PERAKENDE SATIŞ VERİLERİNDE ANOMALİ TESPİTİ

SELİN AVCI (211307026) NİSA AKSOY (211307089)
KOCAELİ ÜNİVERSİTESİ BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ

Özet

Perakende Satış Verilerinde Anomali Tespiti isimli proje için önce bir veri seti seçilmiştir. Seçilen veri seti üzerinde veri temizleme, düzenleme, ön işleme, görselleştirme adımları yapılmıştır. Bu işlemler sonrasında Apache Kafka ve Apache Spark ile verilerin entegrasyonu sağlanmıştır. Makine öğrenmesi ve derin öğrenme teknikleri kullanılarak "Perakende Satış Verilerinde Anomali Tespiti" projesinin test sonuçları elde edilmiştir.

Veri Seti – Online Retail

Online Retail Veri Seti, Birleşik Krallık merkezli bir çevrim içi perakende şirketinin **1 Aralık 2010** ile **9 Aralık 2011** tarihleri arasındaki sekiz aylık alışveriş verilerini içermektedir. Veri seti, e-ticaret davranışlarını, müşteri trendlerini ve satış performansını analiz etmek için zengin bir kaynaktır. Aşağıda veri setine dair detaylar yer almaktadır:

Sütunlar

1. **InvoiceNo (Fatura Numarası)**
    - Her işlem için atanmış benzersiz fatura numaraları.
    - Örnek: 536365
    - Bireysel alışverişlerin takibini sağlar.
2. **StockCode (Ürün Kodu)**
    - Satılan ürünlere atanan kodlar.
    - Örnek: 85123A
3. **Description (Açıklama)**
    - Alınan ürünlerin isimleri veya açıklamaları.
    - Örnek: WHITE HANGING HEART T-LIGHT HOLDER
4. **Quantity (Miktar)**
    - Her üründen satın alınan birim sayısı.
    - Örnek: 6
5. **InvoiceDate (Fatura Tarihi)**
    - İşlemin gerçekleştiği tarih ve saat.
    - Örnek: 12/1/2010 8:26
6. **UnitPrice (Birim Fiyat)**
    - Ürünün bir birim fiyatı (GBP cinsinden).
    - Örnek: 2.55
7. **CustomerID (Müşteri Kimliği)**
    - Her müşteriye atanmış benzersiz kimlik numaraları.

- Örnek: 17850
- Müşteri davranışlarını ve sadakatini analiz etmek için kullanılabilir.

8. **Country (Ülke)**
    - Siparişin verildiği ülke.
    - Örnek: United Kingdom(Birleşik Krallık)

Veri Setinin Önemli Özellikleri

- **İşlem Hacmi**:
  Veri seti toplamda 541.909 işlem kaydı içermektedir. Bu, trend analizi ve modelleme için büyük bir veri kaynağıdır.
- **Ürün Çeşitliliği**:
  4.070 farklı ürün kodu yer almakta olup, geniş bir ürün yelpazesi sunmaktadır.
- **Müşteri Bilgisi**:
  37.000+ benzersiz müşteri bilgisi, müşteri segmentasyonu ve kişiselleştirilmiş pazarlama analizi yapılmasını sağlar.
- **Negatif Miktarlar ve İadeler**:
  Veri setinde negatif miktarlar (iade işlemleri) da yer almaktadır. Bu, hem iade süreçlerini hem de müşteri memnuniyetini analiz etmeye olanak tanır.

Veri Ön İşleme ve Görselleştirme

Veri ön işleme ve görselleştirme adımlarının uygulanabilmesi için önce çeşitli çıkarımlar yapılmıştır.

- **İlk 5 Satırın Gösterimi:** Veri setinin ilk 5 satırı görüntülenmiş.
- **Veri Seti Bilgisi:** pandas.DataFrame.info()metodu kullanılarak veri setindeki sütun adları, veri türleri ve eksik değer sayıları görüntülenmiş.
- **Eksik Değerlerin Tespiti:** Hangi sütunlarda eksik değer olduğu kontrol edilmiş ve bu eksik değerlerin toplamı listelenmiş.
- **Betimsel İstatistiklerin Çıkartılması:** DataFrame.describe()metodu ile sayısal sütunların temel istatistikleri (ortalama, standart sapma, minimum, maksimum vb.) analiz edilmiş. Ayrıca kategorik veriler için unique, topve freqgibi bilgiler elde edilmiş.

