Savunma Bilimleri Dergisi, 21(2): 291-308 (2025)

SAVUNMA BİLİMLERİ DERGİSİ
The Journal of Defense Sciences

https://dergipark.org.tr/tr/pub/khosbd

Denetimli Makine Öğrenmesi Teknikleri ile Anomali Tespiti: Shuttle Uzay Verisi Örneği

Anomaly Detection with Supervised Machine Learning Techniques: The Case of Shuttle Space Data

Ünal DİKBAŞ [1] Meral EBEGİL [2*]

[1]*Gazi Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Ana Bilim Dalı, Ankara, Türkiye*

[2]*Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara, Türkiye*

Makale Bilgisi

Araştırma makalesi
Başvuru: 26.08.2025
Düzeltme: 15.09.2025
Kabul: 24.09.2025

Anahtar Kelimeler

Makine öğrenmesi
Denetimli öğrenme
Anomali tespiti

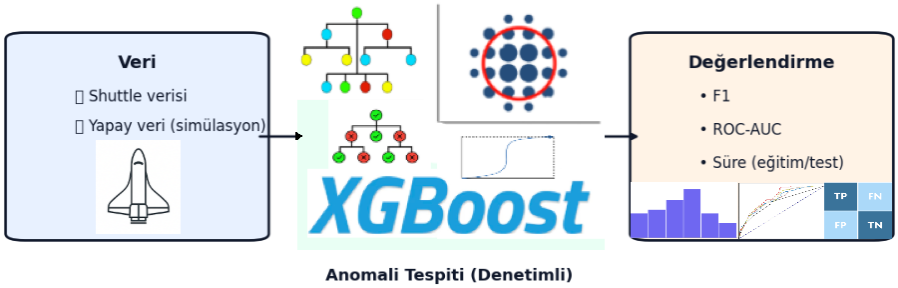
Keywords

Machine learning
Supervised learning
Anomaly detection

Önemli Noktalar

Anomali tespiti, uzay görevlerinde güvenlik, süreklilik ve maliyet etkinliği açısından kritik bir rol oynamaktadır. Shuttle uzay verisi üzerinde gerçekleştirilen bu çalışma, uçuş sırasında oluşabilecek olağandışı durumların erken belirlenmesine katkı sağlayarak görevlerin güvenli şekilde tamamlanmasına yardımcı olmaktadır. Uzay görevlerinin karmaşık ve yüksek riskli yapısı göz önüne alındığında, makine öğrenmesi tabanlı yaklaşımların anomali tespitinde sunduğu imkânlar, hem teknik hem de operasyonel açıdan önemli bir yenilik ve değer yaratmaktadır.

Grafiksel Özet



Özet

Bu çalışmada, anomali tespiti problemine yönelik Lojistik Regresyon, k-En Yakın Komşu, Karar Ağacı, Rastgele Orman ve XGBoost olmak üzere beş farklı denetimli makine öğrenmesi algoritmasının performansları karşılaştırmalı olarak analiz edilmiştir. Hem gerçek dünya verisi hem de yapay olarak oluşturulan simülasyon verisi kullanılarak modellerin F1 skoru, ROC-AUC değeri, eğitim ve test süreleri gibi ölçütler üzerinden değerlendirmeleri yapılmıştır. Gerçek veri üzerinde yapılan analizde sınıf dengesizliği göz önünde bulundurulmasına rağmen bazı modellerin yüksek başarı sağladığı görülmüştür. Simülasyon verileri ise modellerin anomali yapısını öğrenmedeki başarısını daha nesnel şekilde test etme imkânı sunmuştur. Bulgular, özellikle k-En Yakın Komşu, Karar Ağacı ve Rastgele Orman modellerinin yüksek doğrulukla anomalileri tespit ettiğini göstermektedir. XGBoost modelinin ise bazı durumlarda düşük anomali oranını yeterince ayırt edemediği gözlemlenmiştir. Bu kapsamda, parametre optimizasyonunun model başarısında kritik rol oynadığı vurgulanmıştır. Elde edilen bulgular, uzay ve savunma gibi güvenlik odaklı alanlarda karar destek sistemlerinin güçlendirilmesi için önemli görülmektedir.

Abstract

In this study, the performance of five different supervised machine learning algorithms—Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, and XGBoost—was comparatively analyzed for the anomaly detection problem. Both real-world data and artificially generated simulation data were used to evaluate the models based on metrics such as F1-score, ROC-AUC, and training and testing times. Despite considering class imbalance in the real-world dataset, some models achieved high performance. The simulation data provided a more objective means of testing the models' ability to learn anomaly structures. The findings revealed that, in particular, k-Nearest Neighbors, Decision Tree, and Random Forest models were able to detect anomalies with high accuracy, while the XGBoost model, in some cases, struggled to sufficiently distinguish low anomaly rates. In this context, parameter optimization was emphasized as a critical factor in model performance. The findings obtained in this study are considered important for strengthening decision support systems in security-oriented domains such as space and defense.

*Corresponding author, e-mail: mdemirel@gazi.edu.tr

DOI:10.17134/khosbd.1772308

Elektronik/Online ISSN: 2148-1776
Basılı/Printed ISSN: 1303-6831

1. GİRİŞ (INTRODUCTION)

Anomali tespiti, bir veri kümesinde çoğunluktan önemli ölçüde farklılık gösteren sıra dışı gözlemlerin tanımlanmasını amaçlayan bir süreçtir. Finans, sağlık, güvenlik ve uzay araştırmaları gibi birçok alanda kritik öneme sahiptir. Anomalilerin belirlenmesi özellikle savunma ve uzay bilimleri gibi stratejik alanlarda, normal operasyonlar dışındaki beklenmedik ve olağandışı olayların erken tespitinde büyük avantajlar sağlamaktadır. Bununla birlikte, sistem güvenliği, maliyet tasarrufu ve operasyonel verimlilik açısından da önemli görülmektedir.

Günümüzde artan veri hacmi ve karmaşıklığı, geleneksel istatistiksel yöntemlerle anomalilerin tespitini zorlaştırmakta, bu nedenle makine öğrenmesi teknikleri giderek daha fazla kullanılmaktadır. Makine öğrenmesi, büyük veri setlerinden örüntü çıkarma ve sınıflandırma yetenekleri sayesinde, hem denetimli hem de denetimsiz öğrenme modelleri aracılığıyla anomali tespiti süreçlerinde yüksek başarılar elde etmektedir.

Anomali tespiti, literatürde geniş bir uygulama alanına sahip olup finansal dolandırıcılığın önlenmesinden, siber güvenliğe, üretim hatalarının izlenmesinden uzay görevlerine kadar birçok alanda kullanılmaktadır. Bu kapsamda, hem denetimli hem de denetimsiz öğrenme yöntemleri literatürde sıklıkla karşılaştırmalı olarak ele alınmıştır.

Örneğin Schwabacher vd. (2009) dört farklı denetimsiz anomali tespit algoritmasını iki ayrı roket motoru test ortamına ait veriler üzerinde uygulamıştır. Kullanılan algoritmalar farklı anomalilik tanımlarına dayanmakta olup; komşuluk, kümeleme ve kural tabanlı algoritmalar ile tek sınıf destek vektör makineleri (One-Class Support Vector Machine-OCSVM) yöntemi kullanılarak tahminler yapılmıştır. Çalışmada tespit edilen dokuz anomali, bu algoritmaların birbirinden farklı anomalileri saptadığını göstermiştir. Bu durum, tek bir algoritma yerine çoklu algoritma kullanımının daha etkili sonuçlar verebileceğini ortaya koymaktadır [1]. Goldstein ve Uchida (2016), çok değişkenli veri setlerinde denetimsiz anomali tespiti algoritmalarının performanslarını kapsamlı bir şekilde karşılaştırmış ve yüksek boyutlu verilerde bazı yöntemlerin daha başarılı sonuçlar verdiğini göstermiştir [2].

