Büyük Veri Projesi

Proje Özeti

Kafka producer ve kafka consumer yapılarının oluşturulması ve kafka producera satır satır gönderilen test verilerinin kafka consumer olan spark tarafından streaming şekilde alınması.

Daha sonra lojistik regresyon sınıflandırma algoritmasıyla modelin eğitilmesi ve spark streaming ile alınan akış halindeki verinin test verisi olarak kullanılarak model testinin yapılması işlemleri gerçekleştirildi.

Kullanılan Teknolojiler

1. **Spark:** Apache Spark, büyük veri analizi ve işleme için açık kaynaklı bir platformdur. Hızlı, dağıtık ve genel amaçlı bir veri işleme motorudur. Paralel hesaplama yapabilme yeteneğiyle büyük veri setlerini işleyebilir ve genellikle veri analitiği, makine öğrenimi ve gerçek zamanlı işleme gibi alanlarda kullanılır.
2. **Spark Structured Streaming:** Spark Structured Streaming, Apache Spark'ın bir bileşenidir ve gerçek zamanlı veri işleme için kullanılır. Yapılandırılmış verileri işleyebilme yeteneğiyle, veri akışlarını işlemek ve işlemek için bir API sağlar. Veriler akış halindeyken işlenebilir, analiz edilebilir veya depolanabilir.
3. **Apache Kafka:** Apache Kafka, yüksek ölçeklenebilirlikte gerçek zamanlı veri akışı sağlayan bir dağıtık olay akışı platformudur. Üreticilerden (producer) tüketicilere (consumer) kadar olan veri akışını yönetir ve büyük ölçekli veri akışlarını depolamak, işlemek ve entegre etmek için kullanılır.
4. **Scala:** Scala, genel amaçlı bir programlama dilidir ve aynı zamanda Apache Spark gibi büyük veri işleme sistemlerinde sıkça kullanılır. Fonksiyonel ve nesne yönelimli programlama özelliklerini bir araya getirir. Scala, temiz ve esnek bir sözdizimine sahiptir.

Kullanılan Veri Analiz Yöntemleri

1. **VectorAssembler:** VectorAssembler, Apache Spark'ta bulunan bir bileşendir ve genellikle makine öğrenimi modellerine veri hazırlamak için kullanılır. Bu özellik, bir DataFrame içindeki birden fazla sütunu alır ve bunları birleştirerek bir vektör haline getirir. Genellikle, makine öğrenimi modelleri için birden çok özellikten oluşan verileri tek bir vektör içinde kullanmak gerekebilir. Örneğin, bir sınıflandırma modeli için farklı özelliklerden (mesela sayısal özellikler veya kategorik özelliklerin kodlanmış hali) oluşan girdi verileri birleştirilerek tek bir vektör haline getirilebilir. Bu sayede, veri işleme adımları ve model eğitimi için veri hazırlığı kolaylaşır.
2. **Lojistik Regresyon:** Lojistik regresyon, sınıflandırma problemlerinde kullanılan bir istatistiksel modeldir. Sınıflandırma, veri noktalarını belirli kategorilere veya sınıflara atamayı amaçlar. Lojistik regresyon, bir girdi vektörünü alır ve bu vektör üzerinden bir çıktı üretir. Örneğin, ikili sınıflandırma problemlerinde kullanılabilir; yani, belirli bir olayın olasılığını hesaplamak veya bir veri noktasını iki sınıfa ayırmak için kullanılabilir. Model, girdi verilerinin özelliklerini kullanarak bir çıktı sınıfı veya olasılığını tahmin etmeye çalışır.

Kullanılan Verisetinin Tanımlanması

Kullanılan veri seti, Iris veri seti olarak adlandırılan ünlü bir veri setidir. Iris bitkisi türlerine ait özellikleri içeren bu veri seti, istatistiksel analizlerde ve makine öğrenimi algoritmalarının eğitiminde sıkça kullanılır. Veri seti, üç farklı Iris türüne (Iris setosa, Iris versicolor, Iris virginica) ait ölçümleri içerir.

