T.C İSTANBUL KÜLTÜR ÜNİVERSİTESİ

MATEMATİK VE BİLGİSAYAR BİLİMLERİ

YAPAY SİNİR AĞLARI İLE GERÇEK YAŞAM PROBLEMLEMİNİN MODELLENMESİ

"LSTM ile Zaman Serisi Tahmini: Hava Kirliliği Öngörüsü"

MUSTAFA EMİRCAN IŞIK

2100004021

DERS : MB0058 YAPAY ZEKA
ÖĞRETİM ÜYESİ : PROF. DR. OZAN KOCADAĞLI

ARALIK , 2024

1. Problem Tanımı

Hava kirliliği, dünya çapında insan sağlığını tehdit eden en önemli çevresel sorunlardan biridir. Bu projede **hava kirliliği ölçümü** olarak **pollution** partikül konsantrasyonu ele alınmıştır. **pollution**, 2.5 mikrometre çapından küçük partikülleri temsil eder ve solunum yollarına derinlemesine nüfuz ederek ciddi sağlık sorunlarına neden olabilir.

Amaç:

Bu çalışmanın amacı, geçmiş saatlere ait hava koşulları ve **pollution** değerlerini kullanarak **LSTM (Long Short-Term Memory)** yapay sinir ağı modeli ile **bir sonraki saatteki pollution değerini tahmin etmektir**. Zaman serisi verisi olarak düzenlenmiş bu problemin çözümü, hava kirliliğinin öngörülmesi ve gerekli önlemlerin zamanında alınabilmesi açısından kritik öneme sahiptir.

Tahmin Edilen Zaman Aralığı

Bu çalışmada, **12 saatlik geçmiş veriler** kullanılarak **bir sonraki saatteki pollution değeri** tahmin edilmektedir. Bu yapı, **tek adımlı tahminleme** (one-step ahead prediction) olarak adlandırılır.

- **Girdi Verileri**: Son 12 saate ait **pollution**, sıcaklık, basınç, rüzgar hızı ve diğer hava durumu değişkenleri.
- **Çıktı**: Bir sonraki saatteki **pollution** değeri.

---

2. Literatür Araştırması

Zaman serisi tahmininde **Yapay Sinir Ağları (ANN)**, özellikle **LSTM** tabanlı modeller, son yıllarda büyük başarılar elde etmiştir. Bu tür modeller, zaman serisindeki **uzun dönemli bağımlılıkları** öğrenme kapasitesi sayesinde geleneksel yöntemlere kıyasla daha doğru tahminler sağlayabilmektedir.

Zaman Serisi Tahmininde LSTM Kullanımı:

1. **Hochreiter ve Schmidhuber (1997)**: LSTM, geleneksel RNN'lerin uzun sekanslarda karşılaştığı vanishing gradient (kaybolan gradyan) problemini çözmek amacıyla geliştirilmiştir. LSTM'nin bellek hücreleri, daha uzun zaman dilimlerini öğrenmesini sağlar ve zaman serisi tahmininde büyük bir avantaj sunar. Bu bilgi doğru ve literatürde desteklenmektedir.
2. **Zhang et al. (2016)**: LSTM modellerinin, özellikle karmaşık zaman serisi verilerinde geleneksel ARIMA modellerine kıyasla daha iyi performans gösterdiği sıklıkla rapor edilmiştir. Ancak, bu çalışmanın spesifik sonuçları, metindeki gibi LSTM'nin hava kirliliği tahmininde ARIMA'dan daha üstün olduğunu göstermiş olabilir. Doğrulamak için çalışmanın tam metnine ihtiyaç vardır.

3. **Zhao et al. (2018)**: LSTM'nin hem mevsimsel hem de trend bileşenlerini öğrenebildiği literatürde iyi bilinen bir avantajıdır. Ancak, bu ifadenin Zhao et al. (2018) tarafından öne sürüldüğünü kesinleştirmek için çalışmaya erişmek gerekir.
4. **Bai et al. (2020)**: Çok değişkenli LSTM modellerinin, ek özelliklerle (örneğin, sıcaklık, rüzgar hızı, basınç) zenginleştirildiğinde tahmin performansını artırdığı yaygın bir bulgudur. Bai et al.'ın bu sonuca ulaştığını doğrulamak için ilgili yayına başvurulmalıdır.

