NUMUNE TENÖR DEĞERİNDEN LİTOLOJİK TAHMİN İÇİN RASTGELE ORMAN ENTEGRASYONU

INTEGRATION OF RANDOM FOREST FOR LITHOLOGICAL PREDICTION FROM SAMPLE ASSAY VALUE

ONUR MERİÇ TEKNECİ

DOÇ. DR. FIRAT ATALAY

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Egitim-Öğretim ve Sınav Yönetmeliğinin

Maden Mühendisliği Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2025

ÖZET

NUMUNE TENÖR DEĞERİNDEN LİTOLOJİK TAHMİN İÇİN RASTGELE ORMAN ENTEGRASYONU

Onur Meriç TEKNECİ

Yüksek Lisans, Maden Mühendisliği Bölümü

Tez Danışmanı: Doç. Dr. Fırat ATALAY

Haziran 2025, 86 sayfa

Bu çalışma, litolojik kimlik tahmini için makine öğrenmesi temelli yaklaşımla tenör (assay) verisinin girdi olarak kullanılma potansiyelini araştırmaktadır. Bu amaçla, Güney Avustralya'daki Nuckulla bölgesinden alınan 21 sondaja ait veri seti ön işleme ve standardizasyon adımlarından geçirilmiş, eğitim ve test setlerine ayrılmış; Karar Ağacı ve Rastgele Orman algoritmalarıyla sınıflandırma karşılaştırması yapılmış; ardından Rastgele Orman modeli aşırı örnekleme (oversampling) ve hiperparametre optimizasyonu ile iyileştirilmiştir. Model performansları doğruluk (accuracy), hassasiyet (precision), Cohen's Kappa, F1-skor, duyarlılık (recall), kısmi bağımlılık (partial dependence), t-SNE, ROC eğrisi ve PR eğrisi analizleriyle değerlendirilmiştir.

Başlangıçta oluşturulan Karar Ağacı ve Rastgele Orman modelleri benzer sonuçlar vermiş; optimize edilen Rastgele Orman modeli orta düzeyde doğruluk sergilemekle

birlikte zayıf genelleme yeteneği göstermiştir. Elde edilen en iyi dengelenmiş doğruluk skoru 0,60, geçerleme doğruluğu ise 0,32 olmuştur. Sonuçlar, sınıf dengesizliğinin özellikle nadir sınıflarda düşük doğruluk ve F1-skorlarına yol açtığını; aşırı örneklemenin ise modelin aşırı uyum eğilimini tam olarak önleyemediğini ortaya koymuştur. Kısmi bağımlılık analizi, tenör değeri 20'nin altındaki örneklerin daha güvenilir sınıflandığını, yüksek tenörlü örneklerin ise daha fazla özelliğe ihtiyaç duyduğunu göstermiştir. Ek olarak, her sınıfın olasılık tahminleri kullanılarak, sonuçların yeterli güvenilirlikle değerlendirilemeyeceğini göstermiştir. Bu nedenle, modelin pratik kullanım için yeterli güvenilirlik sunmadığı görülmüştür.

Rastgele Orman, Karar Ağacı'na karşı belirgin bir üstünlük ortaya koymasa ve bu görev için güvenilir bir model olarak önerilemese de çalışma, tenör verisinin düşük tenör aralıklarında sınıflandırmaya katkı potansiyelini göstermiştir. Gelecek araştırmalar için alternatif aşırı örnekleme ve hiperparametre senaryolarının denenmesi, ek litolojik özelliklerin girdi olarak kullanılması ve dengeli veri setleri üzerinde benzer metodolojinin uygulanması önerilmektedir. Bu yaklaşımlar, gelecekte daha doğru ve maliyet-etkin litoloji tahmini çalışmalarına zemin hazırlayacaktır.

Anahtar Kelimeler: Tenör Verisi, Makine Öğrenmesi, Karar Ağacı, Litoloji Sınıflandırması, Rastgele Orman

ABSTRACT

INTEGRATION OF RANDOM FOREST FOR LITHOLOGICAL PREDICTION FROM SAMPLE ASSAY VALUE

Onur Meriç TEKNECİ

Master Of Science, Department of Mining Engineering

Supervisor: Assoc. Prof. FIRAT ATALAY

Jun 2025, 86 pages

This study investigates potential of using assay data as an input for lithological identity prediction with machine learning based approach. For that purpose, a dataset of 21 drillholes taken from Nuckulla area in South Australia were preprocessed, normalized, splitted as train and test sets, classified by the usage of Decision Tree and Random Forest algorithm and the Random Forest model was optimized by oversampling and hyperparameter tuning. Model performances were evaluated by using accuracy,precision, Cohen's Kappa, F1-score, recall, partial dependence, t-SNE, ROC curve and PR curve analyses.

Initial Decision Tree and Random Forest models performed similarly and the optimized Random Forest model showed moderate accuracy but poor generalization. Best balanced-accuracy yielded was 0.60 and validation accuracy was 0.32. Results showed that the data imbalance lead to low accuracy and F1-scores especially on rare classes and oversampling did not success sufficiently considering that the model approached to overfitting. Partial dependence reliability revealed that samples with grade values lower than 20 classified more reliably while high grades require more features to classify more reliably. In addition, Class probability estimates presented that the results can not be considered with enough reliability. As a result, model did no show sufficient reliability to be suggested for practical usage.

Although Random Forest did not state a clear advantage over Decision Tree and can not be used a reliable model for this task, study demostrated the promise of assay data contribution for low grade intervals and showed possible future research areas. The specified future research can be made on trying alternative over sampling and hyper parameter scenarios, usage of additional lithological characteristics as input parameter and observation of the same method on a balanced dataset. These approaches provide a foundation for more accurate, cost-effective lithology prediction in future investigations.

Keywords: Assay Data, Machine Learning, Decision Tree, Lithology Prediction, Random Forest

TEŞEKKÜR

Her şeyden önce, bu süreç boyunca bilgi ve tecrübelerinden yararlandığım danışmanım Doç. Dr. Fırat Atalay'a en derin teşekkürlerimi iletmek isterim. Çalışmama dahil olduğu andan itibaren, saha verilerinin temin edilmesinden modelleme aşamalarına, bulguların yorumlanmasından tez yazımına kadar her konuda bana rehberlik ederek sürecin son derece verimli ve hızlı ilerlemesini sağladı. Ayrıca, süreci mümkün olduğunca benim yürütmeme izin veren hoşgörüsü ve anlayışına minnettarım. Özellikle bu araştırma boyunca yaşadığım ülke değişim süreci boyunca bana gösterdiği sabır ve aynı zamanda etkin yönlendirme sayesinde araştırmamı titizlikle kurgulayabildim.

Bu çalışmanın temellerini atan, yüksek lisans hayatımın ilk yıllarında beni yönlendiren ve bu alana olan ilgimi pekiştirmemde rol oynayan eski danışmanım Dr. Öğr. Üyesi Güneş Ertunç'a da şükranlarımı sunmak isterim. Bununla birlikte bana bu sürece girme cesareti veren, makine öğrenimi ve ilgili kavramları anlamamda, bu disiplinin ana iskeletini kafamda oluşturmamda büyük katkısı olan ve bu araştırmada girdiğim birçok çıkmazda bana yardımlarını esirgemeyen sevgili arkadaşım Serkan Demirci'ye de teşekkür etmek istiyorum. Bu sürece yakından şahit olan ve desteğini eksik etmeyen arkadaşım Volkan Satar'a da teşekkür ederim. Ayrıca yüksek lisans sürecimin en kritik aşamasında bilime hevesimi ve inancımı güçlendiren, bu araştırmayı sürdürmem ve bitirmem için bana mutluluk veren İngiltere yıllarıma ve Oasis grubuna teşekkür ederim.

Her zaman yanımda olan arkadaşlarıma, sevgili anneme ve en büyük destekçim babama teşekkür ederim. Son olarak, ilham perim Zeynep İmir Tekneci'ye her zaman bana mutluluk, huzur ve destek verdiği için teşekkür ederim.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

SİMGELER VE KISALTMALAR

Simgeler

n	Toplam Örnek Sayısı
TP	True Pozitif (Gerçek Pozitif)
FP	False Pozitif (Yanlış Pozitif)
FN	False Negative
n_{10}	McNemar tablosunda Rastgele Orman doğru, Karar Ağacı yanlış
n_{01}	McNemar tablosunda Rastgele Orman yanlış, Karar Ağacı doğru
n_{corr}	doğru sınıflandırılmış örnekler
p(j ∣ t)	t 'deki örneklerin j' ye ait olma ihtimali

Kısaltmalar

DT	Karar Ağacı
RF	Random Forest (Rastgele Orman)
SMOTE	Synthetic Minority Over-sampling Technique
t-SNE	t-distributed Stochastic Neighbor Embedding
ROC	Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)
PR	Precision-Recall (Hassasiyet-Duyarlılık)
AUC	Area Under the Curve (Eğri altındaki alan)
OOB	Out-Of-Bag (Torba dışı)
ML	Machine Learning (Makine Öğrenimi)

1. GİRİŞ

Litoloji tahmini, kayaçların litolojik adlarını ve özelliklerini istatistiksel yöntemlerle belirleme sürecidir. Madencilik, petrografi ve çevresel çalışmalarda litoloji verisi daima hayati öneme sahiptir. Keşif sondajları ve kuyu loglaması, litolojiyi tespit etmenin en yaygın uygulamalarıdır. Geleneksel olarak jeoistatistik; son yıllarda ise makine öğrenmesi teknikleri, litolojik özelliklerin öngörüsünde başvurulan başlıca yaklaşımları oluşturur. Bu yöntemlerde kaydedilen ilerlemeler, zaman içinde daha yüksek doğruluk ve performans sağlarken, sondaj sayısını azaltmak maliyet ve zaman tasarrufu açısından avantajlıdır.

Litoloji; bir kayacın makroskobik özellikleri olarak tanımlanır [1]. Litolojik sınıflandırma, tüm sondaj ve ölçüm cihazlarının doğru çalışması için temel oluşturur ve doygunluk, gözeneklilik, geçirgenlik gibi parametrelerin hesaplanmasında kritik rol oynar. Kayaç ve litoloji tespitinin en kesin yolu, doğrudan kuyu örneği almaktır. Bu örneklerin analizinde yaygın olarak kullanılan "kuyu loglaması", bir kuyudaki kaya ve sıvı özelliklerinin derinlik boyunca kaydedilmesi anlamına gelir [2]. SP log ve gama ışını logu yöntemleri kaya türünü belirlemede tercih edilen diğer yöntemler arasındadır [3], ancak örneklerin alım sırasında zarar görmesi veya bileşenlerini kaybetmesi riski, bu yöntemin zorluklarından biridir.

Litoloji tahmininde sıkça başvurulan jeoistatistik, istatistiğin bir alt dalı olarak petrol jeolojisi, jeokimya, hidrojeoloji ve jeometalurji gibi pek çok farklı alanda kullanılır. Veri odaklı bir yöntem olan jeoistatistik, verinin temel yapısını yakalamaya odaklanır ve variogram model seçiminde uzman görüşünü ikinci planda tutar. Bu yaklaşım, birden çok geçerli yorumu aynı anda ayırt etme noktasında sınırlılıklar taşısa da mekânsal verilerin belirsizliklerini rastgele fonksiyonlarla modelleyerek sağlam tahminler sunar.

Yeni keşiflerin yüksek örtü seviyelerine sahip bölgelerde gerçekleştirilmesiyle ilgili artan maliyetler ve zorluklar nedeniyle, daha verimli maden arama yöntemlerine ihtiyaç duyulmaktadır. Makine öğrenmesi algoritmaları, litoloji tahmininde giderek daha iyi sonuçlar vermeye başlamış ve her geçen gün daha fazla kullanılmaya başlanmıştır.

Son yıllarda, bu teknikler litolojik özelliklerin tahmin edilmesi ve haritalanması amacıyla kullanılarak sürecin hızlandırılması ve maliyetlerin düşürülmesi hedeflenmektedir. Yapılan araştırmalar, makine öğrenmesi algoritmalarının litoloji tahmini ve sınıflandırmasında faydalı olabileceğini göstermektedir. Makine öğrenmesinin amacı, bilgisayarların öğrenmesini sağlayacak teknikler geliştirmektir. Bu algoritmaların temel özelliği, ampirik verilerden öğrenebilme yetenekleridir; bu da onları temsil edilen süreçlerin belirsiz, gözlemlenmesi zor veya yetersiz tanımlandığı durumlar için uygun hale getirmektedir [4]. Mekânsal verilerin tahmini ve haritalanmasında en çok kullanılan makine öğrenmesi algoritmaları Yapay Sinir Ağları (ANN), Destek Vektör Makineleri (SVM), Karar Ağaçları (DT) ve Rastgele Orman (RF) yöntemleridir.

RF, özellikle son yıllarda litoloji tahmini ve haritalanması konusunda araştırılmış makine öğrenmesi yöntemlerinden biridir. RF algoritması, her biri aynı dağılıma sahip rastgele örneklenmiş bir rastgele vektörün değerlerine bağlı olan farklı ağaç tahminleyicilerinin birleştirilmesiyle oluşturulur. Ormandaki her bir ağacın gücü ve ağaçlar arasındaki korelasyon, orman içerisindeki sınıflandırıcıların genelleme hatasını belirlemektedir [5]. RF algoritmasıyla jeokimyasal elementlerin tahmini, kayaç gözenekliliği, formasyon ve jeolojik haritalama gibi konularda çok sayıda girdi parametresi incelenmektedir; bunlara yatak tipleri, jeofizik ve uzaktan algılama verileri gibi unsurlar da dahildir.

Tenör (assay) verisi, cevherlerde ve metalürjik maddelerdeki değerli metal içeriğini ifade eder. Cevherin ekonomik değeri açısından kritik öneme sahip olan tenör verileri, jeofiziksel loglama, uzaktan algılama ve jeokimyasal çalışmalarında yaygın kullanılıyor olmasına rağmen litoloji tahmininde kullanımına şimdiye kadar sıkça karşılaşılmamıştır. Bu eksikliği gidermek amacıyla, bu çalışmada assay verilerinin makine öğrenmesi ile kullanılması denenmiştir.

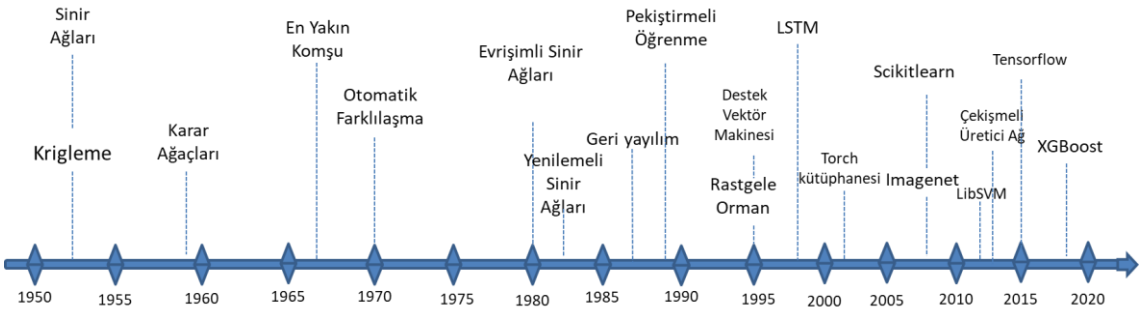
2. GENEL BİLGİLER

2.1. Yerbilimlerinde Makine Öğrenmesine Genel Bakış

Makine öğreniminin son yıllarda artan önemi, biyoloji, kimya, tıp ve eczacılık gibi birçok bilim alanına katkı sağlamıştır. Turing'in 1950 yılında yapay zekâ (AI) kavramını ortaya atmasından bu yana, makine öğrenimi genellikle yapay zekânın bir alt alanı olarak kabul edilmektedir. "Makine öğrenimi" terimi ilk kez 1959 yılında Samuel tarafından kullanılmıştır. Mitchell ve çalışma arkadaşları makine öğrenimini şu şekilde tanımlamıştır:

"Bir bilgisayar programının, T sınıfındaki bazı görevler ve P performans ölçütü dikkate alındığında, E deneyimiyle birlikte bu görevlerdeki başarımının P'ye göre gelişmesi durumunda, bu programın deneyim E'den öğrendiği söylenir." [6]

Gauss süreçleri (Gaussian processes), makine öğreniminde uygulanan matematiksel ve istatistiksel yöntemlerden biridir. Bu süreçlerin kökeni zaman serisi uygulamalarına ve jeoistatistik bilimine dayanmaktadır; bu da onları yerbilimleri bağlamında makine öğrenimi açısından önemli kılmaktadır. "Krigleme" yöntemi, ilk kez altın madenlerinin değerini tahmin etmek için iki boyutlu Gauss süreçleri kullanıldığında tanıtılmış olup, o zamandan bu yana jeoistatistikte yaygın şekilde kullanılmaktadır.



Şekil 2.1. Makine Öğrenimi Zaman Çizelgesi (Dramsch,2020) [7]

Şekil 2.1.'de gösterildiği gibi, 1950 ile 2020 yılları arasında pek çok şey değişmiştir. Günümüzde, hesaplama kaynakları hem donanım hem de yazılım açısından oldukça yaygın hale gelmiş, bulut bilişim sağlayıcıları yüksek performanslı hesaplamayı düşük

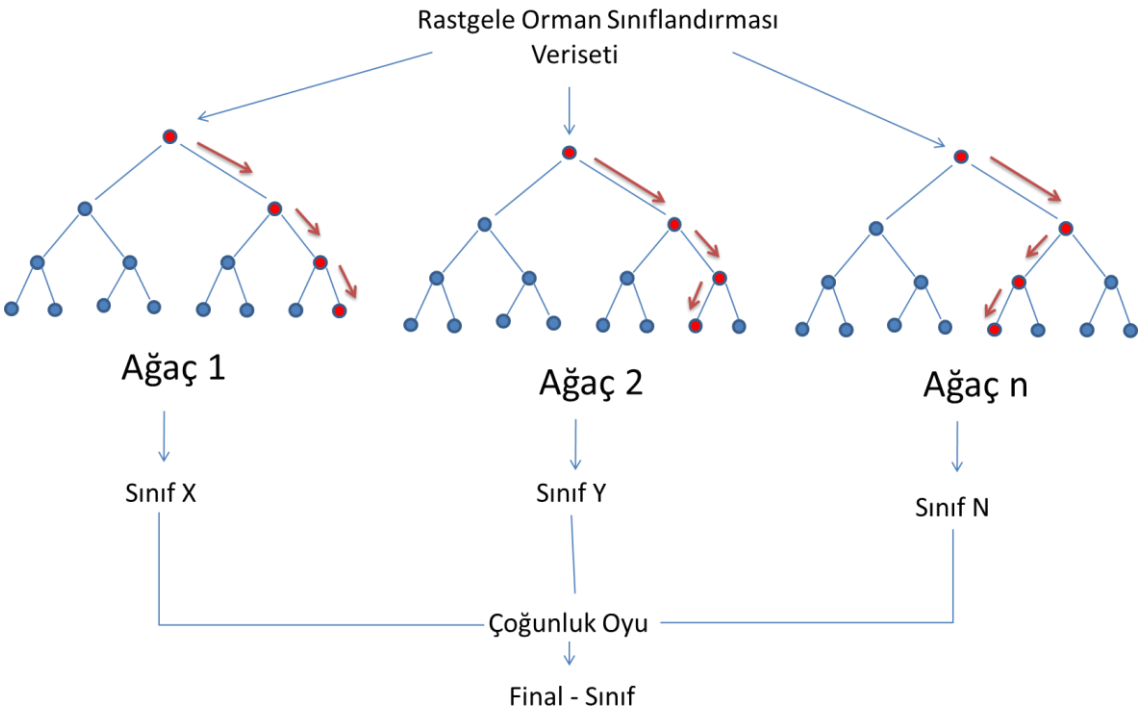
maliyetle herkesin erişimine sunmuştur. Ücretsiz ve açık kaynak yazılım hareketi, yüksek kaliteli makine öğrenimi yazılımlarının yaygın biçimde erişilebilir olmasını sağlamıştır. Ancak, yer bilimlerinde makine öğrenimi araştırmaları yeni bir alan değildir. Makine öğrenimi araştırmaları, ilk ortaya çıktığı günden itibaren yer bilimlerinde ilgiyle takip edilmiş ve güven duyulmuş bir alandır; fakat bu durum finansman açısından olumsuz bir etki yaratmıştır. Yer bilimlerindeki gelişmeler, genellikle uzun yıllar süren bir zaman gecikmesiyle, makine öğrenimine yönelik bu yaygın heyecan ve ilgisizlik döngüsünü takip eder. Bu durumun, araştırma finansmanının mevcut olup olmaması ya da elde edilen bulguları yayımlama isteğindeki değişim gibi birçok nedeni olabilir [7].