Literatürde yer alan çalışmalardan biri de, denetimli ve denetimsiz makine öğrenmesi yöntemlerinin anomali tespiti açısından karşılaştırılmasına odaklanmıştır. Çalışmada, OCSVM, yerel aykırı değer faktörü (Local Outlier Factor-LOF), izolasyon ormanı (Isolation Forest-IF) ve eliptic zarflama (Elliptic Envelope-EE) gibi yaygın denetimsiz öğrenme algoritmaları; F1 skoru ve ROC-AUC gibi performans ölçütleriyle değerlendirilmiştir. Analizlerde Shuttle ve Satellite veri setleri kullanılmıştır. Bu denetimsiz yöntemlerin performansı, destek vektör makineleri (Support vector machine-SVM) ve KNN gibi denetimli sınıflandırma yöntemleriyle karşılaştırılmıştır. Elde edilen bulgular, denetimsiz yöntemlerin, özellikle etiketli verinin sınırlı olduğu durumlarda, denetimli yöntemlere kıyasla eşdeğer hatta üstün performans gösterebildiğini ortaya koymuştur [3]. Bir diğer çalışmada ise Logistic Boosting, RF, SVM algoritmaları,

endüstriyel nesnelerin interneti (Internet of Things-IoT) ortamlarında anomali tespiti amacıyla karşılaştırmalı olarak değerlendirilmiştir. Gerçek dünyadan alınan ve fabrika sensörlerinden elde edilen 15.000 örnek içeren veri seti kullanılarak ROC eğrileri, karmaşıklık matrisleri ve standart performans metrikleri üzerinden analiz gerçekleştirilmiştir. Logistic Boosting modeli yüksek doğruluk oranları ile en başarılı sonuçları vermiştir. Bu modelin özellikle dengesiz veri setleri üzerindeki başarısı dikkat çekmiştir. RF modeli güçlü bir performans sergilerken, SVM yüksek duyarlılık değeri ile öne çıkmıştır [4].

Anomalilerin belirlenme süreci özellikle veri kaydetme ve işleme teknolojisinin gelişmesi ile daha fazla çalışılır hale gelmiştir. Astronomi ve uzay bilimleri bu gelişimin öncü alanlarından birisi olmuştur Bu kapsamda birçok proje üzerinde çalışılmaktadır. Bunlardan biri de büyük ölçekli astronomik taramalarda anomali tespitine odaklanan uluslararası bir proje olan Supernova Anomaly Detection (SNAD) projesidir. Proje kapsamında aktif öğrenme ve çeşitli makine öğrenmesi algoritmaları kullanılarak, gökbilimsel olayların keşfi ve sınıflandırılması sağlanmakta, aynı zamanda astrofizik alanında makine öğrenmesinin uygulama düzeyi derinleştirilmektedir [5].

Giderek büyüyen ve karmaşıklaşan veri ortamında, gözlemlenen yapıların sınıflandırılması ve beklenmeyen anomali durumların tespiti için sağlam ve etkili yöntemler kullanılması kaçınılmazdır. Bu bağlamda, anomali tespiti amacıyla iki farklı denetimsiz makine öğrenmesi algoritmasının kullanıldığı bir başka çalışmada görüntü ve katalog veriler üzerinde Autoencoder ve RF denetimsiz modelleri ile analizler yapılmıştır. Burada, ilk yöntem, etkileşim halindeki galaksiler ve kütle çekimsel mercekler gibi sıra dışı nesnelerin tanımlanmasına olanak tanırken; ikinci yöntem ile, olağandışı renk ve parlaklık değerlerine sahip nesnelerin tespitini mümkün kılarak potansiyel astronomik anomali gözlemlerin belirlenmesine katkı sağlanmaktadır [6].

Son yıllarda, denetimli öğrenme yöntemlerinin de etiketli verilerin mevcut olduğu durumlarda yüksek performans sunduğu gösterilmiştir. Özellikle DT, RF ve XGBoost gibi ağaç tabanlı modeller, karmaşık karar sınırlarını öğrenebilme yetenekleri sayesinde anomali sınıflarını ayırt etmede etkili sonuçlar vermektedir. Ayrıca, parametre optimizasyonu ve model süresi gibi faktörler de literatürde karşılaştırmalı olarak değerlendirilmektedir.

Bu çalışmada, Amerikan Ulusal Havacılık ve Uzay Dairesi (National Aeronautics and Space Administration -NASA) tarafından sağlanan ve uzay mekiği görevlerine ait sensör ölçümlerini içeren shuttle veri seti kullanılarak, denetimli makine öğrenmesi algoritmaları ile anomali tespiti gerçekleştirilmiştir. Veri kümesinde çoğunluğu oluşturan “normal” ve azınlıkta kalan “anomali” sınıflar üzerinden, lojistik regresyon (Logistic Regression -LR), k-en yakın komşu (K-Nearest Neighbors - KNN), karar ağacı (Decision Tree - DT), rastgele orman (Random Forest - RF), aşırı gradyan artırma algoritması (Extreme Gradient Boosting -XGBoost) gibi modeller ile tahminler yapılarak model performansları karşılaştırmalı olarak değerlendirilmiştir. Ayrıca, gerçek veri

kümesine benzer bir yapay veri seti üretilmiş ve bu veriler üzerindeki model davranışları da analiz edilmiştir.

Çalışmanın amacı, farklı makine öğrenmesi algoritmalarının hem simülasyon ile üretilmiş veri setinde hem de Shuttle uzay verisi gibi teknik ve dengesiz veri setlerinde anomali tespiti performansını ölçmek ve bu algoritmaların uygulanabilirliğini karşılaştırmalı olarak ortaya koymaktır. Bu amaç doğrultusunda, hem simülasyon ile üretilmiş veri setinde hem de shuttle veri seti üzerinde anomali gözlemlerin tespitine yönelik olarak beş farklı denetimli makine öğrenmesi algoritması kullanılmıştır. Anomali gözlemlerin belirlenmesinde makine öğrenmesi yöntemlerinin farklı performanslara sahip olabilmesi uygun modellerin ve tekniklerin seçilmesinin oldukça önemli olduğunu ortaya koymaktadır. Çalışmada kullanılan modellerin belirlenmesinde çok kapsamlı bir değerlendirme yapılmıştır. Model seçilirken, denge ve çeşitlilik, performans ve anomali tespit yeteneği, uygulama kolaylığı ve anlaşılabilirlik gibi kriterler göz önünde bulundurulmuştur. Seçilen modeller, hem farklı öğrenme yaklaşımlarını temsil etmesi hem de sınıflandırma performansı açısından literatürde geniş uygulama alanı bulmaktadır. Bu kapsamda, modellerin anomali tespit performansını ölçmek ve bu algoritmaların uygulanabilirliğini karşılaştırmalı olarak ortaya koymak için, parametre optimizasyonu, eğitim ve test süresi karşılaştırmaları, ROC (Receiver Operating Characteristic) eğrileri ve karmaşıklık matrisleri gibi kriterler kullanılarak kapsamlı bir değerlendirme yapılmıştır.

2. YÖNTEMLER (METHODS)

Anomali probleminin karakteristiğine özgü farklı anomali tespit yöntemleri tercih edilmektedir. Yöntemler farklı gereklilikler, varsayımlar ve başka nedenlerden ötürü farklı anomali saptama sürecinde farklı tahmin performansına sahip olabilmektedir.

Bu çalışmada, Shuttle veri seti üzerinde ve Shuttle verisinin dağılım özelliklerini koruyan ve kovaryans yapısına göre oluşturulan yapay veri setinde anomali tespiti amacıyla denetimli makine öğrenmesi algoritmaları uygulanmıştır. Uygulama süreci veri ön işleme, model eğitimi, parametre optimizasyonu ve performans değerlendirmesi olmak üzere dört temel aşamadan oluşmaktadır. Öncelikle çalışmada kullanılan modellerin teorik yapıları verilmiştir.