Bu bilgilere göre aşağıdaki adımlar uygulanmıştır.

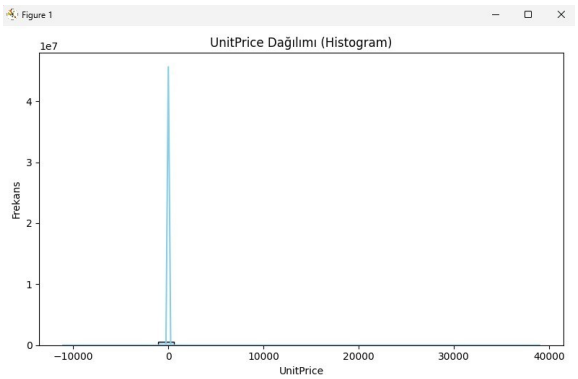
1. Eksik Değerlerin Temizlenmesi

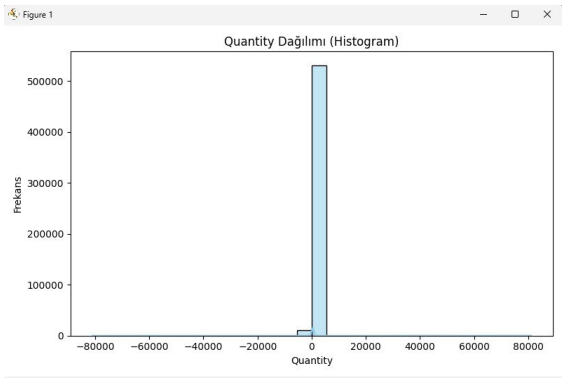
- Eksik değerlere sahip satır veya sütunların temizlendiği belirtilmiş. Bu işlem genellikle dropna()veya eksik değerlerin doldurulması için fillna()kullanılarak yapılır.
- Eksik değer temizliğinden sonra veri seti tekrar görüntülenmiş.

2. Veri Türlerinin Düzenlenmesi

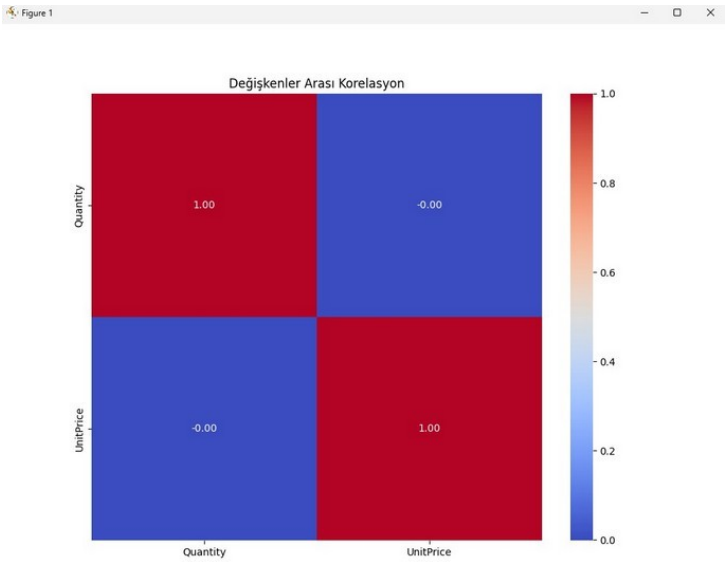
- Bazı sütunların veri türleri (örneğin, InvoiceDate) uygun formatlara dönüştürülmüş olabilir (örneğin, tarih-saat formatı).
- Bu, genelde pd.to_datetime()gibi yöntemlerle yapılır.

3. Veri Görselleştirme

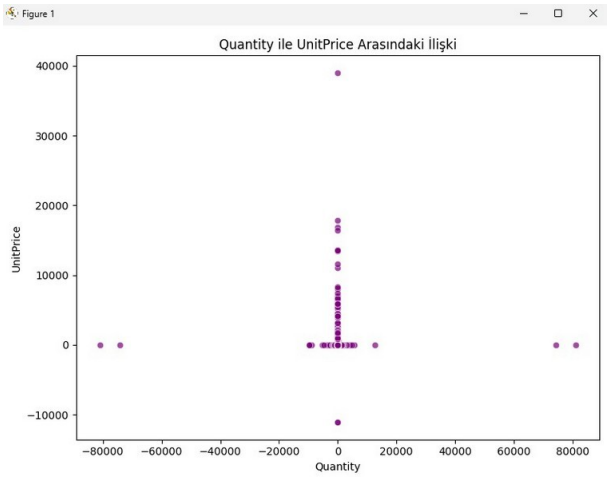




- **Histogram Grafikleri:** Veri setindeki bazı sayısal sütunların dağılımını göstermek için histogramlar oluşturulmuş.
- **Korelasyon Haritası (Heatmap):** Sütunlar arasındaki ilişkiyi göstermek için korelasyon matrisi çıkarılmış ve görselleştirilmiş.