Bu süre zarfında, jeoloji alanında birçok farklı makine öğrenimi (ML) algoritması kullanılmıştır. Yapay Sinir Ağı (Artificial Neural Network) bunlardan biridir. ANN'ler çeşitli biçimlerde bulunabilir. En bilinen ANN türlerinden biri ileri beslemeli (feed-forward) sinir ağıdır. Nöronlar, yani birim ya da düğümler olarak da adlandırılan yapılar, yapay sinir ağlarının temel işlem birimleridir. Bir sinir ağı, bilgilerin giriş katmanından çıkış katmanına, gizli katmandaki birimler aracılığıyla tek yönlü olarak aktığı, birbirine bağlı katmanlardan oluşur. Bir nöron, doğrusal regresyonu doğrusal olmayan bir fonksiyon yardımıyla gerçekleştirebilir. Birden fazla nöron katmanı, ağırlıklarla birbirine bağlanır.

Jeolojide kullanılan bir diğer makine öğrenimi algoritması Destek Vektör Makineleri'dir (Support Vector Machines). Destek vektör makineleri, regresyon ve sınıflandırma için uygun algoritmaları kullanan, denetimli öğrenmeye (supervised learning) dayalı makine öğrenimi yöntemleridir. Vladimir Vapnik ve çalışma arkadaşları tarafından geliştirilen SVM'ler, Vapnik (1982, 1995) ve Chervonenkis (1974) tarafından ortaya konulan VC teorisine dayalı olarak sağlamlıklarıyla tanınırlar. SVM'ler, eğitim verilerinden yola çıkarak yeni örnekleri iki kategoriye ayıran bir model oluşturur; bu eğitim verileri de iki kategoriye ayrılmıştır. Bu model, olasılıksal olmayan ikili (binary) doğrusal bir sınıflayıcı olarak çalışır. İki nokta arasındaki mesafeyi maksimize etmek için SVM'ler veriyi haritalama noktalarına dönüştürürler. Daha sonra, bu uzayda o kategoriye karşılık gelen tarafa düşen yeni örnekler, aynı sınıfa dahil edilir.

Karar ağaçları (Decision Trees) ile yapay sinir ağlarının birleştirilmesi, yer bilimlerinde en yaygın kullanılan makine öğrenimi yöntemlerindendir. Karar ağaçları, okunabilirlikleri (grafiksel olarak gösterilebilme yetileri), basitlikleri ve göreli olarak düşük hesaplama maliyetleri nedeniyle giderek daha popüler hale gelmektedir. Bir karar ağacı, kökten başlayarak en son düğüme veya yaprağa kadar kısıtlamaları ya da sınırlayıcı faktörleri hiyerarşik bir biçimde uygulamaya devam eder. Karar ağaçlarının şeffaflığı ve yorumlama kolaylığı, onları yapay sinir ağlarından ayıran temel özelliklerdir. Karar ağacı, bir veri kümesinden türetilirken, her önemli özelliğe ilişkin bir değerlendirme ölçütü kullanılarak düğümler arası fark en üst düzeye çıkarılmaya çalışılır. Karar ağacı teknikleri, regresyon ağaçları (Regression Trees) ve sınıflandırma ağaçları (Classification Trees) olmak üzere iki ana kategoriye ayrılır.

Rastgele Orman (Random Forest), birçok karar ağacı tekniğinin birlikte kullanılmasıyla bir değişkenin değerini sınıflandırmak ya da tahmin etmek için uygulanan bir regresyon yöntemidir. RF, bir eğitim bölgesine ait çeşitli kanıta dayalı özellikler içeren bir giriş vektörü (x) kullanarak regresyon ağaçları oluşturur (Şekil 2.2.'de gösterildiği gibi). Daha sonra, elde edilen çıktı ortalama alınarak sonuca ulaşılır [8].



Şekil 2.2. Rastgele Orman Sınıflandırma Şeması [9]

RF, torbalama (bagging) ve artırma (boosting) yöntemleri, topluluk öğrenme algoritmaları (ensemble learning) sınıfında değerlendirilir. Bu yöntemlerin ortak özelliği, temel sınıflandırıcıların kümelerini oluşturabilmeleri ve bu sınıflandırıcılara çeşitli eğitim veri setleri vererek çeşitliliklerini garanti altına almalarıdır [10]. Günümüzde, maden arama haritalaması için en popüler makine öğrenimi algoritmalarından (MLA) biri, torbalama metodolojisini kullanan RF algoritmasıdır [11]. Dietterich, T.G. torbalama, artırma ve rastgeleleştirme (randomizing) yöntemlerini karşılaştırmış ve verilerde çok az ya da hiç gürültü (noise) olmadığı durumlarda artırmanın genellikle en iyi sonuçları verdiğini belirtmiştir. Torbalama ve rastgeleleştirme oldukça benzer sonuçlar üretse de, düşük gürültülü ortamlarda rastgeleleştirmenin torbalamadan biraz daha iyi performans gösterdiğine dair bazı bulgular vardır [12].

Regresyon ormanı analizinin (RF) başlıca avantajları arasında şunlar yer alır: basit parametre ayarları, önyargısız (unbiased) içsel tahmin hatası değerlendirmesi, farklı istatistiksel dağılımlara sahip karmaşık verileri işleyebilme kapasitesi, değişkenler arasındaki doğrusal olmayan ilişkilere yanıt verebilme yetisi, kategorik değişkenlerin kullanılabilmesi ve değişken öneminin değerlendirilebilmesi. Dahası, RF tarafından en önemli olarak belirlenen özellikler, mevcut jeolojik beklentilerle uyumludur [13].

2.2. Literatür Taraması

1950 ile 2000 yılları arasında, k-means, Markov zincirleri ve karar ağaçları yöntemleri, yerbilimlerinde Makine Öğrenimi algoritmalarının kullanılmasının başlangıcını oluşturmuştur. Çoğu kişi, kümeleme için kullanılan k-means yöntemine aşinadır ve bu yöntemi kullanmaktadır [7]. Literatürde önerilmiş birçok k-means uzantısı mevcuttur[14]. Preston ve Henderson, stratigrafik korelasyon amacıyla döngüsel tortul birikintileri tanımlamak için k-means yöntemini kullanmışlardır [15]. Shwarzacher ise tortul süreçleri modellemek için yarı-Markov süreçlerini kullanmıştır [16]. Karar ağacı tabanlı tekniklerin erken dönem kullanımları, arama haritalaması ve ekonomik jeoloji alanlarında olmuştur. Bu süreçte Krigleme de bu gelişmelerde yerini almış, ancak yapay zekâ teknolojisi olarak değerlendirilmemiştir [7]. Matheron, Krigleme ile spline enterpolasyonu karşılaştırmalı olarak ele almıştır [17].

1950'lerde geliştirilmiş olmasına rağmen, sinir ağları jeofizikte yaygın olarak 1980'lerin sonlarına kadar kullanılmamıştır. Bu yaygınlaşma, hata düzeltmeli makine öğrenimi için otomatik türev alma ve geri yayılım (backpropagation) algoritmalarının geliştirilmesi sayesinde olmuştur [7]. 1990 yılında Dowla ve arkadaşları, bölgesel sismik fazların mesafe düzeltilmiş spektral verilerini kullanarak, yeraltı nükleer patlamaları ile doğal depremleri ayırt etmek için yapay sinir ağlarının bir uygulamasını incelemişlerdir [18]. Daha sonra, Röth ve Tarantola da benzer şekilde bu ayırt etme işlemine yönelik bir yapay sinir ağı uygulaması incelemiştir [19]. Feng ve Seto, kaya mikro-çatlama süreçleri sırasında zamanla ortaya çıkan akustik emisyon olaylarının genelleme başarımını artırmak için sinir ağı tekniklerini kullanarak gelişmiş bir öğrenme yöntemi önermiştir [20]. Benaouda ve arkadaşları ise, 1999 yılında, delik içi verilerden yararlanarak, kısmi karot (core) alımı olan alanlarda sınıflandırıcıları eğitmek yoluyla, alınamayan karot bölümleri için küresel bir sınıflandırma şeması sağlayan, litolojik yorumlamayı önemli ölçüde geliştiren sinir ağı tabanlı kullanışlı bir yöntem sunmuşlardır [21].

Arka planda, 1990'lar SVM, RF ve belirli bir tür yinelemeli sinir ağı olan uzun kısa süreli bellekler (LSTM) gibi yöntemlerin geliştirilmesine sahne olmuştur [6]. Hermes ve arkadaşları, yakın zamanda tanıtılan destek vektör makinelerinin uzaktan algılama uygulamaları için potansiyelini göstermiştir [22]. 1999'da Benaouda ve arkadaşları, kısmi karot alımı olan deliklerde, küresel bir sınıflandırma şeması oluşturmak için delik içi verilerle sınıflandırıcıları eğiterek, alınamayan karot bölümleri için litolojik yorumlamayı geliştiren bir sinir ağı tabanlı yöntem sunmuşlardır.

1997'de Broadley ve Freidl, çeşitli karar ağacı sınıflandırma teknikleri sunmuş ve bunları üç farklı uzaktan algılama veri setiyle değerlendirmiştir. Sonuçlar, kategorileştirme yapısının şeffaf olması nedeniyle, karar ağaçlarının yer bilimsel uygulamalar için güçlü bir sezgisel çekiciliğe sahip olduğunu göstermiştir. Bir analist, bir karar ağacına bakarak, sınıfları birbirinden ayıran temel öğeleri belirleyebilir; bu da onların yorumlanabilirliğinden kaynaklanır. Karar ağaçları, bu bağlamda keşifsel veri analizi için faydalı araçlardır ve temel sınıflar, özellik uzayı ve girdi verileri arasındaki hiyerarşik ilişkileri incelemek için kullanılabilir. Sonuç olarak, karar ağaçları genellikle oldukça güvenilir ve etkilidir. Özellikle, karar ağacı tahmin teknikleri, genellikle hızlı olmaları ve girdi verilerindeki gürültüye karşı dayanıklı olmaları sayesinde, uzaktan algılama arazi

örtüsü haritalama uygulamalarında bulunan büyük miktarda veriyi sınıflandırmak için önemli bir değere sahiptir [23].

2000'li yıllarla birlikte, makine öğrenimi ve derin öğrenme araştırmaları ve uygulamaları için çeşitli araçlar ortaya çıkmıştır. Özellikle MATLAB ve Python kütüphaneleri bu alanlarda kendini göstermeye başlamıştır [7]. RF'nin uygulanması gecikmiştir çünkü bu terim Leo Breiman tarafından 2001 yılında ortaya atılmıştır [5].

2006 yılında Zhai ve arkadaşları, ileri beslemeli geri yayılımlı sinir ağlarına dayalı sınıflandırıcılar oluşturmak için alternatif sinir ağı mimarilerini kullanmış ve bu ağları toprak dokularını sınıflandırmak üzere eğitmişlerdir. Sonuçlar, verilerin aynı bölgeden toplandığı durumlarda sınıflandırıcıların iyi çalıştığını göstermiştir [24]. Wang ve arkadaşları ise, karmaşık ve derin formasyonlarda sondaj hızını artırmak amacıyla geliştirilmiş bir geri yayılım modeli içeren bir sinir ağı geliştirmiştir [25].

Rodriguez-Galiano ve arkadaşları, bir alanın arazi örtüsünü sınıflandırmada RF performansını incelemişlerdir. Değerlendirme; gürültü, veri seti boyutuna duyarlılık ve haritalama doğruluğu gibi çeşitli faktörlere dayanmıştır. Sonuçlara göre, RF algoritması genel olarak %92 doğrulukta arazi örtüsü sınıflandırmaları üretmiş ve 0.92 Kappa indeksi elde etmiştir. Kappa değerlerinde önemli farklılıklar yalnızca veri azaltımı %50'yi ve gürültü seviyesi %20'yi aştığında gözlemlenmiştir. Bu da RF'nin eğitim verisi azalmalarına ve gürültüye karşı dayanıklı olduğunu göstermektedir. Dahası, RF'nin en önemli olarak belirlediği özellikler, arazi örtüsü sınıflandırması açısından yapılan öngörülerle örtüşmektedir. RF tek bir karar ağacına kıyasla bazı özelliklerdeki aşırı uyumu (overfitting) azaltmıştır ve karar ağacı (CT) tarafından yapılan yanlış sınıflandırmaları düzeltmiştir [26].

2014 yılında Cracknell ve Reading, yaygın olarak erişilebilen ve coğrafî olarak sınırlı uydu kaynaklı jeofizik verileri kullanarak denetimli bir litoloji sınıflandırma çalışmasında beş makine öğrenimi algoritmasını titizlikle karşılaştırmışlardır: Naive Bayes, k-En Yakın Komşu (k-NN), Rastgele Orman (RF), Destek Vektör Makineleri (SVM) ve Yapay Sinir Ağları (ANN). Ayrıca, bu algoritmaları eğitim verilerindeki mekânsal kümelenme düzeyindeki değişikliklere ne kadar duyarlı oldukları ve açık coğrafî bilginin eklenmesi

durumunda nasıl tepki verdikleri açısından değerlendirmişlerdir. Araştırma sahası, Avustralya'nın uzak New South Wales bölgesinde yer almakta ve 13 farklı litolojik sınıf içermektedir. Litoloji sınıflandırma görevine uygunluğu artırmak için, toplanan veri yansıtılan jeofiziksel özelliklere göre çeşitli şekillerde işlenmiş; makine öğrenimi eğitimi ve tahmini için toplam 17 bağımsız girdi kullanılmıştır.

Sonuç olarak önce, eğitim verilerinin coğrafî dağılımındaki değişikliklerin girdi değişkenlerinin göreceli ağırlıklarını ve algoritma parametre seçimini nasıl etkilediği incelenmiştir. Ardından, mekânsal dağılımdaki farklılıklar ve açık mekânsal bilginin eklenmesi göz önünde bulundurularak MLA'ların test istatistikleri karşılaştırılmıştır. Kararlılık, kullanılabilirlik, işlem hızı ve tahmin doğruluğu bakımlarından RF, bu çalışmada test edilen diğer tüm algoritmalardan üstün performans göstermiştir. Parametre değerlerindeki değişikliklere karşı görece duyarsızlığı, RF'nin aşırı öğrenme (overfitting) riskini azalttığını işaret etmektedir. Diğer ML algoritmalarıyla kıyaslandığında, SVM modelleri sınıflandırma modeli eğitimi sırasında yakınsama hatalarına daha yatkın bulunmuştur. RF'nin uyarlanabilir k-NN yaklaşımına benzer öğrenme tekniği, mekânsal olarak sınırlı denetimli sınıflandırma problemlerinde RF'nin üstünlüğünü açıklamaktadır. Ayrıca, çok sayıda ilgili değişkenin bulunduğu durumlarda RF genellikle gürültülü girdilere karşı bağışıktır. Sonuçta, yüksek boyutlu, çok kaynaklı, uzaktan algılanmış ve hazır uydu kaynaklı jeofizik değişkenlerle kombine edildiğinde RF'lerin çoklu sınıf çıkarımı için mükemmel bir birinci tercih algoritması olduğu sonucuna varılmıştır. Bu bağlamda RF modelleri, hesaplama maliyeti düşük, eğitimi basit, geniş bir model parametre aralığında kararlı ve mekânsal dağılımlı eğitim verileriyle çalışırken diğer ML algoritmalarına kıyasla önemli ölçüde daha yüksek doğruluk sunmaktadır. Gürültü ve aşırı öğrenmeye karşı duyarsızlık özellikleri de RF'yi uzaktan algılamada litolojik sınıflandırma uygulamaları için ideal kılmaktadır [27].

2015 yılında Rodriguez-Galiano ve arkadaşları, mineral bulunabilirliği modellemesinde ANN, SVM, RF ve Regresyon Ağaçları'nın (RT) performansını; potansiyel alan tanımındaki doğruluk derecesi, hiperparametre tahminine duyarlılık, eğitim veri hacmine duyarlılık ve model parametrelerinin yorumlanabilirliği açısından karşılaştırmışlardır. Çalışma, güneydoğu İspanya'da, altın yatakları bakımından zengin bir bölgede gerçekleştirilmiş; 46 altın yatağı konumu (işlenmiş yataklar ve tanınmış mineralize

yapılar) ile 59 element içeren jeokimyasal bir çalışma (372 nokta), 330 yer istasyonlu gravitasyon ve manyetik ölçümler ve kırılma/litoloji verilerinden oluşan zengin bir veri tabanı kullanılmıştır. Sonuçlar, SVM modellerinin ortalama ve standart sapma bakımından en yüksek MSE hatalarına ulaştığını ve diğer yöntemlerden daha düşük doğruluk sunduğunu göstermiştir. RF ise en düşük ortalama ve standart sapma MSE değerleriyle son derece kararlı ve dayanıklı bulunmuştur. RF dışındaki algoritmalar, eğitim parametrelerindeki değişikliklere yüksek duyarlılık göstermiş; özellikle ANN'de en iyi hata değerleri, çok spesifik parametre kombinasyonlarında ortaya çıkmıştır. ANN ve RT ise, sırasıyla Kappa = 0,77 (genel doğruluk 0,89) ve Kappa = 0,66 (genel doğruluk 0,83) değerleriyle daha az doğru mineral bulunabilirliği haritaları üretmiş, aşırı öğrenme nedeniyle genelleme kabiliyetlerini yitirmişlerdir. Genel sonuçlara göre RF ve SVM en yüksek sınıflandırma doğruluğuna ulaşmış; Kappa değerleri sırasıyla 0,92 ve 0,87'dir. ANN (Kappa = 0,77) ancak çok belirli parametre ayarlarında makul haritalama doğruluğu göstermiştir. ANN ve SVM, yataklanma dışı alanları aşırı tahmin ederken RF, her iki alanı da dengeli doğrulukla ayırt etmiştir. Minerallerin potansiyelini modellemede RT ve RF yaklaşımları, her kanıt parçasının göreceli ağırlığını belirleyebilmektedir [28].

Aynı yıl içinde Harris ve Grunsky, iki farklı eğitim yöntemiyle RF tabanlı litoloji haritalaması yapmışlardır. Birinci yöntemde göl çökellerinden alınan jeokimyasal örneklerin konumları kullanılarak, her örnek noktasında eski jeoloji haritasından türetilen kaya türü işlenmiştir. İkinci yöntem ise arazi gözlemleriyle elde edilen litoloji verilerine dayanmaktadır. Çalışma, Kanada'nın Churchill İlçesi'nin batı bölümünde yer alan Hearne jeolojik bölgesi güneyinde yürütülmüştür. Tahmine dayalı litoloji haritaları, hava manyetik (düşük çözünürlüklü), hava gama ışını spektrometre ve göl çökelleri jeokimyasal verilerini içeren üç veri türüyle oluşturulmuştur. Özel jeokimyasal ve radyoelement özellikleriyle karakterize Nueltin granitleri, her iki eğitim yönteminde de en güvenilir şekilde sınıflandırılan litolojiyi oluşturmuştur. Göl çökelleri örneklerine dayalı eğitimin üretici doğruluğu (PA) en yüksek litolojiler Nueltin granitleri (%80), diyorit, gabro ve Arkeen gnayslar (%97) olmuştur. Sonuç olarak, RF tabanlı litoloji sınıflandırma yaklaşımının saha haritalama çalışmalarını destekleyebileceği ve Kanada'nın yüksek enlemlerindeki iyi belgelenmemiş bölgeler için birinci derece jeolojik bilgi sağlayabileceği belirtilmiştir. Haritalar öngörücü niteliktedir; gerçek jeolojiyi tam yansıtmayabilir. RF sınıflandırma doğruluğu özellikle daha kararlı tahmin teknikleriyle

kıyaslandığında çok yüksek olmasa da, hiçbir jeolojik veri olmamasından iyidir. RF'nin rastgele doğası gereği farklı eğitim çalıştırmalarında küçük sapmalar olması beklenir, ancak genel sonuç RF'lerin litoloji tahmini için yararlı bir sınıflandırma tekniği olduğunu göstermiştir [29].

Böyle çalışmalarla RF algoritmasının litoloji tahmini ve mineral bulunabilirliği haritalamasındaki kullanımı ivme kazanmıştır. Carranza ve Laborte, Baguio altın bölgesinde RF modellemesini kanıt ağırlıkları(weights-of-evidence), evidential belief ve lojistik regresyon yöntemleriyle karşılaştırmış; farklı yataklanmış ve yataklanmamış eğitim kümeleriyle tutarlı ve tekrarlanabilir sonuçlar elde ederek RF'nin en yüksek başarı oranını sunduğunu göstermiştir. Epithermal altın bulunabilirliği için veri odaklı tahminsel haritalamada RF, başarı oranı bakımından kanıt ağırlıkları modelini geride bırakmış, evidential belief ve lojistik regresyona ise benzer performans sergilemiştir [30].

Roslin ve Esterle, jeofizik tel log verilerinden kömür litotiplerini nesnel ve tekrarlanabilir şekilde ayırt eden yeni bir yöntem sunmuş; gama ışını, yoğunluk, laterolog direnç, mikro direnç ve PEF loglarını birleştirerek parlak/bantlı ile mat kömürü benzer yoğunluklarda yüksek çözünürlükle sınıflandırmıştır [31].

Bhattacharya ve arkadaşları, geleneksel kuyu log verilerine matematiksel teknikler (SVM, ANN, SOM, MRGC) uygulayarak Devoniyen Bakken ve Mahantango-Marcellus Şeyl formasyonlarında sayısal litofasiyesleri modelleyerek çökelme geçmişi ve hidrokarbon potansiyelini değerlendirmiş; jeolojik kurallarla denetlenen SVM'nin en yüksek doğruluğu sunduğunu göstermiştir [32].

