Journal of Quantum Technologies and Informatics Research
JQTAIR

KUANTUM TEKNOLOJİLERİ VE ENFORMATİK ARAŞTIRMALARI
Volume / Cilt: 2, Issue / Sayı: 1, 2024, pp. 31-43
E - ISSN: 3023-4735
URL: https://journals.gen.tr/index.php/jqtair/
DOI: https://doi.org/10.5281/zenodo.10102956

Received / Gelis: 05.01.2024
Acccepted / Kabul: 22.02.2024

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Öğrenen Makineler ve Fasiyes Ayrımı; İlk Sonuçlar

Machines Learning and Facies Discrimination; Prelimineary Results

Ayetullah Ercel[1] **Emin Uğur Ulugergerli**[2]

[1] Çanakkale Onsekizmart Üniversitesi, Mühendislik Fakültesi, Jeofizik Mühendisliği Bölümü, Türkiye, e-mail: ayetullahercel@gmail.com
[2] Çanakkale Onsekizmart Üniversitesi, Mühendislik Fakültesi, Jeofizik Mühendisliği Bölümü, Türkiye, e-mail: emin@comu.edu.tr

Öz

Makine öğrenmesi uygulamalası kuyu loglarından jeolojik istifin ayırtlanması için kullanılmıştır. Kuyu logu verilerinden fasiyesleri tahmin etmek için makine öğrenmesi yöntemi sınıflayıcılarından biri olan değişim arttırıcı türev (gradient boosting) algoritmasından yararlanılarak ağaç tabanlı (tree-based) bir eğitim modeli geliştirilmiştir. Tahmin başarı oranını arttırmak için veri topluluğu üzerinde iyileştirmeler yapılmıştır. Deneme veri topluluğu olarak, Society of exploration geophysics tarafından makine öğrenmesi için önerilen, Kansas (ABD) eyaletindeki kuyu verileri kullanılmıştır. Çalışmada makine öğrenmesi yöntemi olarak değişim arttırıcı türev ile sınıflama algoritması ile tekil deneme kuyusu üzerinde % 57, komşu fasiyes bilgisi ile %88 oranında doğruluk ile elde edilen tahmin sonuçları elde edilmiştir.

Anahtar Kelimeler: makine öğrenmesi, gradient boosting, kuyu logları

Abstract

Machine learning applications have been used to distinguish the geological sequence from well logs. A tree-based training model was developed using the gradient boosting algorithm to predict facies from the well log data set. Testing data were, obtained from Kansas state (USA), recommended for machine learning studies by the Society of exploration geophysics. In the study, the gradient boosting algorithm predicted results with 57% accuracy on the single trial well and with 88% accuracy with using well log information from neighbourhood.

Keywords: machine learning, gradient boosting, well logs

Citation/Atıf: ERCEL, A. & ULUGERGERLİ, E.U. (2024). Öğrenen Makineler ve Fasiyes Ayrımı; İlk Sonuçlar. *Kuantum Teknolojileri ve Enformatik Araştırmaları*. 2(1): 31-43, DOI: 10.5281/zenodo.10102956

Corresponding Author/ Sorumlu Yazar:
Ayetullah Ercel
E-mail: ayetullahercel@gmail.com

Bu çalışma, Creative Commons Atıf 4.0 Uluslararası Lisansı ile lisanslanmıştır.
This work is licensed under a Creative Commons Attribution 4.0 International License.

GİRİŞ

Yeraltı kaynaklarının aranmasında kuyu logları yaygın olarak kullanılmaktadır (örn. Anderson 2001, Pekiner 2002). Yüksek ayrımlılık gücüne karşın karmaşık jeolojik istiflerde hatalı yorumların yapılması en önemli sorunlardan biridir (örn. Dubois vd. 2007). Bu soruna bir çözüm olarak insan tabanlı ve eldeki var olan veriler ile edinilecek bilgi birikiminin kullanımı en yaygın yoldur (URL1). Bu çalışmada, günümüzde her alanda olduğu gibi yeraltının araştırılmasında da geniş bir kullanım alanı bulan yapay zeka uygulamaları (YZU) kuyu loglarından jeolojik istifin ayırtlanması için kullanılmıştır (örn., Ahmadi vd. 2013).

Jeolojik tanımlamada yaygın olarak kullanılan fasiyes terimi, belirgin bir jeolojik süreci ve birikme ortamını ve koşullarını yansıtan tortul birimlerin ve özelliklerinin genel tanımı olarak kullanılagelmektedir. Fasiyeslerin birbirleriyle olan ilişkilerinde kademeli geçişler izlenebilir ve özellikleri birbirine oldukça yakın komşu fasiyesler oluşabilir. Bu ve benzeri sorunlar litolojik olarak benzer fasiyeslerin sınırlarının tanımlanmasında zorluk oluşturmaktadır (Dubois vd. 2007)

Çalışmada kuyu logu verisinden yola çıkarak fasiyesleri tahmin etmek için makina öğrenmesi algoritması geliştirilmiştir. Algoritmada Python programlama dili için geliştirilen "scikit-learn-SCL" açık kaynak makina öğrenmesi kütüphanesi (Pedregosa vd. 2011) kullanılarak "gradient boosting" (GB) algoritması ile bir sınıflayıcı karar ağacı modeli oluşturulmuş ve tahmin başarı oranını geliştirmek için veri topluluğu üzerinde iyileştirmeler yapılmıştır.

Veri olarak, Society of exploration geophysics (SEG) tarafından YZU için önerilen, Kansas (ABD) eyaletinde elde edilmiş halka açık kuyu verileri kullanılmıştır. Anılan veriler hem SEG tarafından düzenlenen YZU çalıştaylarında hemde Kansas Üniversitesinde Sinir Ağları ve Bulanık Sistemler (URL2) üzerinde yapılan çalışmalarda kullanılmıştır (Bohling ve Dubois 2003, Dubois vd. 2007).

İzleyen bölümlerde kuyu logları kısaca tanıtılmış, YZU nun makine öğrenmesi uygulamalarından biri olan GB ile kuyu logu verilerinin bilgisayar tarafından tanımlanması, sınıflayıcı algoritması ve verilerin sınıflandırılması açıklanmıştır.

Kuyu logları

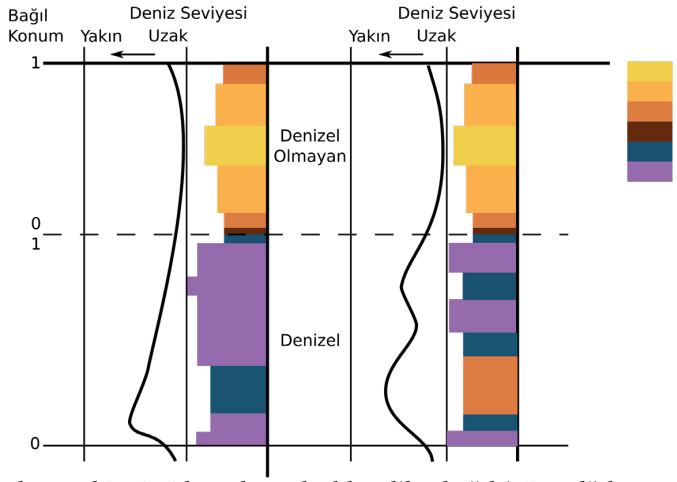
Kuyu logları, jeolojik birimlerin fiziksel özelliklerini doğrudan birim içinde yapılan ölçümler ile belirlemeye çalışan bir jeofizik yöntemdir (örn. Darling 2005). Bu çalışmada beş farklı log verisi ele alınmıştır; gamma ışını (Gamma Ray - GR), özdirenç (Resistivity), fotoelektrik etki (Photoelectric effect - PE), netron-yoğunluk (Neutron - Density - ND) gözeneklilik farkı (Porosity Difference) logları kullanılmıştır (örn. Glover 2000). Bu loglara dahil edilmek üzere iki farklı jeolojik tanımlayıcı koşul (constraint) işlemlere dahil edilmiştir; denizel veya karasal ortam belirleyici (nonmarine-marine indicator - NM_M) ve göreceli konum (relative position- RELPOS).