- **Scatter Plot:** Quantityve UnitPricesütunları arasındaki ilişkiyi göstermek için bir scatter plot (dağılım grafiği) oluşturulmuş.



4. Aykırı Değerlerin Tespiti ve İşlenmesi

- **Quantityve UnitPriceSütunları:** Aykırı değerler tespit edilmiş ve bu değerler listelenmiş.
- Bu işlem genellikle IQR (Interquartile Range), Z-score veya diğer aykırı değer tespit yöntemleriyle yapılır.
- Aykırı değerler ya temizlenmiş ya da yeniden işlenmiş olabilir.

5. Verinin Normalize Edilmesi veya Standartlaştırılması

- **Normalizasyon veya Standartlaştırma:** Sayısal sütunlar (örneğin, Quantityve UnitPrice) belirli bir aralıkta (örneğin, 0 ile 1 arasında) normalize edilmiş.
- Bu işlem genellikle MinMaxScalerveya StandardScalergibi yöntemlerle yapılır.

6. Kategorik Değişkenlerin Sayısallaştırılması (Encoding)

- Countrygibi kategorik değişkenler sayısal değerlere dönüştürülmüş. Örneğin, "United Kingdom" → 36gibi bir dönüşüm yapılmış.
- Bu işlem LabelEncoderveya OneHotEncoderkullanılarak yapılmış olabilir.

7. Temizlenmiş ve İşlenmiş Veri Setinin Hazırlanması

- Bütün işlemlerden sonra elde edilen veri seti ekrana yazdırılmıştır.

Apache Kafka – Apache Spark

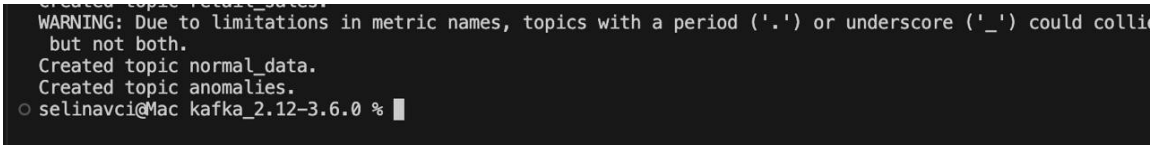
Apache Kafka, yüksek hacimli verilerin hızlı ve güvenilir bir şekilde işlenmesini sağlayan bir mesajlaşma altyapısıdır. Genellikle genellikle gerçek zamanlı veri akışlarının merkezi bir platform olarak kullanılan Kafka'da:

- **Producer (Üretici):** Verileri Kafka'ya gönderir.
- **Consumer (Tüketici):** Kafka'dan verileri alır ve işler.
- **Topic:** Verilerin kategorilere ayrılarak saklandığı birimdir.

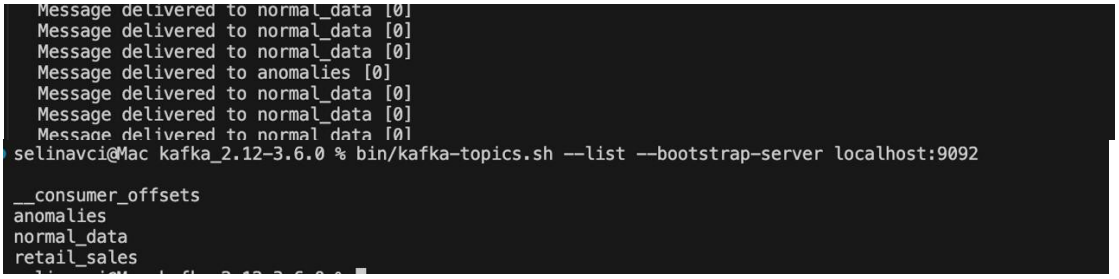
Apache Spark, büyük veri işlemleri için hızlı ve güçlü bir çerçevedir. Spark Streaming modülü, gerçek zamanlı veri akışlarını işlemek için kullanılır ve Kafka ile entegre olarak çalışabilir.
Perakende satış verilerinde anomali tespiti için Spark, anormal desenleri tespit etmek üzere verileri analiz eder ve sonuçları kaydeder.