Ghosh ve arkadaşları, Hindistan'daki Korba Kömür Sahası'ndan alınan karot ve log verileriyle hiyerarşik kümeleme, regresyon ve sinir ağı analizleri kullanarak kömür fasiyeslerini tahmin etmiş; MLFN modeliyle kül ve nem oranı tahmininde regresyondan daha iyi sonuçlar elde etmiştir [33]. Lundberg ve Lee, karmaşık modellerde yorumlanabilirlikle tahmin doğruluğunu dengeleyen SHAP çerçevesini tanıtmış; özelliklere önem değerleri atayarak farklı yöntemleri birleştirmiş ve hem hesaplama performansını hem de insan sezgisiyle uyumu iyileştirmiştir [34].

2017’de Shabankareh ve Hezarkhani bakır potansiyel haritalaması için SVM kullanmışlardır [35]. Blouin ve arkadaşları, sondaj deliklerindeki kaya tiplerini fiziksel özelliklere dayanarak otomatik tanımlamak için makine öğrenimi tekniklerini uygulamış ve fasiyes tahmin doğruluğunu değerlendirmişlerdir [36]. Rouet-Leduc ve ekibi, laboratuvar kayma testleri veri setleriyle depremleri öngören gizli sinyalleri makine öğrenimi ile keşfetmiştir [37]. Bestagini ve arkadaşları ise tel log ölçümlerini zenginleştirilmiş özelliklerle besleyerek RF sınıflayıcı temelli bir fasiyes sınıflandırma iş akışı sunmuşlardır [38].

Bu çalışmalardan bir yıl sonra, ön inceleme (reconnaissance) aşamasında jeokimyasal örnek bulunmadığı için Kuhn ve arkadaşları, tarihi açıdan önemli Junction altın madeninin yakınında az araştırılmış bir alanın litolojisini, makine öğrenmesi algoritması olan Rastgele Ormanlar (RF) kullanarak jeofizik ve uzaktan algılama verileri ile sınıflandırdı. Bilgi entropisini (information entropy) kullanarak bu yeni RF sınıflandırmasıyla ilişkili belirsizliği değerlendirdiler ve güncellenmiş haritada yanlış sınıflandırma alma olasılığı en yüksek olan bölgeleri belirlediler. Araştırma, Batı Avustralya’daki Junction altın madenine 15 km uzaklıkta yapıldı. Sahanın tahmini altın içeriği 300 tondur ve stratigrafi orojenik ve sokulum ilişkili altın yataklarını içermektedir. Ana kaya, mafik-ultramafik volkanik ve intrüzif birimler, volkanoklastik tortular ve felsik intrüzyonlardan oluşmaktadır. Bu çalışmada, on altı jeofizik ve uzaktan algılama veri kümesi kullanıldı ve bu veriler, veri toplama hat aralığının %20–25’ine karşılık gelen bir grid hücre boyutunda enterpole edildi. Her bir değişken yer değiştirilerek, torba dışı (out-of-bag) sınıflandırma doğruluğu üzerindeki etkisi ölçüldü. RF sınıflandırıcısı, her litolojik birimden yüz örnek olmak üzere toplam sekiz yüz örnekle eğitildi. Araştırma sonuçları, RF’nin haritalanmış jeolojiyi doğru bir şekilde tahmin ederek jeolojik haritanın revize edilmiş bir versiyonunu oluşturduğunu gösterdi. Granitik ve bazaltik birimler dahil olmak üzere birkaç grup yüksek doğrulukla tahmin edildi. Bu çalışmada, mevcut yorumlanmış jeolojik haritaya kıyasla yalnızca verilerin %2’si eğitim örneği olarak kullanıldığında RF’nin yaklaşık %76 doğrulukla litoloji sınıflandırması yapabileceği gösterildi. RF’nin stratigrafik ilişkilerden türetilmiş sınıf etiketlerini koruyabilmesi ve eşdeğer litolojileri ayırt edebilmesi önemli bir bulguydu. Burada önemli olan, RF’nin, fiziksel yanıt, yükselti, kaynağa derinlik veya sensör yüksekliği gibi unsurlardan bağımsız olarak,

yalnızca litoloji problemini en isabetli şekilde çözmeye olanak tanıyan değişkenleri kullanmasıdır. Bu çalışma, jeokimyanın bulunmadığı durumlarda RF'lerin keşif türü jeofizik verileri analiz etmek ve güvenilir litoloji tahminleri üretmek için kullanılabileceğini göstermektedir [39].

Ao ve arkadaşları, litoloji olasılığı tahmin doğruluğunu artırmak amacıyla olasılığa dayalı bulanık karakterizasyon yaklaşımı ve olasılıksal RF algoritmasını tanıttı. Gerçek veri setleri üzerinde yapılan karşılaştırmalı deneylerle yöntemin etkinliği gösterildi ve formasyon özelliklerine dair değerli içgörüler sağlama potansiyeli vurgulandı. Bu da rezervuar karakterizasyon inceliklerinin artırılmasını sağladı [40].

Bressan ve arkadaşlarının çalışması, Uluslararası Okyanus Keşif Programı (IODP) kapsamındaki açık deniz kuyusu verilerini kullanarak litolojileri sınıflandırmak için makine öğrenmesi yöntemlerini uyguladı. RF algoritmasıyla %80'in üzerinde doğruluk elde edildi. Bu durum, makine öğrenmesinin farklı bölgelerde jeolojik veri yorumlamada verimlilik ve doğruluk açısından yüksek potansiyele sahip olduğunu vurguladı [41].

Zhong ve arkadaşları, makine öğrenmesini kullanarak, özellikle extreme gradient boosting (XGBoost) algoritması ile sondaj verileri ve sondaj esnasında yapılan ölçümlere (LWD) dayalı sahte yoğunluk logu oluşturmayı başardı. Yöntem, kömür gazı kuyularında %5'ten düşük ortalama hata oranıyla gerçek zamanlı rezervuar karakterizasyonu sağlama kapasitesi gösterdi. Bu sayede, kablolu log veya pahalı LWD yöntemlerine olan ihtiyacın azalması ihtimali yükseldi [42].

Zhenhao ve arkadaşları, kaya görüntüleri ve element verilerini birleştirerek maliyet-verimli litoloji tanımlaması için aşamalı model eğitimi yaklaşımı önerdi. Bu yaklaşım, yalnızca görüntü kullanımına kıyasla daha yüksek doğruluk (%94.62) sağladı ve kaya element verilerine olan bağımlılığı azalttı. Görüntü benzerliği ve küçük litoloji özellikleri görüntü belirleme yönteminin performansını düşüren faktörlerdir. Bu yöntem bunlar gibi zorlukların aşılmasını sağladı ve litoloji belirlenmesinde doğruluğu artırdı [43].

Nugroho ve arkadaşları, yoğun bitki örtüsüne sahip bölgelerde RF algoritmasının hiperparametre ayarlarıyla birlikte, uzaktan algılama verileri kullanarak litolojik sınırları belirleme başarımını değerlendirdi. Eğitim verisinin sınırlı olduğu koşullarda bile, modeller yüksek eğitim doğrulukları (%92–100) elde etti. Yalnızca 50 dengelenmiş eğitim noktasına sahip Model 9 en iyi sınıflandırma sonuçlarını verdi. Bu durum, dengeli eğitim verisiyle RF sınıflandırmasının öngörüsel litoloji haritalamada etkili olabileceğini gösterdi [44]. Martin ve arkadaşları, Birleşik Krallık Kıta Sahanlığı'ndaki 204. Kuadrant'tan alınan açık kaynak karot ve kuyu log verilerini kullanarak santimetre ölçeğinde litoloji ve fasiyes tahmini yapmak için makine öğrenmesi modellerini kullandı. En iyi performans gösteren model, litoloji tahmininde %69 ve kumtaşı ile çamurtaşı sınıflarında %80'in üzerinde doğruluk sağladı. Bu yöntem, karot görüntü verilerini analiz etmek için ölçeklenebilir ve tekrarlanabilir bir iş akışı sundu [45].

Xie ve arkadaşları, kuyu log verilerinde litoloji sınıflandırması için birleştirilmiş bir çerçeve sundu. Aykırı değerlerin tespit edilmesi ve çok sınıflı sınıflandırma gibi zorluklar, aşırı rastgele ağaç bazlı sınıflandırıcıyla ele alındı. Deneysel sonuçlar, modelin özellikle kumtaşı tahmininde temel sınıflandırıcılara göre üstün olduğunu gösterdi [46].

Merambayev ve arkadaşları, Kazakistan ve Norveç'ten alınan çeşitli kuyu log verilerinde litofasiyes sınıflandırmasını makine öğrenmesi yöntemleriyle değerlendirdi. Ek olarak, kuyu log verilerinden jeolojik özellikleri çıkarmak için dalgacık dönüştürümlü modeller oluşturuldu ve RF, KNN, Karar Ağacı, XGBoost ve LightGBM algoritmaları karşılaştırıldı. Doğruluk, Hamming loss ve ceza matrisi gibi ölçütlerle modelin skoru değerlendirildi. RF sınıflandırıcısı 200 ağaç ve maksimum 70 derinlikle çalıştırıldı. RF'nin yüksek doğruluk sağladığı görüldü. RF sınıflayıcısı modelinin yüksek doğruluğunu doğrulamak için bir SHAP paketi oluşturuldu. GR (gamma ray) özelliğinin yüksek SHAP değerlerine sahip olduğu ve kumtaşı, kireçtaşı, tebeşir, halit, anhidrit, kömür sınıflarını etkilediği görüldü. Ancak GR özelliği, şist, marn ve temel kaya gibi litoloji sınıflarında negatif etkiye sahipti. GR, DTC ve RHOB gibi değişkenlerinin çoğunun litoloji sınıfını etkilediği gözlemlendi. RF modeli, diğer algoritmalardan daha iyi performans sergiledi ve çapraz doğrulama yöntemiyle doğrulandı [47].

Gamal ve arkadaşları, 2021 yılında RF ve Karar Ağacı (DT) algoritmalarının yer bilimlerindeki araştırmalarına katkı sundu. Bu çalışmada, zorlu litolojik koşullarda yapılan sondaj operasyonlar sırasında kaya gözenekliliğini gerçek zamanlı tahmin etmek için RF ve DT'ye dayalı iki akıllı model geliştirildi. Modeller, matkap ucunda oluşan ağırlık, tork, boru basıncı, matkap dönüş hızı, delme hızı ve pompa hızı gibi sondaj parametreleriyle eğitildi ve doğrulandı. 3767 veri noktasından oluşan iki veri seti kullanıldı. Modellerin başarımı dört ölçütle değerlendirildi: Determinasyon katsayısı (R2), Ortalama mutlak yüzde hata (AAPE), Açıklanan varyans (VAF) ve a20 indeksi. RF modeli, hem eğitim hem de test aşamalarında genellikle DT'ye göre daha iyi performans gösterdi. RF modeli, %99 ve %90 R2, %1.5 ve %7 AAPE sağladı. DT modeli %94 ve %87 R2, %6.07 ve %9 AAPE ile sonuç verdi. RF modeli 100 ağaç ve maksimum 15 derinlikle çalıştırıldı. Her iki modelin de yüksek doğrulukla tahmin yaptığı doğrulandı [48].

Albert ve Ammar, Tunus'un merkezindeki Jebel Meloussi bölgesinin özgün jeolojik özelliklerine odaklanarak Sentinel 2 uydu sensörü ve MERIT DEM verileriyle kurak bölgelerde jeoloji haritalama yapmak için uzaktan algılamanın uygulanma potansiyelini inceledi. R script'leri ve RF sınıflandırma yöntemiyle uydu görüntülerinden türetilmiş dört litolojik değişken ve iki morfometrik parametre modellendi. Sınıflandırma, Eosen ve Kretase dönemine ait evaporitli seriler, Sidi Khalif Formasyonu'nun piritli argillitleri ve Kuaterner dönemine ait sebka ve kumullar gibi jeolojik birimleri doğru şekilde tanımladı. Model, 26 farklı jeolojik kategorinin dağılımını gerçek harita ile oldukça uyumlu olarak sundu [49].

Kumar ve arkadaşları, Hindistan'ın doğusundaki bir kömür sahasına ait kuyu log verilerini kullanarak litoloji tahmini için makine öğrenmesi tekniklerini uyguladı. Talcher kömür sahasındaki dört sondaj kuyusundan elde edilen verilerle Destek Vektör Makineleri, Karar Ağacı, RF, Çok Katmanlı Algılayıcı (Multi-Layer Perceptron) ve Extreme Gradient Boosting algoritmaları test edildi. Tüm modeller %88'in üzerinde doğruluk sağladı. ROC eğrileri olumlu alan altında kalan değerler gösterdi. Yakındaki kuyulara uygulandığında da %80'in üzerinde doğruluk elde edildi. RF modeli 100 ağaç ve maksimum 10 derinlikle çalıştırıldı. Makine öğrenmesi performansı, karışıklık matrisi, çapraz doğrulama ve hiperparametre ayarlamalarıyla değerlendirildi [50].

Makine öğrenmesi; çevresel kısıtlamaların az olduğu, verinin bol, kararların ucuz olduğu çevrimiçi reklamcılık, alışveriş ve oyun gibi alanlarda hızla ilerleme kaydetmiştir. Öte yandan, yer bilimi tam tersi özellikler taşır: kararlar pahalı, verileri edinmenin zor olduğu ve sık sık hatalı verilerin bulunduğu, ortam ise değişken ve kısıtlıdır. Bu nedenle otomatik sismik yorumlama gibi görevler başlangıçta daha az kısıtlamaya sahiptir. Ancak genel olarak makine öğrenmesindeki gelişmeler yer bilimi, özellikle de jeofizik alanı tarafından aktif olarak takip edilmektedir. Sonuç olarak, yer bilimi alanında makine öğrenmesi uzun bir geçmişe sahiptir.

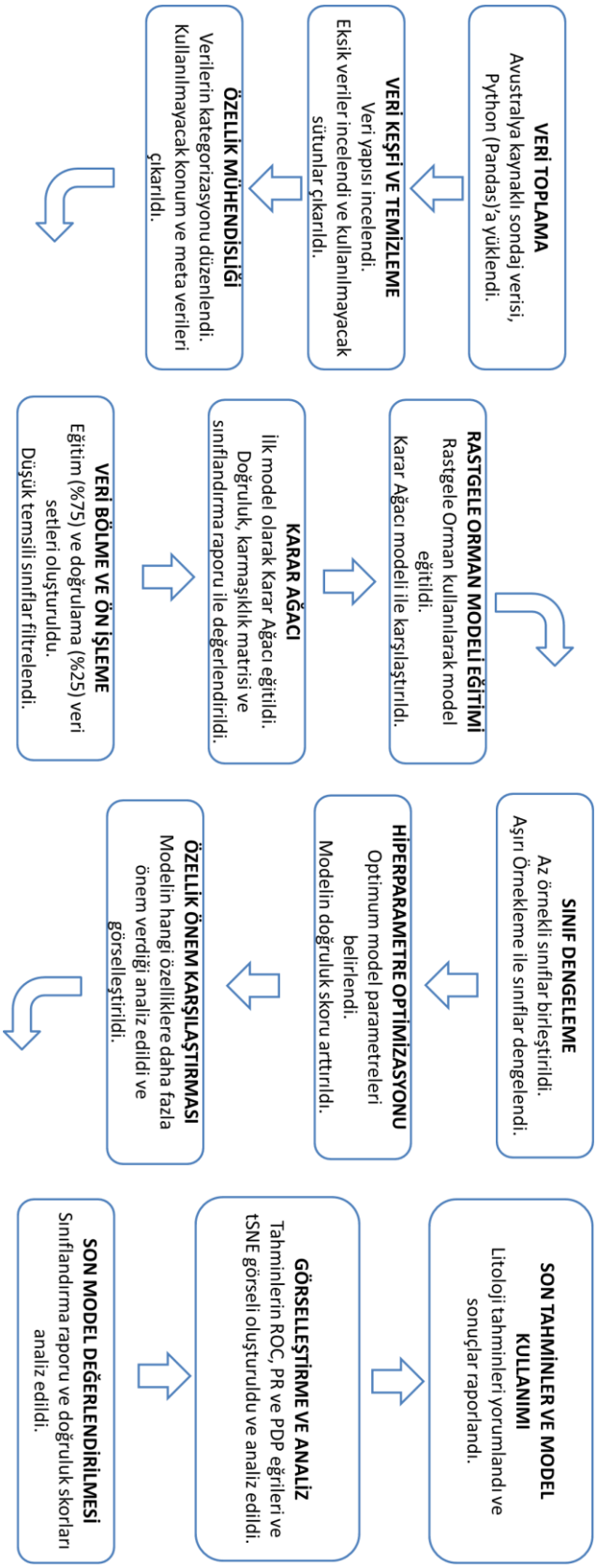
Bu çalışmada, RF sınıflandırıcısı, çeşitli kuyulardan alınan analiz (assay) parametreleri yardımıyla litolojik kimliği tahmin etmek amaçlanarak kullanılmıştır. Elde edilmek istenen temel çıktı, litoloji tahmini için bir veri tabanı oluşturmaya yönelik bir kod geliştirmektir. Bu veri tabanı sayesinde sondaj faaliyetlerinin maliyeti ve zaman tüketimi önemli ölçüde azaltmak amaçlanmaktadır. Bu veri tabanının daha önce hazırlanmış blok modeller ve tenör verileri üzerinden jeolojik model üretme, litolojik veri elde etme ve litolojik ölçümlerde insan hatasının faktörünün önüne geçerek kimyasal verilerin verdiği güven ile teknik bir doğrulama yapmak amaçlarıyla kullanılması öngörülmüştür. Rastgele Orman sınıflandırıcısı özellikle son yıllarda yer bilimleri verileriyle uyumlu tahmin yapabilme kapasitesini, özellikle diğer Makine Öğrenimi algoritmaları ile kıyaslanarak yapılan çalışmalarda da göstermiştir. Çalışma kapsamında Güney Avustralya'da yapılmış ve açık kaynak olarak verilen 21 farklı sondaj deliğinden toplanan veriler ön işleme ve hiper parametre optimizasyonu aşamalarından geçirilmişlerdir. Araştırma sırasında Rastgele Orman modeli ve Karar Ağacı modelinin karşılaştırması yapılmıştır.

3. DENEYSEL ÇALIŞMALAR

3.1. Çalışma Aşamaları

Çalışma, verinin toplanmasından model sonuçlarının yorumlanmasına kadar uzanan kapsamlı bir süreç olarak planlanmıştır. Şekil 3.1'de gösterildiği üzere araştırma, veriye ön işleme uygulanmasından modelin değerlendirilmesine kadar 11 aşamalı bir yöntem izlenerek gerçekleştirilmiştir.

Çalışma öncelikle verinin edinilmesi ve ön işlemden geçirilmesi sonrasında birer karar ağacı ve rastgele orman modeli oluşturulmuş ve iki modelin performansı kararlaştırılmıştır. Ardından rastgele orman modelinin performansını artırmak amacıyla sınıf dengelemesi, hiperparametre optimizasyonu ve aşırı örnekleme yapılmıştır. Son olarak son model değerlendirilmesi yapılmıştır.