Elde edilen verilerin yüzey jeofizik yöntemlerden farkı özdirenç, hız vb. fiziksel büyüklükleri doğrudan birim içinde ölçülmesidir. Ancak, elde edilen veriler algılayıcıların (probe) boyutu ile jeolojik birimin kalınlık ilişkisine bağlı olarak gerçek değerinden saparlar. Bu aşamada gerçek fiziksel değerler elde etmek için abaklar (URL1) veya modelleme çalışmaları (örn. Anderson 2001, Ulugergerli 2017) yapılmalıdır. Her iki yaklaşımda kendi içinde sorunlar barındırır. Bunun başlıca nedenleri jeolojik birimlerin tekdüze olmaması kimyasal içeriğinin yer bağımlı olması vb. nedenler sayılabilir. Makine öğrenmesinin log verilerine uygulanması bu soruna çözüm için farklı bir yaklaşım olarak önerilmektedir.

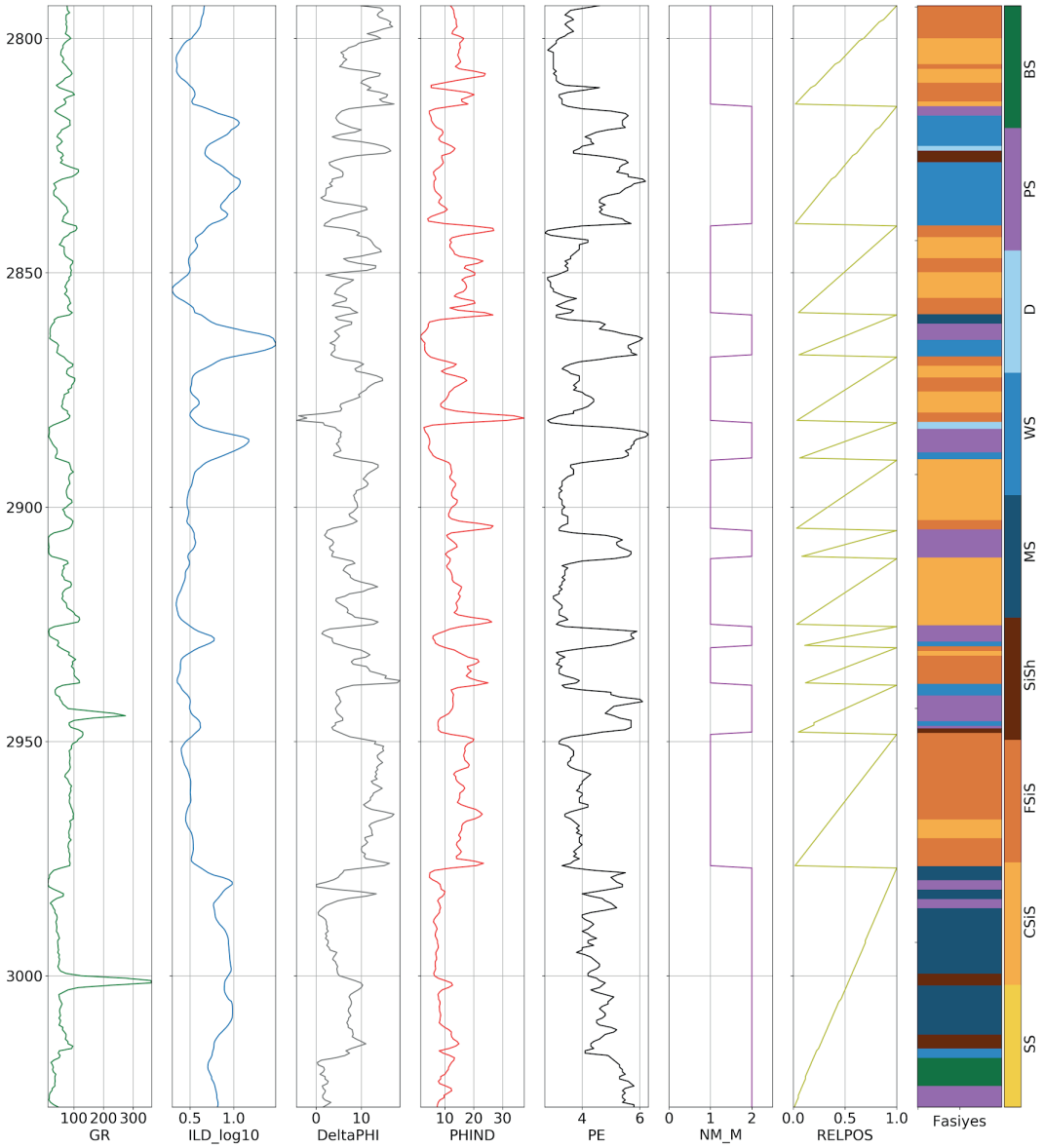
Log verisi stratigrafik olarakta bir anlam taşıdığından bu veri ile ortamın tanımlamasını fiziksel olarak yapmak olasıdır. Bu nedenle, kuyu loglarına göre ayrımlama, kuyu verisinin ön işlemesinin önemli bir adımıdır. Log eğrilerinden elde edilen sonuçların karşılaştırılması, önerilen modelin kuyunun bulunduğu bölgesindeki birimlere uyumu doğruluğunun değerlendirilmesine olanak sağlar.

Öğrenen makinelerin katkısı

Fasiyeslerin sınıflandırılması sedimanların ölçülen belirli fiziksel veya kimyasal özelliklerine



Şekil 1. Sedimanter kayaç döngüsü baz alınarak elde edilen bağıl (göreceli) konum ve ortam belirleyicisinin fasiyeslere göre etkisinin eğri ile ifadesi. (Dubois vd 2007'ten değiştirilerek alınmıştır.)



Şekil 2. Kuyu log verisi örneği

göre bir gruba atanması işlemidir. Sınıflandırma için kuyudan çıkarılan karotların analizlerinden yararlanılır, fakat maliyetinden dolayı yaygın değildir (örn. Bestagini vd. 2017).