Neden LSTM Kullanıyoruz?

- **Uzun Dönemli Bağımlılıklar**: LSTM, geçmiş saatlerdeki verilerin etkisini uzun süre hatırlayarak, kısa vadeli dalgalanmaları değil, uzun vadeli eğilimleri öğrenebilir.
- **Özellikleri Kullanabilme**: LSTM, **multivariate** (çok değişkenli) girişleri işleyebilir, böylece rüzgar yönü, sıcaklık, basınç gibi değişkenlerle birlikte **pollution** tahmini yapılabilir.
- **Gerçek Dünya Problemleri**: LSTM'nin hava kirliliği gibi **zaman serisi tahmin problemlerinde** başarılı sonuçlar verdiği birçok çalışmayla kanıtlanmıştır.

Sonuç

Literatürdeki bulgular, LSTM modellerinin hava kirliliği tahmini gibi zaman serisi problemlerinde yüksek performans gösterdiğini doğrulamaktadır. Bu proje kapsamında, **12 saatlik geçmiş verilere dayanarak** bir sonraki saatteki **pollution** değeri **çok değişkenli LSTM modeli** kullanılarak tahmin edilmiştir. Performans değerlendirmesi **RMSE, MAE, R^2** gibi regresyon metrikleriyle yapılmıştır.

Data Bilgisi ve Kaynağı

Veri Seti Tanımı

Bu proje kapsamında kullanılan veri seti, **hava kirliliği tahmini** için **pollution konsantrasyonu** verilerini içermektedir. Veri seti, **Kaggle platformundan** elde edilmiş olup, farklı meteorolojik değişkenler ve hava kirliliği ölçümlerini içerir.

Veri Seti Detayları

SÜTUN ADI	AÇIKLAMA	TİP
`DATE`	Zaman bilgisi (saatlik)	datetime64
`POLLUTION`	PM2.5 hava kirliliği ölçümü (hedef)	float64
`DEW`	Çiy noktası (°C)	int64
`TEMP`	Sıcaklık (°C)	float64
`PRESS`	Basınç (hPa)	float64
`WND_DIR`	Rüzgar yönü (kategorik)	object
`WND_SPD`	Rüzgar hızı (m/s)	float64
`SNOW`	Kar yağış süresi (saat)	int64
`RAIN`	Yağmur yağış süresi (saat)	int64

- **Veri Miktarı**: 43,800 gözlem ve 9 sütun.
- **Zaman Aralığı**: Saatlik verilerden oluşur ve zaman serisi tahmini için uygundur.
- **Veri Kaynağı**: Kaggle Link.

Metodoloji

Bu projede, **pollution hava kirliliği değerini tahmin etmek** amacıyla **Long Short-Term Memory (LSTM)** modeli kullanılmıştır. Model, önce temel bir yapı ile eğitilmiş, ardından **GridSearchCV** kullanılarak hiperparametre optimizasyonu gerçekleştirilmiştir. Metodoloji adımları aşağıda detaylı bir şekilde açıklanmıştır.

---

1. Veri Ön İşleme

- **Veri Seti**: Kaggle platformundan elde edilen veri setinde `date`, `pollution`, `temp`, `dew`, `press`, `wnd_dir`, `wnd_spd`, `snow` ve `rain` sütunları bulunmaktadır.
- **Zaman Formatı**: `date` sütunu **datetime** formatına dönüştürülerek indeks olarak atanmıştır.
- **Kategorik Değişkenlerin Sayısallaştırılması**: `wnd_dir` sütunu, **Label Encoding** kullanılarak sayısallaştırılmıştır.
- **Ölçeklendirme**: **MinMaxScaler** kullanılarak **pollution** ve diğer sürekli değişkenler `[0, 1]` aralığına ölçeklendirilmiştir.