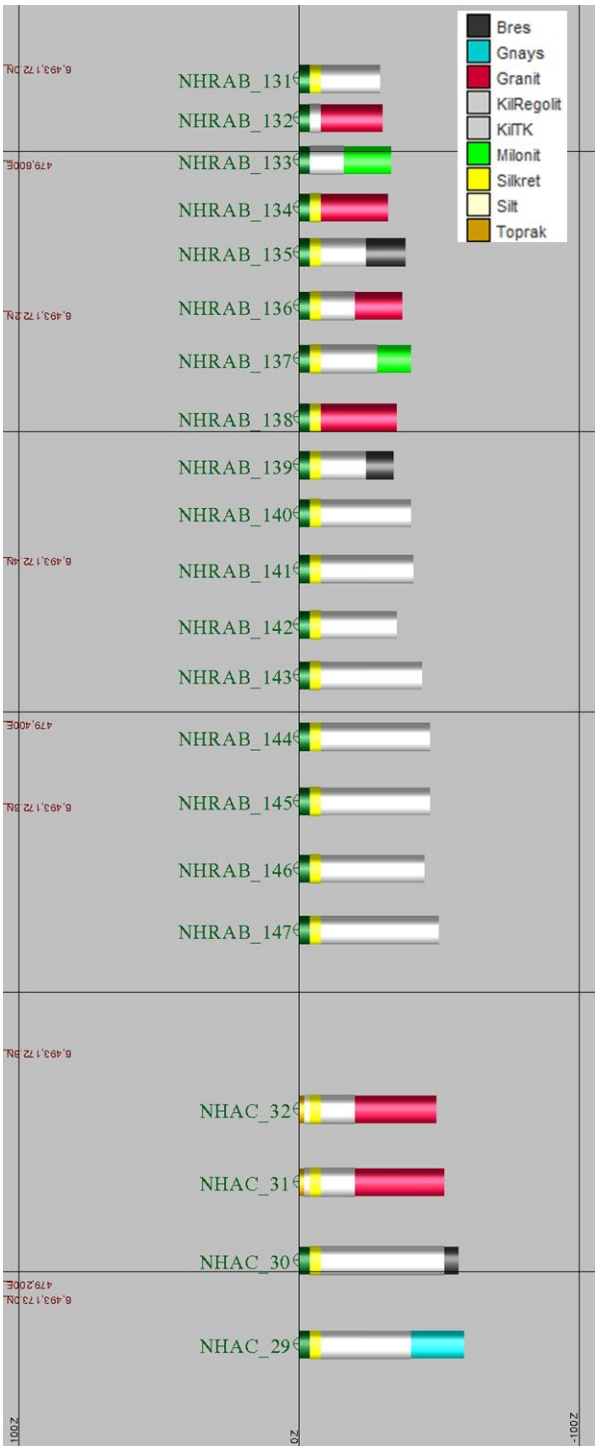


Şekil 3.1. Çalışma Aşamaları

3.2. Saha Çalışması ve Veriler

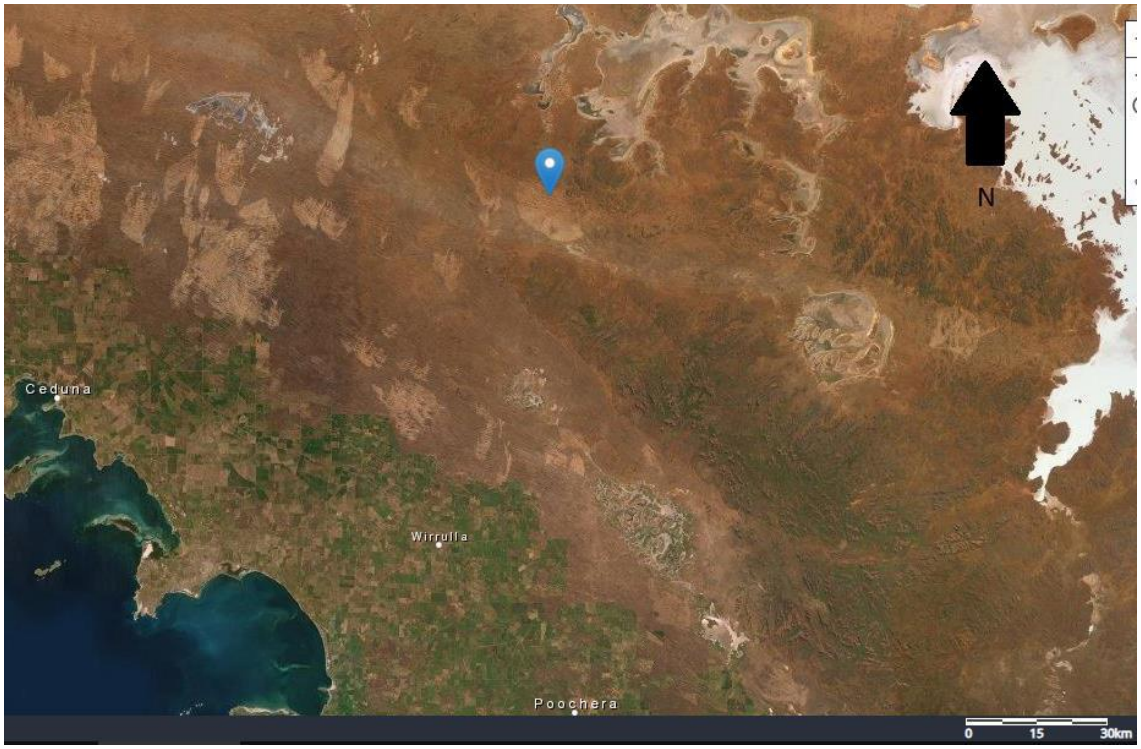
Araştırmada kullanılan veriler SARIG Map sitesinden alınmıştır. SARIG (South Australian Resource Information Gateway), madencilik, petrol ve arama topluluklarına çeşitli açık jeoloji bilimi verileri ve ürünleri sunan bir haritalama portalıdır. Veri Şekil 3.3.'de gösterilen haritada gösterildiği gibi Güney Avustralya'nın Ceduna kasabasına yakın bir bölgedeki Nuckulla isimli bir tepede yapılan sondaj çalışmasından alınmıştır. Toplamda 21 adet sondaj verisi ile çalışılmıştır ve jeokimya sonuçları veri setleri kullanılmıştır.

Veri setleri içinde Numune No, Numune Kaynak Kodu, Numune Kaynağı, Kayaç Grup Kodu, Kayaç Grubu, Litoloji Kodu, Litoloji İsmi, Harita Sembolü, Stratigrafi İsmi, Verinin Kim Tarafından Toplandığı, Verinin Toplandığı Tarih, Delik No, Delik İsmi, Numune Derinliği, Saha No, Doğu Koordinatı ,Kuzey Koordinatı, Enlem (GDA2020 ve GDA94 sistemlerinde), Boylam (GDA2020 ve GDA94 sistemlerinde), Numune Analiz Numarası, Analiz Türü, Analizin Yapıldığı Laboratuvarın Adı , Metal Cinsi, Mineral Tenör Değeri, Tenör Birimi, Kimyasal Metot Kodu, Kimyasal Metot Tanımı bulunmaktadır. Veriler 1995 ve 1996 tarihlerinde Equinox Resources NL. tarafından toplanmıştır. Veri setinde 9 adet litoloji sınıfı ve 7 adet metal sınıfı bulunmaktadır. Litoloji sınıfları: Silkret, kil (regolit), breş, kil (tanımsız köken), gnays, granit, milonit, silt ve toprak. Metal sınıfları: altın, gümüş, çinko, kurşun, bakır, krom ve nikeldir. Derinlik değerleri 4 m ile 59 m arasındadır. Deliklerden ilkinin koordinatı-31.6965993° enlemi ve 134.785041° boylamı olarak verilmiştir. Sondajların 3 boyutlu gösterimi Şekil 3.2.'de gösterilmiştir.



Şekil 3.2. Sondajların 3 boyutlu gösterimi

Veri tabanlarındaki her veri setinin özellikleri arasında tutarsızlıklar, boş değerler, uç değerler, gürültüler bulunur. Veriyi temizleme yöntemleri boş değerlerin doldurulmasını, uç değerlerin tanımlanmasını ve tutarsızlıkların çözülmesini sağlar. Temizlenmemiş veri tarama sürecinde karışıklık yaratabilir ve bazı algoritmalar bu veri temizleme aşamaları olmadan yeterince etkili çalışmayabilir.



Şekil 3.3. SARIG Map (Verilerin alındığı lokasyon mavi nokta ile gösterilmiştir) [51]

Veriler excel formatında aktarılmıştır. Veri setindeki litoloji birimleri hakkındaki bilgiler şu şekildedir:

Silkret: Kuvars tanelerinin ve silisli kütlelerin birleşiminden oluşan bir sertleşmiş sedimenter kayaçtır. Genellikle kuru ve yarı kuru bölgelerde bulunur. Çözünme, çökelme ve yeniden maruz kalma gibi farklı çevresel koşullarda oluşabilir. Avustralya (özellikle Güney Avustralya) ve Afrika'da yaygındır. Silkret aşırı derecede serttir ve ayrışmaya dayanıklıdır [52].

Kil: Hydro alüminyum filo silikat minerali olan kili ana mineral olarak içeren doğal toprak malzemesidir [53]. İnce tanelidir, yüksek yüzey alanına sahiptir ve ıslandığında yüksek oranda plastiktir.

Regolit: Kayaçları çevreleyen bir üst katman olarak bulunan konsolide olmayan, gevşek, mineral ve cam parçalarıdır. Regolit, kaolin (kil mineral formasyonu) dahil olmak üzere ekonomik değeri olan birçok mineral barındırır [54].

Breş: Çeşitli jeolojik ortamlarda endüstriyel minerallere ve cevher yataklarına ev sahipliği yapmakta olan bir sedimenter kayaçtır. Baz metallerin, hidrokarbon rezervlerinin ve metalik olmayan minerallerin potansiyel yolları olarak incelenmektedir [55].

Gnays: Yeraltı mühendislik yapılarında yaygın olarak kullanılan bir metamorfik kayaçtır. Şistten daha yüksek sıcaklık ve basınç altında metamorfik proses geçirir. Granit veya sedimenter kayaçların metamorfozundan oluşur [56].

Milonit: İnce taneli bir metamorfik kayacıdır. Yüksek basınç ya da dinamik stres etkisi altındaki bölgeler boyunca çeşitli kayaçların sürtünmesi ve kırılması ile oluşur. Baskın olarak kuvars, feldspar, kalsit, dolomit, serisit ve klorit içerir [57].

Silt: Boyutu kum ve kil arasında olan bir granüler sedimenttir. Su, buzul ve rüzgar yoluyla taşınabilir ve yataklanabilir [57].

Granit: Bir magmatik kayaçtır. Genellikle kuvars, feldspar, sodyum plagyoklas, biyotit ve hornblend bulundurur [57].

Toprak: Biyolojik olarak aktif ve yer kabuğunun en üst kabuğunu oluşturan gözenekli maddedir [58].

Verilerin boyutu ve içeriğini belirlenmiş, litoloji ismi, kimyasal kod ve mineral tenör değeri harici veriler değerlendirilmeye alınmamıştır. Eksik verileri olan delikler çalışmaya alınmamış ve üzerinde çalışılacak delik sayısı 21'e düşürülmüştür. Bu sayede hedef değişkenler tamsayıya dönüştürülmüştür.

3.3. Yöntem

Bu araştırmanın amacı, jeolojik sondajlardan elde edilen assay (tenör) verisi aracılığıyla litoloji tahmini yapmaktır. Çıktı değişkeni olan litoloji kategorik bir yapıya sahip olduğu için, bu araştırmadaki tahmin bir sınıflandırma problemidir. Tenör verisi girdi değişkeni, litoloji ise ürün değişkeni olarak tanımlanmıştır. Rastgele Orman ve Karar Ağacı

algoritmaları özellikle son yıllarda yer biliminde makine öğrenimi kullanımı denemelerinde en çok kullanılan algoritmalardan olmaları nedeniyle seçilmiştir. Karar Ağacı daha basit bir model olması dolayısıyla modelin nasıl karar aldığının görüntülenmesi açısından, Rastgele Orman ise birçok ağaç kullanması dolayısıyla daha doğru bir sonuç elde edilebilmesini sağlaması açısından değerlidir.

Kullanılan veri seti gerçek sondaj verileridir ve yer altı litolojisinin belirlenmesinde kullanılabilir. Bu nedenle araştırma hem jeolojik olarak hem de makine öğrenimi açısından değerlidir. Modeller, doğruluk (accuracy), hassasiyet (precision), duyarlılık (recall), F1 skoru (F1 score) ve Kappa endeksi gibi performans göstergeleri ile değerlendirilmiştir.

Sınıflandırma doğruluğu (classification accuracy): Doğru sınıflandırmaların genel sınıflandırmaya göre sıklığına denir [10]. Doğruluk literatürde şu şekilde tanımlanır [10]:

$$Doğruluk = \frac{n_{corr}}{\mathrm{n}} \times 100\%$$

n: problemin mümkün olan bütün örnekleri

n_{corr}: doğru sınıflandırılmış örnekler

Doğruluk, rastgele seçilmiş bir örneğin doğru sınıflandırılmış olma ihtimali olarak da yorumlanabilir. Sınıflandırma doğruluğunun hesaplanması için bağımsız bir test veri setinin bulunması önemlidir, çünkü gerçek verinin doğru sınıflandırmasını olası her örnek için belirlemek imkansızdır.

Hassasiyet (Precision): Pozitif sınıflandırılmış örneklerin içindeki doğru sınıflandırılmış örnek oranını kestirir. Hassasiyet literatürde şu şekilde tanımlanır [10]:

$$Hassasiyet = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$$

TP: Doğru pozitif sınıflanmış örnekler,

FP: Yanlış pozitif sınıflanmış örnekler

Duyarlılık (Recall): Pozitif sınıflanması gereken örneklerin ne kadarının doğru pozitif sınıflandırıldığını gösterir.. Duyarlılık literatürde şu şekilde tanımlanır [10]:

$$Duyarlılık = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

FN: Yanlış negatif sınıflanmış örnekler

F1 Skoru: Hassasiyet ve duyarlılığın harmonik ortalamasıdır[59]. F1 skoru, bu iki ölçütün tek bir ölçütte kombine ederek verimliliği arttırmayı amaçlar[60]. F1 skoru literatürde şu şekilde tanımlanır [10]:

$$F1\ Skoru = \frac{2TP}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Doğruluk, hassasiyet, duyarlılık ve F1 skoru bir model için en iyi 1, en kötü ise 0 değerini alabilir.

Kappa Endeksi: temel fikri, doğruluk değerinden şansa bağlı kısmı, yani rastgele bir sınıflayıcının sonuçlandıracağı kısmı "telafi etmek"tir. Bu kısım da beklenen doğruluk olarak adlandırılır. Kappa değeri -1 ve 1 arasında değerleri arasında olabilir ve negatif Kappa endeksleri rastgele tahminden daha düşük kabul edilir. Kappa endeksi literatürde şu şekilde tanımlanır [61]:

$$\text{Kappa endeksi} = \frac{Doğruluk - Beklenen\ Doğruluk}{1 - Beklenen\ Doğruluk}$$

İki modelin birbirinden farkının şansa bağlı olup olmadığını öğrenmek adına McNemar testi uygulanmıştır. Bu testte bir olasılık tablosu oluşturulur. Bu tablo iki farklı model için doğru tahmin edilen ve yanlış tahmin edilen değerlerin ortaklıklarını gösterir. Formülle

hesaplanan McNemar istatistik değeri 1 serbest dereceli x^2 dağılımına göre değerlendirilir. McNemar istatistiği literatürde şu şekilde tanımlanır [62]:

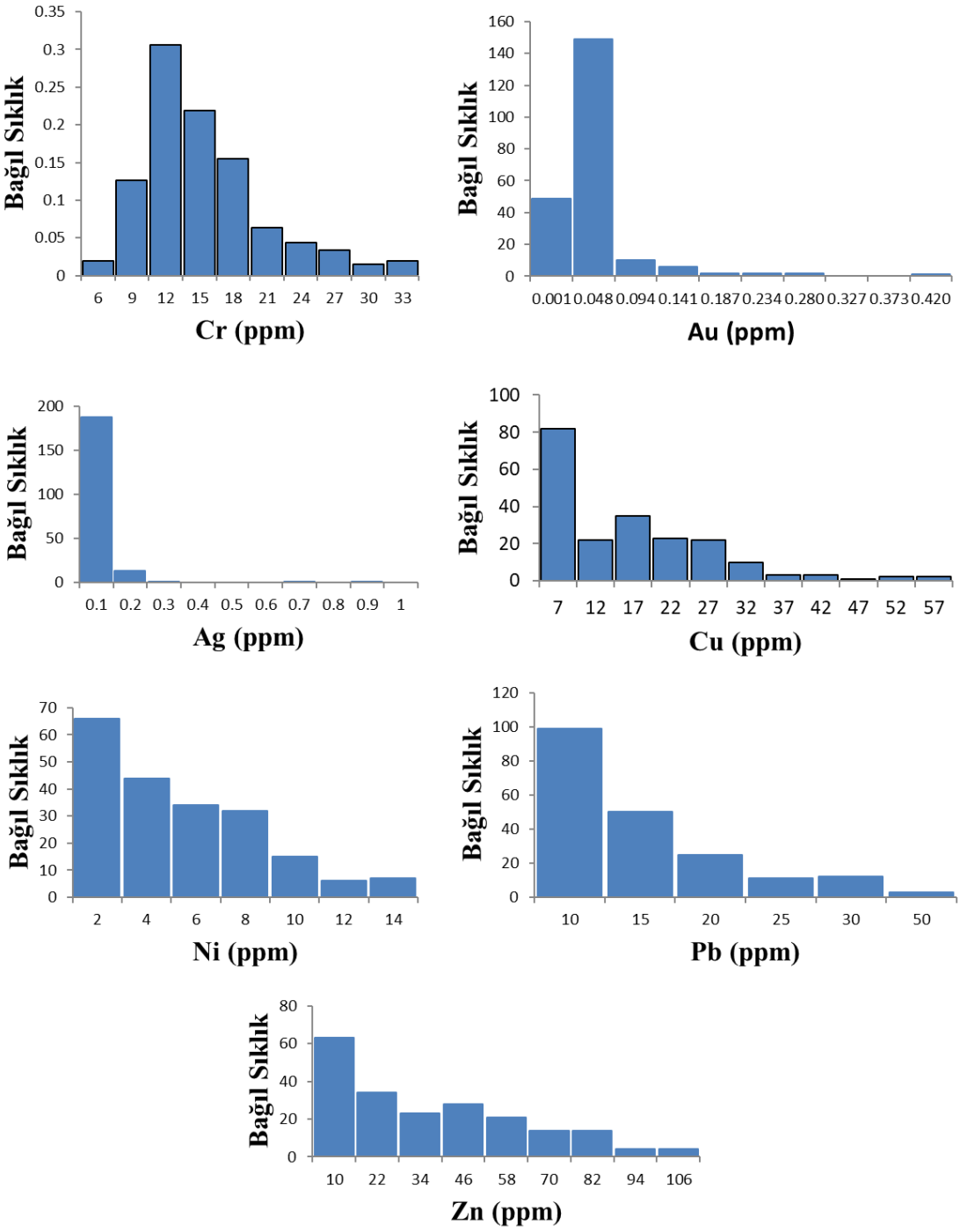
$$\text{McNemar istatisiği} = \frac{((n_{10} - n_{01}) - 1)^2}{n_{10} + n_{01}}$$

n_{10}: McNemar olasılık tablosundaki 2.satır-1.sütun değeri

n_{01}: McNemar olasılık tablosundaki 1.satır-2.sütun değeri

Karar Ağacı, Rastgele Orman ve parametreleri belirli Rastgele Orman algoritmalarının performans değerleri birbirleri ile karşılaştırılmıştır. Tenör verilerinin doğrudan litoloji tahmini için kullanılması bu araştırmanın beklentisidir.

Kullanılan veri setinde bulunan tenor değeri (Value) değişkeninin değer dağılımı bütün metaller için ayrı ayrı kontrol edilmiştir. Bu amaçla Şekil 3.4. ve Çizelge 3.1.'de gösterilen özet istatistikler ve histogramlar incelenmiştir. Histogramlardan görüldüğü üzere dağılımlar Ag dağılımı hariç sağa çarpıktır ve tenör değeri değişkeni 0 ve 20 arasında yoğunlaşmış durumdadçoğu metal sağa çarpık bir dağılım göstermektedir. Bu düşük tenör değerlernin baskın olduğunu ancak çok daha yüksek değerlerin de veri setinde bulunduğunu göstermektedir. Özellikle Cu, Cr, Pb ve Zn dağılımlarında çok yüksek değerlerin de bulunduğu görülmektedir. Bu dağılımın modelin düşük değerleri daha kolay öğrenip yüksek değerleri öğrenmekte zorlanmasına sebep olabileceği ihtimali değerlendirmede göz önünde bulundurulmuştur. Bu uç değerlerin ileri görselleştirme aşamalarında zorluk yaratması ihtimali göz önüne alınarak veri standardize edilmiştir. Bu sayede veri setinin ortalaması 0 ve standart sapması 1 olacak şekilde dönüşmesi ve görselleştirmelerde kolaylık sağlamak amaçlanmıştır.

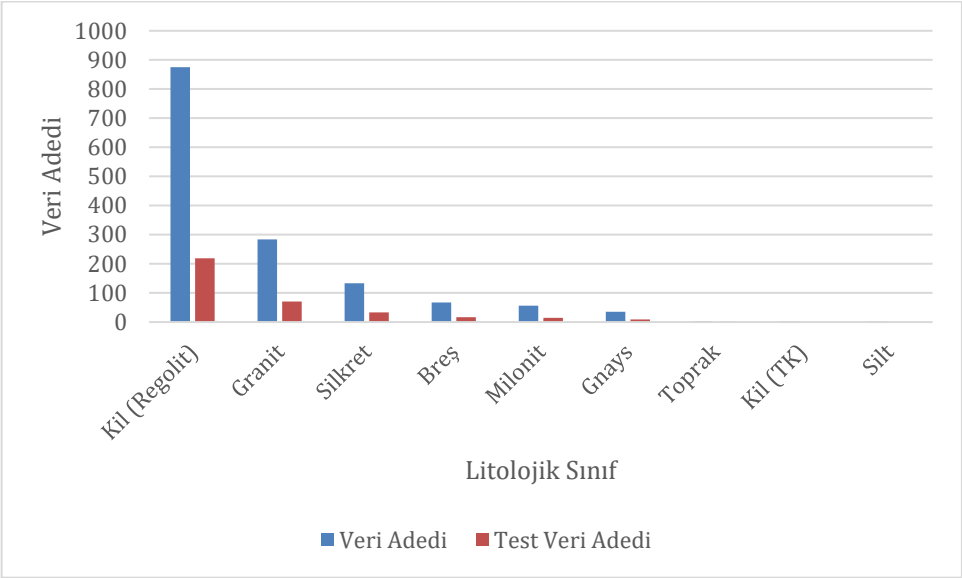


Şekil 3.4. Metal değişkeni histogramları

Çizelge 3.1. Metal değişkeni özet istatistikleri

İstatistik Türü	Au	Ag	Cu	Cr	Ni	Pb	Zn
Minimum	0,001	0,10	1,80	5,00	0,80	2,50	2,80
Maksimum	0,420	1,10	55,20	33,00	21,50	87,10	107,00
Ortalama	0,020	0,12	13,82	14,00	4,60	13,23	31,87
Ortanca	0,004	0,10	11,50	13,00	3,50	10,45	25,35
Standart Sapma	0,049	0,12	10,62	5,41	3,35	11,12	26,00

Modeldeki ilk ve en etkili sorunlardan biri veri setindeki litoloji sınıf dağılımındaki dengesizliktir. Sınıf dağılımı Şekil 3.5.'te gösterildiği üzere yüksek oranda regolit içermektedir ve 10'un altında 3 adet sınıf bulunmaktadır. Bu dengesizlik eğitim setini ve dolayısıyla test setini de dengesiz hale getirmektedir. Bu nedenle aynı grafikte görüldüğü üzere test setinde en az örneği bulunan milonit, gnays, toprak ve silt sınıfları ayrı bir sınıf olarak birleştirilerek "Diğer" adıyla sınıflandırılmıştır.