Proje kapsamında sırasıyla aşağıdaki işlemler yapılır:

- Kafka'dan retail_saleskonusundan veriler alınır (Consumer).



- Negatif değer içeren anomaliler anomalieskonusuna, normal veriler normal_data konusuna gönderilir (Consumer).



- CSV dosyasındaki perakende satış verileri okunur.
- Her veri Kafka'daki retail_saleskonusuna gönderilir (Producer)

SparkSession ve SparkContext Oluşturma:

- Bu adımda, Apache Spark uygulamasını başlatmak için `SparkSession` ve `SparkContext` oluşturulur. Spark, büyük veri analizini paralel bir şekilde gerçekleştirir ve veri işleme işlevlerini sağlar. Burada, Spark'ı "Retail Anomaly Detection" adıyla başlatıyoruz ve veri işlemeyi hızlandırmak için yapılandırmalar ekliyoruz.

Veri Hazırlama

- Verilerin işlenmesi ve model eğitimi için uygun hale getirilmesi gerekmektedir. Bu işlevde:
- - Veriler CSV dosyasından okunur.
- - `Quantity` ve `UnitPrice` sıfırdan büyük olan satırlar seçilir.
- - Eksik değerler (`na.drop()`) temizlenir.
- - `TotalPrice` (toplam fiyat) sütunu, `Quantity` ve `UnitPrice` çarpılarak hesaplanır.
- - Sonrasında, bu veriler bir `VectorAssembler` kullanılarak sayısal özellikler (`Quantity`, `UnitPrice`, `TotalPrice`) tek bir vektöre dönüştürülür.
- - `label` adlı yeni bir sütun eklenir; burada, eğer `Quantity > 10` ve `UnitPrice > 5` ise etiket 1 (normal), aksi takdirde 0 (anormal) olur.

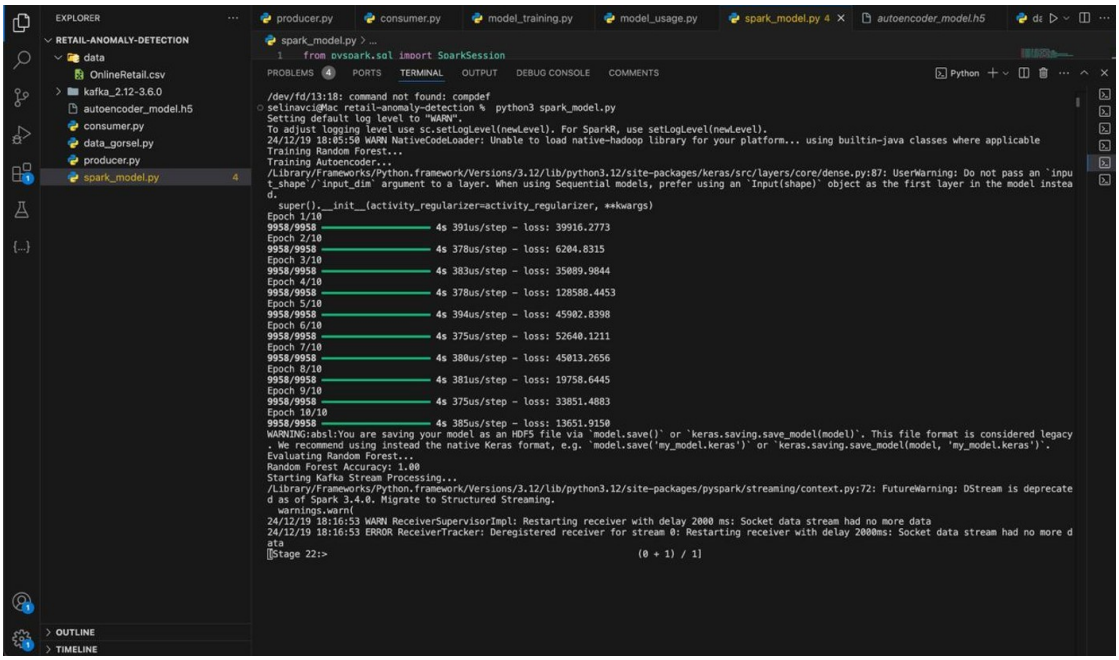
SparkSession ve Random Forest Modeli Eğitimi