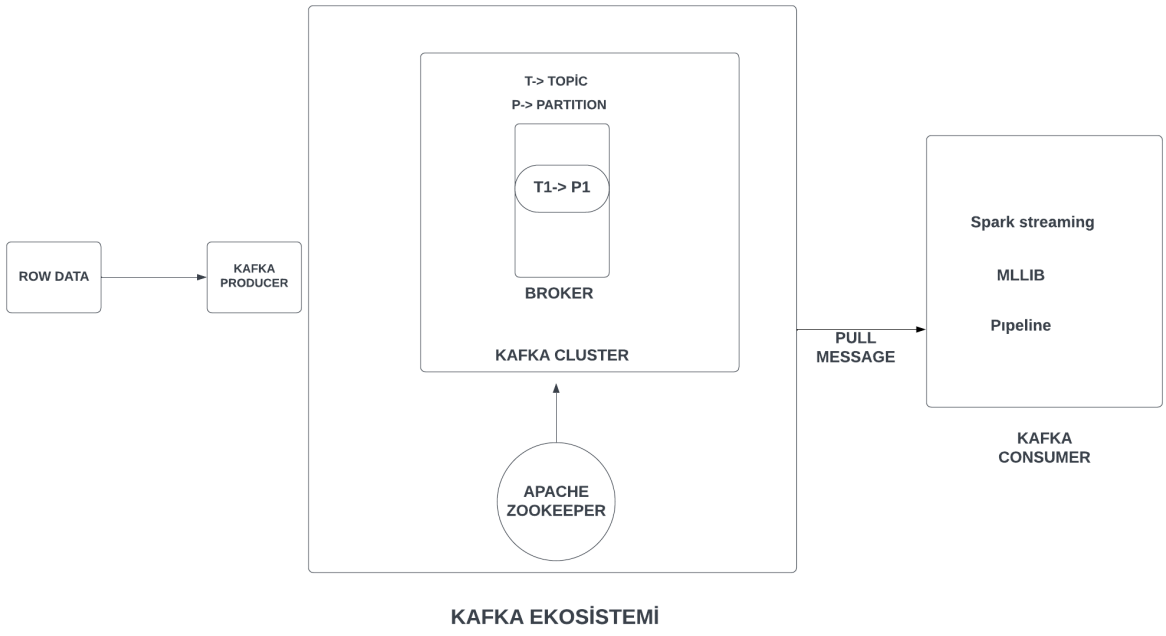
Her bir örnek, bitkinin çanak yaprakları ve taç yaprakları için dört farklı özelliği içerir:

- SepalLengthCm: Çanak yaprağının uzunluğu (cm)
- SepalWidthCm: Çanak yaprağının genişliği (cm)
- PetalLengthCm: Taç yaprağının uzunluğu (cm)
- PetalWidthCm: Taç yaprağının genişliği (cm)

Ayrıca, her bir örnek aynı zamanda bitkinin türünü temsil eden bir "Species" (tür) etiketi ile tanımlanır. Bu etiketler 0, 1 ve 2 olarak kodlanmıştır ve her bir rakam bir Iris bitki türünü temsil eder.