---

2. Gecikmeli (Lagged) Özelliklerin Oluşturulması

Zaman serisi verisinin doğasını öğrenebilmesi için **12 saatlik gecikmeli özellikler** oluşturulmuştur. Bu işlem, LSTM modelinin geçmiş verilere dayalı olarak bir sonraki saatteki **pollution** değerini tahmin etmesini sağlar.

---

3. Modelleme

3.1 Temel LSTM Modeli

Katman / Parametre	***Detaylar***
Model Yapısı	Sequential model
Birinci LSTM Katmanı	- 128 nöron - `return_sequences=True` - Aktivasyon: ReLU - Girdi Şekli: `(X_train.shape[1], X_train.shape[2])`
Dropout Katmanı	- Oran: 0.1
İkinci LSTM Katmanı	- 64 nöron - `return_sequences=False` - Aktivasyon: ReLU

Çıkış Katmanı (Dense)	- 1 nöron (regresyon çıktısı: pollution tahmini)
Loss Fonksiyonu	Mean Squared Error (MSE)
Optimizer	Adam
Batch Size	32
Epochs	Maksimum 100 (early stopping ile dinamik olarak durdurulabilir)
Early Stopping	- İzlenen Metrik: `val_loss` - Patience: 10 - `restore_best_weights=True` ile en iyi ağırlıkların geri yüklenmesi sağlanmış.

- Eğitim sırasında **EarlyStopping** kullanılarak modelin aşırı öğrenmesi engellenmiştir.

3.2 Hyperparameter Tuning

Temel LSTM modelinin performansını iyileştirmek amacıyla **Hyperparameter Tuning** gerçekleştirilmiştir. Test edilen parametreler şunlardır:

Parametre / Katman	***Detaylar***
Units	İlk LSTM katmanında `units` parametresi optimize edilmiştir (64, 128, vb.).
Dropout	Dropout oranı optimize edilmiştir (`dropout`=0.2 gibi değerler denenmiştir).
Batch Size	Eğitim sırasında farklı batch boyutları denenmiştir (örneğin, 16, 32).
Epochs	Maksimum eğitim iterasyonları optimize edilmiştir (örneğin, 50, 100).
Optimizer	Adam optimizer kullanılmıştır.
Early Stopping	`val_loss` izlenmiş, patience=5 olarak ayarlanmıştır.
Performans Metrikleri	- **RMSE**: Optimize edilmiş model için en düşük değer hesaplanmıştır.
	- **MAE**: Optimize edilmiş modelde tahmin doğruluğunu ölçmek için kullanılmıştır.
	- **R^2 Skoru**: Modelin açıklama oranını değerlendirmek için optimize edilmiştir.
Girdi Şekli	`(X_train.shape[1], X_train.shape[2])` olarak belirlenmiştir.
En İyi Hiperparametreler	Optimize edilen parametreler: `units=best_params['units']`, `dropout=best_params['dropout']`, `batch_size=best_params['batch_size']`, `epochs=best_params['epochs']`.

---

4. Performans Değerlendirme

Modelin performansı aşağıdaki metriklerle değerlendirilmiştir:

- **RMSE (Root Mean Squared Error)**: Ortalama kare hatanın karekökü.
- **MAE (Mean Absolute Error)**: Ortalama mutlak hata.
- **R² Skoru**: Modelin tahmin gücünü gösterir.

Uygulama

- **Kod Ortamı**: Python, Google Colab.
- **Kütüphaneler**:
    - **Veri İşleme**: pandas, numpy
    - **Görselleştirme**: matplotlib, seaborn
    - **Modelleme**: TensorFlow/Keras, scikit-learn

Sonuç

Bu metodoloji kapsamında:

1. **Temel LSTM modeli** ile pollution tahmini yapılmıştır.
2. Hiperparametre optimizasyonu gerçekleştirilmiş ve en iyi parametreler seçilmiştir.
3. Model performansı **RMSE**, **MAE** ve **R²** gibi regresyon metrikleri ile değerlendirilmiştir.