Buna karşın dolaylı bir yaklaşım olan kuyu logları da fasiyeslerin sınıflama işlemlerinde kullanılabilir. Ancak log verilerinin geleneksel yöntemler yardımı ile kullanımı oldukça zahmetli ve zaman gerektiren bir süreçtir. Bu nedenle, çeşitli alternatif yaklaşımlar önerilmiştir. İlk çalışmalarda istatistiksel yöntemler geliştirilmiştir (Wolf vd. 1982, Busch vd. 1987). Daha sonra Wolf vd. (1982) ile Busch vd. (1987) sınıflandırma işlemi için yapay sinir ağlarının kullanılmasını önermiştir (Baldwin vd. 1990, Rogers vd. 1992).

Günümüzde özellikle son birkaç yılda bilgisayarların hesaplama güçlerinin artması ve depolama alanlarının genişlemesine koşut olarak gelişmekte olan büyük veri (big data) olgusu ile, makina öğrenmesi teknikleri farklı alanlarda ele alınmaya başlamıştır. Makina öğrenmesi teknikleri günümüzde araştırma grupları tarafından da irdelenmeye başlanmıştır (Smith ve Treitel 2010, Zhang vd. 2014, Zhao vd. 2015, Kobrunov ve Priezhev 2016).

Bu bağlamda Hall (2016) tarafından makina öğrenmesi teknikleri ile basit bir fasiyes sınıflama eğitimi sunmuştur. Eğitimde küçük bir veri topluluğu olarak Kansas'ın güneyindeki Hugoton gaz alanında bulunan 10 farklı kuyudan 7 log bilgisi kullanılmıştır. Bu veri topluluğundaki loglar yorumlanarak ve karotiyer bilgisi kullanılarak fasiyes sınıflaması yapılmıştır (Dubois vd. 2007). Bu veri topluluğundan yola çıkarak fasiyes bilgisi bilinmediği kabul edilen bir deneme kuyu verisinin sonuçları tahmin edilmiştir.

Makina öğrenimi teknikleri veri topluluğunun niteliklerine göre danışmalı öğrenme (supervised learning - SL), danışmasız öğrenme (unsupervised learning -USL) ve türevi olan yöntemlerden oluşur. SL, önceden gözlemlenmiş ve sonuçları bilinen verileri kullanarak bu sınıflamayı tekrardan üretebilecek model parametrelerinin belirlenmesini tanımlar. USL ise giriş çıkış ilişkisini ele almadan verinin kendi içindeki örüntüyü kullanarak tanımlama işlemidir (Nizam ve Akın 2014). Bu çalışma için SL yaklaşımı kullanılmış ve $f_{d,w}$ girdisi ve $y_{d,w}$ çıktısı eşleştirilerek öğrenme (eğitim) sağlanmıştır. Eğitimi izleyen aşamada, eğitilen modelden yola çıkarak bilinmeyen verinin sınıflandırma işlemini algoritmaya yaptırılmaktadır (örn. Alpaydın 2009).

Bu amaçla, kontrol edilmiş ve sınıflanmış veri topluluğundan yola çıkarak özellik vektörünü $f_{d,w}$

$$f(x)_{d,w} = [f_{d,w}^{GR}, f_{d,w}^{OZD}, f_{d,w}^{FE}, f_{d,w}^{Nfark}, f_{d,w}^{Nort}, f_{d,w}^{NM}, f_{d,w}^{GK}] \quad (1)$$

olarak tanımlayabiliriz. Önceki çalışmalarda (Hall 2016, Bohling ve Dubois, 2003, Dubois vd. 2007) sunulan aynı düzenlemeyi varsayarak, ile işaretlenen her bir kuyu için derinlikte yedi farklı sayıl log değeri bulunmaktadır. Bunlar;

- Gamma ray $(f_{d,w}^{GR})$ doğal oluşum radyoaktivitesini,
- Özdirenç $(f_{d,w}^{OZD}$ yerin elektrik akımı karşı gösterdiği direncini
- Fotoelektrik etki $(f_{d,w}^{FE})$ ışık ışınları ile aydınlatılan fasiyeslerin elektron emisyonunu,
- Netron yoğunluk- porozite farklı $(f_{d,w}^{Nfark})$, ortalama netron-yoğunluk porozite : $(f_{d,w}^{Nort})$ fasiyes yoğunluğu ile ilişkisini sunan loglardır. Bunlara ek olarak;
- Karasal /Denizel belirleyici $(f_{d,w}^{NM})$,
- Göreceli konum $(f_{d,w}^{GK})$

veri incelemesi veri topluluğuna dahil edilmiştir. Fasiyes sınıfları (etiketleri)$(y_{d,w})$

$$y_{d,w}^{\square} \in \{SS, CSiS, FSiS, SiSh, MS, WS, D, PS, BS\}$$

Kümesi ile tanımlanmıştır. SL' de $f(x)_{d,w}$ girdisi ile çıktısını birbirine eşleyen bir işlev (sınıflayıcı) aşağıdaki gibi bir görselleştirilebilir.

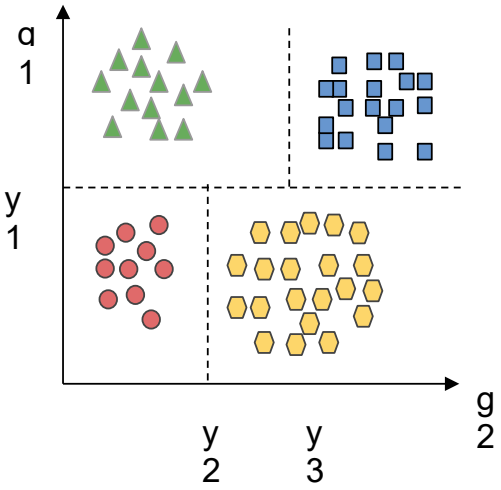
Verilerin Sınıflanması

Eğitimde kullanılacak verilerin sınıflandırılması amacıyla için birçok farklı algoritma geliştirilmiştir (örn. Bishop 2006). Çalışmada kullanılan veri topluluğunun yedi özellikten oluşması ve sınırlı miktarda eğitim verisi olduğunu göz önüne alınarak sınıflandırma işlemi için basit bir karar ağacı topluluğu algoritması kullanılmıştır (Breiman vd. 1984, Friedman 2000). İzleyen bölümde yöntem

hakkında kısa bir bilgi verilecektir. Konunun ayrıntılı açıklaması için Bishop (2006) ve Alpaydın (2009) çalışmalarına bakılabilir.

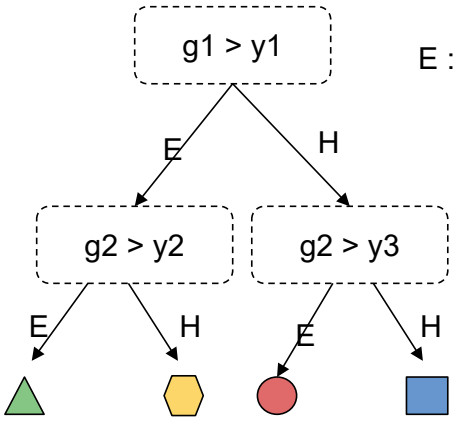
Karar ağaçları verilerin sınıflandırılması amacıyla kullanılan başlıca tekniklerden biridir. Sınıflandırma işlemini veriden elde ettiği kuralları öğrenerek yapar. Bir ağaç yapısını andıran bu yöntemde ağaç düğümleri özellik vektörünün elemanlarını belirtirken, ağacın dalları ise bu elemanlara yapılacak işlemleri göstermektedir (Mitchell 1997, Drucker 1996).