2.1. Lojistik Regresyon (LR) Yöntemi

Çalışmada kullanılan modellerden ilki regresyon amaçlı da kullanılan LR algoritmasıdır. LR lineer bir sınıflandırıcı olup, özellikle iki sınıflı problemler için temel, ancak etkili bir yöntemdir. Model, bağımlı değişkenin belirli bir sınıfa ait olma olasılığını tahmin etmek için lojistik (sigmoid) fonksiyon kullanır. Bağımsız değişken X ile bağımlı değişken Y arasındaki ilişki doğrusal bir kombinasyon ile modellenir ve bu kombinasyon, 0 ile 1 arasında olasılık değeri üreten lojistik fonksiyondan geçirilir. Modelin temel amacı, en çok olabilirlik tahmin (Maximum Likelihood Estimation - MLE) yöntemiyle model parametreleri olan katsayıları tahmin ederek, gözlemlerin doğru sınıflara atanma olasılığını en üst düzeye çıkarmaktır [7]. Lojistik regresyon, sınıflar arasındaki doğrusal ayrımı tahmin eder ancak çıktı olarak olasılık

sunar. Eşitlik (1)' de LR modelinin logaritmasının alınmış halini verilmiştir. Logaritmik dönüşüm olasılığı doğrusal hale getirmektedir.

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + .. + \beta_n X_n \quad (1)$$

Burada, $p(X)$: gözlemin anomali sınıfına ait olma olasılığı, $1 - p(X)$: normal sınıfa ait olma oalsılığı, β_0: sabit terimi, β_i: modeldeki katsayıları ya da modelin bilinmeyen parametrelerini $(i = 1,2, \dots, n)$, X_i: bağımsız değişkenleri $(i = 1,2, \dots, n)$ ifade etmektedir.

2.2. K - En Yakın Komşu (KNN) Yöntemi

KNN modeli, yeni bir gözlemi sınıflandırmak için eğitim kümesindeki en yakın komşuların etiketlerini kullanır. Yöntem, kural tabanlı ve basit bir yöntem olup, küçük veri setlerinde etkili sonuçlar vermektedir. Komşuluk yapısına dayanarak komşu sayısına göre hesaplama yaparak anomali tespiti yapmaktadır. Belirlenecek olan k kadar komşu sayısına, seçilecek bir uzaklık hesaplama yöntemi ile mesafelerin hesaplanmasına dayanmaktadır. Öklid, minkowski ve mahalanobis uzaklık hesaplama yöntemleri en sık kullanılan yöntemlerdir. Eşitlik (2)'de, $d(x_i, x_j)$ mesafesi x_i ile $x_j$noktaları arasındaki öklid mesafesidir. Eşitlikde; n özellik sayısını, x_{ik} : i'inci gözlemin k'ıncı özelliğini, x_{jk}: j'inci gözlemin k'ıncı özelliğini göstermektedir. Buradan elde edilecek değer ile ilgili gözlemin anomali olup olmadığına karar verilir. Yüksek d değeri komşulara uzak olduğunu (anomali), düşük d değeri ise komşulara yakın olduğunu (normal) gösterir. Modelin büyük veri setlerinde hesaplama maliyeti yüksek olsa da parametre sayısının azlığı, basit ve sezgisel yaklaşımı modelin güçlü yönü olarak değerlendirilmektedir hızlıdır [8].

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (2)$$

2.3 Karar Ağacı (DT) Yöntemi

Diğer yöntem ise bütün karar ağacı tabanlı yöntemlerin temel algoritması olan DT yöntemidir. Veriyi özniteliklere göre dallara ayırarak sınıflandıran, yorumlaması kolay bir yöntemdir. Özellikle ayrımın açık olduğu durumlarda etkili sonuçlar vermektedir. Eşitlik (3)'te entropiyi gösteren $H(x)$, karar ağacı algoritmasında bir değişkenin içerdiği bilgi miktarını veya belirsizliğini ölçmek için bilgi kazancı (Information Gain) hesaplamasında kullanılır. Aynı zamanda, veri seti üzerindeki bölmelerin ne kadar anlamlı olduğunu da gösterir. Düşük entropi değerleri, daha saf ya da tek sınıfa ait veri anlamına gelir [9].

$$H(x) = Entropy(S) - \sum_i \frac{|S_i|}{|S|}.Entropy(S_i) \quad (3)$$

Burada, $Entropy(S)$: bölünmeden önceki kümenin entropisi, $Entropy(S_i)$: bölünme sonrası oluşan i'inci alt kümenin entropisi, S_i: bölünme sonrası oluşan i'inci alt küme, $|S_i|$: i'inci alt kümenin eleman sayısı, $|S|$: tüm kümenin eleman sayısını, $i = 1,2, \dots, m$ ve m ise bölünme sonrası oluşan alt küme sayısını göstermektedir. $Entropy$ değerleri Eşitlik (4) ve Eşitlik (5)' deki gibi hesaplanır.

$$Entropy(S) = -\sum_{j=1}^{c} p_j log_2 p_j \quad (4)$$

$$Entropy(S_i) = -\sum_{j=1}^{c} p_{ij} log_2 p_{ij} \quad (5)$$

Burada, p_j: S kümesindeki j sınıfının oranını, p_{ij}: S_i kümesindeki j sınıfı oranını ve c : sınıf sayısını ifade etmektedir. Gerekli hesaplamaların ardında bilgi kazancı elde edilir ve bu anomali tespitinde kullanılır. Bilgi kazancı yüksek olan özellikler, sınıfları normal ve anomali olarak en iyi ayıran özelliklerdir. Yani $H(x)$ değeri ne kadar yüksekse, o özellik anomali tespitinde o kadar etkilidir.

2.4. Rastgele Orman (RF) Yöntemi

RF modeli paralel olarak çalışan birden fazla karar ağacından oluşan bir topluluk öğrenme yöntemdir. Aşırı öğrenmeye karşı dirençlidir ve değişken önem derecelerini belirlemede de kullanılabilir. RF yönteminin anomali tespitindeki gücü, her ağacın bağımsız olarak veri noktalarını değerlendirmesinden ve çoğunlukla aynı sonucu vermeyen anomali noktalarını ayırmasından gelir [10]. Karar ağaçları farklı veri alt kümeleri üzerinde eğitilir ve sonuç olarak, sınıflandırmada çoğunluk oyu, regresyonda ise ortalama alınarak sonuç üretilir. Eşitlik (6)' da özellikle RF modeli veya Boosting gibi çok sayıda modelin çıktısını birleştiren yöntemler için temel karar fonksiyonunu yer almaktadır.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \qquad (6)$$

Burada, x gözlemleri yani özellik vektörünü, $f_t(x)$: t'inci modelin x girdisi için ürettiği tahmini, $\hat{y}$ nihai tahmini, , T ise modeldeki toplam öğrenici sayısını göstermektedir. Buradan, her bir model $f_t(x)$ kendisine verilen gözlem için bir anomali skoru üretir. Tüm modellerin ürettiği bu skorlar bir araya getirilerek ortalaması alınır. Elde edilen bu ortalama değer, gözlemin topluluk anomali skoru olarak adlandırılır. Belirlenen eşik değer aşılırsa gözlem anomali, aksi halde normal olarak sınıflandırılır.

2.5. Aşırı Gradyan Artırma (XGBoost) Yöntemi

XGBoost algoritması, boosting mantığıyla çalışan güçlü bir topluluk öğrenme algoritması olan özellikle son yıllarda oldukça sık tercih edilmektedir. Öğrenme oranı, maksimum derinlik, alt örnekleme gibi parametreleri performansı ciddi ölçüde etkiler. XGBoost, gradyan artırma (gradient boosting) yönteminin gelişmiş bir versiyonudur ve ikinci dereceden gradyan hesaplamalarını kullanarak kayıpları minimize eder. Algoritma, kayıp fonksiyonunu optimize ederken aşırı öğrenmeyi kontrol altında tutar. Hızlı bir şekilde çalışır ve büyük veri kümelerinde yüksek performans sunar. Özellikle karar ağaçları üzerinde çalışarak daha karmaşık modeller elde edilmesini sağlar [11]. Eşitlik (7), XGBoost gibi yöntemlerin tahmin mekanizmasını özetlemektedir.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \qquad (7)$$

Burada, $f_k(x_i)$: k'ıncı ağacın i'inci gözlem için ürettiği tahminini, x_i : i'inci gözlemin özellik vektörünü, K modeldeki toplam öğrenici sayısını ve $\hat{y}$ nihai tahminini gösterir. Model, her bir gözlem için ardışık olarak inşa edilen K adet zayıf modelin çıktılarının toplamını alarak son tahmini üretir. Her bir $f_k(x_i)$ hataları minimize edecek şekilde eğitilir. Böylece, modelin adım adım hatalarını azaltarak güçlü bir tahmin performansı elde etmesi sağlanmış olur. Elde

edilen skorların toplamı, gözlemin anomali olasılığını veya risk puanını belirler. Belirlenen eşik değeri aşıldığında gözlem anomali olarak sınıflandırılır.