Şekil 3.5. Litoloji Sınıfı Veri Dağılımı

Sınıf birleştirme işleminin ardından geriye 5 litoloji sınıfı kalmıştır. Ayrıca eğitim setinde 2'den az örneği olan veriler değerlendirmeye alınmamıştır. Bu nedenle kökeni tanımlanmamış kil de değerlendirmeden çıkarılmıştır. Veri setlerindeki dengesizlikler

azaldıktan sonra geri kalan 5 sınıf ile değerlendirme yapmak için sayıları aşırı örnekleme (oversampling) adı verilen yöntem ile eşitlenmiştir. Aşırı örnekleme verileri kopyalayarak yeni örnekler elde etme işlemidir. SMOTE algoritması orijinal eğitim setini dengeleme amacıyla bir aşırı örnekleme yaklaşımı yürütür. SMOTE'un ana fikri, sentetik veri üretmektir. Yeni örnekler interpolasyon ile üretilir. Bu işlem sınıflar arası 1:1 dağılım elde etmeyi amaçlayarak sınıf sayılarını bir tamsayıda eşitler. Bu çalışmada litoloji sınıfı örnekleri SMOTE aracılığıyla 656'şar örneğe kadar genişletilmiştir. Aşırı örnekleme yapılırken, her örnek yalnızca en yakın komşusuyla interpolasyon yapılarak yeni örnekler üretilmiştir.

Litoloji sınıfının tenör değeri değişkenlerine bağlı istatistikleri Çizelge 3.2.'de verilmiştir. İstatistiklerden anlaşılabileceği gibi Diğer ve Breş sınıfları yüksek standart sapma göstermektedir, bu nedenle örneklerin sınıf içinde çok farklı tenör değerleri aldığı söylenebilir. Granit ve Kil sınıflarının daha çok veriye sahip olması nedeniyle daha güvenilir tahmin edilmesi olasılığı yüksektir.

Çizelge 3.2. Litoloji sınıfı tenör değişkeni istatistikleri

Litoloji	**Veri adedi**	**Ortalama (mean)**	**Standart sapma**
Breş	17	24,89	34,14
Diğer	24	16,55	23,05
Granit	71	11,47	14,44
Kil (Regolit)	219	10,78	14,49
Silkret	33	6,75	7,22

Yerbilim alanında genel olarak başarılı sonuçlar vermesi dolayısıyla bu araştırmada karar ağaçları ve rastgele orman algoritmaları kullanılmıştır. Ayrıca karar ağaçları yorumlanabilirliği yüksek bir algoritma olması ve rastgele orman da yüksek doğruluğu dolayısıyla tercih edilmiştir.

Makine Öğrenimi modelleri eğitim ve test olmak üzere 2 sete ayrılmalıdır. Eğitim seti üzerinde öğrenme ve ardından test seti üzerinde modelin performansı değerlendirilmelidir. Bu uygulamanın amacı modelin yeni veri ile nasıl davranış gösterdiğini test ederek performansını analiz etmektir. Veriler eğitim ve test olmak üzere iki gruba ayrılmıştır. Ayrım %75 eğitim, %25 test grubu olarak yapılmıştır. 'X', tenör ve metal cinsi verisi ile ve 'y'de litoloji verisi ile eşlenmiştir. Dengesizlikleri önlemek amacıyla çok az sayıda olan örnekler sınıflandırmadan çıkarılmıştır.

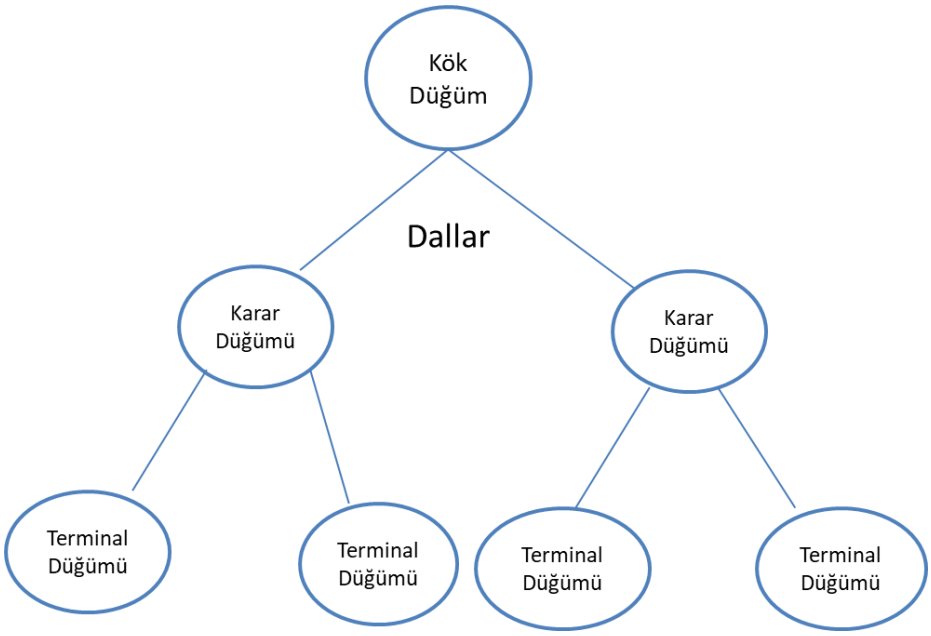
Karar ağaçları Şekil 3.6.'te görüldüğü gibi köklü bir ağacı oluşturan düğümlerden oluşur. Gelen kenarı olmayan düğümlere kök adı verilir. Ayrılacak kenarı bulunan düğümlere "karar düğümü" ve ayrılacak kenarı olmayan düğümlere "terminal düğüm" denir. Çoğu karar ağacında karar düğümleri tek bir değişkenin değerine göre bölünür. Sınıflandırıcı ayrımı yapmak için en iyi niteliği arar. Birçok tek değişkenli ayrım kriteri vardır. Bunlardan biri Gini ölçütüdür ve safsızlık bazlı bir ayrım kriteridir. Hedef nitelik değerlerinin olasılık dağılımları arasındaki sapmayı ölçer[62]. Buna göre kestirilen yanlış sınıflandırma ihtimali Gini ölçütünü verir. Gini ölçütü literatürde şu şekilde tanımlanır [63] :

$$Gini\ \text{Ölçütü} = 1 - \sum_{j}^{n} p^2(\mathrm{j} \mid \mathrm{t})$$

p(j | t): t 'deki örneklerin j' ye ait olma ihtimali

n: sınıf sayısı

Ağaçların büyüme süreci bir durma kriteri tetikleyene kadar devam eder. Durdurma kriterleri uygulamaya budama adı verilir. Ağaçları durdurma kriterleri uygulamak sıkı ve gevşek olma durumuna göre modellerin aşırı uyum veya yetersiz uyum sağlamasına sebep olabilir. Budama metotlarından biri maliyet karmaşık budamasıdır. Bu yöntem inşa edilmiş bir ağacın içinde hata kestirimlerine göre bir budama noktası belirlenmesi şeklinde işler [62] .



Şekil 3.6. Karar Ağacı kısımları [64]

Karar ağacı modeli eğitildikten sonra doğruluk oranı ve sınıflandırma verileri hesaplanmıştır. Performansını değerlendirmek için karmaşıklık matrisi kullanılmıştır. Elde edilen karar ağacı modelinin parametreleri kontrol edilmiştir. Bir terminal düğümü için gereken örnek sayısı minimum 1 olarak korunmuştur. Bir karar düğümü olmak için gereken örnek sayısı 2 olarak korunmuştur. Düğümlerde gereken ayrım stratejisi olarak mümkün olan en iyi ayrımın yapılması seçeneği göz önünde bulundurulmuştur. Geri kalan parametrelerde değişiklik yapılmamış ve herhangi bir maksimum veya minimum sınırlaması kullanılmamıştır.

Rastgele Orman modelinde birden çok karar ağacı büyür. Her ağaç bir sınıflandırma sunar ve ormanda bir "oy" görevi görür. Rastgele orman hem sınıflandırma hem de regresyon problemlerinde kullanılabilir. Büyük boyutta veri ile çalışabilir. Verimli bir şekilde eksik veriyi kestirebilir ve büyük veri ile doğruluk sağlayabilir. Bu özellikler etiketlenmemiş veri için de geçerlidir ve bu nedenle güdümsüz sınıflandırma da yapabilir. Girdi verisini geri koymalı olarak örnekler ve buna bootstrap (geri koymalı) örnekleme denir [65].

Rastgele Orman modeli uygulanmıştır ve karar ağaçlarıyla aynı yöntem kullanılarak, eğitilecek ve test edilecek veriler belirlenmiştir. Doğruluk oranı, sınıflandırma verileri ve performans değerlendirmeleri aynı şekilde yapılmıştır. Modelde geri koymalı örnekleme

kullanılmıştır. Algoritmada 100 adet ağaç bulunmaktadır. Bu iki modelin ardından parametreleri ikinci bir rastgele orman modeli oluşturulmuştur ve bu sefer sınıflandırma başarısını arttırmak amaçlanmıştır. Skorları yükseltmek hedeflenmiştir. Bütün performans değerlendirmeleri bu üçüncü model için de yapılmıştır. Görselleştirmelerin ardından modellerin performanslarını arttırmak amacıyla model iyileştirme uygulamaları yapılmıştır.

İkinci Rastgele Orman modeli için çeşitli veri işleme ve algoritma iyileştirme adımları gerçekleştirilmiştir. Modelin hem dengeli hem de daha yüksek doğrulukla çalışabilmesi amaçlanmıştır. Genelde verimli bir makine öğrenimi modeli oluşturmak, doğru algoritmayı bulmak ve hiperparametre optimizasyonu ile optimal model mimarisini elde etmek amaçlandığı için karmaşık ve zaman alan bir süreçtir [66]. Makine öğrenimi algoritmalarının davranışlarını kontrol eden konfigürasyon değişkenlerine hiperparametre denir [67]. Optimum bir makine öğrenimi algoritması oluşturmak için, olasılıkların çeşitlilikleri keşfedilmelidir. Uygun model mimarisini bir hiperparametre konfigürasyonu ile ayarlama işlemine hiperparametre ayarlama (hyperparameter tuning) denir. Hiperparametre ayarlama, özellikle ağaç bazlı algoritmalarda ve derin sinir ağlarında bir model oluşturmak için kilit bir aşama olarak görülür. Bu süreç farklı makine öğrenimi algoritmaları için farklı işler, çünkü parametreler değişkenlik gösterebilir. Parametreleri manuel olarak test etmek bu sürecin geleneksel bir yoludur. Ancak bu yöntem hiperparametrelerin sayısının çok olması, modellerin karmaşık olması, model değerlendirmelerinin gerektirdiği zaman ve hiperparametrelerin etkileşimlerinin lineer olmaması durumları dolayısıyla verimsizdir. Bu nedenle hiperparametre optimizasyonu (hyperparameter optimization) denilen ve daha yüksek araştırma tekniklerinin kullandığı bir süreç ortaya çıkmıştır. Amaç hiperparametre ayarlama sürecini otomatize etmektir. Bu sayede kullanıcı eforu azalır, modellerin performansları artar ve model ve süreç daha kolay yeniden üretilebilir hale gelir [66].

Hiperparametre optimizasyonunun yöntemlerinden biri uygun verinin tümünü iki sete bölmek ve ilkini bir algoritma için kullanırken diğer seti bekletmektir. Daha karmaşık çözümlerden biri olan çapraz doğrulama (cross-validation), veri setini K adet alt kümeye böldükten sonra her bir model değerlendirmesinin K defa yenilenmesi sonucunda elde edilen ortalamayla tanımlanmasına denir. Her seferinde bir alt küme doğrulama için

ayrılırken, kalan veriler eğitim amacıyla kullanılır [67]. Sklearn kütüphanesinde 'GridSearchCV' optimum hiperparametreleri belirlemek için kullanılabilir. Konfigurasyon alanındaki bütün örnekler değerlendirildiğinde, tanımlanmış arama alanındaki optimum hiperparametre kombinasyonu performans skoru ile sunulur.

Rastgele Orman modelinin belirli parametre seçenekleriyle nasıl en iyi kombine olacağı 5 katlı GridSearchCV ile araştırılmıştır. Bu çapraz doğrulama sonucunda ortaya çıkan ölçütler: dengelenmiş doğruluk (balanced accuracy), geçerleme doğruluğu (validation accuracy) ve sınıflandırma raporudur (classification report). Dengelenmiş doğruluk, Scikitlearn kütüphanesinde bulunan, birden çok sınıfın bulunduğu ve dengesiz veri setleri olan problemler için kullanılan bir ölçüttür [68]. Geçerleme doğruluğu ise modelin test için ayrılmış olan veri seti üzerindeki başarısını gösterir.

Modeller için sınıflandırma raporu değerlendirilmiştir. Sınıflandırma raporu makine öğreniminde bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir araçtır. Hassasiyet, duyarlılık ve F1 skoru gibi ölçütler her sınıf için yer alır.

Scikitlearn kütüphanesi predict_proba adında her bağımsız sınıfın olasılık değerini gösteren bir dizi oluşturan bir yönteme sahiptir [69]. Modelin hangi örneklerde daha yakın skorlar elde ettiğini görmek ve bu sayede güven dağılımını analiz etmek amacıyla bu yöntem aracılığıyla Güven Skoru Dağılımı grafiği çizdirilmiştir.

Görselleştirme, bu araştırmada oluşturulan modellerin daha derinlemesine anlaşılabilmesi, modelin sınıflandırma performansının değerlendirilmesi ve ayrıca özelliklerin önemlerinin karşılaştırılması için de yapılmıştır.

Modellerin tahmin performansları karmaşıklık matrisi kullanılarak görselleştirilmiştir. Karmaşıklık matrisi bir sınıflandırıcının doğru pozitif, yanlış pozitif ve yanlış negatif tahminlerini içeren bir kare matristir [70]. Matrisin köşegeni doğru tahminleri, köşegenin dışındaki noktalar yanlış sınıflandırmaları yansıtır [71]. Bu sayede görsel olarak doğru tahmin ve yanlış tahmin oranını bir ısı haritası yardımıyla görsel yorumlamak kolaylaşır. Karar ağacı modelinin ilk 3 katı görselleştirilmiştir. Bu görselleştirmelerdeki amaç,

modellerin hangi özelliklere ve aralıklara göre ayrım yaptığını ve bu sayede karar verme aşamasını anlamlandırmamıza katkı sağlamıştır. Rastgele Orman modellerinin içlerindeki karar ağaçlarının parametreleri özetlenmiştir ve bir karar ağacı örneği seçilerek görselleştirilmiştir.

Model için kullanılan bir diğer görselleştirme t-dağıtılmış stokastik komşu yerleştirme (t-distributed stochastic neighbor embedding), bir diğer adıyla t-SNE grafiğidir. Bu uygulama benzer örnekleri yakın ve farklı örnekleri birbirinden uzak tutarak kullanılan bir görselleştirme aracıdır. Verilerin düşük boyutlu uzayda görselleştirilmesini sağlar [72]. Bu çalışmada litoloji sınıflarının kümelenme yapılarının analiz edilmesinde fayda sağlayacağı göz önünde bulundurulmuştur.

ROC (Receiver Operating Characteristic) Eğrisi (ROC Curve) sınıflandırıcılar için kullanılan başka bir araçtır. Doğru pozitif duyarlılık ve yanlış pozitif duyarlılık ilişkisini gösterir. Bu grafiklerde sınıflayıcıları karşılaştırmak için AUC skoru kullanılabilir. AUC (area under curve) söz konusu eğrinin altında kalan alanı belirtir. Mükemmel bir sınıflandırıcının AUC skorunun 1 olması gerekir ve tamamıyla rastgele sınıflandırma yapan bir sınıflandırıcı da 0,5 AUC skoruna sahip olur [72].

Hassasiyet-Duyarlılık Eğrisi (Precision-Recall Curve, PR Eğrisi), modelleri oluştururken karşılaşılan ve doğruluk ve hassasiyet arasında bir değiş tokuş (trade-off) olması ve bu değiş tokuş potansiyellerini tamamen görmek için kullanılan bir eğridir. Eğri ne kadar sağ üst köşeye yaklaşırsa, o kadar iyi bir sınıflandırıcı olduğu kabul edilir [69].

Kısmi bağımlılık grafiği (Partial Dependence Plot, PDP eğrisi), bir makine öğrenimi modelinin tahmin edilmiş sonucuna bir veya birden fazla özniteliğin etkisini ölçen bir grafiktir[8]. Bu grafikler özellikle hedefin ilişkisinin monoton, lineer veya daha karmaşık olup olmadığını belirtebilir[73]. Kısmi bağımlılık literatürde şu şekilde tanımlanır [74] :

$$f_S(X_S) = fE_{X_C}(X_S, X_C) = \frac{1}{N}\sum_{i=1}^{N} f(X_S, x_{ic})$$

$f_S(X_S)$: Kısmi bağımlılık

X_S: İlgilenilen girdi değişkeninin alt vektörü

X_C: Tamamlayıcı set

E_{X_C}: X_C'nin marjinal beklentisi (f(X) tahmin fonksiyonunun etkisinde)

f(X): Öğrenilmiş modelin tahmin fonksiyonu

x_{ic}: Eğitim verisinde tamamlayıcı setin aldığı değerler

Son olarak RF modeli sadece Au tenör verileri girdi değişkeni olarak kullanılarak fark gözlenmek istenmiştir.

4. SONUÇLAR VE TARTIŞMA

4.1. Karar Ağacı ve Rastgele Orman Karşılaştırması

Model iyileştirmeleri yapılmadan önce Karar Ağacı ve Rastgele Orman modelleri karşılaştırılmıştır ve Çizelge 4.1.'de sunulmuştur. Bu verilere göre iki modelin doğruluk skorları arasında az bir fark bulunmaktadır ve karar ağacı modeli üstünlük kurmaktadır. Ancak sınıfların hassasiyet, duyarlılık ve F1 skoru değerlerine baktığımızda, 3 litoloji sınıfı hariç (breş, regolit ve granit) skorların 0 olduğu, yani modellerin bu sınıfları tanıma becerisi yansıtamadığı gözlenmiştir. Géron (2019), sınıflandırıcı modellerin değerlendirilmesinde özellikle dengesiz veri setlerinde doğruluk metriğinin yanıltıcı olabileceğini belirtmektedir [72]. Çünkü burada görülebildiği gibi sadece bir sınıf için yapılan tahminler modelin genel doğruluk skorunu şekillendirebilmektedir.

