SAKARYA
ÜNİVERSİTESİ

BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BÜYÜK VERİYE GİRİŞ PROJE RAPORU

B191210007 – Şevin Sena DERE
1/B

1. PROJENİN ÖZETİ

Yaptığım proje, Apache Kafka'nın Java istemcisini kullanarak Kafka sunucusuna mesajlar gönderir.
İlk önce CSV dosyasındaki her satırı okuyup bir Kafka kaydı oluşturarak Kafka'ya gönderir. Ardından, csvProducer.send metodu ile Kafka'ya mesaj gönderilir. Kafka üretici uygulaması konumundadır.
Daha sonra Apache Kafka üzerinden gelen verileri Spark Streaming kullanarak işleyen bir tüketici (consumer) kullanılır. Bu uygulama, Kafka'dan gelen mesajları okur, işler ve ardından işlenmiş verileri bir dosyaya yazarak kaydeder.
Tüm işlenmiş veriler allRecord listesinde biriktirilir ve ardından bu veriler bir csv dosyasına yazdırılır.
Sonuç olarak Kafka'dan gelen veriler burada okunup işlenerek ve ardından bir dosyaya yazarak basit bir veri işleme akışı uygulamaktadır. Uygulama her 20 saniyede bir parti veriyi işleyip sonuçları bir csv dosyasına kaydeder.
Sonra Apache Spark MLlib kütüphanesini kullanarak Lineer Regresyon algoritması kullanılır. Model oluşturulur, eğitimi ve tahmin yapılmasını yönetilir. Spark streamingde oluşturulan Veri seti alınır ve sonuçlar ekrana yazdırılır.

2. KULLANILAN TEKNOLOJİLER

Apache Kafka: Açık kaynaklı bir veri akışı ve olay işleme platformudur. Başlangıçta LinkedIn tarafından geliştirilen Kafka, büyük ölçekli ve dağıtık sistemlerde yüksek performanslı, dayanıklı ve ölçeklenebilir bir şekilde olayları (events) ve veri akışını yönetmek için tasarlanmıştır.
Apache Spark: Geniş veri setlerini paralel olarak işlemek için kullanılan açık kaynaklı bir büyük veri işleme çerçevesidir. Spark, yüksek düzeyde paralelleştirilmiş veri işleme için bir platform sağlar ve veri işleme, sorgulama, makine öğrenimi ve graf analitiği gibi birçok görevi destekler.
Spark MLlib: Apache Spark'ın büyük veri üzerinde makine öğrenimi modelleri oluşturmak ve uygulamak için kullanılan kütüphanesidir. Bu kütüphane sınıflandırma, regresyon, kümeleme, boyut azaltma ve diğer birçok makine öğrenimi algoritmasını içerir.
Spark Streaming: Apache Spark'ın akış verileri üzerinde çalışmak için kullanılan bir modülüdür. Bu, gerçek zamanlı veri akışları üzerinde işlem yapmak için kullanılır.

KULLANILAN VERİ ANALİZİ YÖNTEMLERİ

Veri setinin Lineer Regresyon sınıflandırma algoritmasıyla analizini

yaptım.

Spark Session Oluşturma:
SparkSession oluşturularak Spark bağlantısı sağlanır. Bu bağlamda yerel modda çalışma belirtilmiştir (master("local")).
Veri Setini Yükleme:
sparkSession.read().format("csv") ile bir CSV dosyası okunur.
option("header", "true") ve option("inferSchema", "true") seçenekleri, dosyanın bir başlık satırı içerdiğini ve şema çıkarmanın otomatik olarak yapılmasını sağlar.
StringIndexer ile Kategorik Değişkenleri Dönüştürme:

StringIndexer kullanılarak kategorik değişkenler indekslenir ve sayısallaştırılır.
sütunlara sırasıyla uygulanır.
VectorAssembler ile Özellik Vektörü Oluşturma:
VectorAssembler kullanılarak belirlenen özellik sütunları birleştirilir ve features adlı yeni bir sütun oluşturulur.
Veri Setini Train ve Test Setlerine Bölme:
randomSplit metodu kullanılarak veri seti train ve test setlerine bölünür.
Test Seti Üzerinde Tahmin Yapma:
Eğitilen model kullanılarak test veri seti üzerinde tahminler yapılır.
Sonuçlar predictions adlı bir Dataset içinde tutulur.
Tahmin Sonuçlarını Görüntüleme:
show() metodu kullanılarak tahmin sonuçları görüntülenir.