Seçilen modeller, hem farklı öğrenme yaklaşımlarını temsil etmesi hem de sınıflandırma performansı açısından literatürde geniş uygulama alanı bulmaktadır. Buna ilişkin modellerin güçlü yönleri ve sahip olduğu avantajlar tablo 1'de sunulmuştur.

Tablo 1: Modellerin güçlü yönleri ve avantajları.

Model adı	Modellerin güçlü yönleri ve avantajları	Çıkış yılı
LR	• Basit, yorumlanabilir, küçük verilerde etkili. • Doğrusal sınırlar üzerinden iyi performans sağlamaktadır.	1958
KNN	• Parametre gerektirmez, kolay uygulanabilir. • Örnek tabanlı yaklaşımıyla esneklik sağlamaktadır.	1967
DT	• Karar kuralları sezgisel olarak anlaşılır. • Veri ön işleme gerektirmez, görselleştirilebilmektedir.	1986
RF	• Aşırı öğrenmeye karşı dirençlidir. • Yüksek doğruluk sağlar, değişken önemini ölçebilir.	2001
XGBoost	• Çok yüksek doğruluk, hız ve verimlilik sağlar. • Eksik verilerle başa çıkabilir, büyük veri için uygundur.	2014

KNN algoritmasının temeli çok eskilere dayanmaktadır. Örneklerin benzerliğine dayalı basit ve sezgisel bir yaklaşım sunmaktadır. Özellikle düşük boyutlu verilerde hızlı uygulanabilirliği ve parametrik olmayan yapısı nedeniyle anomali tespiti problemlerinde yaygın olarak kullanılmaktadır. LR modeli ise, özellikle ikili sınıflandırma problemlerinde tercih edilen temel ve yorumlanabilir bir modeldir. Parametrik yapısı sayesinde anomali tespiti sürecinde temel karşılaştırma ölçütü olarak da kullanılmaktadır. Özellikle sınıf olasılıklarının yorumlanabilir olması avantaj sağlamaktadır. DT algoritması ise, veri içerisindeki önemli bölünme noktalarını öğrenerek karar kuralları üretir. Yorumlanabilirlik ve sınıflandırma mantığının görselleştirilebilir olması nedeniyle anomali tespitinde güçlü bir temel modeldir. Regresyon problemlerinde de sıklıkla kullanılmaktadır. Diğer bir model olan RF modeli ise birden fazla karar ağacından oluşan topluluk öğrenme yöntemi olan rastgele orman algoritması, yüksek doğruluk ve aşırı öğrenmeye karşı dayanıklılığı ile öne çıkmaktadır. Veri içindeki karmaşık yapıları öğrenmede oldukça etkilidir. Zayıf öğrenicilerin ardışık olarak eğitilmesi şeklinde ifade edilebilen Boosting temelli güçlü bir topluluk algoritması olan XGBoost modeli ise, özellikle dengesiz sınıf dağılımına sahip verilerde üstün performans

sergileyebilmesiyle bilinmektedir. Ağaç tabanlı modellerin doğruluğunu artırmak için ardışık modelleme yaklaşımı kullanmaktadır.

Bu beş model birlikte ele alındığında; KNN ve LR modelleri gibi hem basit, sezgisel yöntemlerden, hem DT ve RF gibi kural tabanlı açıklanabilir yöntemlerden, hem de XGBoost modeli gibi ileri düzey topluluk modellerinden oluşan dengeli ve kapsamlı bir algoritma portföyü oluşturulmuştur. Böylece, farklı öğrenme biçimlerinin hem shuttle veri seti hem de simülasyon ile üretilmiş veri seti üzerindeki anomali tespit performansı karşılaştırmalı olarak değerlendirilmiştir.

3. UYGULAMA (APPLICATION)

Bu bölümde hem gerçek shuttle verisi üzerinde hem de benzer yapıda üretilmiş yapay bir veri kümesi üzerinde gerçekleştirilen anomali tespiti analizlerine yer verilmiştir. Böylece algoritmaların performansları her iki açıdan değerlendirilmiş olacaktır.

3.1. Veri Ön İşleme ve Parametre Ayarı

Modeller kullanılan parametrelere oldukça duyarlıdır. Bu nedenle tablo 2'de yer alan parametre ve değer aralıkları kümesinden her algoritma için modelin başarımını etkileyen bazı parametreler GridSearchCV yöntemi ile optimize edilmiştir. 3 katlı çapraz doğrulama (k-fold=3) kullanılarak en iyi parametreler belirlenmiş ve bu parametrelerle model test verisi üzerinde yeniden eğitilmiştir.

Tablo 2: Model parametre kümesi.

Yöntem	Parametre ve değer aralıkları
LR	C: {0.01, 0.1, 1, 10}; Penalty: {'l1', 'l2'}
KNN	n_neighbors: {3, 5, 10}; weights: {'uniform', 'distance'}
DT	max_depth: {5, 10, None}; min_samples_split: {2, 5}
RF	n_estimators: {50, 100, 150, 200, 300}; max_depth: {10, None}
XGBoost	n_estimators: {50, 100, 200}; max_depth: {3, 6, 9}; learning_rate: {0.01, 0.1, 0.2}; subsample: {0.8, 1.0}

3.2. Performans Değerlendirme Ölçütleri

Modellerin eğitimi sonrası test verisindeki performans bazı metrikler ile ölçülüp karşılaştırılmaktadır. Anomali veri setleri doğası gereği dengesiz veri setleri olduğundan bireysel modellerin ve modellerin performansları, model tahmin sonuçlarının doğruluğunu daha iyi açıklayan ve sınıf dengesizliğine duyarlı F1 skoru ve ROC AUC metrikleri ile değerlendirilmiştir.

F1 skoru hassasiyet ve duyarlılık dengesini sağlayan bir ölçüttür. Anomali sınıfı genellikle azınlıkta olduğu için, özellikle dengesiz veri kümelerinde, genel başarıyı daha iyi yansıtan F1 skoru önemli bir metriktir. F1 Skoru, hassasiyet (Precision-P) ve duyarlılık (Recall-R) metriklerinin harmonik ortalamasıdır ve Eşitlik (8)' deki gibi hesaplanır.

$$F1 \;=\; 2 * \frac{(P * R)}{(P + R)} \qquad (8)$$

Burada P ve R değerlerinin hesaplanması sırasıyla Eşitlik (9) ve Eşitlik (10)' da verilmiştir. Bu değerler hesaplanırken tablo 3' te verilen karmaşıklık matrisinden yararlanılır. Tablo 3 incelendiğinde; Eşitlik (9) ile hesaplanan P değeri "0" olarak tahmin edilenlerin ne kadarının gerçekten "0" olduğunu göstermektedir. Eşitlik (10) ile hesaplanan R değeri ise, gerçek "0" olanların ne kadarının doğru tahmin edildiğini göstermektedir. Anomali bağlamında değerlendirildiğinde R değeri anomalilerin kaçırılmadan yakalanmasını sağlarken, P ise yanlış pozitiflerin azaltılmasına yardımcı olur.

$$P = TP / (TP + FP) \quad (9)$$

$$R = TP / (TP + FN) \quad (10)$$

Tablo 3: Karmaşıklık matrisi.

	Gerçek Pozitif (0)	**Gerçek Negatif (1)**
Tahmin Pozitif (0)	TP	FP
Tahmin Negatif (1)	FN	TN

TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative

Diğer metrik ise ROC-AUC değerleridir. ROC eğrisi, sınıflandırma modelinin ayırt edici gücünü ortaya koymak için (TP) ile (FP) arasındaki değişimi gösterir. AUC değeri ise, bu eğrinin altındaki alanı ölçerek modelin genel ayırt ediciliğini özetler ve 1'e yaklaştıkça modelin mükemmele yakın olduğunu, 0.5'e yaklaştıkça rastgele tahmin yaptığını gösterir.