Karar ağacı sınıflayıcısı izleyen örnekle açıklanabilir. Eşitlik 1 de verilen özellik vektörü İki sayıl özellik barındırsın, $f = [f(g1), f(g2)]$ sınıf kümesi, y ise dört farklı fasiyes (sınıf, etiket) içersin.



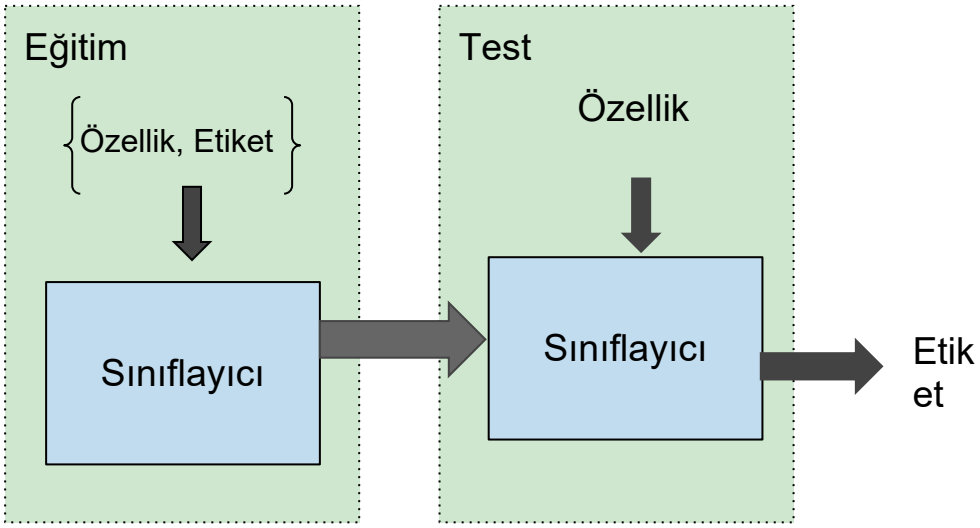
Şekil 4. Özellik vektörünün farklı renkler ve şekil ile tanımlanmış sınıflar için dağılımı

Bu örnekte, (g1, g2) her biri farklı bir sınıfa ait özelliği içeren dört bölüme kolayca ayrılabilir. Eğitim sırasında, karar ağacı sınıflandırıcısı, farklı sınıflara ait özellikleri ayırmak için verileri bölerek eşik değerlerini (örn., y1, y2 ve y3) öğrenir (belirler). Bu işlem uygun bir şekilde tanımlanmış bazı amaç fonksiyonlarını en küçükleyerek yapılır (Rokach ve Maimon 2005). Sınıflayıcı modeli bilinmeyen veriye uygulandığında, veriyi sınıflandırmak için, g1 ve g2 bileşenlerini öğrenilen üç eşikle karşılaştırmak yeterlidir. Bu adım, mantıksal olarak bir ağaç diyagramı olarak gösterilebilir (Şekil 5). Bu nedenle bu tür sınıflandırıcılar karar ağacı olarak bilinir (Breiman vd. 1984).



Şekil 5. Karar ağacı diyagramı

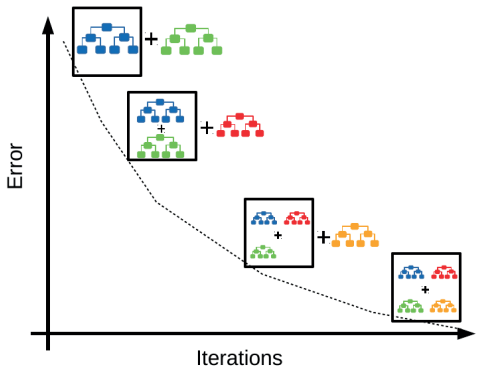
Uygulamada sık rastlanılan sorunlardan biri, algoritmanın eğitim setini "ezberleme" eğilimi göstermesidir. Bunun sonucunda elde



Şekil 3. Önerilen algoritmanın genel hatları, eğitim ve deneme aşamaları.

edilen eşik değerleri başka veri setlerine uyum gösteremezler (Ho, 1998). Bu etkiyi önlemek için, birkaç küçük ağaçtan elde edilen sonuçların birleştirilmesi yönünde çeşitli yaklaşımlar geliştirilmiştir (Ho 1995, 1998, Friedman 2000).

Schapire (1990) tarafından önerilen değişim arttırıcı (boosting) algoritma 2000'li yıllara kadar gelişim göstermiştir (Friedman 2001, Friedman vd. 2000, Freund vd. 1996). Anılan algoritmada amaç zayıf temel öğreniciler ile güçlüleri bir araya getirip güçlü bir sınıflandırıcı oluşturmaktır. Bu amaç için geliştirilmiş birçok algoritma olmakla birlikte anılan çalışma Friedman (2001) tarafından geliştirilen değişim arttırıcı türev sınıflayıcı (Gradient Boosting Classifier – GBC) yaklaşımına dayalıdır ve gözlemleri farklı alt kümelere bölerek çalışır. Her alt kümeden, türev tabanlı hatayı en küçükleme (gradient descent-GD) üzerine kurulu bir algoritmayı çözerek yeni bir karar ağacı oluşturmak için sınırlı oranda özellik seçilir (Friedman, 2001). GBC yöntemi özelliği gereği önceki (doğru olmayan) eğitim sonuçlarından yola çıkarak farklı bir model oluşturur. Her yeni model bir önceki modellerin genel hatasını en küçüklemeye çalışır. Yeni bir özellik alan f vektörünün sınıflandırılması gerektiğinde, ilk önce oluşturulan her farklı ağaçta denenir. Her ağaç bir aday sınıfı sağlar. Tüm ağaçlardan elde edilen sonuçlar daha sonra tek bir kararda birleştirilir (y^t). Böylece karar ağaçları öğrenme tekniği ile eğitilen sistem için bir tahmin modeli üretilmiş olur. Sisteme eklenen her yeni modelin genel hatayı nasıl düşürdüğü Şekil 6'de gösterilmiştir (Kalaycı, 2018).



Şekil 6. Sistemin model eklenerek en küçüklemesi (Kalaycı 2018'dan değiştirilerek alınmıştır.)

Bu bölümden ilk olarak Friedman (2001) tarafından ortaya atılmış öğrenme algoritması GBC'nin basitleştirilmiş mantığını sunacağız. Algoritmaların ve özelliklerinin matematiksel çıkarımı bu makalede ele alınmayacaktır.

SL yönteminde, eğitim seti için f özellik vektörüne karşılık gelen etiket değerleri olmak üzere veri topluluğu uzunluğu için amaç $\hat{f}(x)$ fonksiyonunu bir kayıp (loss) fonksiyonu $L(y, f(x))$ ile $\hat{f}(x)$ fonksiyonu için

$$\hat{f}(x) = min \sum_{xy}^{\square} [L(y, f(x))] \quad (2)$$

biçiminde bir yaklaşımda bulunmaktır. Burada $[L(y, f(x))]$ türevlenebilir bir kayıp fonksiyonunu tanımlar. Friedman (2001) tarafından geliştirilmiş bu yöntem en küçükleme için kayıp fonksiyonu olarak genellikle Eşitlik 3'de bulunan karesel hata ortalamasını kullanır.

