The Journal of International Scientific Researches
2025, 10(3)

Makine Öğrenmesi ile Kentsel Kiralık Bisiklet Verilerinde Anomali Tespiti: Konya İli Örneği

Detection of Anomalies in Urban Bike-Sharing Systems via Unsupervised Machine Learning: Konya Case Study

Özet

Bu çalışmada, şehir içi ulaşım ve bisiklet paylaşım sistemlerinde kalite kontrol ve güvenlik amaçlı veri doğrulama süreçlerine katkı sağlayabilecek bir yaklaşım sunulmaktadır. Bu amaçla, Konya kiralık bisiklet verisi kullanılarak denetimsiz makine öğrenmesi yöntemleriyle anomali tespiti gerçekleştirilmiştir. Analiz sürecinde iki yaygın algoritma olan yerel aykırı değer faktörü ve izolasyon ormanı yöntemleri kullanılmıştır. Veri setinde gözlem sayısı 1069 olup, anomalilerin tahmini oranı %3 olarak belirlenmiştir. Yerel aykırı değer faktörü ve izolasyon ormanı yöntemleri ile 33'er anomali tespit edilmiş, bunlardan 20 tanesi her iki model tarafından ortak olarak belirlenmiştir. Parametre duyarlılık analizleri sonucunda, her iki modelin de belirlenen anomali sayısında farklı parametre değerleri altında anlamlı bir değişiklik göstermediği gözlemlenmiştir. Bulgular, kullanılan yöntemlerin veri seti üzerinde tutarlı sonuçlar verdiğini ve özellikle konum verisi içeren ulaşım kayıtlarında fiziksel olarak mümkün olmayan değerlerin hızlı biçimde tespit edilebildiğini ortaya koymaktadır. Tespit edilen anomalilerin % 55'i hatalı GPS kayıtlarından kaynaklandığı, geri kalan kısmının ise kullanıcı davranış farklılıkları veya istatistiksel olarak nadir görülen durumlar olduğu sonucuna ulaşılmıştır.

Ünal Dikbaş
Gazi Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik ABD, Ankara, Türkiye, unal.dikbas@gazi.edu.tr
Orcid No: https://orcid.org/0000-0002-7851-535X

Meral Ebegil
Prof. Dr., Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara, Türkiye, mdemirel@gazi.edu.tr
Orcid No: https://orcid.org/0000-0003-4798-3422

Article Type / Makale Türü
Research Article / Araştırma Makalesi

Anahtar Kelimeler
Makine öğrenmesi, denetimsiz öğrenme, anomali tespiti, yerel aykırı değer faktörü, izolasyon ormanı

Keywords
Machine learning, unsupervised learning, anomaly detection, local outlier factor, isolation forest

Araştırma ve Yayın Etiği Beyanı
Etik kurul kararı gerektirmemektedir.

JEL Codes: *C38, C55, R41*

Submitted: 01 / 09 / 2025
Accepted: 03 / 11 / 2025

Abstract

This study presents an approach that can contribute to data validation processes for quality control and security in urban transportation and bicycle sharing systems. For this purpose, anomaly detection was performed using unsupervised machine learning methods using Konya rental bicycle data. Two common algorithms, the local outlier factor and isolation forest methods, were used in the analysis process. The number of observations in the data set was 1069, and the anomaly estimate was determined to be 3%. The local outlier factor and isolation forest methods detected 33 anomalies, 20 of which were identified in common by both models. Parameter sensitivity analyses revealed that neither model exhibited a significant change in the number of anomalies identified under different parameter values. The findings demonstrate that the methods used yield consistent results across the data set and that they can quickly detect physically impossible values, especially in transportation records containing location data. It has been concluded that 55% of the detected anomalies originated from erroneous GPS records, while the remaining anomalies were attributed to differences in user behavior or statistically rare occurrences

Giriş

Kent içi ulaşım sistemlerinin sürdürülebilirlik, çevresel etki ve kullanıcı sağlığı açısından yeniden şekillendiği günümüzde, bisiklet paylaşım sistemleri giderek daha yaygın bir kullanım alanı bulmaktadır. Bu sistemler sayesinde kullanıcılar kısa mesafeli ulaşım ihtiyaçlarını çevreci ve ekonomik bir şekilde karşılayabilmekte, belediyeler ise kent trafiğini rahatlatma ve karbon salımını azaltma hedeflerine katkı sağlayabilmektedir. Türkiye'de bu alanda öncü uygulamalardan biri olan

Konya Akıllı Bisiklet Sistemi (KABİS), kullanıcılar ile kamu otoriteleri arasında önemli bir veri akışı oluşturarak kent içi ulaşım analizlerinde kullanılabilecek zengin bir veri kaynağı sunmaktadır.

Büyük veri yapısına sahip bu tür sistemlerden elde edilen işlem kayıtları, kiralama süresi, bisikletle yapılan mesafe, başlangıç ve bitiş konumu gibi birçok değişkeni içermekte ve analiz edilmesi durumunda önemli sonuçlar sunmaktadır. Ancak bu veriler çeşitli nedenlerle hatalı, eksik veya olağan dışı gözlemler de içerebilmektedir. Örneğin, küresel konumlama sistemi (Global Positioning System-GPS) hataları, kullanıcı kaynaklı yanlış kullanımlar ya da teknik arızalar veri setinde anlamlı olmayan değerlerin oluşmasına neden olabilmektedir. Bu durum ise verilerde anomali problemini ortaya çıkarır. Bu tür anomaliler, hem hizmet kalitesinin takibi hem de sistem güvenliğinin sağlanması açısından tespit edilmesi gereken kritik unsurlardır.

Anomaliler, veri madenciliği ve makine öğrenmesi literatüründe, bir veri kümesinde normal kabul edilen örüntülerin dışında kalan ve çoğunluktan belirgin şekilde farklı özellikler sergileyen gözlemler olarak tanımlanmaktadır (Chandola vd., 2009). Bununla birlikte anomaliler bazı yazarlarca normalden büyük ölçüde değişiklik gösteren değerler olarak bazı yazarlarca da veri setinin kalan kısmından önemli ölçüde farklılık gösteren veriler olarak tanımlanmaktadır (Aggarwal, 2017: 1; Mehrotra, 2017: 4). Anomaliler, kimi zaman gerçek bir olağan dışı olayı yansıtırken, kimi zaman da ölçüm hataları, sensör bozuklukları ya da kayıt sistemlerindeki teknik aksaklıklardan kaynaklanmaktadır.

Anomali tespitinin önemi, farklı disiplinlerdeki uygulama alanlarıyla daha da belirginleşmektedir. Finansal piyasalarda olağandışı işlem hareketlerinin saptanması, sağlık verilerinde beklenmeyen ölçüm değerlerinin belirlenmesi, siber güvenlikte saldırı girişimlerinin erken teşhisi ve akıllı ulaşım sistemlerinde GPS verilerindeki uyumsuzlukların ayıklanması, bu yöntemlerin kullanıldığı başlıca alanlardır (Chandola vd., 2009).

Anomali tespiti yalnızca veri işleme sürecinin bir aşaması değil aynı zamanda sistemlerin güvenilirliğini, verimliliğini ve sürdürülebilirliğini artırmaya yönelik stratejik bir adım olarak görülmektedir. Bu nedenle günümüzde, denetimsiz makine öğrenmesi algoritmaları farklı veri yapılarında yaygın olarak uygulanmakta ve literatürde giderek daha fazla önem kazanmaktadır. Literatürde farklı denetimsiz algoritmaların karşılaştırıldığı çalışmalar bulunmaktadır.

