Gerçek Zamanlı Büyük Veri Analitiği ile Anomali Tespiti: Spark ve Kafka Entegrasyonu

Mahmut Satici(211307070), Deniz Ilgın Koca(211307017)

Bilişim Sistemleri Mühendisliği Kocaeli Üniversitesi https://github.com/Denizilgin/BigDataAnalysis

Özet-Bu proje, Kafka ve Spark teknolojilerini kullanarak gerçek zamanlı veri işleme ve anomali tespiti gerçekleştirmeyi amaçlamaktadır. Seçilen veri seti üzerinde eksik verilerin temizlenmesi, normalizasyon, kategorik verilerin dönüştürülmesi gibi ön işlemler yapılmış ve veri dağılımı görselleştirilmiştir. Random Forest, Decision Tree ve Linear Regression gibi makine modelleri kullanılarak öğrenmesi başarı metrikleri değerlendirilmiştir. Kafka ile veri akışı sağlanmış, Spark Streaming ile gerçek zamanlı anomali tespiti yapılarak sonuçlar ilgili Kafka topic'lerine iletilmiştir. Proje çıktıları detaylı bir sekilde analiz edilip raporlanarak GitHub deposunda paylaşılmıştır.

Anahtar Kelimeler – Kafka, Spark Streaming, Anomali Tespiti, Makine Öğrenmesi, Gerçek Zamanlı Veri İşleme, Veri Görselleştirme

I. GİRİŞ

• Proje Amacı:

Bu proje, gerçek zamanlı veri işleme ve anomali tespiti yapmak için Kafka ve Spark Streaming teknolojilerini kullanmayı hedeflemektedir.

• Kapsam:

Veri seti üzerinde ön işleme, görselleştirme ve makine öğrenmesi modellerinin geliştirilmesi. Kafka ile gerçek zamanlı veri akışı sağlanarak Spark ile anomali tespiti gerçekleştirilmesi.

• Uygulanan Yöntemler:

1. Veri Ön İşleme:

Eksik verilerin temizlenmesi, verilerin normalize edilmesi ve kategorik değişkenlerin dönüştürülmesi.

2. Makine Öğrenmesi:

Random Forest, Decision Tree ve Linear Regression gibi algoritmaların kullanımı.

3. Gerçek Zamanlı İşleme:

Kafka ve Spark entegrasyonu ile anlık veri analizi ve tespit edilen anomalilerin Kafka topic'lerine iletilmesi.

II. PROJE ADIMLARI

1. Veri Seti Seçimi

Bu projede, "Automobile.csv" veri seti seçilmiştir. Bu veri seti, otomobillerin teknik özellikleri (motor gücü, ağırlık, hızlanma vb.) ile bazı kategorik bilgilerini (köken, model yılı gibi) içermektedir. Ayrıca, veri seti, belirli bir hedef değişken üzerinden anomali tespiti yapılabilecek yeterli çeşitliliğe sahiptir. Anomali tespiti için hem sürekli hem de kategorik değişkenlere sahip olması ve eksik veri içeriyor olması, bu veri setini projeye uygun hale getirmiştir. Aşağıda seçim nedenleri listelenmektedir:

<u>Ceşitlilik:</u> Veri seti, sayısal (horsepower, weight, acceleration) ve kategorik (origin, model_year) özelliklerin bir kombinasyonunu içerir.

<u>-Eksik Veri İçeriği</u>: Veri setinde bazı sütunlarda eksik değerler bulunması, veri ön işleme adımlarını uygulamak için fırsat sağlamıştır.