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x)_i^t)^2 \quad (3)$$

Burada n veri topluluğu büyüklüğü, y_i $i.$ hedef değeri, $f(x)_i{}^t$ tahmin değeri, olmak üzere $L(y_j, f_j{}^t)$ kayıp fonksiyonunu tanımlar. tane veri için iyileştirilmiş tahmin modeli,

$$\hat{f} = min\{\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))\} \quad (4)$$

şeklinde elde edilir. GB nin amacı, değerlendirilen kayıp fonksiyonu $u^{[k]}$ yı $u^{[k-1]}$ in negatif gradyan vektörüne

$$u_{i,k} = -[\frac{-\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{k-1}(x)} \quad (5)$$

şeklinde kestirmektedir. GB algoritması tahmini zor verilere odaklanır ve yineleme yoluyla yeterince eğitilemeyen model parametrelerinin kestirilebilirliğini arttırır (Mayr vd. 2014).

Karışıklık Matrisi ve Başarı ölçütü

Sınıflandırma yöntemlerinden elde edilen modellerin başarılarını değerlendirmek için doğruluk (Accuracy), duyarlılık (Recall), kesinlik (Precision) ve F1-ölçütleri kullanılır. Başarı, doğru ve yanlış sınıflara atanan örneklerin sayı nicelikleri ile ilgilidir.

Sınıflandırıcının tahmin yeteneğini değerlendirmek amacıyla hedef değişken için tahmin değeri ile gerçek değeri karşılaştırılır. Ulaşılan sonuçların başarı bilgileri karışıklık matrisi ile ifade edilir. Bu matriste satırlar değerlendirilen verideki örneklere ait gerçek sayıları, kolonlar ise tahmin değerlerini ifade eder (Tablo 2).

2x2 örnek olarak verilen matriste DN, YP, YN ve DP modelin gözlem değerlerini belirtmektedir. DN ve DP köşegeni doğru tahmin değerlerini, YP ve YN köşegeni yanlış tahmin değerlerini göstermektedir.

Model başarısının ölçülmesinde kullanılan doğruluk oranı, doğru sınıflandırılmış örneklerin toplam örnek sayısına oranı ile elde edilir.

$$Doğruluk = \frac{DP+DN}{DP+DN+YP+YN} \quad (6)$$

Kesinlik değeri, doğru olumlu (pozitif) olarak sınıflandırılan gözlem sayısının, tahmin değerleri olumlu sınıfı olan tüm gözlemlere oranı şeklinde hesaplanır.

Tablo 1. GBC algoritmasi (Friedman (2000) den değiştirilerek alınmıştır).

Algoritma: Türev Takviyeli Sınıflayıcı Algoritması

Girdiler

Girdi verisi $(f(x),y)_{i=1}^{n}$
Yineleme sayısı M
kayıp fonksiyonu $L(y,f)$
Bir dizi temel öğrenici $h_1(x_1),..h_p(x_p)$ özelleştirilir.

Algoritma

1. Modeli bir tahmini $\hat{f}_0$ ile başlat
2. Döngü $k = 1,\ldots,n$ yap
3. Negatif gradyeni hesapla $u_k(x)$
   $u_{i,k} = -[\frac{-\partial L(y_i,f(x_i))}{\partial f(x_i)}]_{f(x)=f_{k-1}(x)}$
   $i = 1,\ldots n$
4. Her bir temel öğrenici $h_j$ye u_k negatif gradyan vektörü uyarlanır.
   $u_k \rightarrow^{temelöğrenici} \rightarrow h_{k,j}(x_j)$❑ $\quad j = 1,..,p$
5. en iyi azalan türev değerini ρ_t için :
   $\rho_t = min_\rho \sum_{i=1}^{n} L[y_i, \hat{f}_{k-1}(x_i) + \rho h_{k,j}(x_i)]$
6. kestirim fonksiyonunu güncelle
   $\hat{f}_k \leftarrow \hat{f}_{k-1} + sl.\rho_k h_{k,j}(x_j)$ Burada sl (küçük adım uzunluğu) öğrenme oranıdır. 0<sl<=1
7. Döngü sonu

Tablo 2. karışıklık matrisi

	Tahmin edilen olumsuz	Tahmin edilen olumlu
Gerçek olumsuz	Doğru olumsuz (DN)	Yanlış olumlu (YP)
Gerçek olumlu	Yanlış Olumsuz (YN)	Doğru olumlu (DP)

$$Kesinlik = \frac{DP}{DP+YP} \quad (7)$$

Duyarlılık, doğru sınıflandırılmış olumlu örnek sayısının toplam olumlu örneklere oranı olarak elde edilir.

$$Duyarlılık = \frac{DP}{DP+YN} \quad (8)$$

F1-ölçütü, kesinlik ve duyarlılık değerlerinin harmonik ortalamasıyla elde edilir. 0 ile 1 arasında olan bu ölçüt değeri 1'e yaklaşması modelin başarısını gösterir.

$$F1 = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık} \quad (9)$$

Uygulama

Kullandığımız veri topluluğu (Hall 2016), fasiyes tipleri etiketlenmiş 10 farklı kuyudan (4149 örnek) elde edilen yedi farklı log kaydından oluşmaktadır (Şekil 7). Birçok makine öğrenme problemi için bu, küçük bir eğitim seti olarak kabul edilebilir.

Sınıflandırma yöntemlerinin veriye doğrudan uygulanması yerine, verinin önceden çeşitli yöntemlerle sayısal değere dönüşümü ile daha fazla başarı sağlandığı bilinmektedir. Özellik vektörünün elemanları arasında büyük ölçek farkı varsa, hepsinin aynı ölçeğe getirilmesi hem öğrenmeyi hemde öğrenme başarısını arttırır. İzleyen bölümlerde veri topluluğuna sınıflandırma algoritmalarını uygulamadan önce ölçeklendirme işlemi yapılmıştır. Ölçekleme için SCL kütüphanesindeki "StandardScaler" ölçekleyicisini kullanılmıştır (Pedregosa vd. 2011).