Falcão vd. (2019), beş farklı veri seti üzerinde 12 denetimsiz anomali tespit algoritmasını karşılaştırarak kapsamlı bir deneysel analiz yapmıştır. Sonuçlar, izolasyon ormanı (Isolation Forest - IF) ve tek sınıf destek vektör makineleri (One-Class Support Vector Machine- OCSVM) gibi sınıflandırma tabanlı yöntemlerin yüksek doğruluk sağladığını, yerel aykırı değer faktörü (Local Outlier Factor - LOF) ve bağlantılılık tabanlı aykırı değer faktörü (Connectivity-Based Outlier Factor - COF) gibi yoğunluk tabanlı yöntemlerin ise parametre ayarlarına duyarlı olmakla birlikte farklı veri türlerinde etkili olabildiğini göstermiştir. Nowak-Brzezińska ve Horyńa (2020) ise LOF, COF ve K-ortalamalar (K-Means) algoritmalarını bilgi tabanlarında aykırı kural tespiti için karşılaştırmış ve COF algoritmasının bazı durumlarda daha tutarlı sonuçlar ürettiğini göstermiştir. Bu bulgu, farklı veri setlerinde farklı algoritmaların öne çıkabileceğini, dolayısıyla yöntem seçiminin bağlama göre uyarlanması gerektiğini ortaya koymaktadır.

Petrariu vd. (2022), LOF'un da aralarında bulunduğu denetimsiz algoritmaları karşılaştırmış, LOF'un yerel anomalilerde daha başarılı olduğunu histogram tabanlı aykırılık skoru (Histogram-Based Outlier Score – HBOS)'nun ise hız açısından öne çıktığını belirtmiştir. Yazarlar, tek bir algoritmanın tüm veri türleri için en uygun çözümü sunmadığını, bu nedenle farklı yöntemlerin birlikte değerlendirilmesinin önemli olduğunu vurgulamıştır. Benzer biçimde, Adesh vd. (2024), LOF algoritmasını büyük veri platformu olan bir sistem üzerinde uygulamış ve "Improved LOF" adını verdikleri geliştirilmiş bir versiyon önermiştir. Çalışmada, LOF'un özellikle hiperparametre seçimlerine karşı hassasiyet taşıdığı ve farklı veri çoğaltmalarında performans değişiklikleri gösterebildiği tespit edilmiştir.

Bu çalışmaların tümü, anomali tespitinde tek bir yönteme bağlı kalmanın yetersiz olabileceğini, algoritmaların veri yapısına göre farklı avantajlar sunduğunu ve özellikle LOF gibi yerel yoğunluk tabanlı yöntemlerin ve IF gibi ağaç tabanlı bir modelin farklı veri türlerinde etkili olduğunu göstermektedir. Özellikle denetimsiz öğrenme (unsupervised learning) yaklaşımları, veri

etiketlerinin bulunmadığı durumlarda otomatik olarak olağan dışı gözlemleri tespit edebilme kapasitesine sahiptir. Bu bağlamda, LOF ve IF algoritmaları, yüksek doğrulukla anomali belirleyebilen iki güçlü denetimsiz öğrenme tekniği olarak öne çıkmaktadır.

Bu çalışmada, Konya kiralık bisiklet sistemine ait 2022–2023 yılına ilişkin gerçek kullanıcı verileri üzerinde, LOF ve IF algoritmaları kullanılarak anomali tespiti gerçekleştirilmiştir. Modeller yalnızca kiralama süresi ve bisikletle yapılan mesafe bilgileri üzerinden eğitilmiş olup, etiketli veri ihtiyacı olmadan doğrudan ham veriye uygulanmıştır. Çalışmanın amacı, verideki olağan dışı kiralama kayıtlarının belirlenmesi ve bu kayıtların sistemsel ya da kullanıcı kaynaklı olası nedenlerinin yorumlanmasıdır. Ayrıca, iki farklı denetimsiz modelin çıktıları karşılaştırılarak anomali tespitindeki örtüşmeler ve farklılıklar değerlendirilmiştir.

1. Metodoloji

Bu çalışmada, denetimsiz anomali tespiti yöntemlerinden LOF ve IF algoritmaları uygulanmıştır. Her iki yöntem de veri üzerinde herhangi bir ön tanımlı sınıf etiketi gerektirmeden çalışabilmekte ve olağan dışı gözlemleri istatistiksel veya yapısal farklılıklarına dayanarak tespit etmektedir. Bu kısımda çalışmada kullanılan LOF ve IF yöntemleri açıklanmıştır.

1.1. LOF yöntemi

LOF yöntemi, her bir gözlemin bulunduğu çevredeki diğer gözlemlerle olan yerel yoğunluğunu karşılaştırarak anomali skoru üretir. Komşu örneklerin yoğunluğu ile kendi yoğunluğu arasındaki fark ne kadar fazlaysa, ilgili gözlemin anomali olma ihtimali o kadar yüksektir. Eşitlik (1)'de modelin temel formülü yer almaktadır.

$$LOF(x_i) = \frac{\sum_{x_j \in N_k(x_i)} \frac{lrd(x_j)}{lrd(x_i)}}{|N_k(x_i)|} \quad (1)$$

Burada, x_i: i'inci gözlem, x_j: j'inci gözlem, k : komşu sayısı, $N_k(x_i)$: x_i'nin k en yakın komşularının oluşturduğu kümeyi, $|N_k(x_i)|$: bu kümenin içerdiği eleman sayısı, $lrd(x_i)$: x_i'nin yerel yoğunluk oranını, $lrd(x_j)$: x_j'nin yerel yoğunluk oranını, $\frac{lrd(x_j)}{lrd(x_i)}$: komşu x_j'nin komşu x_i'ye oranını göstermektedir (Breunig vd., 2000). Burada yer alan yerel yoğunluk $lrd(x_i)$ ise eşitlik (2) yardımıyla hesaplanmaktadır.

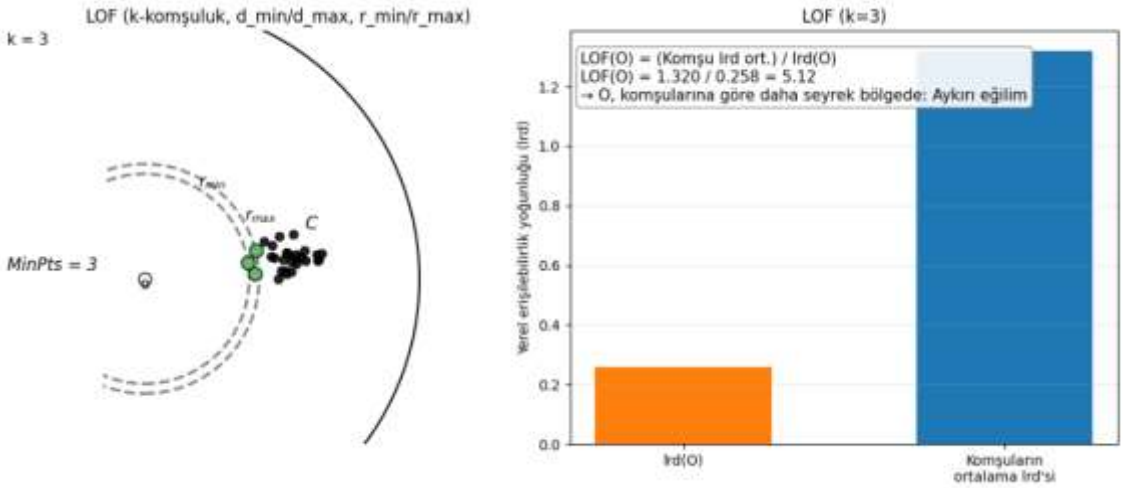
$$lrd(x_i) = \frac{\sum_{x_j \in N_k(x_i)} reach - dist(x_i, x_j)}{|N_k(x_i)|} \quad (2)$$

Eşitlik (2)'de yer alan $reach - dist(x_i, x_j)$ ifadesi x_i gözlemi ile x_j gözlemi arasındaki erişilebilirlik uzaklığını göstermekte olup eşitlik (3) ile hesaplanır.

$$reach - dist(x_i, x_j) = \max\{dist(x_j), d(x_i, x_j)\} \quad (3)$$

Eşitlik (3)'te yer alan $d(x_i, x_j)$ ifadesi ise x_i ve x_j gözlemleri arasındaki gerçek uzaklığı göstermektedir.

