T.C.

SAKARYA ÜNİVERSİTESİ

BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BIG DATA
PROJE RAPORU

B191210044 - BURAK KOZLUCA

1-B Grubu

NUR BANU OĞUR

BITCOIN FİYAT TAHMİNİ

1. Projenin Özeti

Projenin ana amacı Apache kafka ile aldığımız streaming veriyi Apache Spark MLLIB kütüphanesiyle analiz ederek buna göre ekrana son durum göstermedir. Ben projede bitcoin fiyat tahmini yaptım. Bitcoin'in son 5 yıldaki günlük fiyatlarını analiz ederek fiyat tahmini yaptırdım.

2. APACHE KAFKA

Yüksek performanslı ve ölçeklenebilir bir veri akışı platformudur.

-Brokerlar: Kafka'nın temel işlevselliğini sağlayan sunuculardır. Brokerlar, veri akışlarını depolar ve iletimini sağlarlar.

-Producer: Veri üreten uygulamalardır. Producerlar, Kafka'ya veri mesajlarını gönderir.

-Consumer: Veriyi alan uygulamalardır. Consumer'lar, Kafka'dan veri mesajlarını alır ve işlerler.

-Topic: Veri mesajlarının kategorize edildiği birimlerdir. Producer'lar, mesajları belirli bir konuya gönderir ve Consumer'lar bu konulardan veriyi alır.

-Zookeeper: Kafka'nın yapılandırma bilgilerini ve brokerların durumunu yöneten bir koordinasyon servisidir.

Projede kafka yardımıya BTC.csv dosyasındaki günlük bitcoin fiyatlarını producer ederek localhost 9092 portuna ilettim. Daha sonrasında consumer tarafında bu porttan bu verileri alacağım.

3. APACHE SPARK

Büyük veri işleme ve analizi için açık kaynaklı bir veri işleme çerçevesidir. Hadoop MapReduce modelini geliştirmek ve performansı artırmak amacıylaortaya çıkan Spark, daha hızlı ve kullanımı daha kolay bir alternatif sunar. Apache Spark, geniş bir veri işleme ekosistemi ile entegre olabilen, paralel ve dağıtık hesaplama için bir platform sağlar.

Projemde Apache spark ile akış halindeki btc verisini mllib kütüphanesi yardımıyla fiyat tahmininde kullandım.

4. Makine Öğrenmesi

Bilgisayar sistemlerinin belirli bir görevi veya problemi, deneyim ve veri kullanarak öğrenme yeteneğini ifade eden bir yapay zeka dalıdır. Geleneksel programlama yaklaşımlarından farklı olarak, makine öğrenmesi modelleri, belirli bir görevi yerine getirmek için doğrudan programlanmazlar. Bunun yerine, algoritmalar, veri setlerinden öğrenme süreci ile görevi gerçekleştirmek için genel bir model oluştururlar.

Kullanılan Veri Analiz Yöntemi

Makine öğrenmesi modeli olan Lineer Regresyon modelini kullanarak Bitcoin fiyat tahminleme uygulaması gerçekleştirdim.

İlk olarak veri, eğitim ve test setlerine ayırdım. Bu sayede modelin performansı daha sonra test edilebilir.

Sonrasında Lineer regresyon modelini eğitim seti üzerinde eğittim.

```
model = LinearRegression()

pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('model', model)
])
```

