SAKARYA
ÜNİVERSİTESİ

BİLGİSAYAR VE BİLİŞİM BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Büyük Veriye Giriş Ödevi

Hazırlayan:

B201210024 – Ahmet Furkan SÖĞÜTCÜ

1/A Grubu

Dersi Veren:

Arş. Gör. Dr. Nur Banu Oğur

Ödevin Özeti

Bu ödevde bizden büyük veri konusu ile alakalı küçük ölçekli bir veri analitiği sistemi yapmamız istenildi. Yapılması istenen sistemde kafka, spark ve makine öğrenmesi teknolojileri ile projenin gerçeklenmesi gerekiyordu. Ödevin senaryosunda belli bir konu hakkında veri setimiz olacaktı. Bu veri setini ilk olarak kafka ile produce edecektik. Ardından da kafka topic'inde bulunan verileri spark streaming yapısı ile de consumer'a yollayacaktık. Veriler streaming yapısı ile akış halinde geldikten sonra da son olarak verilerimizi bir makine öğrenmesi algoritması ile işleyecektik. Yapılacak olan bu işlemleri ben java programlama dili ile IntelliJ editörü üzerinden gerçekleştirdim.

Kullanılan Teknolojiler

1. Kafka

Kafka verilerin depolanmasına ve analiz edilmesine izin vermek için mesajlaşma, depolama ve akış işlemeyi birleştirir. Performanslı bir şekilde bir sistemden diğer sisteme neredeyse gerçek zamanlı olarak veri transferini gerçekleştirir.

Ben de kafkayı Producer kodumda bir kafkaProducer oluşturarak topic'lere koyulması gereken verileri koydurttum.

2. Spark

Apache Spark, büyük veri kümeleri üzerinde işlem yapılmasını sağlar. Verileri işlerken diskten veri okumadan veya diske veri yazmadan verileri RAM'de tutarak daha hızlı işlem yapılır.

Spark streaming yapısı da gerçek zamanlı veriler ile işlemler ve analizler yapabilmemizi sağlar. Normalde 1.000.000 veri ile hesaplama uzun sürecekken spark streaming mimarisiyle verileri örneğin 1000'e bölerek 1000'lik setler halinde hesaplayarak zamandan kazanırız.

Ben de bu yapıyı verileri consumer ile alma aşamasında belli zaman aralıklarıyla topic'e koyulan verileri çekmesini sağladım. Veriler her topic'e atıldığında streaming yapısı yeni devrini gerçekleştirdiğinde güncel olarak gelen verileri işliyor.

3. Makine Öğrenmesi

Makine öğrenimi, aldığı verilere göre öğrenen ya da performansı iyileştiren ve bu şekilde doğru kararları ve tahminleri yapmaya çalışır.

Ben de makine öğrenmesini bireyin diyabete sahip olup olmadığını hesaplamak için kullandım. Sınıflandırma için ise Naive Bayes modelini kullandım.

Kullanılan Veri Analiz Yöntemleri

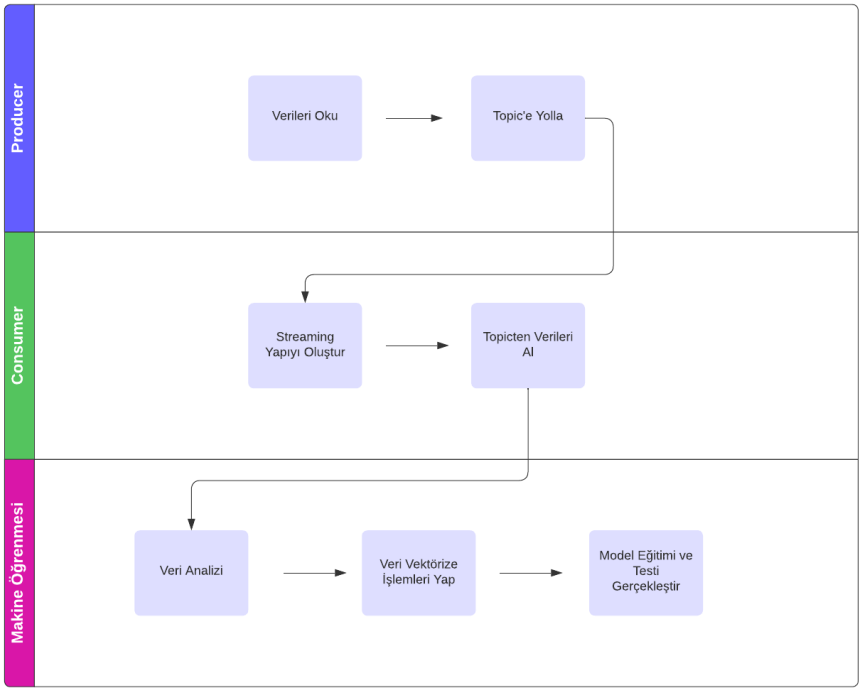
Consumer üzerinden gelen girdi verileri üzerinde makine öğrenmesi yapılacktı lakin ben consumer üzerinden verileri alamadığım için var olan bir csv üzerinden analizlerimi gerçekleştirdim.

İlk olarak csv dosyamı okudum. Ardından headerList() ile her sütunun header'larını belirttim. Ve bunları sırası ile atadım sütunlara. Ardından çıktımın olduğu sütuna label etiketi atadım. Sonrasında da vektörizasyon işlemlerini gerçekleştirerek girdileri features adı altında bir sütunda topladım. Veriler düzenlendikten sonra yüzde 70'e 30'luk bir oranla eğitim ve test veri setlerini böldüm. Veri setleri de ayarlandıktan sonra naive bayes algoritmasını kullanacağımı belirterek onunla eğitim ve test işlemlerini gerçekleştirdim. Bütün işlemler bittikten sonra programın doğruluk oranını ekrana yazdırdım.

Veri Setinin Tanımlanması

Veri seti seçimimi makine öğrenmesinde sınıflandırma algoritmasını tercih etmek istediğimden dolayı diyabet hastalığı verileri olarak yaptım. Kaggle'dan bulduğum bu veri setinde bazı gereksiz girdileri de kendim silerek sadeleştirdim. Girdiler genel olarak diyabet hastalığı olan veya olmayan bir bireyin günlük hayatında neler yaptığı, neler yiyip içtiği veya ne sıklıkla hastaneye gittiği gibi veriler. Bu verilerin yanında da çıktı olarak diyabet hastası olup olmadığı bulunmakta.

Akış Şeması



Zamanlama Şeması

Elde Edilen Bulgular

Tüm programın çalışmasından sonra veriler producer ile alındı ve topic'e koyuldu ardından consumer streaming yapısı yardımı ile de o verilere ulaşıldı. Son olarak da ulaşılan veriler ile de makine öğrenmesi uygulaması kalmıştı lakin streaming yapı ile makine öğrenmesini birleştiremediğimden dolayı makine öğrenmesini dosyadan aldığı veri seti ile yaptı. Doğruluk oranı da yüzde 94 çıktı.