LOF algoritmasının temel çalışma mantığı şekil 1'de açıkça gösterilmektedir.



Şekil 1. LOF Yönteminin Görseli

Sol taraftaki şekilde, $k = 3$ (MinPts = 3) için seçilen LOF skoru hesaplanacak O gözleminin en yakın komşuları d_min, d_max ve bu komşulara olan mesafe sınırları r_min, r_max ile gösterilmektedir. Burada O gözlemine ait konumun komşularına kıyasla daha seyrek bir bölgede bulunduğu dikkat çekmektedir. Sağ taraftaki çubuk grafikte ise O gözleminin yerel erişilebilirlik yoğunluğu (lrd) ile komşularının ortalama lrd değeri karşılaştırılmıştır. O gözlemi için LOF skoru Eşitlik (4)'deki gibi hesaplanır.

$$LOF(0) = \frac{Komsu\ lrd\ ort}{lrd(0)} = \frac{1.320}{0.258} \approx 5.12 \quad (4)$$

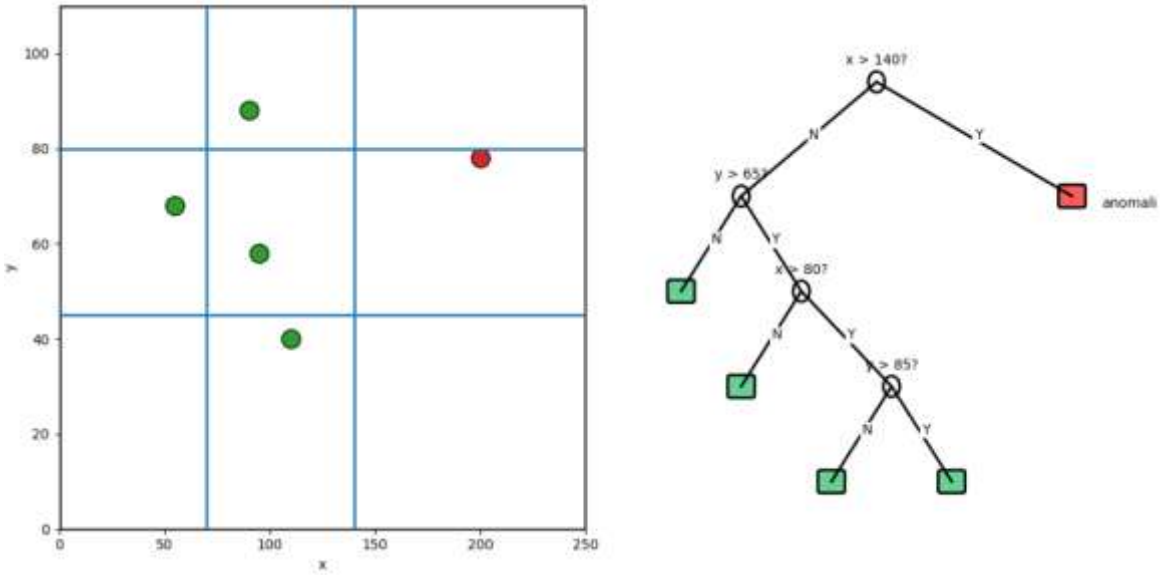
Buna göre, O gözlemi için LOF skoru 5.12 olarak elde edilir. LOF skorunun 1'e yakın olması normal gözlemleri, 1'den büyük değerler ise artan şekilde anomali olasılığını göstermektedir. Burada bu değer 1'den belirgin şekilde büyük olduğundan, O noktasının komşularına göre daha seyrek bölgede yer aldığı ve aykırı eğilim sergilediğini göstermektedir. Bu analiz, LOF yönteminin yalnızca mutlak mesafelere değil, yerel yoğunluk farklarına odaklanarak anomalileri tespit edebildiğini açıkça göstermektedir.

1.2. IF yöntemi

IF yöntemi, gözlemleri rastgele seçilen öznitelikler üzerinden bölerek izole etmeye çalışır. Normal gözlemler, çok sayıda bölmeye ihtiyaç duyar ancak anormal gözlemler ise kısa sürede izole edilebilir. Model, bu izolasyon derinliği üzerinden anomali skoru üretir. Eşitlik (5)'te modelin anomali skorunun elde edilmesini gösteren temel denklem yer almaktadır.

$$S(x_i) = 2^{\frac{-E(h(x_i))}{c(n)}} \quad (5)$$

Burada n toplam veri noktaları sayısı olmak üzere, x_i: i'inci gözlem, $S(x_i)$: x_i gözlemi için anomali skorunu, $h(x_i)$: x_i gözleminin bir izolasyon ağacındaki yol uzunluğunu, $E(h(x_i))$: tüm ağaçlardaki beklenen ortalama yol uzunluğunu ya da ortalama bölünme sayısını, $c(n)$: n örnek sayısına bağlı normalleştirme katsayısını göstermektedir. IF yönteminde, normalize edilmiş $S(x_i)$ skorlarının belirlenmiş bir eşik değeri ile kıyaslanmasıyla bir gözlemin anomali olup olmadığına karar verilir (Liu vd., 2008). Şekil 2'de, modelin iki boyutlu veri üzerinde anomali tespit süreci görselleştirilmektedir.



Şekil 2. IF Yönteminin Görseli

Sol tarafta, veri noktalarının x ve y eksenlerinde dağılımı ile karar sınırları gösterilmektedir. Yeşil noktalar normal gözlemleri, kırmızı nokta ise model tarafından anomali olarak etiketlenen gözlemi temsil etmektedir. Sağ tarafta ise karar ağacının yapısı yer almakta, her düğümde belirli bir eşik değerine göre (örneğin x>140, y>65) veri ayrılmakta ve en sonunda yaprak düğümlerde sınıflandırma yapılmaktadır. Anomali olarak belirlenen kırmızı yaprak, ilgili gözlemin karar ağacındaki yolculuğunu ve modelin onu diğerlerinden ayırma mantığını açıkça göstermektedir. Anomali gözlemin en hızlı şekilde ayrılarak izole edilmesi modeli oldukça etkin ve güçlü kılmaktadır. Bu görselleştirme, IF modeline temel teşkil eden karar ağacı algoritmasının ayrıştırma mantığını sezgisel olarak anlamak açısından oldukça faydalıdır.