-Anlamlı Anomaliler: Motor gücü, ağırlık gibi özelliklerde olabilecek aykırı değerler, otomobillerin performans analizine dair çıkarımlar yapmayı sağlar.

```
horsepower, weight, acceleration, model_year, origin
name, mpg, cylinders,
chevrolet chevelle malibu, 18,8,307, 130,3504, 12,70, usa
buick skylark 320,15,8,350,165,3693,11.5,70,usa
plymouth satellite,18,8,318,150,3436,11,70,usa
amc rebel sst,16,8,304,150,3433,12,70,usa ford torino,17,8,302,140,3449,10.5,70,usa
ford galaxie 500,15,8,429,198,4341,10,70,usa chevrolet impala,14,8,454,220,4354,9,70,usa
                                  215,4312,8.5,70,us
plymouth fury iii,14,8,
pontiac catalina,14,8,4
                                55, 225,4425, 10,70, usa
                                   0, 190, 3850, 8.5, 70, usa
amc ambassador dpl,15,8,3
dodge challenger se,15,8,383,170,3563,10,70,usa
plymouth 'cuda 340,14,8,34
                                    160,3609,8,70,usa
chevrolet monte carlo,15,8,400,150,3761,9.5,70,u
buick estate wagon (sw),14,8,455,225,3086,10,70,u
toyota corona mark ii,24,4,113,95,2372,15,70,japa
plymouth duster,22,6,198,95,2833,15.5,70,usa amc hornet,18,6,199,97,2774,15.5,70,usa
ford maverick,21,6,200,85,2587,16,70,usa datsun pl510,27,4,97,88,2130,14.5,70,jap
volkswagen 1131 deluxe sedan, 26, 4, 97, 46, 1835, 20.5, 70, europe
peugeot 504,25,4,110,87,2672,17.5,70,europe
audi 100 ls,24,4,
                            90,2430,14.5,70,eu
saab 99e,25,4,104,95,2375,17.5,70,europe
                       113,2234,12.5,70,e
```

(Veri setine ait bir görüntü)

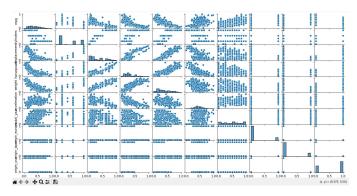
2. Kodda Kullandığımız Kütüphaneler ve Uygulanan İşlemler

Kütüphaneler:

- <u>pandas:</u> Veri işleme ve tablo manipülasyonu için kullanıldı.
- <u>numpy:</u> Eksik verilerin doldurulması gibi matematiksel işlemler için kullanıldı.
- <u>seaborn ve matplotlib:</u> Veri görselleştirme (korelasyon matrisleri, dağılım grafikleri) için kullanıldı.
- <u>sklearn.preprocessing.MinMaxScaler:</u> Verilerin normalize edilmesi için kullanıldı.



(Korelasyon Matrisine ait bir resim)



(Oluşturulan grafiklere ati bir resim)

Uygulanan Adımlar:

Eksik Verilerin Doldurulması:

horsepower sütununda eksik değerler bulunduğu tespit edildi ve bu sütun eksik değerlerin ortalaması ile dolduruldu.

• Kategorik Değişkenlerin Kodlanması: origin sütunu, one-hot encoding yöntemiyle sayısal değerlere dönüştürüldü. Örneğin, origin_usa, origin_europe ve origin japan gibi sütunlar oluşturuldu.

• Normalizasyon:

Tüm sayısal sütunlar Min-Max Scaling yöntemiyle [0, 1] aralığına getirildi. Bu, farklı ölçeklerdeki verilerin makine öğrenmesi algoritmaları tarafından daha iyi işlenmesini sağlar.

• Görselleştirme:

Korelasyon matrisi, veri setindeki özellikler arasındaki ilişkileri ortaya koymak için görselleştirildi.

Dağılım grafikleri ile verinin genel dağılımı incelendi.

• İşlenmiş Verinin Kaydedilmesi:

Ön işlemden geçirilen veri, Processed_Automobile.csv adıyla yeni bir dosya olarak kaydedildi.

3. Yapay Zeka Modeli Geliştirme

Bu adımda, işlenmiş veriler kullanılarak makine öğrenmesi modelleri geliştirilmiş ve performansları karşılaştırılmıştır. Hedef, otomobillerin silindir sayısını tahmin edebilecek en iyi modeli belirlemektir. Silindir sayısı (cylinders), hedef değişken olarak belirlenmiş ve diğer tüm özellikler tahmin için kullanılmıştır.

(Modelleri oluşturan ve skorları üreten kod parçacığı)

• Veri Bölme:

İşlenmiş veri, modelin öğrenmesi ve test edilmesi için iki bölüme ayrılmıstır:

- -Eğitim Verisi (Training Data): Modelin öğrenmesi için kullanılan veri (%80).
- -Test Verisi (Test Data): Modelin performansını değerlendirmek için ayrılan veri (%20).