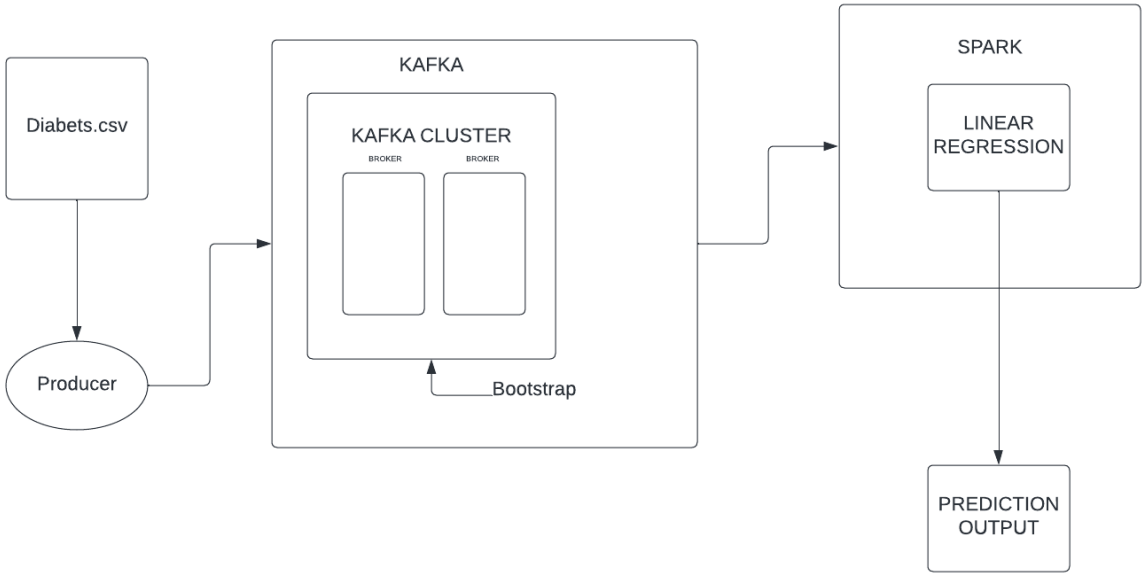
4.KULLANILAN VERİ SETİNİN TANIMLANMASI

Kullandığım veri seti, diyabet teşhisi konulmuş veya konulmamış bireylerin klinik ve demografik özelliklerini içeren bir tıbbi veri setidir. Her bir örnek, bir kişinin hamilelik sayısı, glukoz seviyeleri, kan basıncı, cilt kalınlığı, insülin seviyeleri, vücut kitle indeksi (BMI), diyabet soyağacı fonksiyonu, yaş ve sonuç (1: diyabet teşhisi konulmuş, 0: konulmamış) gibi çeşitli özelliklerini içerir.

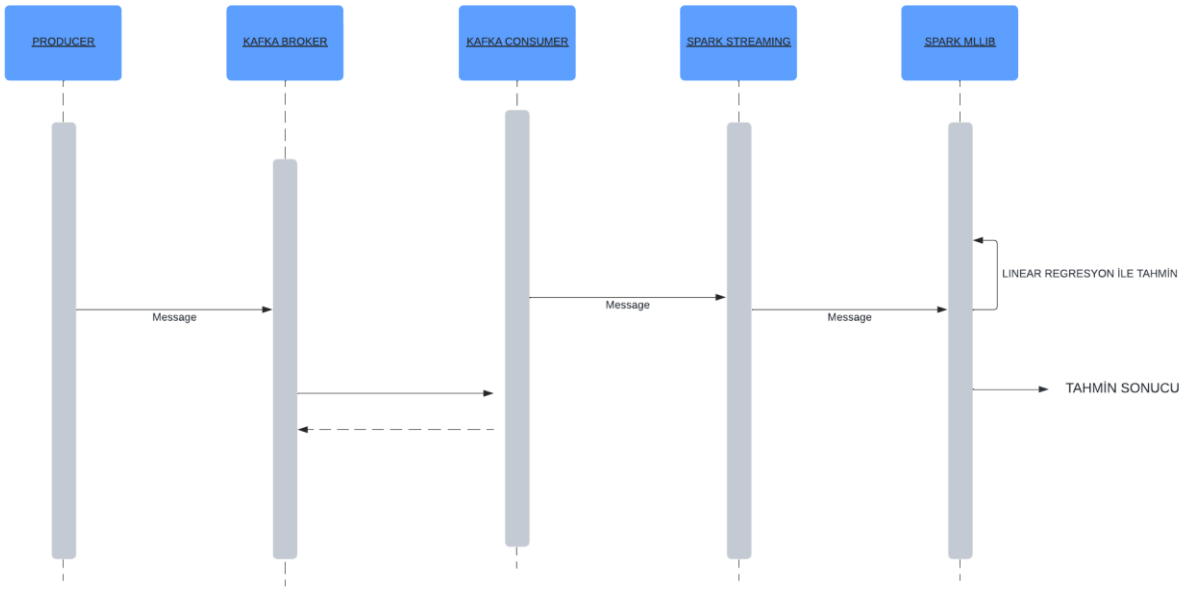
Başka bir deyişle, veri seti, bir kişinin diyabet olup olmadığını belirlemede kullanılabilecek potansiyel öngörücülerin bulunduğu bir öğrenme amacı taşır. Bu özellikler, özellikle diyabetin teşhisinde önemli olan klinik ölçümlerdir.

Veri Setinin Boyutu:
Her bir satırda 9 kolon var.
Veri Türleri:
Bütün veri türleri sayısaldır.

5. AKIŞ ŞEMASI



6. ZAMANLAMA ŞEMASI



7. ELDE EDİLEN BULGULAR

```
Received message: {"Pregnancies": 10, "Glucose": 90, "BloodPressure": 85, "SkinThickness": 32, "Insulin": 0
, "BMI": 34.9, "DiabetesPedigreeFunction": 0.825, "Age": 56, "Outcome": 1}
Predicted Price: 0.4644994281122651, Actual Price: 1
Accuracy: 99.46449942811226%
```