Derin öğrenme (Deep Learning) projesi

Mevsimsel Hastalık Tahmini

Dataset: <https://www.kaggle.com/datasets/sandeep2812/time-series-disease-prediction>

Bu projenin amacı, geçmiş yıllara ait günlük hastalık verilerini kullanarak, belirli bir hastalığın (örneğin **Fever**) yılın hangi mevsimlerinde artış gösterdiğini analiz etmek ve **gelecek mevsimlerde** bu hastalığın vaka sayısını **zaman serisi derin öğrenme modelleri** ile tahmin etmektir. Böylece, sağlık hizmetleri ve kamu yönetimi için **erken uyarı** ve **kaynak planlaması** yapılması hedeflenmektedir.

**Veri Seti İncelemesi:**

Proje, Kaggle'da bulunan "Time Series Seasonal Disease Prediction" veri setini kullanmaktadır. Bu veri seti, belirli hastalıkların zaman içindeki görülme sıklığını içeren zaman serisi verilerini içermektedir. Veri setinin sütunları genellikle şu şekildedir.

* **Tarih**: Verinin kaydedildiği tarih.
* **Hastalık Adı**: İlgili hastalığın adı.
* **Vaka Sayısı**: Belirli bir tarihte kaydedilen vaka sayısı.
* **Bölge**: Verinin toplandığı coğrafi bölge.

Bu sütunlar, hastalıkların zaman içindeki dağılımını ve mevsimsel eğilimlerini analiz etmek için kullanılmaktadır.

**Proje Aşamaları:**

1. **Veri Ön İşleme**: Eksik verilerin doldurulması, tarih formatlarının standartlaştırılması ve gerekirse bölgesel verilerin gruplanması.
2. **Keşifsel Veri Analizi (EDA)**: Hastalıkların zaman içindeki dağılımını görselleştirmek için çizgi grafikleri ve mevsimsel eğilimleri analiz etmek için kutu grafikleri kullanılır.
3. **Zaman Serisi Modelleme**: ARIMA, SARIMA veya Prophet gibi modeller kullanılarak hastalıkların gelecekteki eğilimleri tahmin edilir.
4. **Model Değerlendirme**: Tahminlerin doğruluğunu ölçmek için RMSE (Root Mean Square Error) veya MAE (Mean Absolute Error) gibi metrikler kullanılır.
5. **Sonuçların Yorumlanması**: Elde edilen tahminler, halk sağlığı yetkilileri için politika önerileri geliştirmek amacıyla yorumlanır.

**Gerçek Hayat Uygulamaları:**

Bu tür projeler, grip gibi mevsimsel hastalıkların yayılma dönemlerini önceden tahmin ederek aşı kampanyalarının zamanlamasını optimize etmek veya sağlık hizmetleri kaynaklarını daha etkin bir şekilde tahsis etmek için kullanılabilir.

Eğer bu projeyi geliştirmek veya daha spesifik bir hastalık üzerine odaklanmak isterseniz, veri setini detaylı bir şekilde analiz ederek başlayabilir ve yukarıda belirtilen adımları takip edebilirsiniz.

**Hedef:**

Belirli hastalıkların (örneğin sıtma, grip vb.) **mevsimsel döngülerine** göre **gelecekteki vaka sayılarını tahmin etmek**.

**Kullanılabilecek Derin Öğrenme Modelleri**

**LSTM (Long Short-Term Memory)**

* **Neden?** Zaman serisi verilerinde geçmiş bilgileri uzun vadeli tutabilir.
* **Avantajı**: Mevsimsel dalgalanmalarda geçmiş yılların verisini hatırlayıp etkili tahmin yapabilir.
* **Kullanım**: Giriş olarak örneğin son 12 ayın verisini verip bir sonraki ayın vaka sayısını tahmin etmek.

**GRU (Gated Recurrent Unit)**

* **LSTM’e alternatif**: Daha hızlı çalışır ve daha az parametre içerir.
* **Kullanım**: Özellikle veri boyutu küçükse GRU, LSTM’den daha verimli olabilir.

**CNN-LSTM (Karma Model)**

* **Neden?** CNN bölümü, giriş verisini öznitelik olarak çıkarırken, LSTM zaman boyutundaki ilişkiyi modelleyebilir.
* **Kullanım**: Eğer elimizde çok boyutlu zaman serisi varsa (örn. farklı şehirlerdeki hastalıklar), bu yaklaşım uygundur.