Bu ayrım, modelin gerçek hayatta hiç görmediği veriler üzerinde nasıl bir performans göstereceğini anlamak için yapılmıştır.

```
Eksik Değer Sayısı:
                0
cylinders
                0
displacement
horsepower
                6
weight
acceleration
                0
model year
                0
origin
                0
dtype: int64
İşlenmiş Veri Kümesi:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 398 entries, 0 to 397
Data columns (total 10 columns):
    Column
                    Non-Null Count Dtype
                    398 non-null
                                     float64
0
    mpg
    cylinders
                    398 non-null
                                     float64
                    398 non-null
                                     float64
    displacement
    horsenower
                    398 non-null
                                     float64
                    398 non-null
                                     float64
    weight
                                     float64
    acceleration
                    398 non-null
    model year
                    398 non-null
                                     float64
                                     bool
    origin europe
                    398 non-null
    origin japan
                    398 non-null
                                     boo1
                    398 non-null
    origin usa
                                     boo]
```

(main.py dosyasının çıktısı)

• Kullanılan Modeller:

Üç farklı model kullanılarak tahminleme yapılmıştır:

• Random Forest Regressor:

Karar ağaçlarının bir topluluğunu kullanarak tahmin yapar. Ağaç sayısı (n_estimators) 100 olarak belirlenmiştir. Daha karmaşık ilişkileri öğrenme kapasitesine sahiptir ve aşırı öğrenme riskini azaltır.

• Decision Tree Regressor:

Veriyi bir karar ağacı mantığıyla bölerek tahmin yapar. Hızlı bir şekilde çalışır, ancak aşırı öğrenmeye (overfitting) yatkındır.

• Linear Regression:

Özellikler ve hedef değişken arasındaki ilişkiyi doğrusal bir modelle tahmin eder.

Daha basit ilişkilerde etkili, ancak doğrusal olmayan karmaşık ilişkilerde yetersizdir.

• Performans Ölçütleri:

Model performansını değerlendirmek için aşağıdaki metrikler kullanılmıştır:

Mean Squared Error (MSE):

Tahminlerin hedef değişkenden ne kadar sapma gösterdiğini ölçer. Daha düşük bir MSE, daha iyi bir model anlamına gelir.

R² (R-Kare) Skoru:

Modelin veriye ne kadar iyi uyduğunu gösterir. 1.0'a ne kadar yakınsa, model o kadar iyi demektir.

```
C:\Users\st_fa\Desktop\big data 2\Proje>p
Random Forest - MSE: 0.0027
Random Forest - R^2 Score: 0.9785
Decision Tree - MSE: 0.0020
Decision Tree - R^2 Score: 0.9839
Linear Regression - MSE: 0.0130
Linear Regression - R^2 Score: 0.8953

C:\Users\st_fa\Desktop\big data 2\Proje>
```

(Skor sonucları)

- Sonuçlar:
- ✓ Random Forest:

MSE: 0.0027 R²: 0.9785

Çok düşük hata oranı ve yüksek doğrulukla en iyi sonucu vermiştir.

✓ Decision Tree:

MSE: 0.0020 R²: 0.9839

Random Forest'a benzer şekilde başarılı sonuçlar vermiştir.

✓ Linear Regression:

MSE: 0.0130 R²: 0.8953

Daha yüksek bir hata oranı ve daha düşük doğrulukla diğer modellere kıyasla daha zayıf bir performans göstermiştir.

• Yorum:

Karar ağaçlarına dayalı modeller (Random Forest ve Decision Tree), karmaşık ilişkileri daha iyi öğrenmiş ve Linear Regression modeline göre çok daha iyi sonuçlar vermiştir. Random Forest modeli, karar ağaçlarının topluluk yöntemiyle çalışması sayesinde aşırı öğrenme riskini azaltarak daha genellenebilir sonuçlar sağlamıştır. Bu nedenle, veri seti üzerindeki en uygun model olarak değerlendirilmiştir.

4. Zookeeper, Kafka ve Spark Yükleme ve Kullanma Aşamaları

I. Zookeeper ve Kafka Yükleme

Zookeeper ve Kafka, gerçek zamanlı veri akışı işlemleri için birlikte çalışır. Kafka, mesajların üretildiği ve tüketildiği bir sistemdir; Zookeeper ise Kafka'nın koordinasyonu için gereklidir.

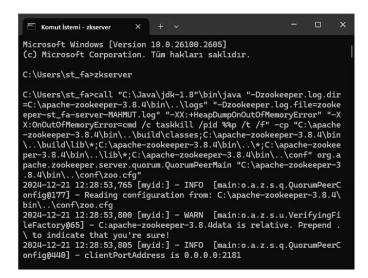
❖ Zookeeper ve Kafka'nın İndirilmesi:

Apache Kafka'nın resmi web sitesinden uygun Kafka sürümü indirilir.

Zookeeper indirilirken Kafka ile uyumlu sürümünün indirilmesi gerekir.

Zookeeper'ın Çalıştırılması:

Zookeeper'ın başlatılması için bin/zookeeper-server-start.sh dosyası kullanılır.