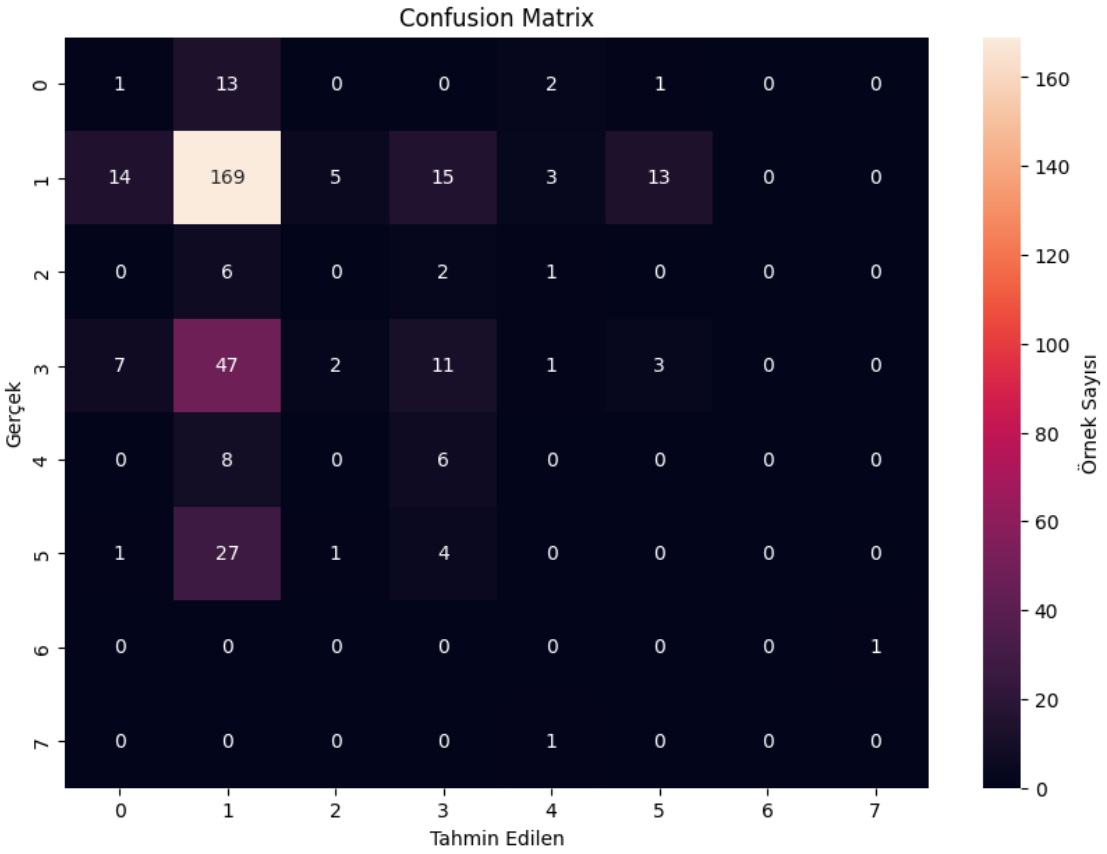
Çizelge 4.1. Sınıflandırma Raporu karşılaştırması (Python)

	Karar Ağacı			**Rastgele Orman**		
	Hassasiyet	Duyarlılık	F1 Skoru	Hassasiyet	Duyarlılık	F1 Skoru
Breş	0,04	0,06	0,05	0,06	0,06	0,06
Kil (Regolit)	0,63	0,77	0,69	0,62	0,75	0,68
Gnays	0,00	0,00	0,00	0,00	0,00	0,00
Granit	0,29	0,15	0,20	0,26	0,15	0,19
Milonit	0,00	0,00	0,00	0,07	0,07	0,07
Silkret	0,00	0,00	0,00	0,00	0,00	0,00
Silt	0,00	0,00	0,00	0,00	0,00	0,00
Toprak	0,00	0,00	0,00	0,00	0,00	0,00
Kappa Skoru			0,044			0,035
Doğruluk			0,4959			0,4850
Aritmetik Ortalama	0,12	0,12	0,12	0,12	0,12	0,12
Ağırlıklı Ortalama	0,43	0,50	0,46	0,42	0,48	0,45

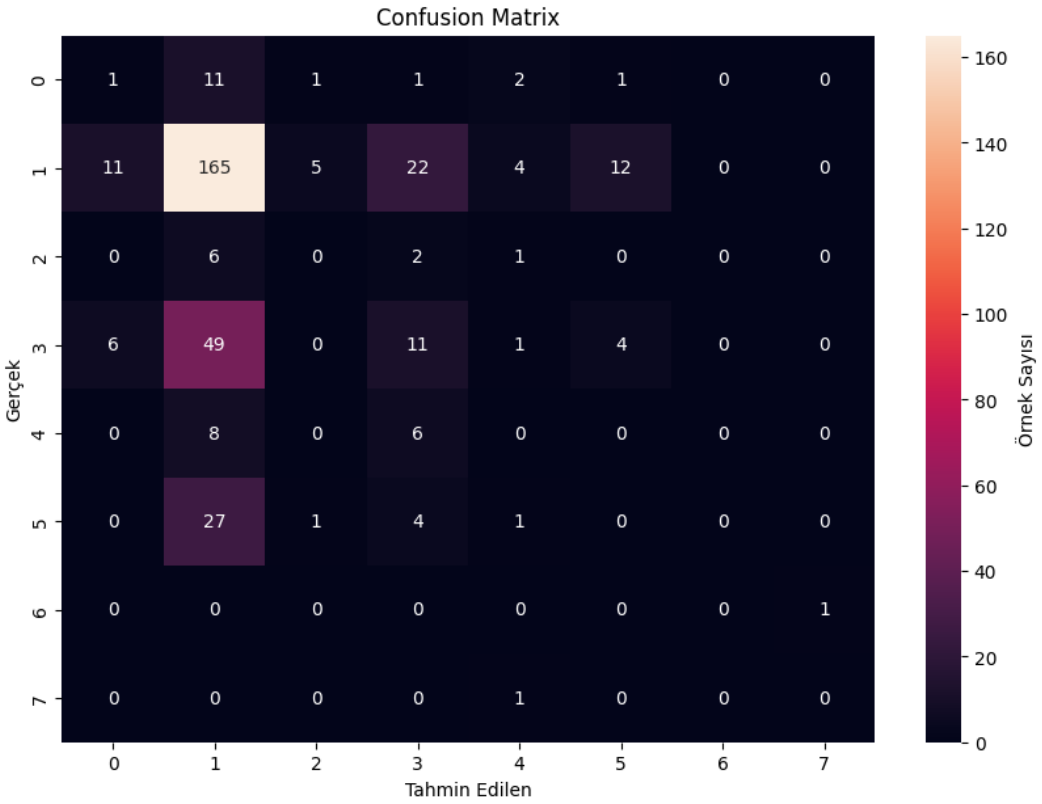
Tabloya göre hassasiyet, duyarlılık ve F1 skoru değerlerinde karar ağacı ve rastgele orman kıyaslaması göz önüne alındığında önemli bir fark görülmemektedir. Bu aritmetik ortalama değerlerinden daha net anlaşılmaktadır. Ağırlıklı ortalama için ise her sınıf için precision, recall ve F1 skoru hesaplanır ve bu değerler o sınıfın veri setindeki ağırlığı (veri adedi) ile çarpılır. Bu sebeple ağırlıklı ortalamanın sınıf dengesizliği durumunda daha doğru bir fikir verebilecek bir ölçüt olmasına karşın, bu ölçüt konusunda da iki farklı

modelde gözle görünür bir fark yoktur. İki model için de en belirgin özellik, sayıca üstün olan sınıfların ölçüt değerlerinin diğerlerine kıyasla çok daha yüksek olduğudur.

Sınıf dengesizliği, özellikle azınlık sınıfların örnek sayısının yetersiz olduğu durumlarda sınıflandırma modellerinin performansını ciddi şekilde etkileyebilir. Fernandez ve arkadaşlarının (2018) belirttiği üzere bir sınıflandırıcılar dengesiz veri setleri üzerinde sistematik olarak sadece baskın olan sınıfı atayarak %99 doğruluk elde edebilir. Bunun gibi durumlarda Cohen'in Kappa endeksi yardımcı olabilir.



Şekil 4.1. Karar Ağacı karmaşıklık matrisi (Python, scikitlearn kütüphanesi)



Şekil 4.2. Rastgele Orman karmaşıklık matrisi (Python, Scikitlearn kütüphanesi)

İki modelin karmaşıklık matrisleri de Şekil 4.1. ve Şekil 4.2.'de gösterilmiştir. Karmaşıklık matrislerinin satırları gerçek veriyi, sütunları ise tahmin edilmiş veriyi yansıtır[72]. Bu nedenle matrisin sol üst kenarından başlayan köşegen doğru tahminleri gösterir. Söz konusu iki modelin matrisleri incelendiğinde, sınıflandırma raporuna benzer şekilde, azınlık sınıfların doğru tahmin edilemediği görülmektedir. Karar ağacı ve rastgele orman modelleri arasında da değerlendirme yapılmasına imkân vermeyecek kadar az fark görülmektedir.

Landis ve arkadaşlarının belirttiği Kappa endeksi sınırları göz önüne alındığında, iki modelin de 0,2'den düşük olması dolayısıyla "düşük (slight)" başarılı olarak adlandırılması mümkündür [76]. Bu aynı zamanda iki modelin Kappa değerlerinin modeller arasında bir fark yaratamadığının da göstergesidir. Ancak Çizelge 4.1. genel olarak değerlendirildiğinde, karar ağacı modelinin biraz daha üstün olduğu söylenebilir. Bu fark rastgele orman algoritmasının karar ağacından farklı olarak bir birleşim algoritması olması ve rastgele özellik seçimi sebebiyle ortaya çıkmış olabilir.

Bu noktada kullanışlı olabilecek yöntemlerden biri McNemar testidir ve Çizelge 4.2.'de gösterildiği gibi olasılık tablosu oluşturulmuştur. Eğer 〖P(|McNemar istatistiği|>x〗 _(1;0,95)^2)<0.05, modellerin hata oranları arasında istatistiksel olarak anlamlı bir fark olduğu kabul edilir; aksi halde, gözlenen farkın şansa bağlı olduğu sonucuna varılır[62]. Değerlendirme sonucunda McNemar istatistik sonucu 0,450 ve p değeri 0,502 gelmiştir ve buna göre modeller arasındaki farkın şansa bağlı olduğu sonucuna varılabilir. Bu modeller arasında istatistiksel olarak anlamlı bir fark olmadığını gösterir.

Çizelge 4.2. McNemar olasılık çizelgesi (Python)

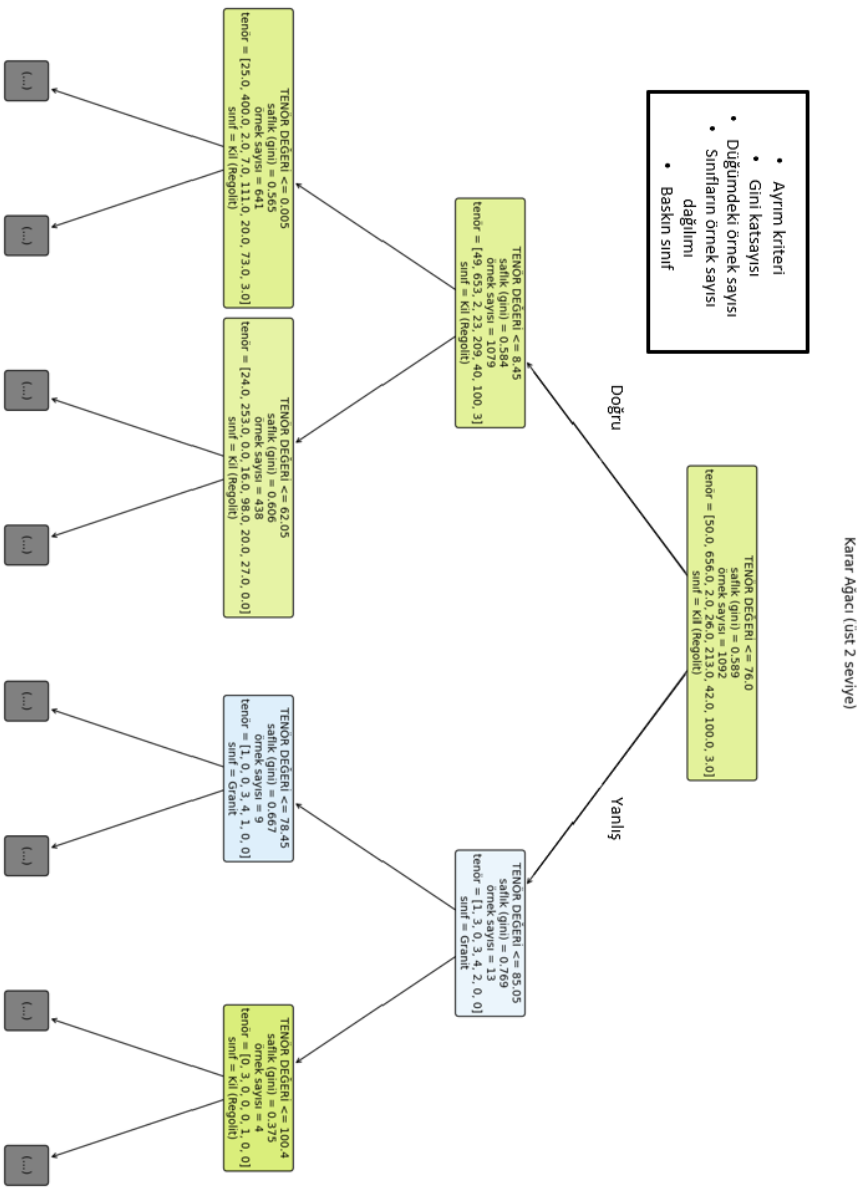
	Rastgele orman yanlış tahmin	**Rastgele orman doğru tahmin**
Karar ağacı yanlış tahmin	174	10
Karar ağacı doğru tahmin	13	168

4.2. Ağaç Görselleştirmesi

Karar ağacı modeli, hem karar ağacı algoritmasının bu sınıflandırma ölçeğinde nasıl davrandığını görmek hem de rastgele orman algoritmasının da nasıl çalıştığı konusunda fikir oluşturması amaçlanarak ağacın sadece ilk iki katını gösterecek şekilde Şekil 4.3.'te verilmiştir. Her bir kutu, bir düğümü simgeler. Düğümlerin içindeki satırlar sırasıyla ayrım koşulunu oluşturan özellik ve değerini, ayrım kriterini, ayrılmış (düğüme girdiği sıradaki) örnek sayısını, farklı sınıflara ait örnek sayısını dağılım olarak gösterimini (sırasıyla breş, regolit, kökeni tanımsız kil, gnays, granit, milonit, silkret, silt, toprak) ve söz konusu düğümde verilerin çoğunluğunun hangi sınıfa ait olduğunu gösterir.

Breiman'ın belirttiği üzere, kök düğümde yapılacak ayrım tüm aday bölünmeler arasından en yüksek safsızlık azalmasını (impurity decrease) sağlayacak şekilde seçilir [63]. Bu nedenle, karar ağacında kök düğümden başlayan ilk ayrımlar, modelin sınıfları ayırt etme yetisinin temelini oluşturur. Bu yüzden, bu görselleştirme karar ağacı algoritmasının tenör verisini kullanarak nasıl çalıştığı yönünde önemli ipuçları verebilir

ve rastgele orman modelinin çalışma prensibine de ışık tutabilir. Bu görsele göre de anlaşılacağı üzere, ayrımların çoğu Regolit (Sınıf 1) üzerinden yapılmıştır. Bu, veri setindeki dengesizliğin veya bu sınıfın daha belirgin özelliklere sahip olabileceğini düşündürür. Şekilde görüldüğü üzere ilk ayrım tenör değerinin 76'dan küçük ve eşit olma koşulunu sağlayan örnekler için sol düğüme, sağlamayan örnekler için ise sağ düğüme gidecek şekilde yapılmıştır. Sağ düğümde en çok örneği bulunan sınıf, kutucuğun en altında belirtildiği gibi Regolit sınıfıdır (653). Buna göre sağ daldaki örnek sayısının çok düşük olduğu, bu nedenle ayrımın güven vermediği görülmektedir.



Şekil 4.3. Karar Ağacı modelinin ilk 2 kat ağaç görselleştirmesi (Python, scikitlearn kütüphanesi)

4.3. Hiperparametre Optimizasyonu ve Son Model Değerlendirmesi

Aşırı örnekleme işlemi sonrası rastgele orman modelinin üzerinde hiperparametre optimizasyonu yapılmış ve optimum parametreler Çizelge 4.3.'te verilmiştir.

Çizelge 4.3. Hiperparametre Optimizasyonu en iyi model parametreleri (Python)

Rastgele Orman Parametreleri	Açıklama	Tercih
Bootstrap	Geri koymalı örnekleme kullanımı kararı	Kullanılmıştır
class_weight	Herhangi bir sınıfa ağırlık verilip verilmeyeceği	Dengeli (balanced)
criterion	Düğümlerin ayrım kriteri	Gini
max_features	En iyi ayrım aranırken göz önüne alınacak özellik sayısı	"sqrt": toplam özellik sayısının karekökü ile sınırlanmıştır
min_samples_leaf	Terminal düğüm olmak için gereken minimum örnek sayısı	4
min_samples_split	Bir düğümün ayrılması için gereken minimum örnek sayısı	2
n_estimators	Modeldeki ağaç sayısı	50
oob_score	Genelleştirme skorunu tahmin etmek için çanta dışı örneklerin kullanılıp kullanılmayacağı	Kullanılmadı
min_weight_fraction_leaf	Bir düğüm olmak için bütün girdi örneklerinin toplam ağırlıklarının minimum ağırlıklı oranı	Bütün örnekler aynı ağırlığa sahip olarak kabul edildi
warm_start	Model yeniden eğitilirken önceki durumu koruyup koruyamayacağını belirler	Her çağırıldığında model sıfırlandı

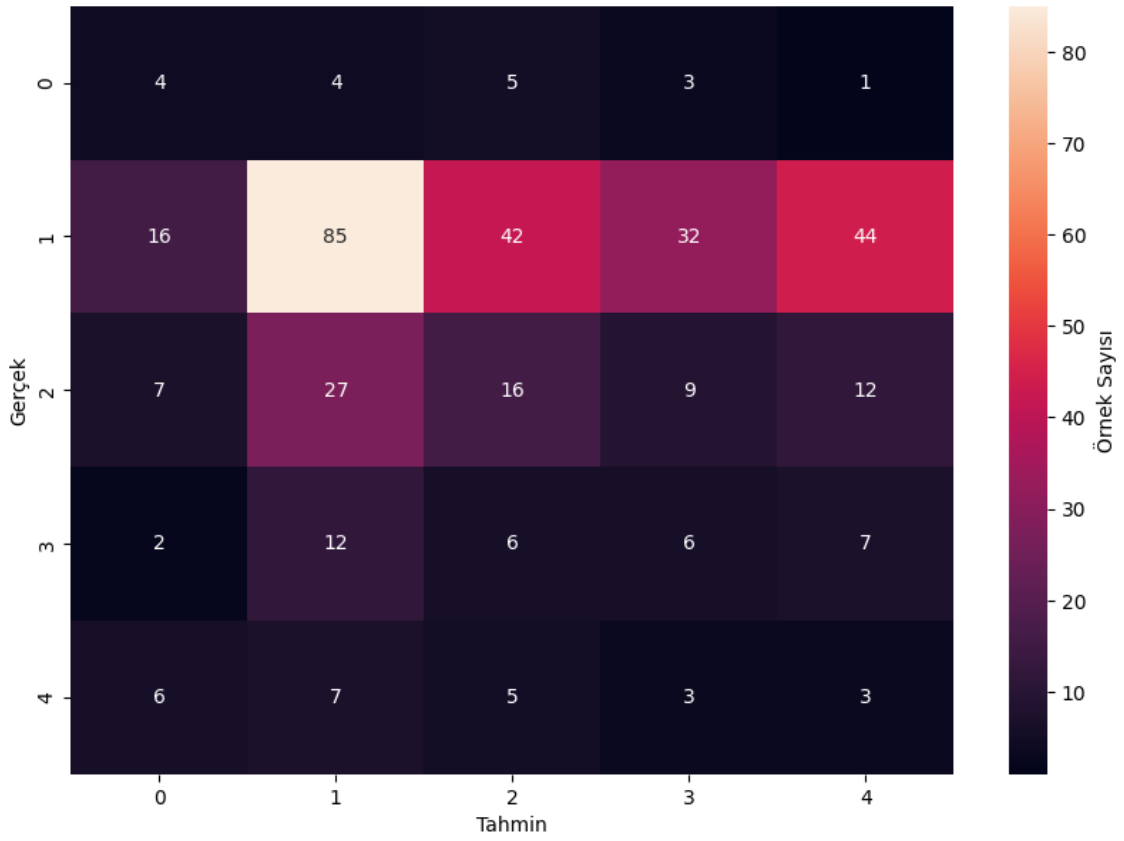
GridSearch Parametreleri		
Refit	En iyi parametrelerle modeli tekrar eğitmek	True: (varsayılan) kullanılmıştır

Grid Search Sonuçları	
Dengelenmiş doğruluk	0,5966
Geçerleme doğruluğu	0,3214

Bütün model iyileştirme denemelerinin ardından dengelenmiş doğruluk ve geçerleme doğruluğunun yanında Çizelge 4.4.'te gösterilen sınıflandırma raporu ve Şekil 4.4.'te gösterilen karmaşıklık matrisi de değerlendirilmiştir. Hiperparametre optimizasyonu, aşırı örnekleme, sınıf birleşimi ve veri standardizasyonu yapıldıktan sonraki "en iyi" modelin bütün sınıflarda tahmin yapabildiği görülmektedir. Fernandez ve arkadaşlarının (2018) belirttiği üzere, sınıf oranları her ne kadar dengesiz olsa da azınlık sınıfların yeterli sayıda örnekle temsil edilmesi durumunda model bu sınıflara ait örüntüleri daha sağlıklı şekilde öğrenebilir [61]. Sınıflandırma raporunda yine baskın sınıfın (regolit) belirgin şekilde diğer sınıflardan daha yüksek hassasiyet, duyarlılık ve F1 skoru çıkardığı gözlenmektedir. Ancak bu sınıfta özellikle duyarlılığın iyileştirme yapılmamış rastgele orman modeline göre (Çizelge 4.1.) çok düştüğü gözlemlenmektedir. Bu sınıfın sonuçları, diğer sınıfların duyarlılık değerlerinin yükselmesinden de anlaşılacağı gibi Regolit sınıfında yanlış pozitif olarak atanan örneklerin azaldığı, ancak doğru pozitiflerin gerektiği gibi artmadığı sonucunu getirebilir.

