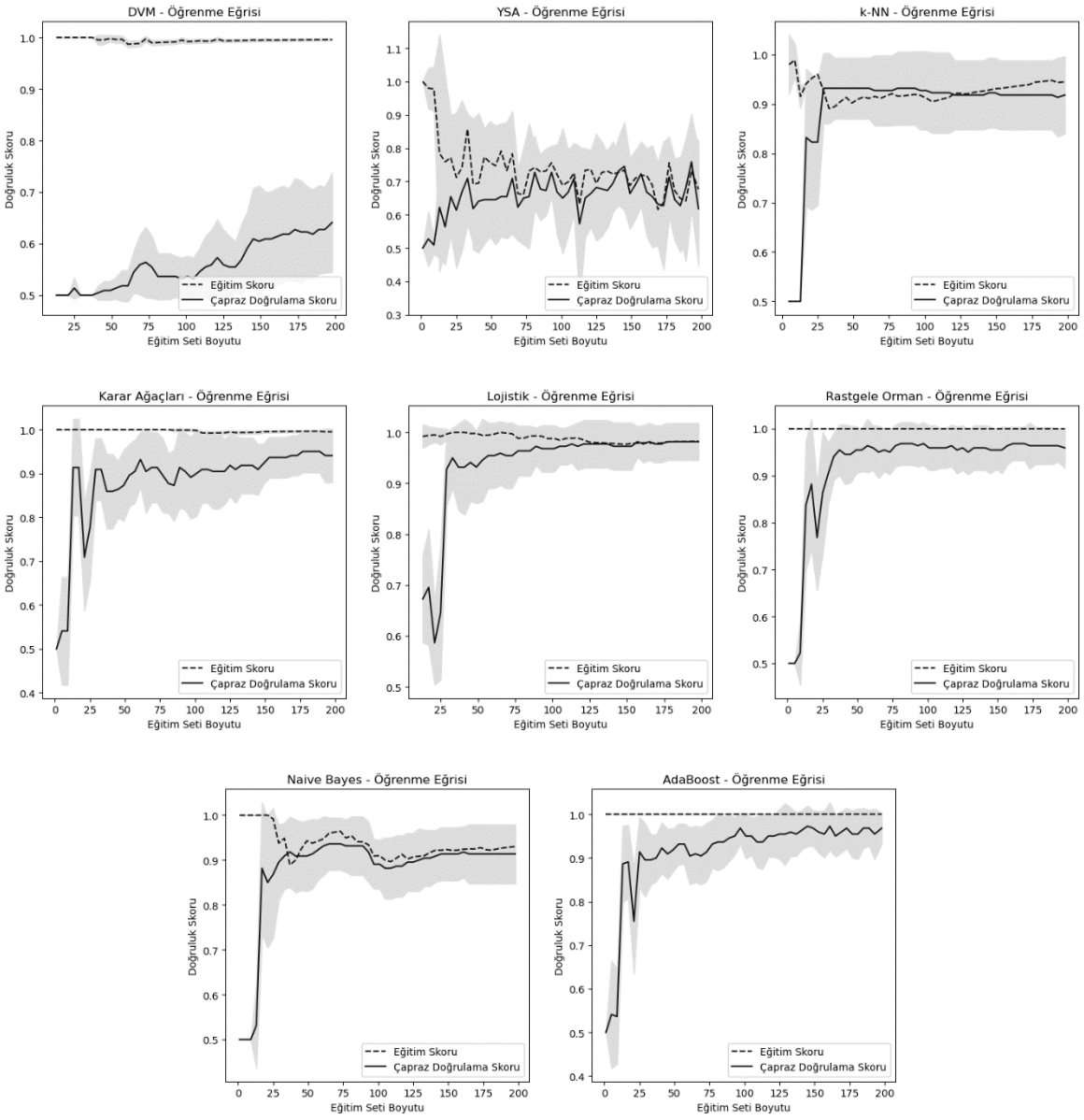
Her bir modelin doğruluğunu değerlendirmek için karışıklık matrisinin yanında Alıcı İşlem Karakteristikleri (Receiver Operating Characteristics, ROC) eğrileri de oluşturulmuştur. ROC eğrisi, bir sınıflandırma modelinin tüm sınıflandırma eşiklerindeki performansını gösterir. ROC eğrisi, dikey eksen üzerinde DP ve yatay eksen üzerinde YP oranlarının yer aldığı bir eğridir. ROC Eğrisi

ise temelde DP oranının artması durumunda yani 1'e yakınsaması durumunda YP oranının alacağı değerleri gösteren eğridir. Burada DP oranının 1'e yakınsaması durumu istediğimiz bir durum fakat bu YP oranının 1'e yakınsaması durumunun yanında YP oranının düşük kalması istenmektedir (Feng ve ark., 2019). Toprak horizonlarının tahminlerinden elde edilen ROC eğrileri, Şekil 5'de verilmiştir. Rastgele Orman ve Adaboost için ROC eğrisinin altında kalan oran %100'dur. Bu sonuçlar yapılan analiz sonuçlarının doğruluğunun oldukça yüksek olduğunu göstermiştir.