(Zookeeperin başlatılması)

Kafka Broker'ın Başlatılması:

Kafka, bir broker üzerinden çalışır ve veri akışı buradan yönetilir.

Kafka'yı başlatmak için bin/kafka-server-start.sh komutu kullanılır ve Kafka yapılandırma dosyasına bağlı olarak çalıştırılır.

```
C:\Windows\System32\cmd.e X
Microsoft Windows [Version 10.0.26100.2605]
(c) Microsoft Corporation. Tüm hakları saklıdır.
C:\kafka_2.13-3.7.2\bin\windows>kafka-server-start.bat ..\..\config
[2024-12-21 12:29:24,465] INFO Registered kafka:type=kafka.Log4jCon
troller MBean (kafka.utils.Log4jControllerRegistration$)
[2024-12-21 12:29:24,799] INFO Setting -D jdk.tls.rejectClientIniti
atedRenegotiation=true to disable client-initiated TLS renegotiatio
n (org.apache.zookeeper.common.X509Util)
[2024-12-21 12:29:24,924] INFO starting (kafka.server.KafkaServer)
[2024-12-21 12:29:24,925] INFO Connecting to zookeeper on localhost
 2181 (kafka.server.KafkaServer)
[2024-12-21 12:29:24,945] INFO [ZooKeeperClient Kafka server] Initi
alizing a new session to localhost:2181. (kafka.zookeeper.ZooKeeper
Client)
[2024-12-21 12:29:24,955] INFO Client environment:zookeeper.version
-
=3.8.4-9316c2a7a97e1666d8f4593f34dd6fc36ecc436c, built on 2024-02-1
2 22:16 UTC (org.apache.zookeeper.Zookeeper)
[2024-12-21 12:29:24,955] INFO Client environment:host.name=mahmut
 org.apache.zookeeper.ZooKeeper
[2024-12-21 12:29:24,956] INFO Client environment:java.version=1.8.
```

(Kafkanın başlatılması)

II. Kafka'nın Yapılandırılması

Kafka, mesajların akışı için topic adlı veri kanallarını kullanır. Veri üretmek için Producer, tüketmek için ise Consumer kullanılır.

* Topic Oluşturma:

Kafka üzerinde mesajların aktarılacağı bir topic oluşturulur. Bu topic, verinin depolandığı bir kanal gibi çalışır. Topic oluştururken partition ve replication gibi parametreler ayarlanır. Bu parametreler, veri paralelizmini ve güvenliğini etkiler.

* Kafka Producer ve Consumer Kullanımı:

Producer, topic'e veri gönderir. Veriler, gerçek zamanlı olarak veya belirli aralıklarla Kafka'ya aktarılır.

Consumer, bu verileri ilgili topic'ten okur ve işleme alır. Kafka'nın tasarımı sayesinde aynı topic'i birden fazla consumer grubu dinleyebilir.

❖ Anomali Tespiti:

Buradaki anomali tespitinde beygir gücü 200'den fazla veya litre hacmi 10'dan küçük olan arabaların tespiti yapılmakta ve işaretlenmektedir.

```
C:\kafka_2.13-3.7.2\bin\windows>kafka-console-consumer --bootstrap-server localhost:9092 --topic araba_i --from-beginning {"name": "buick skylark 320", "mpg": 15.0, "cylinders": 8, "displacement": 350.0, "horsepower": 165.0, "weight": 3693, "acceleration": 11.5, "model_year": 70, "origin": "usa"} {"name": "amc rebel sst", "mpg": 16.0, "cylinders": 8, "displacemen t": 304.0, "horsepower": 150.0, "weight": 3433, "acceleration": 12.0, "model_year": 70, "origin": "usa"} {"name": "plymouth 'cuda 340", "mpg": 14.0, "cylinders": 8, "displacemen t": 340.0, "horsepower": 160.0, "weight": 3609, "acceleration": 8.0, "model_year": 70, "origin": "usa"} {"name": "ford maverick", "mpg": 21.0, "cylinders": 6, "displacemen t": 200.0, "horsepower": 85.0, "weight": 2587, "acceleration": 16.0, "model_year": 70, "origin": "usa"} {"name": "datsun pl510", "mpg": 27.0, "cylinders": 4, "displacemen t": 97.0, "horsepower": 88.0, "weight": 2130, "acceleration": 14.5, "model_year": 70, "origin": "japan"} {"name": "puegeot 564", "mpg": 25.0, "cylinders": 4, "displacement": 110.0, "horsepower": 87.0, "weight": 2672, "acceleration": 17.5, "model_year": 70, "origin": "europe"} {"name": "toyota corona", "mpg": 25.0, "cylinders": 4, "displacement": 110.0, "horsepower": 87.0, "weight": 2672, "acceleration": 17.5, "model_year": 70, "origin": "europe"} {"name": "toyota corona", "mpg": 25.0, "cylinders": 4, "displacement": 110.0, "horsepower": 87.0, "weight": 2672, "acceleration": 17.5, "model_year": 70, "origin": "europe"} {"name": "toyota corona", "mpg": 25.0, "cylinders": 4, "displacement": 17.5, "model_year": 70, "origin": "europe"} {"name": "toyota corona", "mpg": 25.0, "cylinders": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "displacement": 4, "di
```