Diğer bir performans kriteri ise modellerin eğitim ve test süresidir. Eğitim süresi, modelin öğrenme aşamasındaki hesaplama yükünü, test süresi ise modelin yeni veriler üzerindeki tahmin hızını ifade eder. Modellerin zamanlama açısından karşılaştırılmasını sağlar. Büyük boyutlu veri kümesi kullanıldığında işlem süresi önemli hale geldiğinden özellikle model seçimi ve maliyeti gibi kararlarda kritik bir değerlendirme ölçütüdür.

Bu çalışmada, veri setindeki sınıf dengesizliği yapay olarak dengelenmemiştir. Bunun temel nedeni, anomali problemlerinin doğası gereği bu tür dengesizliğin doğal olarak yer alması ve gerçek dağılımın korunmasının daha anlamlı sonuçlar vereceği yönündeki yaklaşımdır. Ayrıca, kullanılan makine öğrenmesi algoritmaları dengesiz veri setlerinde de etkin performans gösterebilme özelliğine sahiptir. Değerlendirme sürecinde dengesizliğe duyarlı F1 skoru ve ROC-AUC metriklerin de kullanılması ile modellerin başarısı adil bir şekilde ölçülmüştür. Bu nedenle, SMOTE (Synthetic Minority Over-sampling Technique) gibi sentetik örnekleme yöntemlerine ihtiyaç duyulmamıştır.

3.3. Shuttle Verisi Analizi

Shuttle verisi, NASA tarafından sağlanan uzay mekiği görevlerine ilişkin çok boyutlu sensör ölçümlerini içeren bir gerçek dünya verisidir. Veri seti, University of California Irvine (UCI) Makine Öğrenmesi Veri Havuzundan alınmıştır [12]. Veri kümesindeki sınıf dağılımı anomali veri setlerine benzer şekilde; normal gözlemler %92.5 oranında olup, geri kalan %7.5'luk kısım

anomalilerden oluşmaktadır. Tablo 4'de veri setinin içeriği verilmiştir.

Tablo 4: Veri setinin içeriği.

Değişken adı	Rolü	Tipi	Kayıp değer
Time	Feature0	Integer	Mevcut değil
Rad Flow	Feature1	Integer	Mevcut değil
Fpv Close	Feature2	Integer	Mevcut değil
Fpv Open	Feature3	Integer	Mevcut değil
High	Feature4	Integer	Mevcut değil
Bypass	Feature5	Integer	Mevcut değil
Bpv Close	Feature6	Integer	Mevcut değil
Bpv Open	Feature7	Integer	Mevcut değil
class	Target	Integer	Mevcut değil

Tablo 4'de yer alan bilgiler incelendiğinde veri setinde sekiz bağımsız, bir bağımlı değişken vardır. Veri seti toplamda 49.097 gözlem ve her biri uçuş performansını yansıtan 9 sayısal özellikten oluşmaktadır. Bu gözlemlerin 3511' i anomali olarak, kalan 45586' sı ise normal olarak etiketlenmiştir. Test verisi tüm verinin % 30'u olduğundan, test verisinde 14730 gözlem vardır. Bunların da 1053 adedi anomali olarak etiketlidir. Veri setinde herhangi bir kayıp değer bulunmamaktadır. Bağımlı değişken olan class verisi etiketlenmiş olup, normal gözlem (class=0) ve anomali gözlem (class=1) olarak kodlanmıştır. Ardından tüm özellikler StandardScaler fonksiyonu ile normalize edilmiştir. Veriler %70 eğitim ve %30 test olarak stratify=y parametresiyle dengeli şekilde bölünmüştür.

Yapılan analizler sonucunda yöntemlerin performanslarını karşılaştırmak amacıyla, Tablo 5' te çalışmada kullanılan algoritmaların test kümesi üzerindeki F1 skorları, ROC-AUC değerleri, eğitim ve test süreleri ile en iyi parametre değerleri verilmiştir.

Tablo 5: Modellerin performans karşılaştırma sonuçları.

Modeller	F1 skoru	ROC-AUC değeri	Eğitim süresi (saniye)	Test süresi (saniye)	En iyi parametreler
LR	0.97	0.98	13.64	0.00	{'C': 10, 'penalty': 'L1'}
KNN	0.99	0.99	13.05	3.03	{'n_neighbors': 3, 'weights': 'distance'}
DT	0.99	0.99	0.57	0.00	{'max_depth': None, 'min_samples_split': 2}
RF	0.99	1.00	30.14	0.36	{'max_depth': 10, 'n_estimators': 200}
XGBoost	0.99	1.00	13.72	0.01	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.8}

Tablo 5' teki sonuçlar incelendiğinde, XGBoost ve RF modelleri, ROC-AUC değerinin 1.00 olmasıyla mükemmel sınıflandırma başarısı göstermiştir. XGBoost modeli aynı zamanda düşük test süresi ile öne çıkarak hem doğruluk hem de zaman açısından en iyi model olmuştur. RF modeli benzer şekilde yüksek doğruluk sağlamasına karşın, en uzun eğitim süresine sahiptir. DT modeli ise, hızlı eğitim ve test süresi ile birlikte 0.99 F1 ve ROC-AUC değeri ile oldukça verimli çalışmıştır. KNN modeli, yüksek başarıya rağmen test süresinin uzun olması nedeniyle özellikle gerçek zamanlı yüksek boyutlu veri uygulamaları açısından dikkatle değerlendirilmelidir. LR modeli ise diğer yöntemlere kıyasla daha düşük bir F1 ve ROC-AUC değerleri üretmiştir.

Optimizasyon sonucunda elde edilen parametreler de dikkat çekicidir. LR modelinde C=10 değeri, daha az regularizasyon anlamına gelmekte ve modelin karmaşık sınırları yakalamaya çalıştığına işaret etmektedir. Ceza parametresi için penalty='L1' seçimi ise modelin sadece önemli değişkenlere odaklanarak sadeleşmesini sağlamaktadır. KNN modelinde n_neighbors=3 ve weights='distance' parametreleri, yakın komşuların mesafeye göre ağırlıklandırılmasını sağlayarak kararların daha hassas alınmasına olanak tanımıştır. DT modeli için max_depth=None, modelin maksimum derinlik sınırı olmadan tüm veriye tam olarak uyum sağlamasına izin verirken, min_samples_split=2 değeri ise her düğümün kolaylıkla bölünmesini sağlamıştır. RF modelinde ise max_depth=10 aşırı öğrenme riskini azaltırken, n_estimators=200 değeri modeli daha genellenebilir hale getirmiştir. XGBoost modeli için seçilen learning_rate=0.1, max_depth=3, n_estimators=100 ve subsample=0.8 parametreleri ise modelin öğrenme sürecini dengeli hale getirerek hem aşırı uyumu önlemiş hem de güçlü performans sağlamıştır.

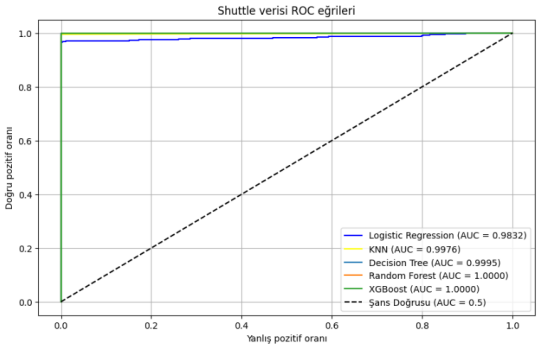
Tablo 6: Modellerin karmaşıklık matrisleri.

Modeller		Gerçek Pozitif (0)	Gerçek Negatif (1)
LR	Tahmin Pozitif (0)	13676	1
	Tahmin Negatif (1)	59	994
KNN	Tahmin Pozitif (0)	13677	5
	Tahmin Negatif (1)	10	1043
DT	Tahmin Pozitif (0)	13675	2
	Tahmin Negatif (1)	1	1052
RF	Tahmin Pozitif (0)	13677	0
	Tahmin Negatif (1)	2	1051
XGBoost	Tahmin Pozitif (0)	13677	0
	Tahmin Negatif (1)	1	1052

Her bir model için Tablo 6'da yer alan karmaşıklık matrisleri incelendiğinde tüm modellerin genel olarak oldukça başarılı bir performans sergilediği görülmektedir. XGBoost ve RF modelleri en yüksek başarıyı göstermiştir.