- Bu fonksiyon, eğitim verilerini kullanarak bir Random Forest sınıflandırıcı modeli oluşturur. Random Forest, birçok karar ağacından oluşan ve sınıflandırma yapabilen bir makine öğrenmesi modelidir. Burada modelin özellikleri:
- - `featuresCol` olarak özelliklerin bulunduğu sütun belirlenir (`features`).
- - `labelCol` olarak etiketlerin bulunduğu sütun belirlenir (`label`).
- - 100 ağaçla model eğitilir ve model geri döndürülür.
- 4. Autoencoder Modeli Eğitimi
- Autoencoder, verinin sıkıştırılması ve yeniden yapılandırılması temeline dayanan bir yapay sinir ağı modelidir. Burada:

- - Autoencoder, giriş verisini düşük boyutlu bir uzaya sıkıştırıp tekrar eski haline getirmeye çalışır.
- - Modelin girdi boyutu, veri setinin özellik sayısına (özellik vektörü uzunluğu) göre belirlenir.
- - Modelin yapısı, 64, 32 ve 64 nörondan oluşan 3 katmandan meydana gelir ve çıktı, orijinal veriyi yeniden oluşturmayı hedefler.
- - Model, `mse` (mean squared error) kayıp fonksiyonu ile eğitilir ve eğitim tamamlandığında `autoencoder_model.h5` dosyası olarak kaydedilir.

Kafka Entegrasyonu

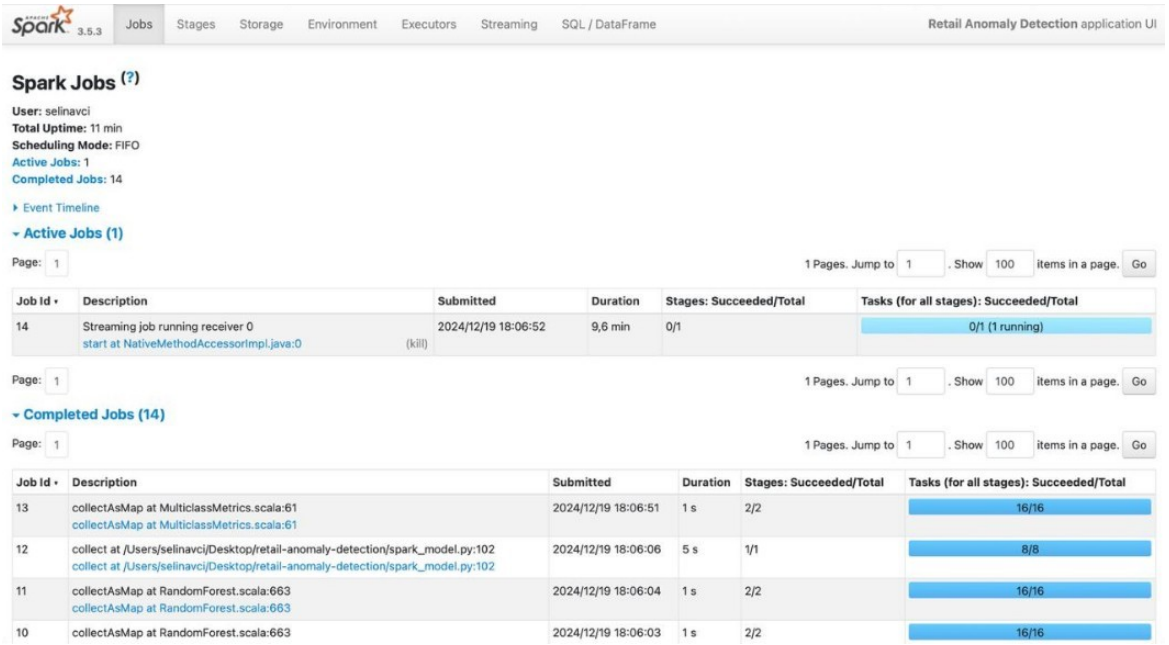
- Bu adımda, bir veri akışının işlenmesi için Apache Kafka kullanılır. Kafka, gerçek zamanlı veri akışlarını işlemek için kullanılan bir platformdur.
- - `process_stream` fonksiyonu, verileri Kafka'dan alır, gerekli veriyi çıkarır ve bu veriler üzerinde işlem yapar.
- - Veriler önce JSON formatında alınır, sonra özellikler `VectorAssembler` ile birleştirilir.
- - Random Forest modelinden tahminler yapılır.
- - Autoencoder ile verinin yeniden yapılandırılması gerçekleştirilir ve MSE (mean squared error) hesaplanarak anomali tespiti yapılır. MSE değeri belirli bir eşik değeri aşarsa, veri anormal olarak sınıflandırılır.