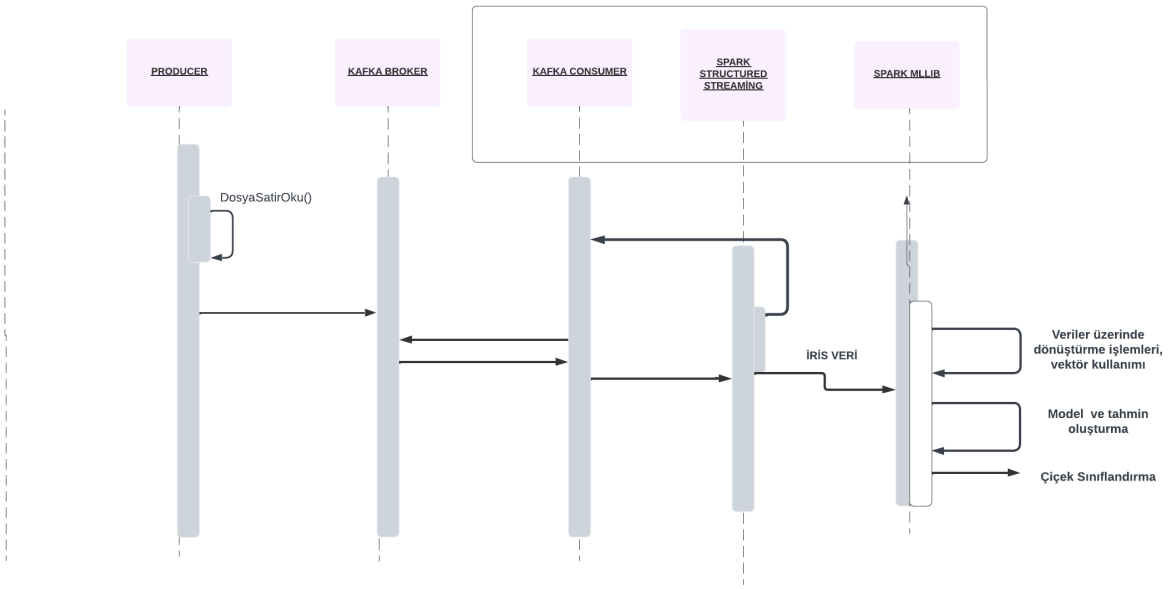
Bu veri seti, makine öğrenimi algoritmalarının sınıflandırma veya kümeleme gibi görevlerde performanslarını değerlendirmek veya eğitmek için sıklıkla kullanılır. Özellikle, örneklerin özellikleri (SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm) kullanılarak bitki türlerinin tahmin edilmesi gibi sınıflandırma problemleri için ideal bir veri setidir.

Projenin genel bir akış şeması



KAFKA EKOSİSTEMİ

Zamanlama şeması



Elde edilen bulgular

1. **Model Performansı**: Iris veri seti üzerinde yapılan denemeler sonucunda, belirli bir model (lojistik regresyon gibi) akış halindeki veriyi doğru bir şekilde sınıflandırabiliyor.

- Makine öğrenmesinde tahmin edilmesi beklenen değişken **Species** ve **Prediction** değerleri

```
//iris setosa :0
//Iris-versicolor : 1
//Iris-virginica : 2
```

```
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|Species|            features|       rawPrediction|         probability|prediction|
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
|          4.4|         2.9|          1.4|         0.2|    0.0|[4.40000009536743...|[66.9341296108201...|[1.0,8.9819672529...|       0.0|
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
```

```
Batch: 26
-------------------------------------------
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|Species|            features|       rawPrediction|         probability|prediction|
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
|          5.7|         2.8|          4.5|         1.3|    1.0|[5.69999980926513...|[-55.193039402163...|[8.47439030554540...|       1.0|
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
```

```
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|Species|            features|       rawPrediction|         probability|prediction|
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
|          4.9|         2.5|          4.5|         1.7|    2.0|[4.90000009536743...|[-75.726504459243...|[2.25777369838696...|       2.0|
+-------------+------------+-------------+------------+-------+--------------------+--------------------+--------------------+----------+
```

2. **Veri Miktarı ve Dağılımı:** 30 satır test verisi ve 120 satır eğitim verisi kullanılarak bu sonuçlar elde edildi. Bu, modelin küçük bir veri seti üzerinde iyi performans gösterdiğini gösterir ancak gerçek dünya senaryoları için daha geniş veri setlerini kullanmak gerekir.
3. **Gerçek Zamanlı Model Uygulaması:** Spark streaming kullanılarak gerçek zamanlı veri alımı ve işlenmesi sağlandı. Bu, canlı veri üzerinde modelin nasıl performans gösterebileceğini ve uygulanabilirliğini gösteriyor olabilir.
4. **Modelin Gelecekteki Kullanımı:** Bu tür bir yapı, gerçek zamanlı veri akışlarını işlemek ve canlı sistemlerde kullanmak için etkili bir yöntem olarak görülebilir. Özellikle ölçeklenebilirlik ve hızlı yanıt verme becerileri göz önüne alındığında, bu tür bir modelin endüstriyel uygulamalarda değerli olabileceği düşünülebilir.