Uygulamada kullanılan yöntemlerin sadece teknik işleyişleri değil, aynı zamanda güçlü ve sınırlı yönlerinin de değerlendirilmesi önem arz etmektedir. Bu çalışmada kullanılan LOF ve IF yöntemlerinin güçlü ve sınırlı yönleri de göz önünde bulundurulmuştur. LOF yöntemi, verideki yerel yoğunluk farklılıklarını dikkate alarak özellikle küçük kümelerdeki aykırılıkları başarılı bir şekilde ortaya çıkarabilmektedir. Bununla birlikte, parametre duyarlılığı nedeniyle bazı veri setlerinde istikrarsız sonuçlar üretebilmesi dezavantaj olarak değerlendirilmektedir. IF yöntemi ise, yüksek boyutlu ve büyük ölçekli veri setlerinde hızlı ve verimli sonuçlar üretmesi, ayrıca parametre ayarlarına karşı görece daha az hassas olması yönleriyle öne çıkmaktadır. Ancak rastgele ağaç yapılarının kullanılması, özellikle küçük örneklemlerde sonuçların tutarlılığı açısından sınırlılıklar yaratabilmektedir. Dolayısıyla her iki yöntem, güçlü yönleri sayesinde kentsel bisiklet verilerinde anomalilerin tespitinde tamamlayıcı nitelikte katkılar sunmaktadır.

2. Uygulama

Bu çalışmada analiz edilen veri kümesi, Konya Büyükşehir Belediyesi tarafından 2022–2023 yılları arasında sunulan kiralık bisiklet hizmetine ilişkin işlem kayıtlarını içermektedir. Veriler Konya Büyükşehir Belediyesi Açık Veri Portalı'ndan alınmıştır (Konya Büyükşehir Belediyesi, 2025). Veri setinde her bir gözlem, bir bisikletin kiralandığı bir seansı temsil etmektedir. Kayıtlar arasında bisiklet kimliği, kiralama başlangıç ve bitiş tarih/saat bilgileri, kullanım süresi (Duration), ödenen ücret (Fare), başlangıç ve bitiş noktalarının coğrafi koordinatları ile tahmini gidilen mesafe (Distance Traveled) gibi değişkenler yer almaktadır.

Veri ön işleme aşamasında, süre değişkeni saniye cinsine çevrilmiş (Duration_seconds), ayrıca hatalı ya da eksik değerler temizlenmiştir. Mesafe bilgisi "Distance Traveled (m)" sütunu üzerinden doğrudan kullanılmıştır. Kullanım süresi ve kat edilen mesafe anomali tespiti açısından anlamlı görüldüğünden, analiz sürecinde temel öznitelikler olarak değerlendirilmiştir. Veri setinden beş

gözlemin örnek olarak gösterildiği tablo 1'de ön işlemeye tabi tutulmuş ve analizde kullanılacak veri setine ilişkin bilgiler yer almaktadır.

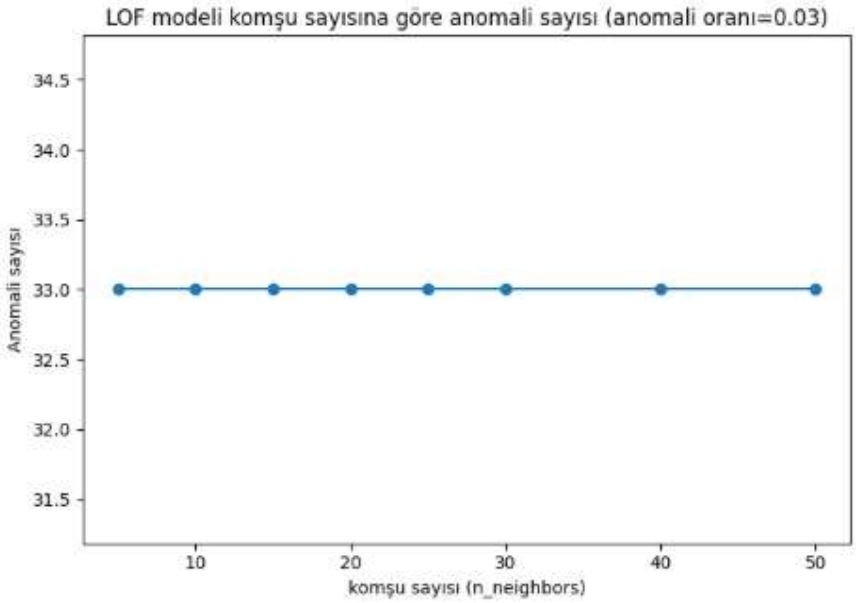
Tablo 1. Veri Setine İlişkin Bilgiler

Bisiklet kimlik numarası	Kiralama başlangıç zamanı	Kiralama bitiş zamanı	Süre (sn)	Mesafe (metre)
1071	2023-12-13 10:01:00	2023-12-13 10:12:00	649	1671
1850	2023-12-13 09:42:00	2023-12-13 09:48:00	384	937
1387	2023-12-13 08:10:00	2023-12-13 08:17:00	470	878
1613	2023-12-13 07:26:00	2023-12-13 07:32:00	383	1174
1666	2023-12-13 07:20:00	2023-12-13 07:29:00	567	2145
….	….	….	….	….

Veri kümesinde ücret bilgisi, başlangıç ve bitiş istasyonları gibi diğer değişkenler analiz dışında bırakılmış, modeller yalnızca süre ve mesafe değerleri üzerinden eğitilmiştir. Bu seçim, denetimsiz öğrenme algoritmalarının yapısına uygun olarak gözlemlerin yalnızca sayısal özellikleriyle öğrenilmesine dayanmaktadır.

Makine öğrenme algoritmaları tahminde kullanılan parametrelere oldukça duyarlıdır. Doğru parametreler ile daha doğru ve isabetli sonuçlar elde edilmektedir. Bu kapsamda modellerin önemli görülen parametrelere karşı tepkisini incelemek amacıyla gerekli parametre ayarlaması analizi yapılmıştır. Her iki model için kullanılacak en önemli parametrelerden biri veri setindeki anomali oranıdır. Denetimsiz öğrenmede etiketli veri olmadığından bu oran kullanıcı tarafından belirlenmektedir. Bu veri seti için anomali oranı %3 olarak belirlenmiştir. Anomaliler tanımı gereği nadir rastlanan gözlemlerdir. Literatürde farklı veri türlerinde olduğu gibi, finansal ve ulaştırma verileri gibi zaman serilerinde ve işlem kayıtlarında da anomaliler oldukça düşük sıklıkla ortaya çıkmaktadır (Chandola vd., 2009). Çok yüksek bir oran seçilmesi durumunda normal gözlemlerin yanlışlıkla anomali olarak etiketlenmesi riski artarken, çok düşük oran seçimi gerçek anomalilerin gözden kaçmasına yol açabilmektedir. %3 seviyesi, hem hatalı pozitif hem de hatalı negatif risklerini makul düzeyde tutan dengeleyici bir oran olarak tercih edilmiştir.