Şekil6. ROC eğrileri (ROC curves)

Makine öğrenmesi modellerine ait öğrenme eğrileri, kesikli çizgi eğitim ve düz çizgi doğrulama verisinin performansı olmak üzere Şekil 6'da gösterilmiştir. Şekil 6'da görüldüğü üzere yüksek doğruluk oranına sahip RO ve AdaBoost modellerinin yüksek varyans ve yüksek sapmalı olduğu görülmektedir. Eğitim verisi ile doğrulama verisi için gelen hata değerlerinin öğrenme eğrilerinde paralel gitmesi ve aralarındaki farkın çok olmaması modelin aşırı uyumlu olduğu göstermektedir.



Şekil 7. Öğrenme eğrileri (Learning curves)

. Herhangi bir makine öğrenimi modelindeki temel adım, modelin doğruluğunu değerlendirmektir. Bu çalışmada sınıflandırma modellerinin performansları değerlendirmek için determinasyon katsayısı (R2), Ortalama Kare Hatası (Mean Squared Error, MSE), Kök Ortalama Kare Hatası (Root Mean squared, RMSE), Ortalama Mutlak Hata (Mean Absolute Error, MAE) performans ölçütleri kullanılmıştır. Tablo 3'de verilen yöntemlerin performanları değerlendirildiğinde özellikle RO ve Adaboost diğer yöntemlere göre büyük üstünlük sağlamıştır.

MSE (Mean Square Error – Ortalama Hata Karesi): Veri setine bağlıdır. Ortalama mutlak farkın karesi alınarak hesaplanır, kısaca hata karelerinin ortalamasıdır da denilebilir.

$$MSE = \frac{1}{n} . \sum_{i=1}^{n} (y_i - y')^2 \quad (5)$$

RMSE (Root Mean Square Error - Ortalama Karekök Sapması): MSE' nin kare köküne eşittir.

$$RMSE = \sqrt{\frac{1}{n} . \sum_{i=1}^{n} (y_i - y')^2} \quad (6)$$

MAE (Mean Absolute Error - Ortalama Mutlak Hata): Veri setinden gözlenen değer ile tahmin edilen değerin farkının mutlak değerlerinin toplamının ortalaması ile hesaplanır.

$$MAE = \frac{1}{n} . \sum_{i=1}^{n} |y_i - y'| \quad (7)$$

R^2 (R-Squared – R Kare**):** Açıklayıcılık katsayısı olarak söyleyebiliriz. Bağımsız değişkenlerden tahmin edilebilen bağımlı değişkendeki varyasyonun oranıdır.

$$R^2 = \frac{\sum(y_i - y')^2}{\sum(y_i - y^-)^2} \quad (8)$$

Tablo 3. Modellerin performans karşılaştırması

Yöntem	MSE	RMSE	MAE	R^2-Score
DVM	0.28	0.52	7.58	0.70
YSA	0.25	0.50	6.06	0.76
k-NN	0.25	0.50	6.06	0.76
KA	0.17	0.42	3.03	0.88
LR	0.21	0.46	4.55	0.82
RO	0.18	0.42	7.00	0.94
NB	0.21	0.46	4.55	0.82
AdaBoost	0.12	0.35	1.50	0.88

Çalışmada tüm yöntemlerin performansı doğruluk, kesinlik, geri çağırma ve F1-Skor ölçütleri ile Tablo 4'de verildiği üzere %100 ile en başarılı sonuçların RO ve AdaBoost ile elde edildiği görülmüştür.

Tablo 4. Modellerin Sınıflandırma başarıları

Yöntem	Accuary	Precision	Recall	F_1-Score
DVM	92.42 %	93.24 %	93.19 %	93.18 %
YSA	93.94 %	92.83 %	91.83 %	91.76 %
k-NN	93.94 %	93.50 %	93.19 %	93.17 %
KA	96.97 %	90.90 %	90.02 %	89.94 %
LR	95.45 %	96.39 %	96.37 %	81.75 %
RO	98.48 %	97.31 %	97.28 %	97.28 %
NB	95.45 %	93.40 %	92.72 %	92.67 %
AdaBoost	96.97 %	96.20 %	95.92 %	95.91 %