Bu modeller sırasıyla 1 ve 2 yanlış sınıflandırma yaparak hem gerçek pozitif hem de gerçek negatif sınıflarda neredeyse hatasız sonuçlar üretmişlerdir. DT modeli de yalnızca 3 gözlemi yanlış sınıflayarak güçlü bir performans ortaya koymuştur. KNN modeli, 15 gözlemde hata yaparak diğer güçlü modellerin biraz gerisinde kalmıştır. LR modeli ise 60 gözlemde yanlış sınıflama yaparak en çok hata yapan model olmuştur. Bu durum, LR modelinin doğrusal yapısı nedeniyle karmaşık veri yapılarına yeterince uyum sağlayamadığına işaret etmektedir.

Her modelin AUC değerinin de yer aldığı ROC eğrileri şekil 1'de yer almaktadır. Eğrinin sol üst köşeye yakın olması, modelin yüksek ayırt etme gücüne sahip olduğunu, AUC değerinin 0.5'ten büyük olması ise modelin rastgele tahminden daha iyi olduğunu göstermektedir. Şekil 1'deki ROC eğrisi sonuçları genel olarak değerlendirildiğinde, tüm modellerin güçlü performans sergilediğini ancak özellikle ağaç tabanlı modellerden sırasıyla XGBoost, RF ve DT modellerinin anomali tespitinde öne çıktığını ve yüksek ayırt ediciliğe sahip olduğunu göstermektedir.



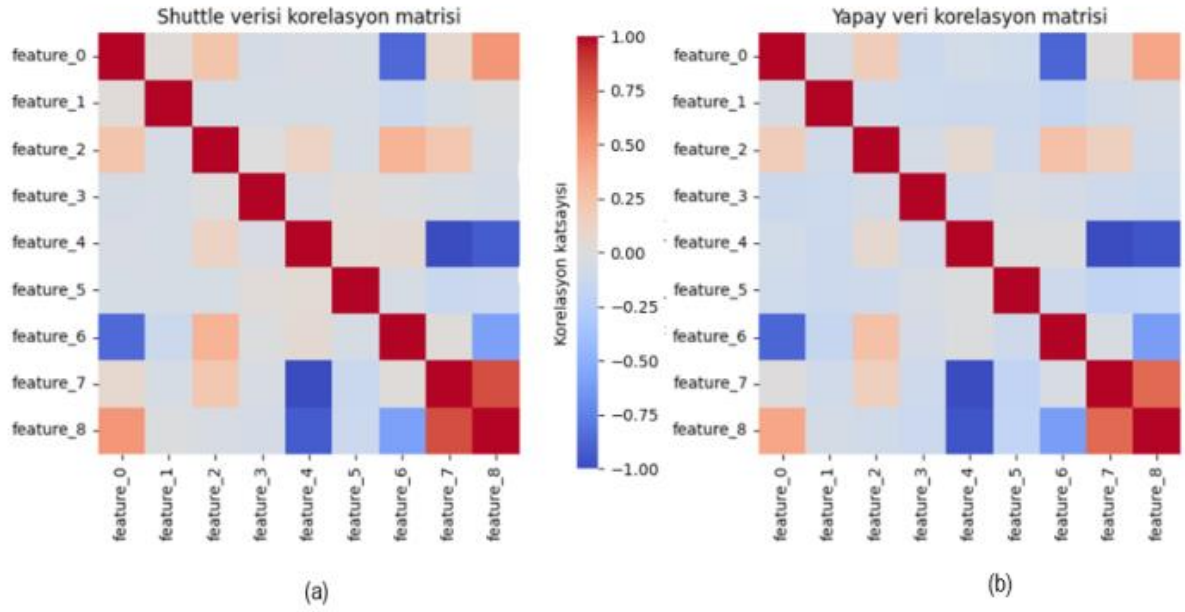
Şekil 1: ROC-AUC eğrileri

3.4. Yapay Veri Analizi

Gerçek shuttle verisinin dağılım özelliklerini koruyan ve kovaryans yapısına göre oluşturulan 10000 gözlemli yapay veri kümesi oluşturulmuştur. Yapay veri, gerçek veriye ait ortalama vektörü ve kovaryans matrisi kullanılarak python kütüphanesinden {numpy.random.multivariate_normal()} fonksiyonu ile üretilmiştir. Bu veri kümesinde rastgele seçilen 100 gözlem satırında rastgele seçilen öznitelik değerlerine ortalamadan 10 standart sapma uzaklıkta yapay bir uç değer atanarak anomali oluşturulmuş ve bu satırlar anomali (1) olarak etiketlenmiştir.

Şekil 2'de korelasyon matrisleri verilmiştir. Şekil 2 (a)'da gerçek shuttle veri setindeki, Şekil 2 (b)'de ise yapay veri setindeki öznitelikler arasındaki ilişkiler gösterilmektedir. Görüldüğü üzere yapay veri seti, korelasyon yapısı bakımından gerçek veriye oldukça benzer bir yapı sergilemektedir. Yapay veri üretim sürecinin gerçek veri yapısına ne kadar yakın olduğunu tespit etmek oldukça önemlidir. Çünkü üretilen verinin korelasyon yapısı gerçek veriye ne kadar benzerse, bu veri üzerinde yapılan model performans değerlendirmeleri de o kadar anlamlı ve genellenebilir olmaktadır. Korelasyon

matrisi görsellerinde renk tonları değişkenler arasındaki korelasyon katsayısının yönünü ve büyüklüğünü temsil etmektedir. Kırmızı tonlar pozitif korelasyonu, mavi tonlar negatif korelasyonu, beyaz veya açık renkler ise düşük ya da önemsiz düzeyde korelasyonu göstermektedir.



Şekil 2: Gerçek ve yapay veriye ilişkin korelasyon matrisleri.

Yapılan analizler sonucunda yöntemlerin performanslarını karşılaştırmak amacıyla, Tablo 7' de modellerin yapay veriye ait test kümesi üzerindeki F1 skorları, ROC-AUC değerleri, eğitim ve test süreleri ile en iyi parametre değerleri verilmiştir.

Tablo 7: Modellerin performans karşılaştırma sonuçları.

Modeller	F1 skoru	ROC-AUC değeri	Eğitim süresi (saniye)	Test süresi (saniye)	En iyi parametreler
LR	0.81	0.99	0.23	0.00	{'C': 10, 'penalty': 'L1'}
KNN	1.00	1.00	1.17	0.33	{'n_neighbors': 3, 'weights': 'uniform'}
DT	1.00	1.00	0.42	0.00	{'max_depth': 10, 'min_samples_split': 2}
RF	1.00	1.00	27.86	0.02	{'max_depth': 10, 'n_estimators': 50}
XGBoost	0.56	0.99	10.08	0.02	{'learning_rate': 0.01, 'max_depth': 9, 'n_estimators': 50, 'subsample': 0.8}

Yapay veri seti 10.000 gözlem içerip yalnızca 100 anomaliden oluştuğu için, test kümesinde 3.000 gözlem ve yaklaşık 30 anomali vardır. Bu ciddi dengesizlik, özellikle anomali tespiti için oldukça zorlayıcı bir durumdur. Buna rağmen KNN, DT ve RF modelleri test kümesinde hem F1 skoru hem de ROC-AUC değeri açısından 1.00 ile mükemmel sonuçlar vermiştir. Bu durum, modellerin az sayıda anomaliyi bile başarıyla tanıyabildiğini ve dengesiz sınıflar arasında etkin ayrım yapabildiğini göstermektedir. LR modeli ise ROC-AUC değeri 0.99 ile yüksek genel ayırt ediciliğe sahip olmasına rağmen, F1 skorunun 0.81 olması modelin anomali sınıfını eksik tespit ettiğini göstermektedir. XGBoost modelinin ROC-AUC değeri 0.99 olsa da F1 skorunun 0.56'da kalması, modelin görece küçük veri gruplarında öğrenmede zayıf kaldığına ve yanlış sınıflandırmalara eğilimli olduğuna işaret etmektedir.