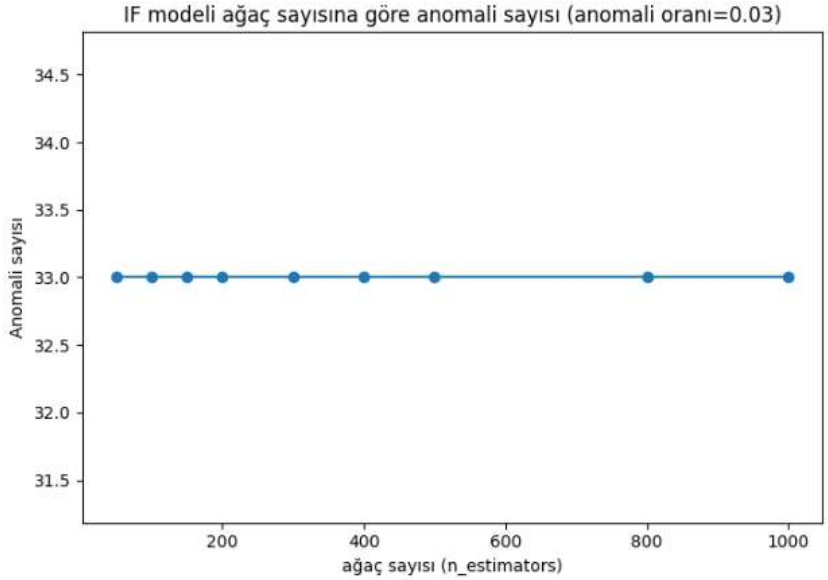
LOF modeli anomali skorları üretirken anomali oranı dışında kullandığı en önemli parametre komşu sayısıdır (n_neighbors) (Alghushairy vd., 2021). Bu nedenle en uygun sayının belirlenmesi gerekmektedir. Şekil 3, LOF modelinde komşu sayısı parametresinin, tespit edilen anomali sayısına etkisini göstermektedir. Analizde, farklı komşu sayılarında (5 ile 50 arasında) anomali sayısının 33'te sabit kaldığı görülmektedir. Bu durum, kullanılan veri setinde LOF modelinin komşu sayısı değişiminden etkilenmediğini ve modelin istikrarlı bir şekilde aynı gözlemleri anomali olarak tanımladığını ortaya koymaktadır. Parametreye karşı bu tür bir duyarsızlık, modelin belirli veri yapılarında parametre optimizasyonuna düşük bağımlılıkla çalışabildiğini göstermesi açısından önemlidir.



Şekil 3. OF Modeli Komşu Sayısı ve Tespit Edilen Anomali Sayısı

Bu çalışmada LOF modeli 20 komşu (n_neighbors=20) üzerinden çalıştırılmış ve %3 oranında anomali tespiti yapılması hedeflenmiştir (contamination=0.03). Komşu sayısı ve anomali oranının yanı sıra modelin diğer parametreleri ise her bir yaprak düğümünde en fazla kaç gözlemin olacağını belirleyen yaprak düğümü 30 (leaf size= 30), noktalar arasındaki mesafe ölçümü için minkowski (metric= minkowski) olarak belirlenmiştir.

IF modelinde ise tahmin skorlarını etkileyen en önemli parametre ağaç sayısıdır (n_estimators). Şekil 4, IF modelinde kullanılan ağaç sayısının tespit edilen anomali sayısına etkisini göstermektedir. Anomali oranının sabit olarak %3 tutulduğu bu çalışmada, ağaç sayısı 100'den 1000'e kadar artırıldığında tespit edilen anomali sayısının 33 olarak sabit kaldığı görülmektedir. Bu durum, modelin belirli bir anomali yüzdesinde ağaç sayısından bağımsız olarak istikrarlı sonuçlar ürettiğini ve parametre değişiminin anomali sayısını etkilemediğini ortaya koymaktadır. Dolayısıyla, ağaç sayısının artırılması bu senaryoda modelin tespit gücünü değiştirmemekte, yalnızca hesaplama süresini etkileyebilmektedir.



Şekil 4. IF Modeli Komşu Sayısı ve Tespit Edilen Anomali Sayısı

Bu çalışmada IF modeli de yine %3 anomali oranı ve 100 ağaç sayısı ile çalıştırılmıştır. Modelin diğer parametreleri de bootstrap=False, max features=1.0, max_samples=auto olarak belirlenmiştir. Bu ayarlar, modelin örnekleme sürecinde tüm özelliklerin kullanılmasını ve örneklerin otomatik olarak seçilmesini sağlamaktadır

Her iki model için Python programlama dili ve scikit-learn kütüphanesi kullanılarak analizler yapılmış ve elde edilen model tahmin sonuçları bulgular kısmında paylaşılmıştır.

3. Bulgular

Model tahmin sonuçları tablo 2'de paylaşılmıştır. Analizde toplam 1069 gözlem değerlendirilmiş ve anomali oranı parametre değeri %3 olarak belirlenmiştir. Bu orana göre beklenen anomali sayısı yaklaşık 32'dir.

Tablo 2. Analizde Kullanılan Veriler ve Model Tahmin Sonuçları

Ölçüt	Değer
Analizde kullanılan toplam gözlem	1069
Belirlenen anomali yüzdesi	0.03
Beklenen anomali sayısı	32
LOF modeli tarafından tespit edilen anomali sayısı	33
IF modeli tarafından tespit edilen anomali sayısı	33
Her iki modelce tespit edilen ortak anomali sayısı	20
Sadece LOF modelince tespit edilen anomali sayısı	13
Sadece IF modelince tespit edilen anomali sayısı	13

Elde edilen sonuçlara göre hem LOF hem de IF modelleri ayrı ayrı 33 anomali tespit etmiştir. Çalışmada kullanılan her iki model için varsayılan anomali oranı %3 olarak belirlenmiştir. Bu oran, 1069 gözlem içeren temizlenmiş veri kümesi üzerinde yaklaşık 32 anomali gözlem öngörmektedir. Gerçekleşen sonuçlar, bu varsayımla büyük oranda örtüşmektedir.

Bu anomalilerden 20 tanesi her iki model tarafından ortak olarak belirlenmiş olup, bu değer modellerin tespit ettiği anomalilerde belirli bir ölçüde örtüşme olduğunu göstermektedir. Öte yandan, LOF ve IF modelleri tarafından ayrı ayrı tespit edilen farklı 13 anomali gözlem bulunmaktadır. Bu durum, her iki modelin istatistiksel öğrenme yaklaşımındaki farklılıkların, bazı gözlemlerin yalnızca bir model tarafından anomali olarak değerlendirilmesine yol açtığını göstermektedir. LOF, gözlemleri komşuluk yoğunluklarına göre değerlendirerek yerel farklılıklara odaklanmaktadır. Bu nedenle bazı gözlemler yalnızca LOF modeli tarafından aykırı olarak işaretlenmiştir. Buna karşılık IF modeli ise örnekleri rastgele ağaç bölmeleriyle izole etme mantığına dayandığından, daha çok kolay izole edilebilen değerleri anomali gözlemler olarak tespit etmiştir.

Genel olarak, iki modelin de beklenen 32 anomali gözlem sayısına oldukça yakın sonuçlar üretmesi, anomali yüzdesinin uygun seçildiğini ve tespitlerin veri setinin genel yapısıyla uyumlu olduğunu ortaya koymaktadır. Ortak anomalilerin sayısının yüksek olması, modellerin birlikte kullanımının güvenilir sonuçlar sağlayabileceğini düşündürmektedir. Buna karşılık, yalnızca bir modelin tespit ettiği anomaliler, farklı algoritmik yaklaşımların güçlü yönlerinin tamamlayıcı şekilde kullanılmasının önemine işaret etmektedir.

Bu sonuç, LOF ve IF algoritmalarının hem anomali yüzdesi parametresine duyarlı şekilde çalıştığını hem de veri setindeki olağan dışı örüntüleri yüksek oranda benzer şekilde tanımladığını göstermektedir. Modellerin örtüşmeyen 13'er anomali gözlemleri ise, her bir algoritmanın farklı anomali tanımına sahip olmasından kaynaklanmakta olup analiz için tamamlayıcı niteliktedir. Örneğin, LOF ve IF modelleri tarafından aynı anda anomali olarak işaretlenen 20 gözlemden 10 tanesi tablo 3'te sunulmuştur. Model çıktıları incelendiğinde, hem LOF hem de IF algoritmalarının birçok ortak anomali gözlemi tespit ettiği görülmüştür. Özellikle bazı gözlemler, imkânsız seviyede yüksek mesafe ya da çok uzun süreli kiralama gibi yapısal bozukluklar içermekte ve her iki yöntem tarafından da açık şekilde anomali olarak tespit edilmiştir.