Veri Setinin Tanımlanması

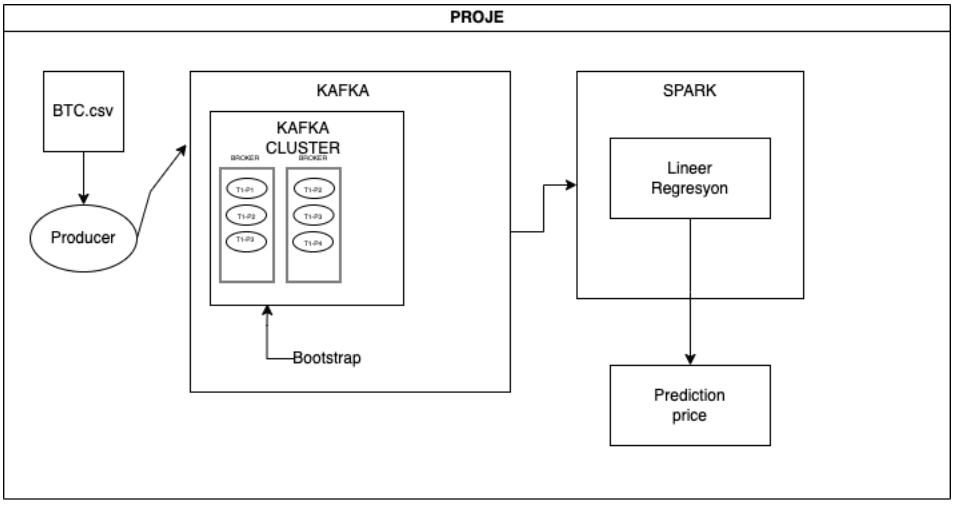
Veri Setini yahoo.finance sitesinden 5 yıllık bitcoin fiyatını .csv formatında indirdim. Bu veri setinin kolonları Tarih, Açılış fiyatı, En yüksek fiyat, en küçük fiyat, Kapanış fiyatı, adj kapanış fiyatı ve o günkü hacimden oluşmaktadır.

Aldığım veri setindeki değerleri istediğim tipe dönüştürerek Lineer Regresyon modelinde işledim.

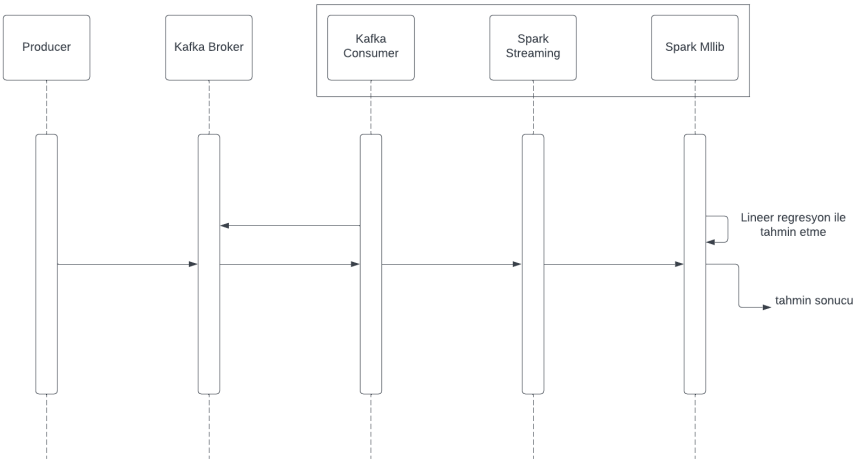
```
new_data = json.loads(msg.value().decode("utf-8"))
new_data_df = pd.DataFrame([new_data], columns=['Open', 'High', 'Low'])

predicted_price = pipeline.predict(new_data_df)[0]
actual_price = new_data['Close']
```

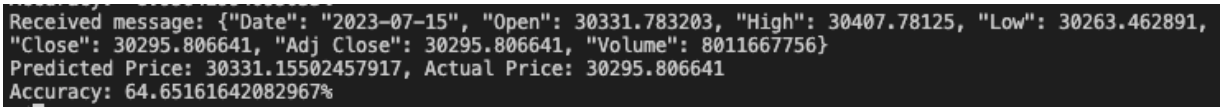
Akış Şeması



Zamanlama Şeması



Elde Edilen Bulgular



Referanslar

[1] https://finance.yahoo.com/quote/BTC-USD/history?period1=1700611200&period2=1703203200&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

[2] https://www.conduktor.io/kafka/how-to-install-apache-kafka-on-mac/

[3] https://medium.com/beeranddiapers/installing-apache-spark-on-mac-os-ce416007d79f