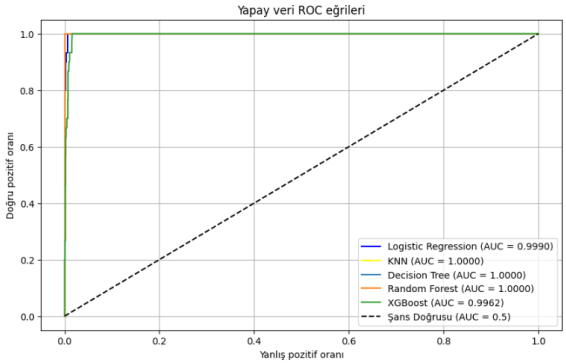
Parametre sonuçları incelendiğinde ise KNN modeli, 3 komşu ve uniform ağırlıklandırma ile yüksek başarı sağlamıştır. Bu da verinin yapısının güçlü ayrım içerdiğini ve KNN için uygun bir yapı sunduğunu göstermektedir. DT ve RF modellerinde kullanılan max_depth=10 sınırlaması, modelin aşırı öğrenmeden kaçınmasına yardımcı olurken, RF modelinin yalnızca 50 ağaç ile yüksek doğruluk sağlaması dikkat çekicidir. Bu, modelin sınıflar arasındaki ayrımı çok net yapabildiğini göstermektedir. LR modelinde kullanılan C=10 ve L1 ceza (penalty='l1') ise modelin bazı parametreleri önemsiz olarak değerlendirdiği böylece modelin sadeleşmesini sağlamış olabileceği fikrini uyandırmaktadır. Ancak bu durum anomali sınıfına duyarlılığı azaltmış ve gerçek anomaliyi tanıyamama riskini artırmıştır. XGBoost modeli için belirlenen düşük öğrenme oranı (0.01) ve max_depth=9 gibi değerler modelin öğrenme hızını düşürmüş olduğu sonucuna işaret etmektedir. Dolayısıyla, dengesiz veri setinde anomali örneklerini öğrenmede yetersiz kaldığı görülmüştür. F1 skorundaki düşüklük, modelin özellikle anomalilere karşı yeterli duyarlılığı geliştiremediğini ortaya koymaktadır.

Tablo 8: Modellerin karmaşıklık matrisleri.

Modeler		Gerçek Pozitif (0)	Gerçek Negatif (1)
LR	Tahmin Pozitif (0)	2968	2
	Tahmin Negatif (1)	8	22
KNN	Tahmin Pozitif (0)	2970	0
	Tahmin Negatif (1)	0	30
DT	Tahmin Pozitif (0)	2970	0
	Tahmin Negatif (1)	0	30
RF	Tahmin Pozitif (0)	2970	0
	Tahmin Negatif (1)	0	30
XGBoost	Tahmin Pozitif (0)	2964	6
	Tahmin Negatif (1)	16	14

Tablo 8'deki karmaşıklık matrisleri incelendiğinde, sınıf dengesizliğine rağmen modellerin anomali sınıfını oldukça yüksek doğrulukla tespit ettiği görülmektedir. KNN, DT ve RF modelleri, 30 anomalinin tamamını doğru şekilde tanımlamış ve hiçbir yanlış pozitif üretmemiştir. Bu da bu üç modelin yüksek duyarlılık sergilediğini göstermektedir. LR modeli ise 30 anomali gözlemden 22'sini doğru sınıflarken, 8'ini yanlış sınıflamış ve 2 normal gözlemi ise yanlış sınıflayarak anomali olarak etiketlemiştir. Bu da modelin neden daha düşük F1 skoruna sahip olduğunu açıklamaktadır. XGBoost modeli ise 30 anomali gözlemden 14'ünü doğru sınıflarken, 16'sını yanlış sınıflamış ve 6 normal gözlemi ise yanlış sınıflayarak anomali olarak etiketlemiştir. Bu durum, XGBoost modelinin az sayıda anomali gözlemlerin olduğu senaryoda fazla temkinli ya da yetersiz öğrenen bir davranış sergilediğine işaret etmektedir.

Şekil 3'teki ROC eğrileri incelendiğinde, özellikle KNN, DT ve RF modellerinin anomali tespitinde mükemmele yakın bir performans sergilediği görülmektedir. Bu üç model için AUC değerleri tam olarak 1.00 olup, modelin ideal performansa ulaştığını göstermektedir. LR ve XGBoost modelleri ise 0.99 AUC değeriyle oldukça başarılı bir sonuç elde etmiştir. XGBoost modeli ise 0.99 AUC değeriyle diğer modellerden biraz daha düşük bir performans sergilemiş, bu durum modelin karmaşıklık matrisinden de gözlemlenmiştir.



Şekil 3: ROC-AUC eğrileri.

4. BULGULAR ve TARTIŞMA (FINDINGS and DISCUSSION)

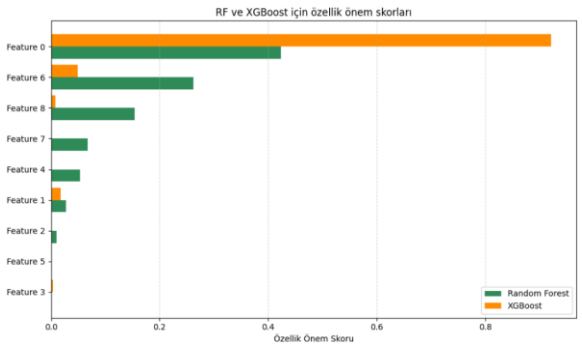
Bu çalışmada, hem gerçek shuttle uzay veri seti hem de bu veri setine benzer yapay olarak üretilmiş bir veri kümesi üzerinde denetimli makine öğrenmesi algoritmaları ile anomali tespiti gerçekleştirilmiştir.

Gerçek shuttle verisi üzerinde yapılan analizlerde, özellikle KNN, DT, RF ve XGBoost modelleri yüksek doğruluk, F1 skoru ve ROC-AUC değerleriyle dikkat çekmiştir. LR modeli de tatmin edici bir performans göstermesine rağmen, diğer modellerin gerisinde kalmıştır. Gerçek verideki sonuçlar, bu modellerin az sayıda anomali gözlemine rağmen oldukça isabetli tahminler yapabildiğini ortaya koymuştur.

Yapay veri analizinde KNN, DT ve RF modelleri anomali yapısını güçlü şekilde öğrenmiş ve yüksek F1 ve ROC-AUC skorlarına ulaşmıştır. LR modeli burada da yüksek başarı gösterse de gerçek veri sonucuna göre geride kalmıştır. XGBoost modeli karmaşık parametre yapısına rağmen özellikle F1 skorunda düşük performans sergilemesi dikkat çekmiştir. Bu durum, XGBoost modelinin düşük oranlı anomali sınıfını yeterince iyi ayırt edemediğini, dolayısıyla parametre ayarlarının dikkatle

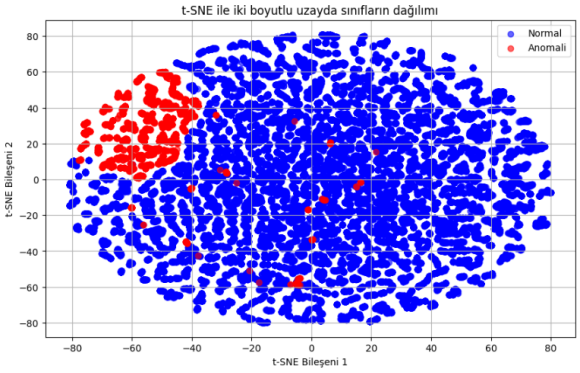
yapılması gerektiğini ortaya koymaktadır. Sonuç olarak, gerçek veri analizi özellikle ağaç tabanlı yöntemlerin yüksek başarı gösterdiği, yapay veride ise bazı modellerin başarılarının değiştiği gözlemlenmiştir.