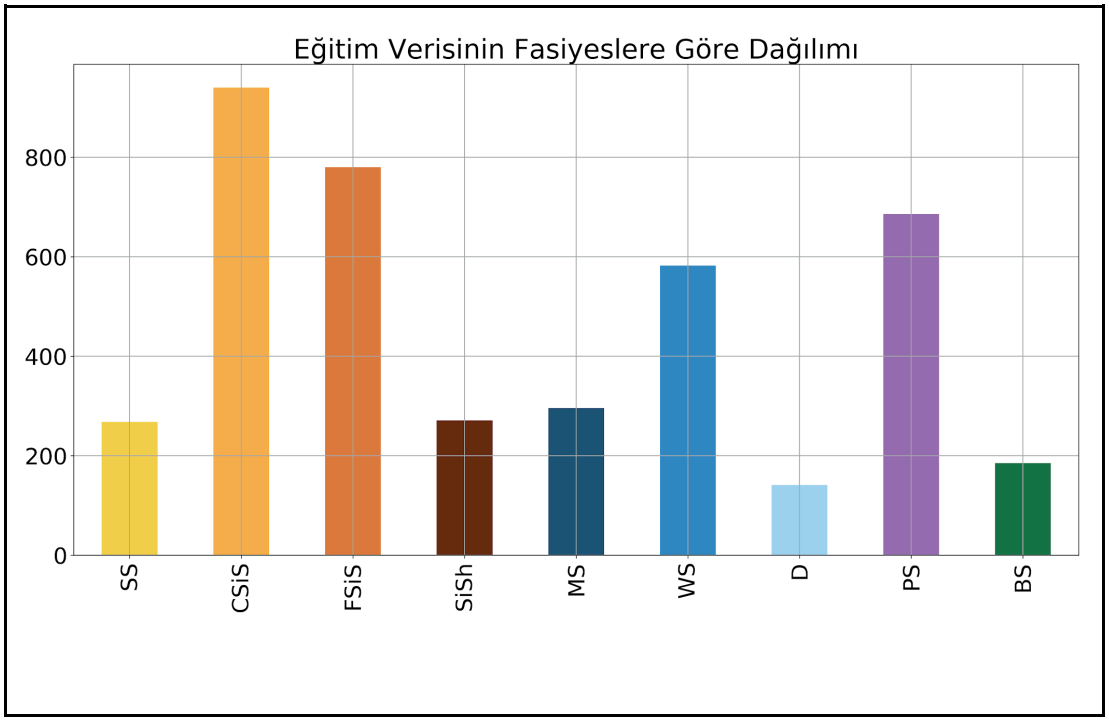
Tabloda 3 çalışmaya konu olan fasiyeslerin kısaltılmış etiketleri ve komşuları tanımlanmıştır. Fasiyes ile ilgili bilgiler karotiyer ile örnek alınarak ve sondaj sırasında yüzeye ulaşan kırıntılardan elde edilmiştir. Üç fasiyes karasal kökenli (SS, CSiS, FSiS) ve altı tanesi denizel kökenli (SiSh, MS, WS, D, PS ve BS). Fasiyelerin birbirlerine yakınlığı karışmalarına neden olur. Fasiyesler ve bunların yakın komşuları, Tablo 3 de kısaltmalar ve tutarlı renklerle belirtilmiş ve işaretlenmiştir.

Fasiyesler içinde kireçtaşı türü olan "bafflestone" (BS) kısıtlı sayıda yer almaktadır. Bu nedenle fasiyesin sınıflama yetkinliğini, diğer bir deyişle tahmin edilebilirliğini, arttırmak için eldeki veri setine ek olarak, tamamı BS taşından oluşan bir kuyu daha eklenmiştir (Dubois et al. 2007).

PE log değeri veri topluluğundaki 4149 örnekten sadece 3232 örnekte bulunmaktadır. Bu problemi ortadan kaldırmak için eksik kalan PE değerleri için sıfır değeri atanmıştır.

Komşu katmanlardaki fasiyeslerin ilişkili olduğunu göz önünde bulundurarak, eğitim sonrası sınıflandırmayı iyileştirmek için izleyen işlem yapılmıştır; Bütün fasiyesin tek bir birimden oluştuğu varsayılırsa, buna bağlı olarak farklı tek bir fasiyes bu hakim fasiyesin içinde yer alırsa bunun tek sebebinin yanlış sınıflandırma hatasından kaynaklandığı şeklinde yorumlanabilir. Bu tür hataları gidermek için, elde edilen sınıf tahminlerine () bir medyan süzgeç uygulanabilir. Bu süzgeç, ara katmanlardaki fasiyesleri en fazla sayıdaki komşu fasiyesle değiştirir.

Giriş verisi olarak veri topluluğunun %25i doğrulama verisi %75i eğitim verisi olarak seçilmiş ve çok-sınıflı (9 farklı fasiyes) bir problemle başa çıkmak için bir seçim stratejisi olan bire-bir (One-vs-One) işlemi uygulanmıştır (Bishop, 2006). İlk adım olarak deneme yanılma yolu ile GBC algoritmasının en iyi çözün sağladığı, yineleme sayısı, öğrenme oranı, özellik sayısı vb parametreleri tanımnıştır. Eğitim sonucunda, eğitimin başarısı önceki bölümde tanımlanan başarı ölçütleri çerçevesinde değerlendirilmiştir. Şekil 7a'da, önerilen yöntemle elde edilen karışıklık matrisini göstermektedir. Köşegenleri baskın bir dizey iyi bir sınıflayıcı sonucuna işaret etmektedir. Daha sonra, eğitimli modeli test etmek için fasiyes bilgisi içermeyen iki kuyu seti (474'ün STUART ve 356'nın CRAWFORD'a ait 830 örnek) elde edilen modelin başarısını gözlemlemek için kullanılmıştır. Son olarak, Hall (2016) tarafından önerildiği gibi, algoritmayı f-ölçüsü açısından değerlendirilmiştir.



Şekil 7. Veri topluluğunun fasiyeslere göre dağılımı

Tablo 3. Kuyu log analizlerinden elde edilen fasiyesler komşuluk ilişkileri ve veride karşılık geldikleri etiket ve renk tanımlamaları.

Fasiyes Renk Temsili	Fasiyes No.	Etiket	Fasiyes	Komşu fasiyesler	
	1	SS	Nonmarine sandstone (Tr) Karasal kumtaşı	2	Denizel olmayan
	2	CSiS	Nonmarine coarse siltstone İri taneli silttaşı	1,3	
	3	FSiS	Nonmarine fine siltstone Karasal taneli ince silttaşı	2	
	4	SiSh	Marine siltstone and shale Denizel silt taşı ve şist	5	Denizel
	5	MS	Mudstone (limestone) Çamurtaşı	4,6	
	6	WS	Wackestone (limestone) Kireçtaşı	5,7	
	7	D	Dolomite Dolomit	6,8	
	8	PS	Packstone-grainstone (limestone) Taneli Kireçtaşı	6,7,9	
	9	BS	Phylloid-algal bafflestone (limestone) Algil Engeltaşı	7,8	

BULGULAR VE TARTIŞMA

Bazı fasiyelerin neden diğerlerinden daha iyi belirlendiğini açıklamak için, Şekil 7b'de her bir kuyuda hangi birimden toplam kaç adet gözlemin bulunduğuna ilişkin bir tablo verilmiştir. Burada her bir fasiyes için farklı sayıda örneğin olduğu görülmektedir, bu durum sınıflandırma problemini dengesiz hale getirir. Örneğin, çamurtaşı (MS), veri topluluğunda iri taneli silttaşına (CSiS) nazaran daha az örnek vardır, bu nedenle, sınıflandırıcının bu fasiyesi yüksek doğrulukla tahmin etmesi beklenemez. Bu durumun aksine, denizel olmayan iri taneli silttaşı (CSiS) örneği daha yaygındır. Bundan dolayı en iyi tanımlanan fasiyelerden biridir.