Çizelge 4.4. İyileştirilmiş model sınıflandırma raporu (Python)

	Hassasiyet	Duyarlılık	F1 Skor
Breş	0,13	0,24	0,17
Kil (Regolit)	0,64	0,40	0,49
Granit	0,20	0,21	0,21
Silkret	0,12	0,18	0,14
Diğer	0,06	0,17	0,09
Aritmetik Ortalama	0,23	0,24	0,22
Ağırlıklı Ortalama	0,45	0,32	0,36



Şekil 4.4. İyileştirilmiş Model Karmaşıklık Matrisi (Python, scikitlearn kütüphanesi)

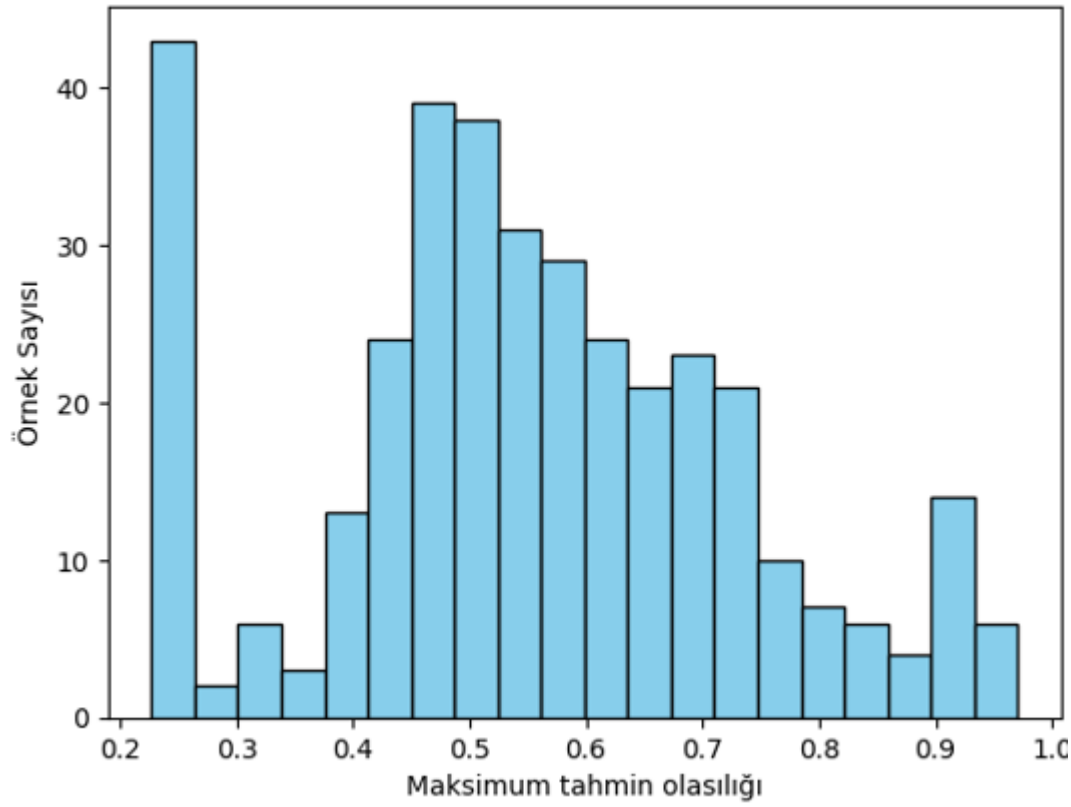
Karmaşıklık matrisi değerlendirildiğinde, sınıflandırma raporunun gösterdiği verilere benzer olarak, regolit sınıfının sınıflandırmalarında bir kayıp yok iken, nadir sınıfların doğru sınıflandırılmış örneklerinin oluşmaya başladığı görülebilir. Ancak özellikle nadir sınıfların hala belirgin bir şekilde yanlış sınıflandırılmış örnekleri olduğunu da söylemek mümkün.

Hiperparametre optimizasyonunun sonunda dengelenmiş doğruluk ve geçerleme doğruluğu değerleri elde edilmiş ve Çizelge 4.3.'te gösterilmiştir. Dengelenmiş doğruluk, doğru pozitif duyarlılık ve yanlış pozitif duyarlılıkların aritmetik ortalamasıdır. Eğer kullanılan veri seti dengeli ise, dengelenmiş doğruluk standart doğruluğa eşittir. Ancak dengesiz ve çarpık veri setlerinde doğruluk baskın sınıfın etkisinde kalarak yüksek görünebilir ve güvenilmez bir sonuç çıkarabilir. Bunun gibi durumlarda dengelenmiş doğruluk sınıfların başarı oranını ayrı ayrı ölçüp hesapladığı için bu etkiden korunacaktır ve daha "gerçek" bir sonuç olacaktır [61]. Ancak dengelenmiş doğruluğa göre geçerleme doğruluğu çok düşüktür (0,32). Bu modelin eğitim verisi ile iyi performans gösterdiğini,

ancak çapraz doğrulama metrikleri ile zayıf bir genelleme yaptığını göstermektedir ve bu durum aşırı uyum (overfitting) göstergesi olabilir [72]. Alkhawaldeh ve arkadaşları aşırı örnekleme yöntemlerinin yaygın bir şekilde kullanılmasına rağmen, yol açtıkları sınırlamalara dikkat etmenin önemli olduğunu vurgulamışlardır. Çünkü aşırı örnekleme azınlık sınıfları yeterince temsil etmeyen veriler üreterek aşırı uyuma yol açabilir ve gerçek dünya sorunlarına yanıltıcı sonuçlar üretebilir [77].

4.4. Güven Değerlendirmesi

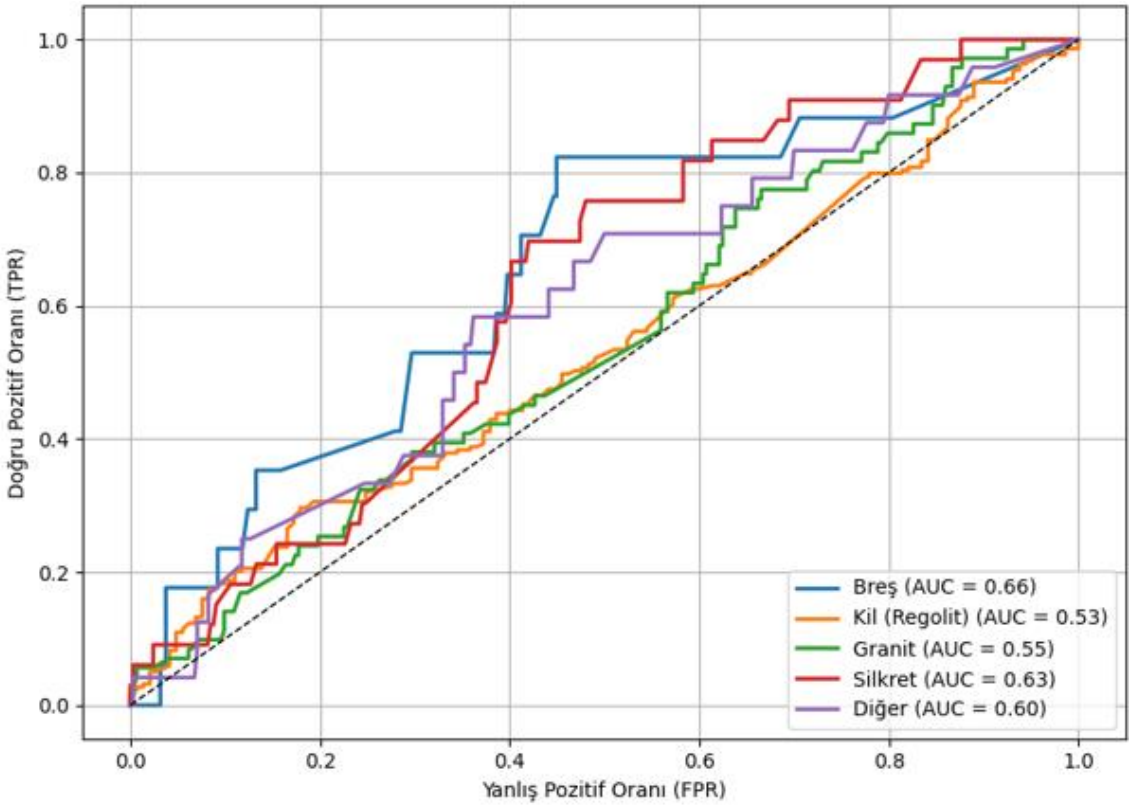
Scikit-learn kütüphanesinin predict_proba yöntemi ile üretilen tahmin güven dağılımı grafiği Şekil 4.5.'de gösterilmiştir. Her ne kadar karmaşıklık matrisi ve sınıflandırma raporu gibi ölçütler modelin tamamladığı tahminler üzerinden bir değerlendirme yapabilme imkânı sunsa da, modeller yapılan tahminlerin verdiği güven oranını da değerlendirir ve bu modelin yaptığı tahminde ne kadar kesinlik olduğunu gösterir [61]. Grafikte görüldüğü gibi, en yüksek tahmin olasılığı genellikle 0,5 civarında yoğunlaşmış ve 0,8-0,9 civarında olan yüksek güvenli tahminler 10'u geçmemektedir. Bu modelin yaptığı tahminlerde aslında ne kadar kararsız olduğunu da göz önüne sermektedir.



Şekil 4.5. Tahmin güven dağılımı grafiği (Python, Scikitlearn kütüphanesi, predict_proba)

4.5. ROC Eğrisi

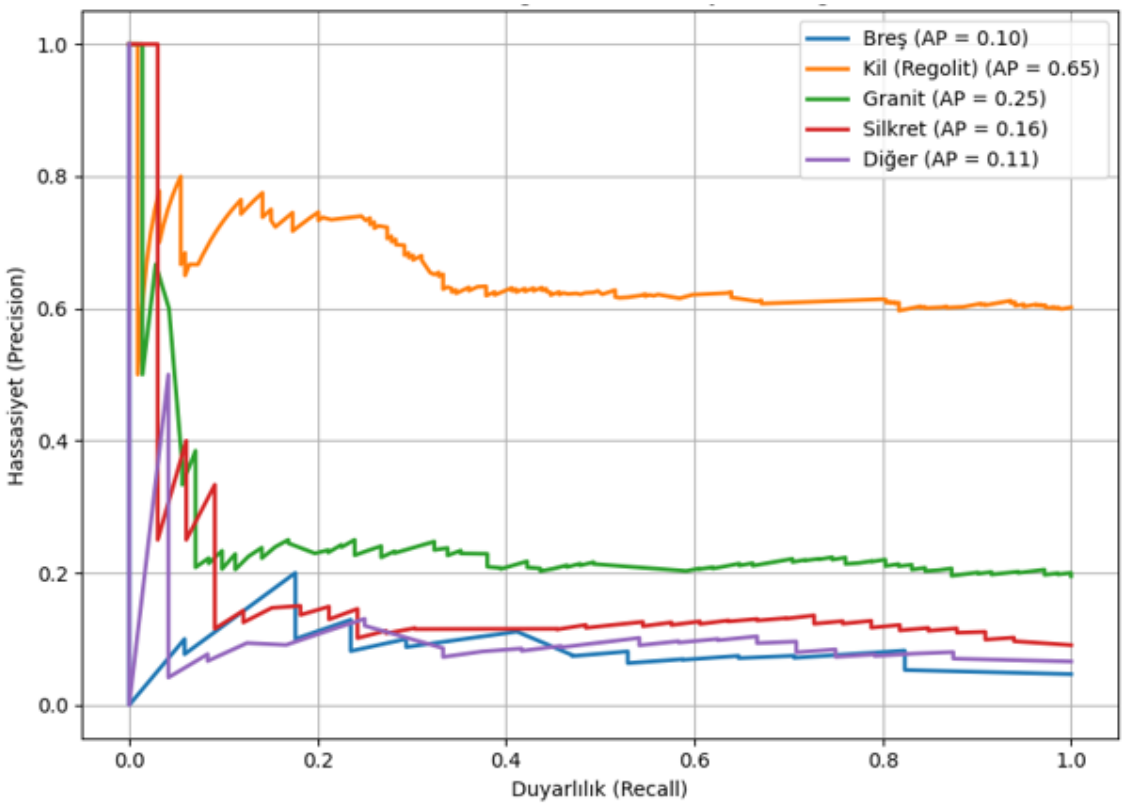
Eğri Şekil 4.6.'daki grafikte gösterilmiştir. Grafiğin ortasından geçen kesikli çizgi son derece rastgele bir sınıflandırıcıyı simgeler. Buna göre iyi bir sınıflandırıcının bu kesikli çizgiden olabildiğince uzakta ve tercihen sol üst köşeye dayanacak kadar yakın olması istenir. AUC skorlarına göre, Breş ve Silkret sınıfları diğer sınıflardan daha iyi bir performans gösteriyor. Kil (Regolit) ise en zayıf skor elde eden sınıf konumunda.



Şekil 4.6. Çok sınıflı ROC Eğrisi (Python, scikitlearn kütüphanesi)

4.6. Hassasiyet-Duyarlılık Eğrileri

Eğri Şekil 4.7.'deki grafikte gösterilmiştir. Regolit sınıfı burada nispeten daha iyi hassasiyet ve duyarlılık kombinasyonu sonucu vermektedir. Granit ve Silkret gibi sınıflarda ise duyarlılığı arttırmak hassasiyetin düşmesine sebep olmaktadır.

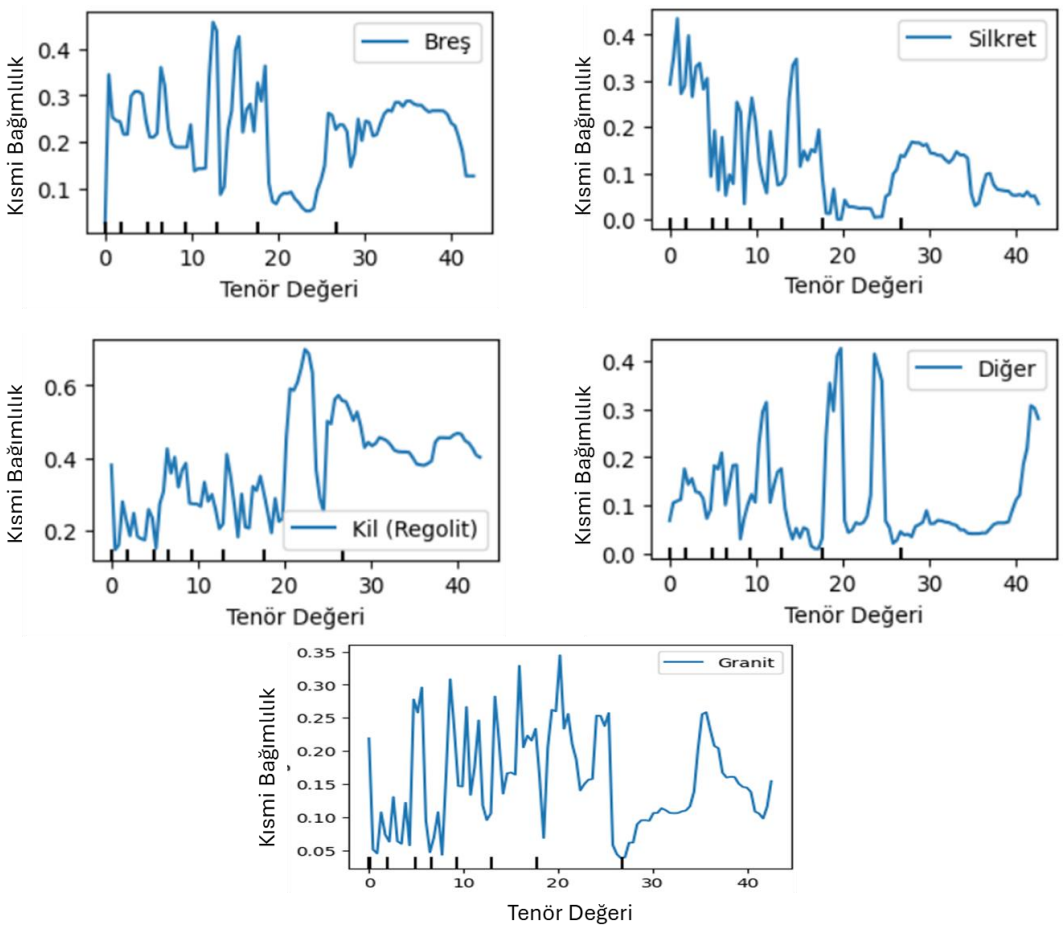


Şekil 4.7. Hassasiyet-Duyarlılık Grafiği (Python, scikitlearn kütüphanesi)

ROC eğrisi göz önünde bulundurulduğunda nadir sınıflar ile ilgili daha olumlu bir fikir edinmek mümkün olmakla birlikte, ROC eğrisi ve PR eğrileri arasında karar verilmesine yardımcı olabilecek bir yöntem mevcuttur. Pozitif sınıfın daha nadir olduğu durumlarda PR eğrileri tercih edilmelidir. Buna göre nadir sınıflar için daha doğru sonuç verdiği söylenebilir.

4.7. Kısmi Bağımlılık Grafikleri

Kısmi bağımlılık grafikleri Şekil 4.8.'de gösterilmiştir.

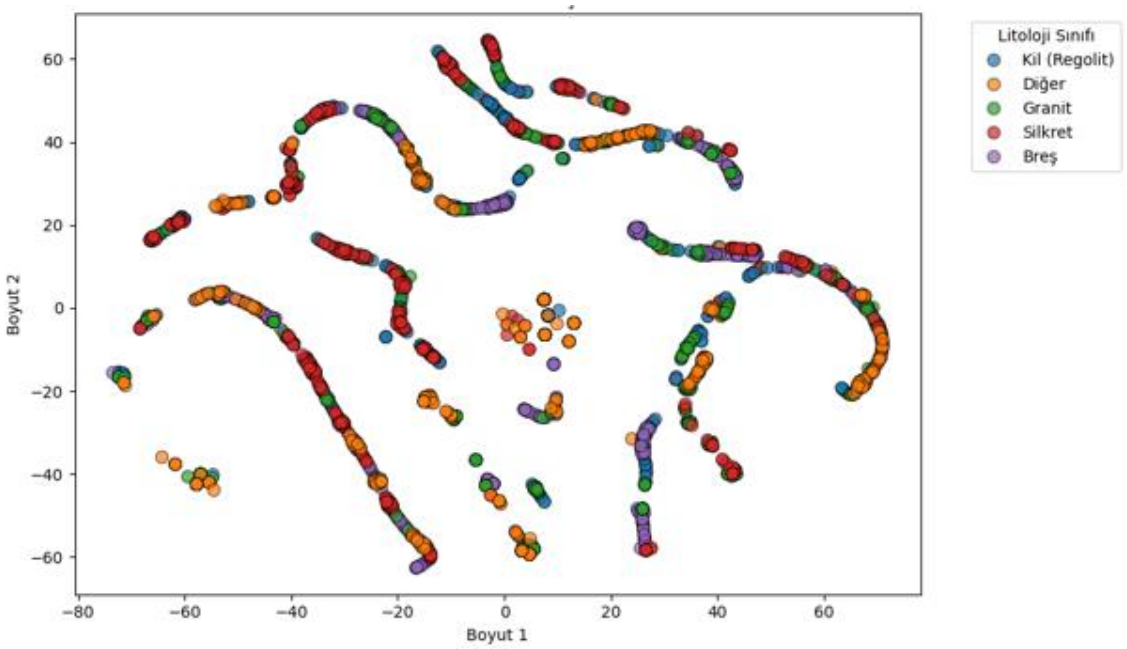


Şekil 4.8. Kısmi bağımlılık grafikleri

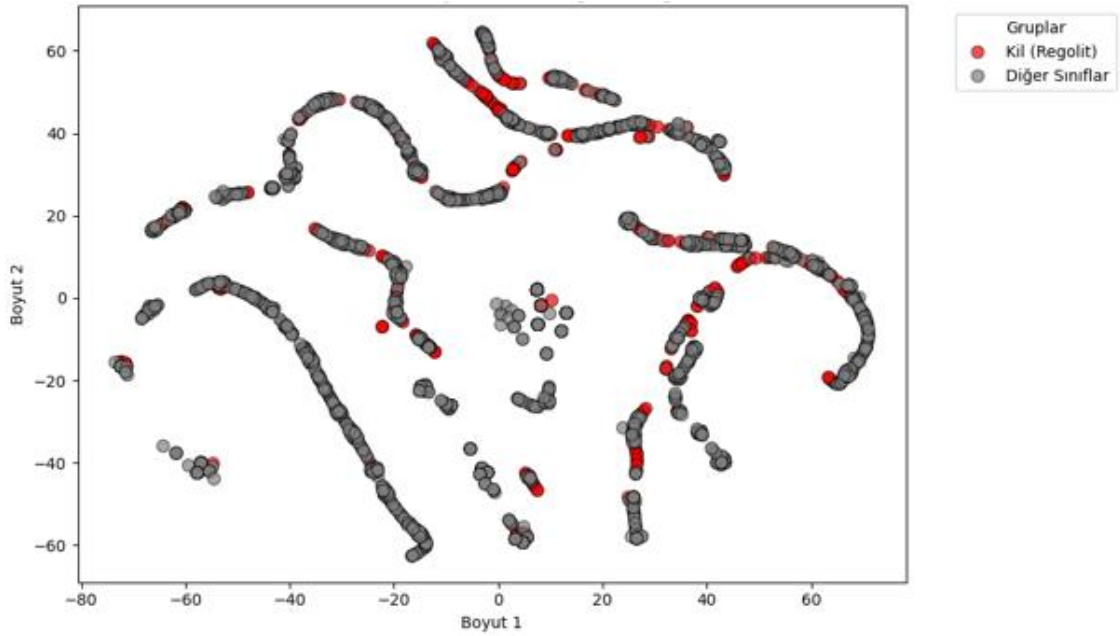
Belirli bir sınıfın belirli özellik için hangi değerler aldığını görüntülemek için her sınıf için ayrı bir grafik hazırlamak uygun bir yöntemdir [73]. Regolit ve granit başta olmak üzere sınıfların neredeyse hepsi için özellikle tenör 20 değerine gelene kadar modele çok güçlü bir şekilde etki edebildiği gözlenmiştir. Tenörün 20'den daha yüksek olması durumunda modelin daha çekimser davrandığı gözlenmiştir.