Ağaç tabanlı yöntemlerden RF ve XGBoost için karar verme süreçlerinde hangi özniteliklere ağırlık verdiklerini belirlemek amacıyla şekil 4'te özellik önem skorları hesaplanmıştır. Şekil 4 incelendiğinde her iki modelin de "Feature 0" olarak adlandırılan özniteliğe en yüksek önemi atfettiği gözlemlenmiştir. Özellikle XGBoost yöntemi kararlarının büyük çoğunluğunu bu tek öznitelik üzerinden şekillendirirken, RF yöntemi daha dengeli bir dağılım sergilemiş ve Feature 6, Feature 8 ve Feature 7 gibi özniteliklere de anlamlı katkı yüklemiştir. Bu farklılık, XGBoost yönteminin daha seçici ve yoğunlaştırılmış bir bilgi kullanımı benimsediğini, RF yönteminin ise daha geniş tabanlı bir özellik setinden yararlanarak karar verdiğini göstermektedir. Bu durum, yüksek boyutlu ve karmaşık veri setlerinde model seçiminde dikkate alınması gereken önemli bir faktör olarak öne çıkmaktadır.



Şekil 4: Özellik önem skorları.

Verilerdeki sınıf dengesizliğine rağmen bazı modeller sınıfları doğru şekilde ayırmakta çok başarılı olmuştur. Bu durum, modelin aşırı öğrenme yapmadan karar sınırlarını doğru öğrenebildiğini göstermektedir. Şekil 5'te t-SNE (t-Distributed Stochastic Neighbor Embedding) yöntemiyle oluşturulan iki boyutlu görselleştirme bu bulguları desteklemektedir. t-SNE yöntemi yüksek boyutlu verileri iki boyutlu düzleme indirgemeyi sağlayan güçlü bir boyut indirgeme yöntemidir. Şekilde, normal sınıfa ait gözlemler mavi, anomali sınıfına ait gözlemler ise kırmızı ile gösterilmiştir. Şekil 5 incelendiğinde, anomalilerin veri kümesinin belirli bir bölgesinde yoğunlaştığı ve normal örneklerden ayrıştığı gözlemlenmektedir. Bu durum, uygulanan modellerin ayırt edici bilgi taşıyan örüntüleri başarılı şekilde öğrenebildiğini göstermektedir.



Şekil 5: t-SNE ile iki boyutta sınıf dağılımı.

Eğitim ve test süresi açısından değerlendirildiğinde ise, DT en hızlı model olarak öne çıkmıştır. RF her iki veri setinde de en uzun eğitim süresine sahip model olmuştur. KNN modeli ise test aşamasında en yavaş çalışan model olmuştur. LR ve XGBoost modelleri eğitim süreleri açısından dengeli performans göstermiştir. Gerçek zamanlı uygulamalarda hızlı tahmin yapan DT ve LR modelleri daha avantajlı iken, RF ve XGBoost

gibi modeller yüksek doğruluk sunmalarına rağmen daha uzun eğitim süresi gerektirmektedir.

5. SONUÇLAR (RESULTS)

Bu çalışmada, makine öğrenmesi algoritmaları kullanılarak hem gerçek hem de yapay olarak üretilmiş veriler üzerinde anomali tespiti gerçekleştirilmiş ve modellerin başarıları karşılaştırmalı olarak değerlendirilmiştir. Shuttle uzay verisi örneği, yüksek doğruluk oranı ve sınıf dengesizliği içeren yapısıyla, anomali tespiti uygulamaları için oldukça iyi bir örnek konumundadır.

Gerçek veri üzerinde yapılan analizlerde özellikle DT, RF ve XGBoost modelleri gibi ağaç tabanlı yöntemlerin LR ve KNN modellerine göre üstün performans sergilediği görülmüştür. Bu modeller, karmaşık karar sınırlarını öğrenmede ve anomali sınıfını doğru ayırt etmede oldukça başarılı olmuştur. LR ve KNN gibi daha temel yöntemler de uygun parametrelerle kullanıldığında tatmin edici sonuçlar üretmiştir.

Yapay veri analizinde ise modellerin genelde yüksek başarı gösterdiği, ancak parametre seçimine duyarlılıkları nedeniyle bazı modellerin performansının düştüğü tespit edilmiştir. Bu durum, özellikle yapay anomalilerin dağılım yapısına ve varyansına duyarlı modellerin dikkatli şekilde yapılandırılması gerektiğini ortaya koymaktadır.

Gerçek dünya verilerindeki sınıf dengesizliğine rağmen bazı makine öğrenmesi modelleri yüksek performans sergileyerek anomali tespitinde başarılı sonuçlar ortaya koyabilmektedir. Bununla birlikte, simülasyon yoluyla üretilen yapay veriler, algoritmaların performanslarını daha tarafsız ve kontrollü bir ortamda değerlendirme imkânı sunmaktadır. Elde edilen bulgular, özellikle parametre optimizasyonunun model başarısı üzerinde doğrudan etkili olduğunu göstermektedir. Bu nedenle, model seçimi yapılırken yalnızca algoritmanın yaygınlığı değil; aynı zamanda veri setinin yapısı, örneklem büyüklüğü ve anomali oranı gibi temel özellikler de göz önünde bulundurulmalıdır.

Son olarak, gerek bu çalışmanın gerekse ilerde yapılacak çalışma ve analizlerin savunma, uzay, havacılık ve benzeri alanlarda erken uyarı mekanizmalarının geliştirilmesine katkı sunabileceği değerlendirilmektedir.

TEŞEKKÜR (ACKNOWLEDGMENTS)

Bu araştırma hiçbir dış finansman almamıştır.

YAZAR KATKILARI (AUTHOR CONTRIBUTIONS)

Ünal DİKBAŞ: Kavramsal tasarım, Yazma, Veri düzenleme; Analiz

Meral EBEGİL: Metodoloji, Gözden geçirme ve Düzenleme

ÇIKAR ÇATIŞMALARI (CONFLICTS OF INTEREST)

Yazarlar, herhangi bir çıkar çatışması olmadığını beyan eder.

KAYNAKLAR (REFERENCES)

[1] Schwabacher, M., Oza, N., & Matthews, B. (2009). Unsupervised anomaly detection for liquid-fueled rocket propulsion health monitoring. Journal of Aerospace Computing, Information, and Communication, 6(7), 464–473. https://doi.org/10.2514/1.42783

[2] Goldstein, Markus and Uchida, Seiichi A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data, April 2016 PLoS ONE 11(4):e0152173 DOI:10.1371/journal.pone.0152173

[3] S. Shriram and E. Sivasankar, "Anomaly Detection on Shuttle data using Unsupervised Learning Techniques," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2019, pp. 221-225, doi: 10.1109/ICCIKE47802.2019.9004325

[4] Aly, M., Behiry, M.H. Enhancing anomaly detection in IoT-driven factories using Logistic Boosting, Random Forest, and SVM: A comparative machine learning approach. Sci Rep 15, 23694 (2025). https://doi.org/10.1038/s41598-025-08436-x

[5] Volnova, A.A. et al. (2024). Exploring the Universe with SNAD: Anomaly Detection in Astronomy. In: Baixeries, J., Ignatov, D.I., Kuznetsov, S.O., Stupnikov, S. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2023. Communications in Computer and Information Science, vol 2086. Springer, Cham. https://doi.org/10.1007/978-3-031-67826-4_15

[6] D'Addona, M., Riccio, G., Cavuoti, S., Tortora, C., Brescia, M. (2021). Anomaly Detection in Astrophysics: A Comparison Between Unsupervised Deep and Machine Learning on KiDS Data. In: Zelinka, I., Brescia, M., Baron, D. (eds) Intelligent Astrophysics. Emergence, Complexity and Computation, vol 39. Springer, Cham. https://doi.org/10.1007/978-3-030-65867-0_10

[7] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–242.

[8] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

[9] Quinlan, R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251

[10] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), 785–794. https://doi.org/10.1145/2939672.2939785

[12] Statlog (Shuttle) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5WS31.

[13] Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A Comparative Evaluation of Outlier Detection Algorithms: Experiments and Analyses. Pattern Recognition, 74, 406–421. https://doi.org/10.1016/j.patcog.2017.09.037