**Encoder-Decoder LSTM**

* **Neden?** Giriş olarak uzun bir zaman aralığını alıp çıktıda belirli bir zaman aralığını tahmin etmek için uygundur (örnek: 12 aylık geçmiş → 3 aylık tahmin).
* **Avantajı**: Çoklu adım (multi-step) zaman serisi tahmini için idealdir.

**Transformer (Time Series Forecasting için Uyarlanmış)**

* **Neden?** Özellikle uzun sekanslı verilerde paralel işlem gücü ile avantaj sağlar.
* **Kullanım**: Seasonal (mevsimsel) etkileri güçlü veri setlerinde self-attention yapısı iyi sonuç verir.
* **Projem, geçmiş yıllara ait bulaşıcı hastalık verilerini analiz ederek, mevsimsel desenleri dikkate alıp, gelecek dönemlerdeki vaka sayılarını tahmin etmeye odaklanan bir zaman serisi tahmin projesidir.** Bu çalışmada derin öğrenme algoritmalarından özellikle LSTM ve GRU

**Uygulama Süreci ve Sonuçların Karşılaştırmalı Değerlendirmesi**

**1. Başlangıç Yöntemi: Temel LSTM ile Mevsimsel Tahmin**

Çalışmanın ilk aşamasında, yalnızca geçmiş sezonlardaki vaka sayıları (total\_cases) dikkate alınarak bir **LSTM (Long Short-Term Memory)** modeli oluşturulmuştur. Modelin eğitilmesi için zaman serisi kaydırma yöntemi (sliding window) uygulanmış ve 3 sezonluk veriyle bir sonraki sezonun vaka sayısı tahmin edilmiştir.

**Elde edilen test performansı:**

* MAE: 935.62
* RMSE: 1281.33
* R² Skoru: 0.6608

Bu sonuç, modelin belirli sezonlardaki mevsimsel örüntüleri öğrenebildiğini göstermiştir. Ancak veri sayısının azlığı ve bazı sezonların hiç vaka içermemesi (örneğin bahar ayları) nedeniyle sapmalar gözlemlenmiştir.

**2. Geliştirilmiş Yöntem: prev\_year\_same\_season Özelliğinin Eklenmesi**

İkinci aşamada, modele **"geçmiş yılın aynı sezonundaki vaka sayısı"** (prev\_year\_same\_season) ek özellik olarak dahil edilmiştir. Amaç, modelin mevsimsel tekrarlılığı ve yıllar arası benzerlikleri daha iyi öğrenmesini sağlamaktır.

Modelin girdi boyutu 1'den 2'ye çıkarılmış, LSTM katmanı yeniden eğitilmiştir. Eğitim sürecinde erken durdurma (EarlyStopping) uygulanarak modelin en iyi doğrulama kaybında durması sağlanmıştır.

**Geliştirilmiş modelin test performansı:**

* MAE: 1698.99
* RMSE: 1766.35
* R² Skoru: 0.3554

Sonuçlara göre model, genel eğilimi takip etmekte başarılı olsa da test setindeki bazı ani değişimleri öngörmekte zorlanmıştır. Bunun temel sebebi, geçmiş sezonlardaki sıfır vaka bilgileri ve sınırlı yıllık veri derinliğidir.

**3. Karşılaştırmalı Sonuç**

| **Model Versiyonu** | **MAE** | **RMSE** | **R² Skoru** |
| --- | --- | --- | --- |
| Temel LSTM (1 öznitelik) | 935.62 | 1281.33 | 0.6608 |
| LSTM + Mevsimsel Özellik (2 öznitelik) | 1698.99 | 1766.35 | 0.3554 |

**4. Yorum**

İlk model daha düşük hata üretse de, ikinci model mevsimsel döngüleri daha iyi açıklama potansiyeline sahiptir. Ancak yeterli veri ve dengeli örneklenme olmadığında eklenen özellikler modelin kararsızlığını artırabilmektedir. Daha fazla yıl verisi veya normalizasyon gibi yöntemler eklenerek ikinci modelin performansı daha da artırılabilir.