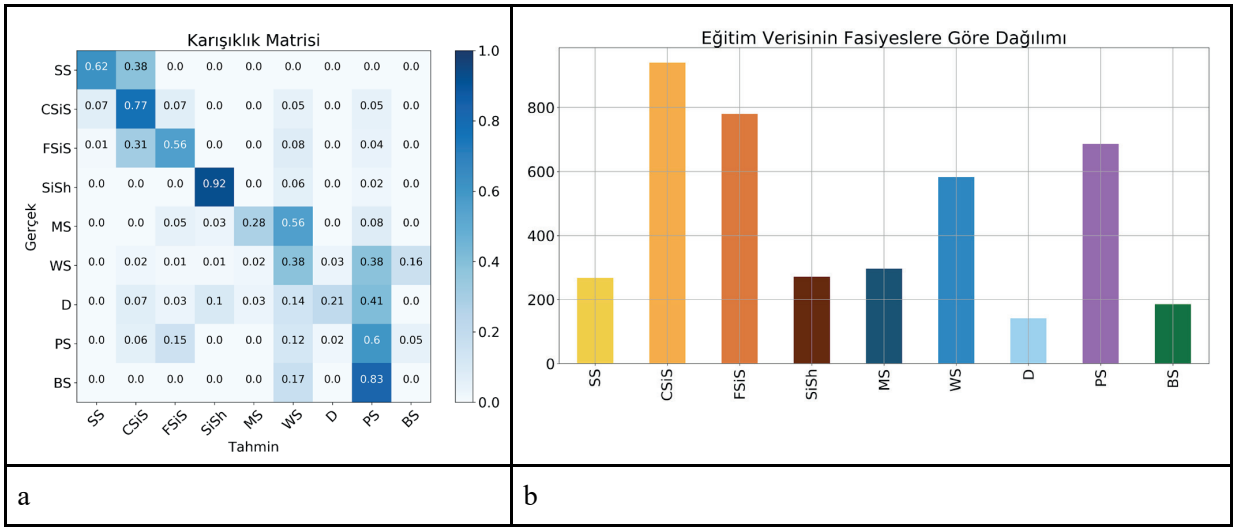
Buna göre, fasiyeslerin doğruluk oranları Şekil 7a'de görüldüğü üzere, bazı fasiyesleri sınıflandırmanın diğerlerinden daha kolay olduğunu görülmüştür. Örnek olarak, Karasal iri silt taşının (CSiS) %77'si doğru şekilde tespit edilmiştir. Bunun aksine, çamurtaşı (MS) ile (WS) birbirine yakın değerler ürettiğinden (0.28, 0.38) doğru tahmin edilebilmeleri zorlaşmıştır.

Fasiyes sınıflandırmasına önerilen yaklaşımın uygulanması, denizel ve denizel olmayan fasiyes için önemli sonuçlar göstermiştir. Jeolojik olarak kısıtlayıcı bir girdinin eklenmesi, modelin tahmin başarısını artırmakla birlikte alana özgü heterojenliği ve değişkenliği tespit etmekte etkili olduğu gözlenmiştir. Deneysel sonuçlar, denizel ve denizel olmayan bölgeleri NM_M eğrisi ile doğrulanan bir şekilde sınıflandırıldığını göstermektedir. (Şekil 8a)

Şekil 8. a) Eğitim modeli ile fasiyes bilgisi tespit edilen kuyunun sınıflandırma sonucu. b) Karşılaştırılan bir diğer algoritma sonucu

Hall (2017), sınıf dengesizliği nedeniyle tahmin sonucu 0,16 başarı değeri civarında olduğunu vurgulamıştır. Algoritmamız, 0.57 f başarı değerine ulaşmaktadır. Fasiyeslerin komşuluk ilişkileri baz alınarak değerlendirildiği takdirde bu sonuç 0,87 başarı değerine kadar yükselmektedir.

Kullandığımız veri topluluğu sınırlı sayıda log verisi içermektedir. Veri topluluğunda yeterince temsil edilmeyen fasiyesler için daha düşük doğrulukta tahminler elde edilmiştir. Gelecekteki çalışmalarda modelin (algoritmanın) doğruluk oranını iyileştirmek amacıyla ek log verisi (yoğunluk, Vs, Vp, vb.) veri topluluğuna dahil edilebilir. Örneğin Vp/Vs oranı doğrudan poisson oranı ile ilişkidir ve bu da fasiyes biriminin yanal ve eksenel gerilmesini tanımlar (Gercek, 2007). Bu bilgi de fasiyes sınıflamasında bir başka belirgin özellik olarak kullanılabilir. İlave jeolojik özelliklerin (örn. her fasiyenin mineralojisi) log verileri ile birlikte kullanımının, karotlu kuyuların fasiyes sınıflandırmasında önemli bir gelişmeye yol açacağı öngörülmektedir.