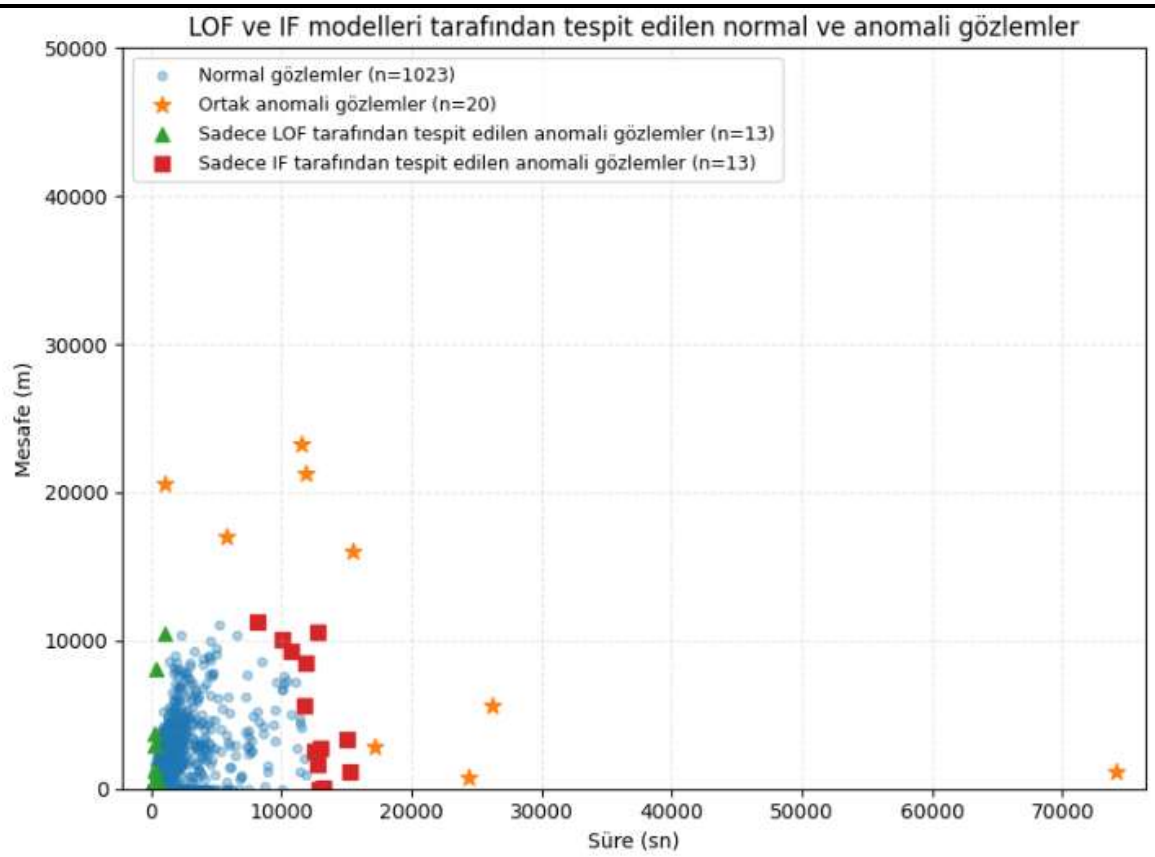
Tablo 3. Her İki Yöntem Tarafından Tespit Edilen Anomaliler

Bisiklet kimlik numarası	Süre(sn)	Mesafe (metre)	LOF	IF	Muhtemel sebep
1452	1003	14563512	✓	✓	imkansız mesafe (muhtemel GPS/kayıt hatası)
1459	1209	14566115	✓	✓	imkansız mesafe (muhtemel GPS/kayıt hatası)
1071	17134	2880	✓	✓	istatistiksel anomali (LOF & IF)
1602	26197	5653	✓	✓	istatistiksel anomali (LOF & IF)
1427	547	14561747	✓	✓	imkansız mesafe (muhtemel GPS/kayıt hatası)
1511	1046	20544	✓	✓	istatistiksel anomali (LOF & IF)
1076	1053	14560382	✓	✓	imkansız mesafe (muhtemel GPS/kayıt hatası)
1375	15491	16004	✓	✓	istatistiksel anomali (LOF & IF)
1458	24376	774	✓	✓	çok uzun süre / düşük mesafe
1557	1344	14558539	✓	✓	imkansız mesafe (muhtemel GPS/kayıt hatası)

Tablo 3 incelendiğinde kayıtların %50'den fazlası, "imkânsız mesafe (muhtemel GPS /kayıt hatası)" kategorisinde toplanmıştır. Örneğin, 1452, 1459, 1427, 1076 ve 1557 numaralı bisikletler, kısa sürelerde olağanüstü uzun mesafeler kaydetmiş olup bu durum veri toplama sürecinde GPS hatası veya kayıt bozukluğuna işaret etmektedir.

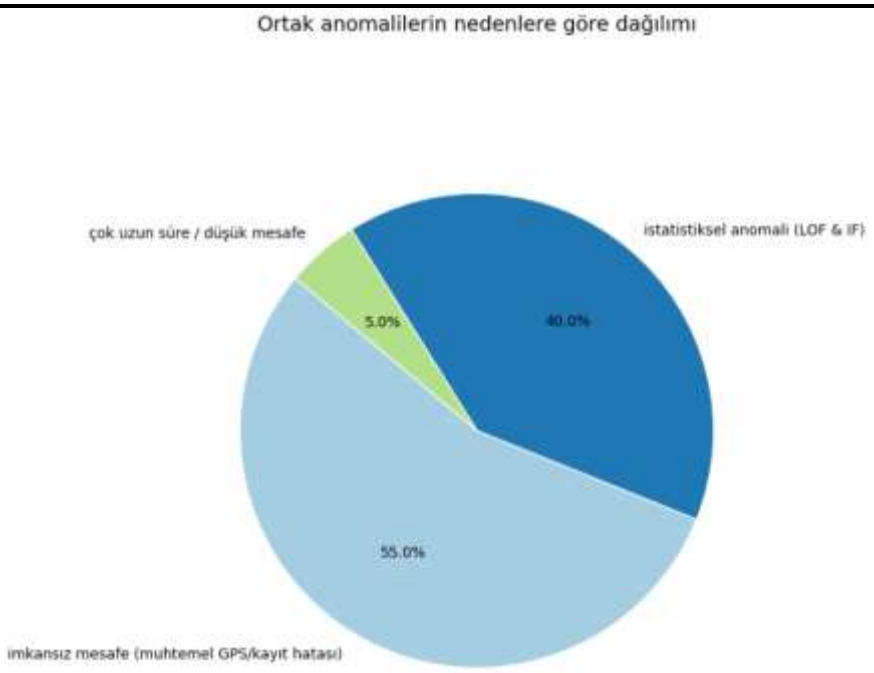
Diğer bir grup, istatistiksel anomali (LOF & IF) olarak sınıflandırılmıştır. Tablo 3'te bu kategoride yer alan 1071, 1602, 1511, 1375 numaralı bisikletler veri dağılımının istatistiksel yapısından belirgin şekilde sapma göstermektedir (Ayrıntı için bkz. ek tablo 4). Bu tür anomalilerin, kullanıcı davranışlarındaki olağandışı örüntülerden (örneğin uzun süreli kullanımda düşük mesafe gibi) veya küçük sistemsel tutarsızlıklardan kaynaklanabileceği düşünülmektedir.