4.8. t-SNE Görselleştirmesi

T-SNE görselleştirmesi Şekil 4.9.'da gösterilmiştir. Görüldüğü gibi sınıflar aralarındaki kümeleşme çok dengesizdir. Belirli noktalarda regolit ve granit kümelenmeleri olsa da bu kümelenmeleri silkret ve breş doldurmaktadır. Diğer, silkret ve breş sınıflarının birbirlerinden çok uzak konumlarda da kümelendiği ve bu nedenle benzerliklerinin düşük olduğu da gözlenmiştir. Gözlemi kolaylaştırmak için regolit için bireysel görselleştirme de yapılmıştır. Şekil 4.10.'de verilen görsele göre regolitin diğer sınıflara kıyasla birbirine daha yakındır.



Şekil 4.9. t-SNE ile 2D Görselleştirme (Python, scikitlearn kütüphanesi)



Şekil 4.10. Regolit t-SNE görselleştirmesi (Python, scikitlearn kütüphanesi)

4.9. Yorumlar

Bu tez çalışmasında, tenör verisinin litoloji tahmini yapma potansiyelini değerlendirmek ve litoloji sınıflandırma ve keşif aşamalarının verimi artırmak amaçlanmıştır. Bu amaç doğrultusunda, veri kümesi üzerinde ön işleme, Karar Ağacı ve Rastgele Orman algoritması eğitimi, veri dengeleme pratikleri, hiperparametre optimizasyonu ve modellerin görsel değerlendirmesi gerçekleştirilmiştir.

Karar Ağacı ve Rastgele Orman modelleri karşılaştırıldığında Rastgele Orman modelinin beklenen üstünlüğü sağlayamadığı ve iki modelin herhangi bir hiperparametre ayarı, özellik mühendisliği veya aşırı örnekleme yapılmadan önce benzer başarı seviyeleri gösterdiği saptanmıştır. Rastgele Orman'ın beklendiği kadar yüksek skorlar elde edememesinin olası nedenlerinden biri veri dengesizliğinin etkisidir. Bu bulgu, tenör verisi ile litoloji tayini yapılması araştırmalarında veri dengesizliğinin Rastgele Orman algoritmasının beklenen avantajını sınırlayabileceğini göstermektedir.

İyileştirme aşamalarının ardından elde edilen "en iyi" modelin tenör değeri bilgisini kullanarak litolojik kimlik tayin etmek konusunda güvenilir olmadığı görülebilir. Ancak çalışmanın sonuçlar ileri aşamalar için fikir verme kapasitesine sahiptir. Modelin dengelenmiş doğruluğunun (balanced accuracy) söz konusu iyileştirme aşamalarının ardından %59.66'ya ulaşması veri setinin dengelenmesinin tahminlerin eğitim setini iyi öğrendiği ve isabet oranında fayda sağladığını gösteriyor. Bununla birlikte, geçerleme doğruluğunun (validation accuracy) % 32.14 ile dengelenmiş doğruluktan oldukça düşük kalması, modelin aşırı öğrenmeye eğilimini göstermektedir. SMOTE ve hiperparametre optimizasyonu sonrasında modelin nadir sınıfları öğrenme potansiyelinin oluşması, hiperparametrelerin farklı kombinasyonlarının ve olası özellik mühendisliğinin modeli geliştirmeye açık olduğunu gösteriyor. Sadece Au tenör değerlerinin girdi değişkeni olarak kullanıldığı senaryonun benzer sonuçlar verdiği (0,36 geçerleme doğruluğu ve 0,59 dengelenmiş doğruluk) gözlenmiştir.

Modelin tenör değeri 20'den düşük olduğu örneklerde ayırt etme kapasitesinin yüksek olduğu gözlenmiştir. Bu nedenle, tenör değeri 20'den yüksek olan örnekler için mineraloji parametreleri gibi ek özellik kullanılması gelecek araştırmalar ve operasyonel

denemeler için faydalı kabul edilebilirken tenörün 20'den düşük olduğu örnekler için süreç hızlanabilir.

Granit, yeterli örnek sayısı bulunmamasına rağmen kıyaslamalı olarak tutarlı bir F1 skoru ve PR eğrilerinde diğer sınıflara göre yüksek performans ve PDP grafiklerinde tenör 25 değerine kadar duyarlı bir hareket göstermektedir. Bu nedenle, diğer sınıflara kıyasla Granit sınıfının veri miktarının dengesizliğin önüne geçerek bir tahmin yapmaya uygun olduğu söylenebilir.

Genel olarak bu çalışmada, tenör değeri özniteliğinin kullanımları ile litoloji tahmini yapmanın zorlukları ve fırsatlarına ışık tutuldu. Sonuçlara göre Rastgele Orman beklenen avantajını tam olarak sergileyemediğini, SMOTE ve hiperparametre optimizasyonunun ise eğitim verisi üzerinde sınıflandırma gücünü arttırmalarına rağmen gerçek doğrulamada kayba yol açtığını gösterdi. Öte yandan modelin tenör 20 değerinden düşük aralıkta iken ayırt ediciliğinin yüksek, 20'den yüksek iken ayırt ediciliğinin düşük olduğunu ve ek parametrelere gereksinim duyduğunu gösterdi. Ayrıca, Granit sınıfının veri miktarından bağımsız olarak tahmin etme potansiyelinin varlığını yansıttı. Ek olarak, güven dağılım grafiklerinin belirttiği üzere modelin farklı davranış sergileme ihtimali mevcuttur. Bütün bu bulguların ışığında, söz konusu Rastgele Orman modelinin kullanılmasının yüksek güven oranı ile önerilmesi mümkün değildir.

Bundan sonraki adımda,

1. Aşırı örneklemenin etkilerini test etmek adına aşırı örnekleme öncesi hiperparametre optimizasyonu yapılarak modellerin test edilmesi,
2. Girdi parametresi olarak yeni bir litolojik özniteliğin ek olarak kullanımı (kayaç grubu, tabaka ismi gibi),
3. Hiperparametre optimizasyonu sırasında farklı parametre senaryolarının denenmesi,
4. Farklı yöntemlerinin denenmesi ve
5. Nadir sınıflardan daha fazla veri elde edilerek veya başka bir veri seti ile araştırmanın dengeli veri ile nasıl davrandığının gözlenmesi

önerilmektedir. Bu sayede çalışmaya olan güven artacak ve potansiyel daha kullanışlı modeller ortaya çıkacaktır.

6. KAYNAKLAR

[1] M. Allaby, Ailsa;Allaby, *A dictionary of earth sciences*. Oxford; New York; Oxford University Press, **1999**.

[2] N. J. Hyne, *Dictionary of Petroleum Exploration, Drilling & Production*. PennWell Publishing Company, **1991**.

[3] J. R. Fanchi, *Shared Earth Modeling*, vol. i. Gulf Professional Publishing, **2002**.

[4] M. Kanevski, A. Pozdnoukhov, and V. Timonin, *Machine learning for spatial environmental data: Theory, applications and software*. **2009**.

[5] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, **2001**.

[6] T. M. Mitchell, *Machine Learning*, vol. 9, no. 3. **1997**.

[7] J. S. Dramsch, "70 Years of Machine Learning in Geoscience in Review," *Advances in Geophysics*, vol. 61. pp. 1–55, **2020**.

[8] Y. Liu, "Machine Learning in Geology : Challenges and Prospects," *Highlights Sci. Eng. Technol.*, vol. 44, pp. 14–21, **2023**.

[9] M. Y. Khan, A. Qayoom, M. S. Nizami, M. S. Siddiqui, S. Wasi, and S. M. K. U. R. Raazi, "Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques," *Complexity*, vol. 2021, **2021**.

[10] I. Kononenko, M. Kukar, *Machine learning and data mining: introduction to principles and algorithms*, vol. 45, no. 07. **2008**.

[11] P. Josso, A. Hall, C. Williams, T. Le Bas, P. Lusty, and B. Murton, "Application of random-forest machine learning algorithm for mineral predictive mapping of Fe-Mn crusts in the World Ocean," *Ore Geol. Rev.*, vol. 162, p. 105671, **2023**.

[12] T. G. Dietterich, "Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, **2000**.

[13] V. F. Rodriguez-Galiano, M. Chica-Olmo, and M. Chica-Rivas, "Predictive modelling of gold potential with the integration of multisource information based

on random forest: a case study on the Rodalquilar area, Southern Spain," *Int. J. Geogr. Inf. Sci.*, vol. 28, no. 7, pp. 1336–1354, **2014**.

[14] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, **2020**.

[15] F. W. Preston and J. Henderson, "Fourier series characterization of cyclic sediments for stratigraphic correlation." **1964**.

[16] W. Schwarzacher, "The Semi-Markov Process as a General Sedimentation Model," pp. 247–268, **1972**.

[17] G. Matheron, "Splines and kriging: their formal equivalence," *Syracuse Univ. Geol. Contrib.*, vol. 8. pp. 77–95, **1981**.

[18] F. U. Dowla, S. R. Taylor, and R. W. Anderson, "Seismic discrimination with artificial neural networks: preliminary results with regional spectral data," *Bull. - Seismol. Soc. Am.*, vol. 80, no. 5, pp. 1346–1373, **1990**.

[19] G. Roth and A. Tarantola, "Neural networks and inversion of seismic data," *J. Geophys. Res. Solid Earth*, vol. 99, no. B4, pp. 6753–6768, **1994**.

[20] X. T. Feng and M. Seto, "Neural network dynamic modelling of rock microfracturing sequences under triaxial compressive stress conditions," *Tectonophysics*, vol. 292, no. 3–4, pp. 293–309, **1998**.

[21] D. Benaouda, G. Wadge, R. B. Whitmarsh, R. G. Rothwell, and C. MacLeod, "Inferring the lithology of borehole rocks by applying neural network classifiers to downhole logs: An example from the Ocean Drilling Program," *Geophys. J. Int.*, vol. 136, no. 2, pp. 477–491, **1999**.

[22] L. Hermes, D. Frieauff, J. Puzicha, and J. M. Buhmann, "Support vector machines for land usage classification in landsat TM imagery," *Int. Geosci. Remote Sens. Symp.*, vol. 1, no. February 1999, pp. 348–350, **1999**.

[23] C. E. Brodley and M. A. Friedl, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 399–409, **1997**.

[24] Y. Zhai, J. A. Thomasson, J. E. Boggess, and R. Sui, "Soil texture classification with artificial neural networks operating on remote sensing data," *Comput. Electron. Agric.*, vol. 54, no. 2, pp. 53–68, **2006**.

[25] K. Wang and L. Zhang, “Predicting formation lithology from log data by using a neural network,” *Pet. Sci.*, vol. 5, no. 3, pp. 242–246, **2008**.

[26] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, “An assessment of the effectiveness of a random forest classifier for land-cover classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 67, no. 1, pp. 93–104, **2012**.

[27] M. J. Cracknell and A. M. Reading, “Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information,” *Comput. Geosci.*, vol. 63, pp. 22–33, **2014**.

[28] V. Rodriguez-galiano, M. Sanchez-castillo, M. Chica-olmo, and M. Chica-rivas, “Machine learning predictive models for mineral prospectivity : An evaluation of neural networks , random forest , regression trees and support vector machines,” *Ore Geol. Rev.*, vol. 71, pp. 804–818, **2015**.

[29] J. R. Harris and E. C. Grunsky, “Predictive lithological mapping of Canada’s North using Random Forest classification applied to geophysical and geochemical data,” *Comput. Geosci.*, vol. 80, pp. 9–25, **2015**.

[30] E. J. M. Carranza and A. G. Laborte, “Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm,” *Ore Geol. Rev.*, vol. 71, pp. 777–787, **2015**.

[31] A. Roslin and J. S. Esterle, “Electrofacies analysis for coal lithotype profiling based on high-resolution wireline log data,” *Comput. Geosci.*, vol. 91, pp. 1–10, **2016**.

[32] S. Bhattacharya, T. R. Carr, and M. Pal, “Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA,” *J. Nat. Gas Sci. Eng.*, vol. 33, pp. 1119–1133, **2016**.

[33] S. Ghosh, R. Chatterjee, and P. Shanker, “Estimation of ash, moisture content and detection of coal lithofacies from well logs using regression and artificial neural network modelling,” *Fuel*, vol. 177, pp. 279–287, **2016**.

[34] S. M. Lundberg and S. I. Lee, “A Unified Approach to Interpreting Model

Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 4766–4775, **2017**.

[35] M. Shabankareh and A. Hezarkhani, "Journal of African Earth Sciences Application of support vector machines for copper potential mapping in Kerman region , Iran," *J. African Earth Sci.*, vol. 128, pp. 116–126, **2017**.

[36] M. Blouin, A. Caté, L. Perozzi, and E. Gloaguen, "Automated facies prediction in drillholes using machine learning," *79th EAGE Conf. Exhib. 2017 - Work.*, no. June, **2017**.

[37] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, and P. A. Johnson, "Machine Learning Predicts Laboratory Earthquakes," *Geophys. Res. Lett.*, vol. 44, no. 18, pp. 9276–9282, **2017**.

[38] P. Bestagini, V. Lipari, and S. Tubaro, "A Machine Learning Approach to Facies Classification Using Well Logs," *SEG Tech. Progr. Expand. Abstr.*, no. August 2017, pp. 2137–2142, **2017**.

[39] S. Kuhn, M. J. Cracknell, and A. M. Reading, "Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia," *Geophysics*, vol. 83, no. 4, pp. B183–B193, **2018**.

[40] Y. Ao, L. Zhu, S. Guo, and Z. Yang, "Probabilistic logging lithology characterization with random forest probability estimation," *Comput. Geosci.*, vol. 144, p. 104556, **2020**.

[41] T. S. Bressan, M. Kehl de Souza, T. J. Girelli, and F. C. Junior, "Evaluation of machine learning methods for lithology classification using geophysical data," *Comput. Geosci.*, vol. 139, no. October 2019, p. 104475, **2020**.

[42] R. Zhong, R. Johnson, and Z. Chen, "Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost)," *Int. J. Coal Geol.*, vol. 220, no. July 2019, p. 103416, **2020**.

[43] Z. Xu, H. Shi, P. Lin, and T. Liu, "Integrated lithology identification based on images and elemental data from rocks," *J. Pet. Sci. Eng.*, vol. 205, p. 108853, Oct. **2021**.

[44] K. W. S. B. A. S. Hary Nugroho and K. W. S. B. A. S. Hary Nugroho, "Lithological

Boundaries Identification in Dense Vegetation Area Based on Satellite Data Using Rare Training Data," *J. Hunan Univ. Nat. Sci.*, vol. 48, no. 10, **2021**.

[45] T. Martin, R. Meyer, and Z. Jobe, "Centimeter-Scale Lithology and Facies Prediction in Cored Wells Using Machine Learning," *Front. Earth Sci.*, vol. 9, p. 491, **2021**.

[46] Y. Xie, C. Zhu, R. Hu, and Z. Zhu, "A Coarse-to-Fine Approach for Intelligent Logging Lithology Identification with Extremely Randomized Trees," *Math. Geosci.*, vol. 53, no. 5, pp. 859–876, Jul. **2021**.

[47] T. Merembayev, D. Kurmangaliyev, B. Bekbauov, and Y. Amanbek, "A Comparison of Machine Learning Algorithms in Predicting Lithofacies: Case Studies from Norway and Kazakhstan," *Energies*, vol. 14, no. 7, pp. 1–16, **2021**.

[48] H. Gamal, S. Elkatatny, A. Alsaihati, and A. Abdulraheem, "Intelligent Prediction for Rock Porosity while Drilling Complex Lithology in Real Time," *Comput. Intell. Neurosci.*, vol. 2021, **2021**.

[49] G. Albert and S. Ammar, "Application of random forest classification and remotely sensed data in geological mapping on the Jebel Meloussi area (Tunisia)," *Arab. J. Geosci.*, vol. 14, no. 21, pp. 1–13, **2021**.

[50] T. Kumar, N. K. Seelam, and G. S. Rao, "Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India," *J. Appl. Geophys.*, vol. 199, p. 104605, **2022**.

[51] "South Australian Resources Information Gateway," *Government of South Australia*. [Online]. Available: https://map.sarig.sa.gov.au/.

[52] M. del C. Gutiérrez-Castorena and W. R. Effland, "Pedogenic and Biogenic Siliceous Features," in *Interpretation of Micromorphological Features of Soils and Regoliths*, **2010**.

[53] and R. L. S. W. W. Olive, A.F. Chleborad, C.W. Frahme, Julius Schlocker, R.R. Schneider, "Swelling clays map of the conterminous United States," **1989**.

[54] L. P. Keller and D. S. McKay, "The nature and origin of rims on lunar soil grains," *Geochim. Cosmochim. Acta*, vol. 61, no. 11, pp. 2331–2341, **1997**.

[55] P. Mambwe, R. Swennen, J. Cailteux, C. Mumba, S. Dewaele, and P. Muchez,

"Review of the origin of breccias and their resource potential in the central Africa Copperbelt," *Ore Geology Reviews*, vol. 156. Elsevier B.V., **2023**.

[56] U. of A. School of Environment, "Gneiss. Rocks and Minerals," **2005**. [Online]. Available: https://rocksminerals.flexiblelearning.auckland.ac.nz/rocks/gneiss.html.

[57] *Introduction to Mineralogy and Petrology*. **2020**.

[58] G. Sposito, "soil," *(**2025**, March 19). Encyclopedia Britannica*. .

[59] Y. Sasaki, "The truth of the F-measure," pp. 1–5, **2007**.

[60] D. M. W. Powers and N. L. Processing, "What the F-measure doesn ' t measure …."

[61] A. Fernández, S. García, M. Galar, and R. C. Prati, *Learning from Imbalanced Data Sets*. .

[62] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications, 2nd Edition*, vol. 81. **2014**.

[63] A. L. S. Chemex *et al.*, *Classification and Regression Trees by Leo Breiman*, no. January. **1999**.

[64] J. A. S. Sá, A. C. Almeida, B. R. P. Rocha, M. A. S. Mota, J. R. S. Souza, and L. M. Dentel, "Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy," **2016**.

[65] "Machine Learning For Beginners Algorithms," *Decis. Tree Random For. Introd.*, pp. 1–198, **2017**.

[66] A. Yang, Li , Sami, "On hyperparameter optimization of machine learning algorithms," *Neurocomputing*, **2019**.

[67] S. Citation *et al.*, "Hyperparameter Optimization in Machine Learning," **2024**.

[68] scikit-learn, "balanced_accuracy_score." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html.

[69] S. Guido, *Introduction to Machine Learning with Python*. .

[70] S. Raschka, *Python Machine Learning*. **2015**.

[71] J. T. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, “An Introduction to Statistical Learning with Applications in Python,” **2023**.

[72] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow_ Concepts, Tools, and Techniques to Build Intelligent Systems*. O’Reilly, **2019**.

[73] Christoph Molnar, “Interpretable Machine Learning, Christoph Molnar,” *https://christophm.github.io/interpretable-ml-book/*, **2022**. .

[74] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” **2009**.

[75] Deepchecks, “Decision Boundary.” [Online]. Available: https://www.deepchecks.com/glossary/decision-boundary/.

[76] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, **1977**.

[77] I. M. Alkhawaldeh, I. Albalkhi, and A. J. Naswhan, “Challenges and limitations of synthetic minority oversampling techniques in machine learning,” *World J. Methodol.*, vol. 13, no. 5, **2023**.