4. Sonuç

Toprak tüm tarımsal faaliyetler için doğal bir kaynak olmakla birlikte bu faaliyetler için temel oluşturan karmaşık ve heterojen bir yapıya sahiptir. Bu çalışmada toprak özelliklerine göre toprak horizonlarının sınıflandırmalarını tahmin etmek için makine öğrenmesi temelli modellerin performanslarını karşılaştırılmıştır. Modellerin karşılaştırılmasında karışıklık matrisi göz önüne alınmıştır ve elde edilen karşılaştırma sonuçları olarak AdaBoost sınıflandırıcı yöntemi için Doğruluk Oranı (%100), Kesinlik (%100.00), Geri Çağırma (%100.00), F1-Skor (%100.00) ve ROC-AUC Oranı (%100.00) değerleri elde edilmiştir. Sınıflandırma probleminde kullanılan sekiz tahmin modelinde en iyi sonucu veren yöntemler AdaBoost ve ikinci olarak ta RO yöntemi olmuştur.

Literatürde bu konuda yapılmış çalışma araştırması yapıldığında; tam anlamıyla çalışmamızda yapmaya çalıştığımız Mississippi nehri havzasının orta kısmındaki toprak özelliklerine göre horizon tahmin etmeye yönelik bir akademik çalışmaya rastlanamamıştır. Bu sebeple hem akademik hem de gerçek hayatta çalışmanın etkinliği ve özgünlüğü oldukça açıktır.

Bu çalışmada sunulan öğrenme modelleri, toprak profili ile ilgili araştırmalar sırasında katmanları sınıflandırmak için araştırmacılar ve mühendisler tarafından kullanılabilir. Bu çalışmanın sonuçları toprak derinliği dağılım haritası, özellikle toprak derinliği verisi olmayan yerlerde horizon sınıflandırılmasına yönelik yapılacak sonraki çalışmalara ışık tutması hedeflenmiştir.

Kaynakça

Ayhan, S., & Erdoğmuş, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, 9(1), 175-201.

Anisha, P. R., Reddy, K. K. R. C., Apoorva, K., & Mangipudi, M. C. (2021). Early diagnosis of breast cancer prediction using random forest classifier. IOP Conf. Series: Materials Science and Engineering, 1116.

Arabameri, A., Pradhan, B., & Rezaei, K. (2019). CBS'de kesinlik faktörü ve rastgele orman modelleri ile entegre coğrafi ağırlıklı regresyon kullanarak oyuntu erozyonu bölgelendirme haritalaması. Çevre yönetimi dergisi, 232, 928-942.

Ayhan, S., & Erdoğmuş, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, 9(1), 175-201.

Basu, T., & Pal, S. (2018). Identification of landslide susceptibility zones in Gish River basin, West Bengal, India. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 12(1), 14-28.

Bhargavi, P., & Jyothi, S. (2009). Applying naive bayes data mining technique for classification of agricultural land soils. International journal of computer science and network security, 9(8), 117-122.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, 144-152.

Chandrasekaran, S., Sankar, V., Muttil, N., Suganya, K., Suji, S., Selvi, m. T., Selvi, R., & Sudha, S. J. (2013). Monthly flow dorecast for Mississippi River basin using artificial neural networks. Neural Computing and Applications, 24(7-8).

Coopersmith, E. J., Minsker, B. S., Wenzel, C. E., & Gilmore, B. J. (2014). Machine learning assessments of soil drying for agricultural planning. Computers and electronics in agriculture, 104, 93-104.

Côté, M., Osseni, M. A., brassard, D., Carbonneau, È., Robitaille, J., Vohl, M.-C., Lemieux, S., Lavioolette, F., & Lamarche, B. (2022). Are machine learning algorithms more accurate in predicting vegetable and fruit concumption than traditional statistical models? An exploratory analysis. Front. Nutr., 9.

Dolp, C., Cho, S. J., & Finlay, J. C. ve ark. (2022). Predicting high resolution total phosphorus concentrations for soils of the upper mississippi river basin using machine learning. Research Square.

Esra, G. (2020). Toprak bilgisi. Ankara Üniversitesi, Ziraat Fakültesi Toprak Bilimi ve Bitki Besleme Bölümü ders notları. Web Site: 30 Aralık 2021 tarihinde https://acikders.ankara.edu.tr/pluginfile.php/181796/mod_resource/content/1/6.%20Hafta.pdf adresinden erişildi.

Feng, K., Hong, H., Tangi K., & Wang, J. (2019). Decision making with machine learning and ROC curves. SSRN Electronic Journal, arXiv:1905.02810v1.

Foody, G. M., Boyd, D. S., & Sanchez-Hernandez, C. (2007). Mapping a specific class with an ensemble of classifiers. International Journal of Remote Sensing, 28(8), 1733-1746.

Goldhaber, M., & Banwart, S. A. (2015). Soil formation. In book: Soil carbon: Science, management and policy for multiple benefits (81-97).

Guo, Q. H., Kelly, M., & Graham, C. H. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling, 182, 75-90.