(Şekil1. Producer topic)

```
rigin":"usa", "anomalies":"High horsepower"}

":70, "origin":"usa", "anomalies":"High horsepower"}

gin":"usa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

"usa", "anomalies":"High horsepower"}

"usa", "anomalies":"High horsepower"}

"usa", "anomalies":"High horsepower"}

year":73, "origin":"usa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

sa", "anomalies":"High horsepower"}

"origin":"usa", "anomalies":"High horsepower"}

"origin":"usa", "anomalies":"High horsepower"}

""europe", "anomalies":"High horsepower"}

""europe", "anomalies":"High horsepower"}

""europe", "anomalies":"High horsepower"}

""europe", "anomalies":"High horsepower"}

"""usa", "anomalies":"High horsepower"}
```

(Şekil 1.2 Consumer Topic)

III. Spark Yükleme

Apache Spark, büyük veri işleme platformudur ve Kafka ile entegre çalışarak gerçek zamanlı veri akışı sağlar.

Spark'ın İndirilmesi ve Kurulumu:

Apache Spark'ın resmi web sitesinden Spark sürümü indirilir. Spark, bağımsız bir sistemde çalışabileceği gibi, Hadoop gibi bir altyapı ile entegre edilerek de kullanılabilir.

Spark Ortamının Hazırlanması:

Spark, bir yerel modda (local mode) çalıştırılabilir veya daha büyük veri işleme için bir küme (cluster) modunda kullanılabilir. Spark, veri işlemek için RDD, DataFrame veya Dataset gibi veri yapıları sunar.

Spark ve Kafka Entegrasyonu:

Spark Streaming, Kafka'dan gelen verileri gerçek zamanlı olarak işler. Kafka ile Spark entegrasyonu, Spark Streaming modülünün Kafka bağlayıcılarını kullanarak gerçekleştirilir. Bu entegrasyon ile Kafka'dan gelen veriler işlenip sonuçlar diğer sistemlere aktarılabilir.

IV. Spark ve Kafka ile Gerçek Zamanlı İşlem Spark ve Kafka entegrasyonu, verinin üretilip tüketilmesi için kullanılır. Bu süreçte:

Kafka'dan gelen veri, Spark Streaming üzerinden alınır. Spark, bu veriyi işler ve sonuçlar tekrar Kafka'ya veya bir başka hedef sisteme gönderilir. Gerçek zamanlı veri analitiği ve anomali tespiti gibi işlemler bu süreçte gerçekleştirilir.

KAYNAKLAR

- [1] M. King, B. Zhu, and S. Tang, "Optimal path planning," Mobile Robots, vol. 8, no. 2, pp. 520-531, March 2001.
- [2] H. Simpson, Dumb Robots, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.
- [3] M. King and B. Zhu, "Gaming strategies," in Path Planning to the West, vol. II, S. Tang and M. King, Eds. Xian: Jiaoda Press, 1998, pp. 158-176.
- [4] B. Simpson, et al, "Title of paper goes here if known," unpublished.
- [5] J.-G. Lu, "Title of paper with only the first word capitalized," J. Name Stand. Abbrev., in press.
- [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Translated J. Magn. Japan*, vol. 2, pp. 740-741, August 1987 [Digest 9th Annual Conf. Magnetics Japan, p. 301, 1982].
- [7] M. Young, The Technical Writer's Handbook, Mill Valley, CA: University Science, 1989.
- [8] J. K. Author, "Title of paper," in *Unabbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), year, pp xx-xxx.
- [9] R. J. Vidmar. (1994). On the use of atmospheric plasmas as electromagnetic reflectors [Online]. Available FTP: atmnext.usc.edu



BvaqProje.pdf

Directory: pub/etext/1994 File: atmosplasma.txt