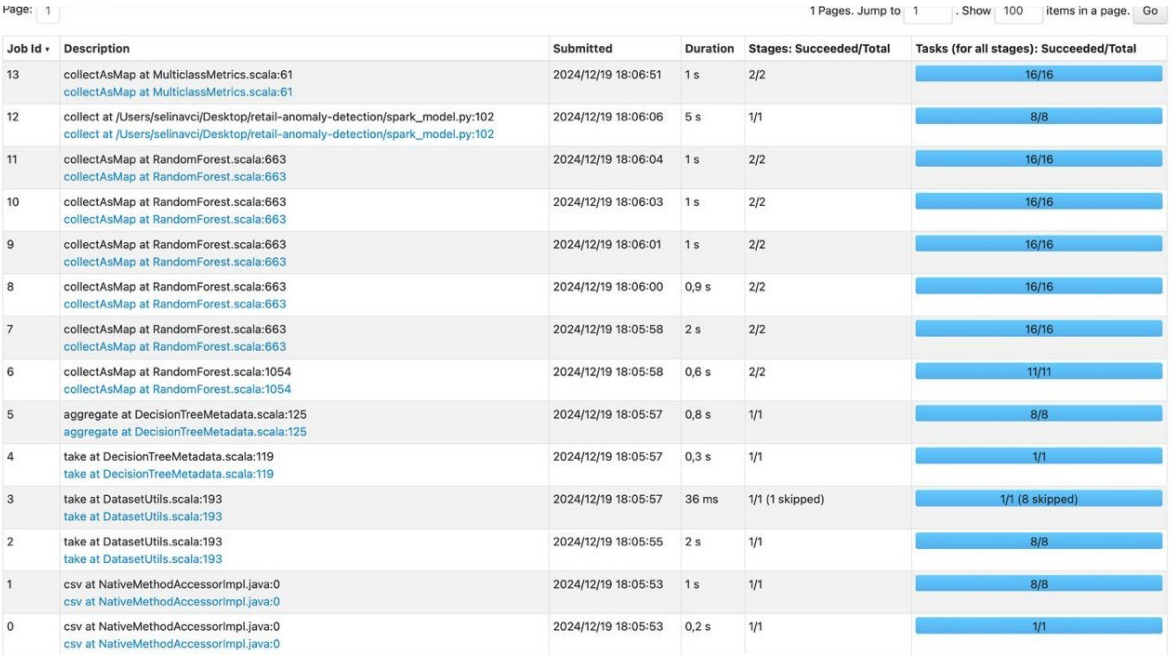
Main İşlevi

- - Veriler yüklenir ve eğitim ile test setlerine ayrılır.
- - Random Forest modeli eğitilir ve sonuçları değerlendirir.
- - Autoencoder modeli eğitilir.
- - Son olarak, Random Forest modelinin doğruluğu değerlendirilir.

- - Kafka akış işleme fonksiyonu başlatılır ve gerçek zamanlı veri işlenmeye başlanır.



- Bu işlem hattı, perakende verileri üzerinde hem sınıflandırma (Random Forest) hem de anomali tespiti (Autoencoder) yapar, ve sonuçları Kafka üzerinden işlemeye olanak tanır. Bu, özellikle alışveriş verileri gibi büyük veri setleriyle çalışırken yararlı olabilir.



Perakende Satış Verilerinde Anomali Tespiti Github Link

- https://github.com/SelinAVCI09/big-data-project.git

- https://www.kaggle.com/datasets/vijayuv/onlineretail

-

Kaynakça:

- https://spark.apache.org/downloads.html
- https://kafka.apache.org/downloads
- https://www.python.org/