Gültepe, Y. (2019). Makine öğrenmesi algoritmaları ile hava kirliliği tahmini üzerine karşılaştırmalı bir değerlendirme. Avrupa Bilim ve Teknoloji Dergisi, 16, 8-15.

Gültepe, Y., & Gültepe, N. (2020). Preliminary study for the evaluation of the hematological blood parameters of seabream with machine learning classification methods. Israeli Journal of Aquaculture-Bamidgeh, 72.

Gültepe, Y. (2021). Performance of lung cancer prediction methods using different classification algorithms. CMC_Computers Materials & Continua, 67(2), 2015-2028.

Gültepe, Y. (2022). Analysis of Alburnus tarichi population by machine learning classification methods for sustainable fisheries. SLAS Technology, 27(2), 1-6.

Hartemink, A. E., Zhang, Y., & Bockheim, J. (2020). Soil horizon variation: A review. Advances in Agronomy, 160, 125-185.

Japkowicz, N. (2011). Performance evaluation for learning algorithms. Cambridge University Press, Cambridge 2011.

Khoi, D. N., Quan, N. T., & Linh, D. Q ve ark. (2022). Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. Water, 14, 1552.

Nelson, A. M., Ashwell, N. E. Q., & Delhom, G. D. Ve ark. (2022). Leveraging big data to preserve the mississippi river valley alluvial aquifer: A blueprint for the national center for alluvial aquifer research. Land, 11(11), 1925.

Pandith, V., Kour, H., Singh, S., Manhas, J., & Sharma, V. (2020). Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. Journal of Scientific Research, 64(2), 394-398.

Pham, V., Weindorf, D. C., & Dang, T. (2021). Soil profile analysis using interactive visualizations, machine learning, and Deep Learning. Computers and Electronics in Agriculture, 191, 106539.

Podgorelec, V., & Zorman, M. (2015). Decision Tree Learning. In: Meyers, R. (eds) Encyclopedia of Complexity and Systems Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27737-5_117-2.

Raschka, S., Patterson, J., & Nolet, C. (2020). Machien learning in Python: Main developments and technology trends in data science. Machine learning, and artificial intelligence, Information, 11(4), 193.

Raunak, J. (2018). Applying naive bayes classification technique for classification of improved agricultural land soils. International Journal for Research in Applied Science and Engineering Technology, 6(5), 189-193.

Rizal, N. N. M., Hayder, G., Mnzool, M., Elnaim, B. M. E., Mohammed, A. O. Y. M., & Khayyat, M. M. (2022). Comparison between regression models, support vector machine (SVM), and artificail neural network (ANN) in river water quality prediction. Process, 10, 1652.

Priya, R., Ramesh, D., & Khosla, E. (2018, September). Crop prediction on the region belts of India: a Naïve Bayes MapReduce precision agricultural model. In 2018 international conference on advances in computing, communications and informatics, 99-104.

Sarker, I. H. (2022). AI-Based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. SN Computer Science, 3.

Sarmadian, F., Keshavarzi, A., Rooien, A., Zahedi, G., Javadikia, H., & Iqbal, M. (2014). Support vector machines based modeling of land suitability analysis for rainfed agriculture. J Geosci Geomatics, 2, 165-71.

Weitzman, J. N., Groffman, P. M., Martel, L., & Kohler, C. (2018). Central Mississippi River Basin LTAR Dataset: NFARM, Inorganic N, & C Production. https://data.nal.usda.gov/dataset/central-mississippi-river-basin-ltar-dataset-nfarm-inorganic-n-c-production-2016-2018.

Whitledge, G. W., Knights, B., & Vallazza, J. ve ark. (2019). Identification of Bighead Carp and Silver Carp early-life environments and inferring Lock and Dam 19 passage in the Upper Mississippi River: insights from otolith chemistry. Biological Invasions, 21(3).

Wong, Y. K. (2021). Machine learning and deep learning technogies. 2nd International Conference on Machine Learning, IOT and Blackchain (MLIOB, 2021).

Ramesh, V., & Ramar, K. (2011). Classification of agricultural land soils: A data mining approach. Agricultural Journal, 6(3), 82-86.

Yang, G. R., & Wang, X. J. (2020). Artificial neural networks for neuroscientists: A primer. Neuron, 107(6), 1048-1070.

Zakaria, M., Al-Shebany, M., & Sarhan, S. (2014). Artificial neural network: a brief overview. International Journal of Engineering Research and Applications, 4(2), 7-12.

Wang, H., Zhang, L., Zhao, J., Hu, X., & Ma, X. (2022). Application of Hyperspectral Technology Combined With Bat Algorithm-AdaBoost Model in Field Soil Nutrient Prediction. IEEE Access, 10, 100286-100299.

https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e.