Son olarak, 1458 numaralı bisiklet için "çok uzun süre / düşük mesafe" anomalisi gözlenmiştir. 24376 saniye (yaklaşık 6.8 saat) kullanım süresine rağmen yalnızca 774 metre yol alınması, bisikletin uzun süre hareketsiz kaldığını veya kullanım amacına uygun olmayan bir şekilde işletildiğini göstermektedir. Bu bulgular, hem sensör/veri kaynağı hatalarını hem de kullanıcı davranışlarından kaynaklı olağandışı durumları ortaya koymaktadır. Dolayısıyla şehir içi bisiklet kullanım verilerinde anomali tespitinin, hem teknik hem de operasyonel açıdan kritik önem taşıdığını göstermektedir. Bu gözlemler arasında çok kısa sürede yüksek mesafe kat eden, ya da çok uzun süre kiralanmasına rağmen neredeyse hiç hareket etmeyen bisikletler gibi dikkat çeken başka örnekler de vardır.



Şekil 5. LOF ve IF Modelleri Tarafından Tespit Edilen Normal ve Anomali Gözlemler

Her iki model tahmin sonuçları ve normal gözlemlerin yer aldığı şekil 5 incelendiğinde, IF algoritmasının daha belirgin anomalileri yakalama konusunda etkin olduğu, LOF modelinin ise daha yerel örüntüleri algılamada başarılı olduğu gözlemlenmiştir. Bununla birlikte, hem LOF hem de IF modeli tarafından ortaklaşa anomali olarak işaretlenen gözlemlerin kümelenmenin dışında yer aldığı ve süre–mesafe ilişkisi açısından belirgin sapmalar sergilediği görülmektedir. Sadece LOF tarafından belirlenen anomaliler daha çok kısa süreli ancak göreli olarak yüksek mesafeli gözlemleri işaret ederken, sadece IF modelinin tespit ettiği anomaliler ise özellikle uzun süreli kullanımlarda ortaya çıkan tutarsızlıklardır. Bu durum, iki yöntemin farklı duyarlılıklara sahip olduğunu ve anomalilerin doğasına ilişkin tamamlayıcı bilgiler sunduğunu göstermektedir. Dolayısıyla, iki modelin birlikte kullanımı hem yapısal hem de istatistiksel anomalilerin daha güvenilir şekilde belirlenmesini sağlamaktadır.



Grafik 1. Ortak Anomalilerin Nedenlere Göre Dağılımı

Grafik 1'de LOF ve IF modelleri tarafından ortak olarak tespit edilen anomalilerin nedenlere göre dağılımı incelendiğinde üç temel kategori ortaya çıkmıştır. En yüksek paya sahip olan "imkânsız mesafe (muhtemel konum bilgisi/kayıt hatası)" grubu, tüm anomalilerin %55'ini oluşturmaktadır. Bu bulgu, sistemde yer alan bazı kayıtların olağan dışı yüksek mesafe değerleri içerdiğini göstermekte olup, büyük olasılıkla GPS sinyal hataları, veri giriş hataları veya sistemsel kayıt sorunlarından kaynaklanmaktadır.

İkinci sırada yer alan "istatistiksel anomali (LOF & IF)" kategorisi, anomalilerin %40'ını oluşturmaktadır. Bu grup, fiziksel olarak tutarsızlık göstermeyen, ancak her iki modelin de normal kullanım örüntülerinden belirgin şekilde sapma olarak değerlendirdiği gözlemlerden oluşmaktadır. Bu durumun, kullanıcı davranışlarındaki alışılmadık örüntülerden, nadir görülen kullanım biçimlerinden veya veri setindeki doğal değişimlerden kaynaklandığı düşünülmektedir. Dolayısıyla bu istatistiksel anomaliler, hem veri kalitesi sorunlarının hem de kullanıcı davranış çeşitliliğinin bir yansıması olarak değerlendirilmektedir.

En düşük orana sahip olan "çok uzun süre / düşük mesafe" kategorisi ise toplam anomalilerin %5'ini kapsamaktadır. Bu tür anomaliler, genellikle bisikletin kiralanmasına rağmen kullanılmadan uzun süre elde tutulması veya park halinde bırakılması ile ilişkilendirilmektedir. Bu durum, sistem kaynaklarının etkin kullanımını olumsuz yönde etkileyebilecek operasyonel bir sorun olarak değerlendirilebilir.

Genel olarak değerlendirildiğinde, tespit edilen anomalilerin önemli bir kısmının veri kaynağı veya sensör hatalarından kaynaklandığı, geri kalan kısmının ise kullanıcı davranış farklılıkları veya istatistiksel olarak nadir görülen durumlar olduğu anlaşılmaktadır. Bu bulgular, veri toplama ve kayıt süreçlerinin doğruluğunun artırılması, GPS sistemlerinin iyileştirilmesi ve uzun süreli boş kullanım vakalarının önlenmesine yönelik operasyonel düzenlemelerin önemini ortaya koymaktadır.

Sonuç ve Değerlendirme

Bu çalışmada, Konya ilinde 2022–2023 yılları arasında gerçekleştirilen kiralık bisiklet kullanımına ait veriler üzerinde anomali tespiti gerçekleştirilmiştir. Denetimsiz öğrenme yöntemlerinden LOF ve IF algoritmaları, yalnızca süre ve mesafe bilgilerine dayanarak olağan dışı kiralama örneklerini belirlemiştir.

Elde edilen bulgular, bazı gözlemlerde hatalı GPS verileri nedeniyle ortaya çıkan imkânsız mesafelerin, bazı durumlarda ise çok uzun süreli kullanıma rağmen neredeyse hiç hareket edilmemiş olmasının anomali kullanım örüntüleri arasında yer aldığını göstermiştir. LOF ve IF modelleri çoğu durumda benzer gözlemleri anomali olarak etiketlemiştir. Ancak bazı örneklerde farklı kararlar vererek modellerin birbirini tamamlayıcı nitelikte olduğunu ortaya koymuştur. Tespit edilen anomalilerin büyük kısmının hatalı GPS kayıtları veya kullanıcı davranış farklılıklarından kaynaklanması, bakım süreçlerinin planlanarak operasyonel verimliliğin artırılmasında ve veri kalitesinin iyileştirilmesinde dikkate alınması gereken kritik unsurlar olarak öne çıkmaktadır. Sonuçlar ayrıca belediyelerin ve sistem işletmecilerinin daha güvenilir hizmet sunabilmeleri için hem teknik altyapı hem de kullanıcı odaklı önlemleri güçlendirmesi gerektiğini ortaya koymaktadır. Gelecek çalışmalarda farklı şehirlerden elde edilecek benzer bisiklet paylaşım verilerinin karşılaştırmalı analizleri, yöntemlerin genellenebilirliğini test etme fırsatı sunacaktır. Böylece, hem şehir içi ulaşım politikaları hem de akıllı ulaşım sistemlerinin sürdürülebilirliği açısından daha bütüncül bir bakış açısı geliştirilebilir.

Kaynakça

Adesh, S., Shobha, G., Shetty, N., & Xu, S. (2024). Improved LOF for anomaly detection on big data using HPCC Systems. *Journal of Parallel and Distributed Computing, 179*, 150–163. https://doi.org/10.1016/j.jpdc.2024.01.00

Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). Cham: Springer.

Alghushairy, O., Hussain, F. K., Hussain, O. K., & Alfarhood, S. (2021). A review of Local Outlier Factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing, 5*(1), 1–22. https://doi.org/10.3390/bdcc5010001

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM SIGMOD Record, 29*(2), 93–104. https://doi.org/10.1145/342009.335388

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Falcão, A. E. Z., Rocha, R. L. S., Falcão, D. S. C., & Rocha, A. R. C. (2019). Quantitative comparison of unsupervised anomaly detection algorithms for intrusion detection. *Journal of Information Security and Applications, 46*, 1–12. https://doi.org/10.1016/j.jisa.2019.02.013

Konya Büyükşehir Belediyesi – KABİS Veri Servisi. (2023). Bisiklet park alanları ve istasyonlarının konumları. Erişim adresi: https://acikveri.konya.bel.tr/tr/dataset/bisiklet-park-alanlari-ve-istasyonlarinin-konumlari (20.08.2025)

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422). https://doi.org/10.1109/ICDM.2008.17

Mehrotra, K. G. (2017). *Anomaly detection principles and algorithms*. Cham: Springer.