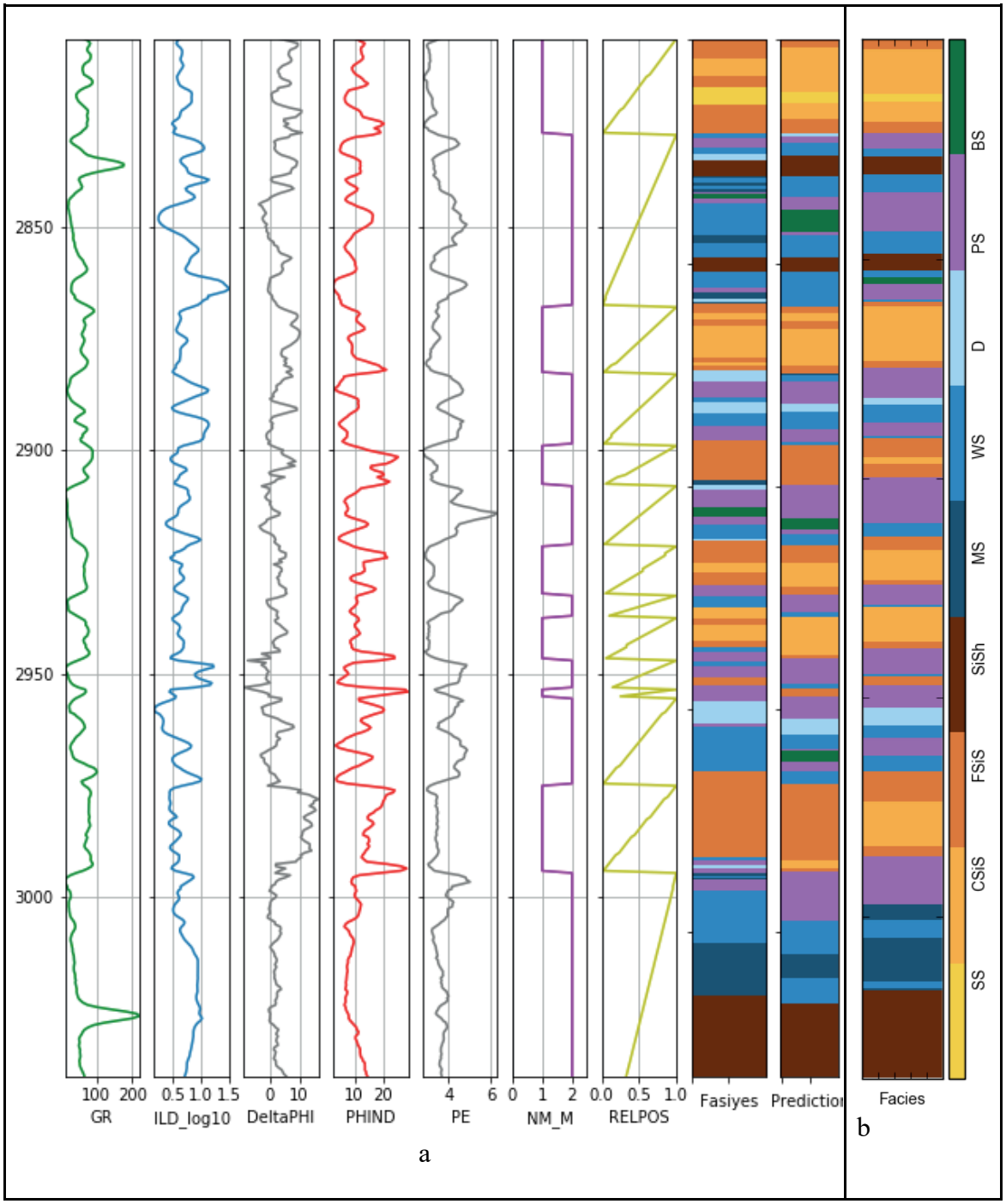


Şekil 7. a) Test verisi için elde edilen karışıklık matrisi b) Eğitim verisinin fasiyeslere göre dağılımı

SONUÇLAR

Bu çalışmada fasiyes sınıflandırma problemine makine öğrenmesi ile bir yaklaşım önerilmiştir. Önerilen algoritma, değişim arttırıcı türev sınıflamasına dayanmaktadır. GBC algoritması ile elde edilen tahmin sonuçları, yorumlanmış log verileri ile benzer olduğu görülmektedir.

Sınıflandırıcı model yaklaşımı başarı gösterip, algoritmanın kendisine öğretilen yorumları kısmi bir şekilde öğrendiğini ve uyguladığını göstermektedir. On kuyudan oluşan veri topluluğu, geliştirilen stratejisinin olumlu bir etkisi olarak, önerilen yaklaşımı doğrulamaktadır. Ayrıca fasiyes bilgisi içermeyen kuyu verisinden elde edilen sonuçlar, makine öğrenmesi algoritmasının yeni veriler için genelleme yapma yeteneğini göstermektedir.



Şekil 8. a) Eğitim modeli ile fasiyes bilgisi tespit edilen kuyunun sınıflandırma sonucu.
b) Karşılaştırılan bir diğer algoritma sonucu

GBC algoritması karmaşık jeolojik yapıların litolojik olarak tanımlanması için uygundur. Önerilen yaklaşım, kuyuların analizinin bir ön aşaması olarak düşünülebilir. Genel olarak özetlemek gerekirse, makine öğrenmesine dayalı olarak önerilen modelin değerlendirme sonuçları, gelecekteki araştırmalar ve fasiyes tanımlamaları için yararlı olacaktır.

Kaynaklar

Ahmadi, A., M., Zendehboudi, S., Lohi, A., Elkamel, A., & Chatzis, I. (2013). Reservoir permeability prediction by neural networks combined with hybrid genetic algorithm and particle swarm optimization. Geophysical Prospecting, 61(3), 582-598.

Alpaydin, E. (2009). Introduction to machine learning. MIT press. https://www.cmpe.boun.edu.tr/~ethem/i2ml/i2ml-figs.pdf (ET Ocak 2024)

Anderson, B. I. (2001). Modeling and Inversion Methods for the Interpretation of Resistivity Logging Tool Response, Delft University Press.

Bailly J. S., Amaud M. and Puech C., (2007). Boosting: a classification method for remote sensing, Int. J. Remote Sensing, 28(7), 1687-1710

Bestagini, P., Lipari, V., & Tubaro, S. (2017). A machine learning approach to facies classification using well logs. In SEG Technical Program Expanded Abstracts 2017 (pp. 2137-2142). Society of Exploration Geophysicists.

Bishop, C. M., (2006). Pattern recognition and machine learning (information science and statistics): Springer-Verlag New York, Inc.

Bohling, G.C., & Dubois, M.K. (2003). An Integrated Application of Neural Network and Markov Chain Techniques to Prediction of Lithofacies from Well Logs (Kansas Geological Survey Open File Report 2003-50).

Breiman, L., J. Friedman, R. Olshen, and C. Stone, (1984). Classification and Regression Trees: Wadsworth and Brooks.

Dubois, M. K., Bohling, G. C., & Chakrabarti, S. (2007). Comparison of four approaches to a rock facies classification problem. *Computers & Geosciences*, *33*(5), 599-617.

Darling, T. (2005). Well logging and formation evaluation. Elsevier.

Drucker, H., & Cortes, C. (1996). Boosting decision trees. In 'Advances in neural information processing systems', pp. 479-485.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. International Conference on Machine Learning, Vol. 96, pp. 148-156.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Gercek, H., (2007). Poisson's ratio values for rocks: International Journal of Rock Mechanics and Mining Sciences, 44, 1-13.

Glover, P. W. (2000). Petrophysics. *University of Aberdeen, UK*.

Hall, B., (2016). Facies classification using machine learning. Lead. Edge 35, 906–909.

Hall, M., and B. Hall, (2017). Distributed collaborative prediction: Results of the machine learning contest: The Leading Edge, 36, 267–269.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition,* Vol. 1, pp. 278-282.

Ho, T. K., (1998). The random subspace method for constructing decision forests: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 20, 832–844.

Kalayci, S. (2018). Makine öğrenmesi yöntemleri ile kredi risk analizi. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi.

Mayr, A., Binder, H., Gefeller, O., Schid, M., (2014). The evolution of Boosting algorithms from Machine Learning to statistical modelling. Methods of Information Medicine, 53(6):419-27.

Mitchell, T.M. (1997). Machine learning. McGraw-Hill

Nizam, H., Akın, S., (2014). Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması. XIX. Türkiye'de İnternet Konferansı Bildirileri. Yaşar Üniversitesi, İzmir. 129-136.

Pekiner, Y. (2002). Kuyu Logları Tekniğiyle Yeraltının Keşfi, Seçkin Yayıncılık.

Rokach, L., and O. Maimon, (2005). Top-down induction of decision trees classifiers - a survey: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (TSMCC), 35, 476–487.

Schapire, R. E. (1999). A brief introduction to boosting. In IJCAI (Vol. 99, pp. 1401-1406).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Ulugergerli, E. U. (2011). Two dimensional combined

inversion of short- and long-normal dc resistivity well log data. Journal of Applied Geophysics, 73 (2011) 130–138.

URL1: https://www.spec2000.net/ (E.T., Aralık 2019)

URL2: Resources for EECS 833, Neural Networks and Fuzzy Systems http://www.people.ku.edu/~gbohling/EECS833/ (E.T., Aralık 2019)