Nowak-Brzezińska, A., & Horyńa, M. (2020). Comparison of algorithms for outlier detection: LOF, COF, and K-Means. *Applied Sciences, 10*(18), 1–18. https://doi.org/10.3390/app10186468

Petrariu, A. I., Moscaliuc, M., Turcu, C. E., & Gherman, L. D. (2022). A comparative study of unsupervised anomaly detection algorithms. *International Journal of Advanced Computer Science and Applications, 13*(9), 536–543. https://doi.org/10.14569/IJACSA.2022.0130963

Extended Abstract

Aim and Scope

The aim of this study is to detect anomalies in urban bicycle sharing data using unsupervised machine learning algorithms. The scope of the research focuses on the bike-sharing system in the city of Konya, Türkiye, which provides a relevant and practical case for analyzing urban mobility data. Anomalies in such datasets may arise due to GPS or recording errors, unusual user behaviors, or statistically rare patterns. Detecting these anomalies is important both for improving the reliability of transportation systems and for ensuring the efficient use of resources in urban mobility planning. This study contributes to the literature by applying two well-known unsupervised

anomaly detection algorithms, Local Outlier Factor (LOF) and Isolation Forest (IF), and by evaluating their performance on real-world bike-sharing data.

Methods

The dataset used in this study consists of 1069 observations from the Konya bike-sharing system. Preprocessing was performed to ensure data consistency, and the main features analyzed were rental duration (in seconds) and distance traveled (in meters). Since the dataset does not include anomaly labels, unsupervised learning methods were employed. Both LOF and IF algorithms were implemented with a contamination parameter of 0.03 (3%), representing the assumed anomaly ratio in the dataset. LOF detects anomalies based on local density comparisons with neighboring observations, while IF isolates anomalies through random partitioning using decision trees. The results from both models were compared in terms of the number of anomalies detected, overlapping results, and unique anomalies identified by each model. In addition, parameter sensitivity analyses were conducted to evaluate the robustness of the models under different settings.

Findings

The analysis revealed that both LOF and IF detected 33 anomalies each, with 20 anomalies identified in common. The expected anomaly count based on the contamination parameter was approximately 32, confirming that both models provided results consistent with the assumption. The anomalies detected were classified into categories such as impossible distances (likely GPS or recording errors), very long duration with low distance and statistical anomalies. The majority of anomalies were related to impossible distances, indicating that technical recording errors are a significant issue in the dataset. The parameter sensitivity analysis showed that varying the number of neighbors in LOF and the number of trees in IF did not lead to significant changes in anomaly detection outcomes. This suggests that both models are relatively stable for this dataset.

Conclusion

This study demonstrates that unsupervised anomaly detection methods such as LOF and IF can effectively identify anomalies in urban bike-sharing data. The results indicate that using multiple algorithms in combination provides more reliable insights, as models agree on a core set of anomalies but also highlight unique patterns individually. The findings align with previous studies in the literature that emphasize the complementary strengths of different algorithms in anomaly detection tasks. Moreover, the dominance of GPS-related anomalies highlights the importance of improving data collection and recording mechanisms in urban mobility systems. In conclusion, this study contributes to both the methodological literature on anomaly detection and the practical domain of urban transportation management, offering implications for enhancing the quality, safety, and reliability of bike-sharing services.

Ek

Tablo 4 LOF ve IF modelleri tarafından ortak olarak tespit edilen 20 anomali gözlemleri ve sebeplerini içermektedir.

Tablo 4. LOF ve IF Modelleri Tarafından Ortak Olarak Tespit Edilen Tüm Anomaliler ve Sebepleri

Bisiklet kimlik numarası	Kiralama başlangıç zamanı	Kiralama bitiş zamanı	Süre (sn)	Mesafe (metre)	Muhtemel sebep	LOF Skoru	IF Skoru
1452	2023-12-12 21:00:00	2023-12-12 21:17:00	1003	14563512	imkansız mesafe (muhtemel GPS/kayıt hatası)	1322.951	0.148
1459	2023-12-11 08:08:00	2023-12-11 08:28:00	1209	14566115	imkansız mesafe (muhtemel GPS/kayıt hatası)	1323.057	0.148
1071	2023-12-09 12:34:00	2023-12-09 17:20:00	17134	2880	istatistiksel anomali (LOF & IF)	1.654	0.048
1602	2023-12-08 12:08:00	2023-12-08 19:25:00	26197	5653	istatistiksel anomali (LOF & IF)	3.245	0.103
1427	2023-12-07 19:07:00	2023-12-07 19:17:00	547	14561747	imkansız mesafe (muhtemel GPS/kayıt hatası)	1322.879	0.159

1511	2023-12-06 18:53:00	2023-12-06 19:10:00	1046	20544	istatistiksel anomali (LOF & IF)	5.016	0.025
1076	2023-12-05 19:47:00	2023-12-05 20:05:00	1053	14560382	imkansız mesafe (muhtemel GPS/kayıt hatası)	1322.823	0.148
1375	2023-12-05 12:27:00	2023-12-05 16:45:00	15491	16004	istatistiksel anomali (LOF & IF)	2.564	0.12
1458	2023-12-05 10:02:00	2023-12-05 16:49:00	24376	774	çok uzun süre / düşük mesafe	3.058	0.095
1557	2023-12-04 14:50:00	2023-12-04 15:12:00	1344	14558539	imkansız mesafe (muhtemel GPS/kayıt hatası)	1322.747	0.15
1584	2023-12-04 14:37:00	2023-12-05 11:13:00	74162	1125	istatistiksel anomali (LOF & IF)	13.223	0.124
1644	2023-12-04 14:18:00	2023-12-04 15:16:00	3469	14570517	imkansız mesafe (muhtemel GPS/kayıt hatası)	1323.237	0.17
1712	2023-12-04 13:15:00	2023-12-04 13:37:00	1339	14563069	imkansız mesafe (muhtemel GPS/kayıt hatası)	1322.933	0.148
1789	2023-12-03 16:57:00	2023-12-03 16:59:00	133	14558535	imkansız mesafe (muhtemel GPS/kayıt hatası)	1322.747	0.162
1160	2023-12-03 13:00:00	2023-12-03 13:26:00	1516	14569720	imkansız mesafe (muhtemel GPS/kayıt hatası)	1323.205	0.165
1286	2023-12-03 11:47:00	2023-12-03 13:14:00	5235	14560416	imkansız mesafe (muhtemel GPS/kayıt hatası)	1322.824	0.165
1406	2023-12-02 21:52:00	2023-12-02 23:29:00	5839	17024	istatistiksel anomali (LOF & IF)	3.181	0.038
1970	2023-12-02 12:51:00	2023-12-02 16:03:00	11505	23257	istatistiksel anomali (LOF & IF)	3.95	0.105
1415	2023-12-02 12:44:00	2023-12-02 16:03:00	11917	21253	istatistiksel anomali (LOF & IF)	3.225	0.111
1486	2023-12-01 18:21:00	2023-12-01 18:42:00	1288	14564864	imkansız mesafe (muhtemel GPS/kayıt hatası)	1323.006	0.148