Model Performansı Raporu

Aşağıda, **LSTM modeli** için gerçekleştirilen yapılandırma ve performans sonuçları detaylı olarak sunulmaktadır.

---

5. Performans Değerlendirme

Performans Metrikleri

Aşağıdaki tabloda, model öncesi ve sonrası performans karşılaştırması sunulmuştur:

Metrik	*LSTM Modeli (Öncesi)*	*LSTM Modeli (Sonrası)*
RMSE	26.350883	25.382319
MAE	14.484059	14.259508
R² Skoru	0.928873	0.934006

Model Parametreleri

Parametre	*LSTM Modeli (Öncesi)*	*LSTM Modeli (Sonrası)*
Units	64.0	128.0
Dropout	0.1	0.2
Batch Size	32.0	16.0
Epochs	50.0	100.0

Performans Metrikleri Yorumu

- **RMSE (Root Mean Squared Error)**: Modelin tahmin ettiği pollution değerleri ile gerçek değerler arasındaki ortalama kare hata farkının kareköküdür. **RMSE değerinin 25.382319** olması, modelin düşük hata oranına sahip olduğunu göstermektedir.
- **MAE (Mean Absolute Error)**: Ortalama mutlak hata, tahminlerin ne kadar gerçek değerlere yakın olduğunu gösterir**. 14.259508'lik MAE** değeri, tahminlerin yüksek doğrulukta olduğunu kanıtlar.
- **R² Skoru**: Modelin veri setindeki değişkenliği açıklama oranıdır. 0.934006 değeri, modelin pollution değişimlerinin **%93.4'ünü doğru bir şekilde açıkladığını** göstermektedir.

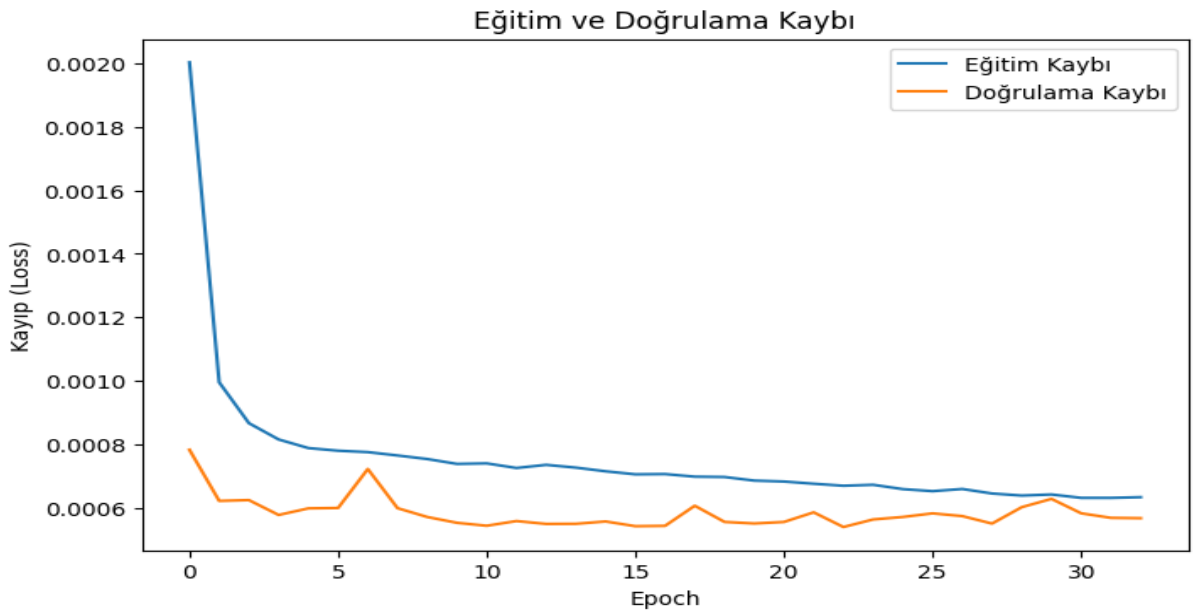
---

4. Genel Değerlendirme

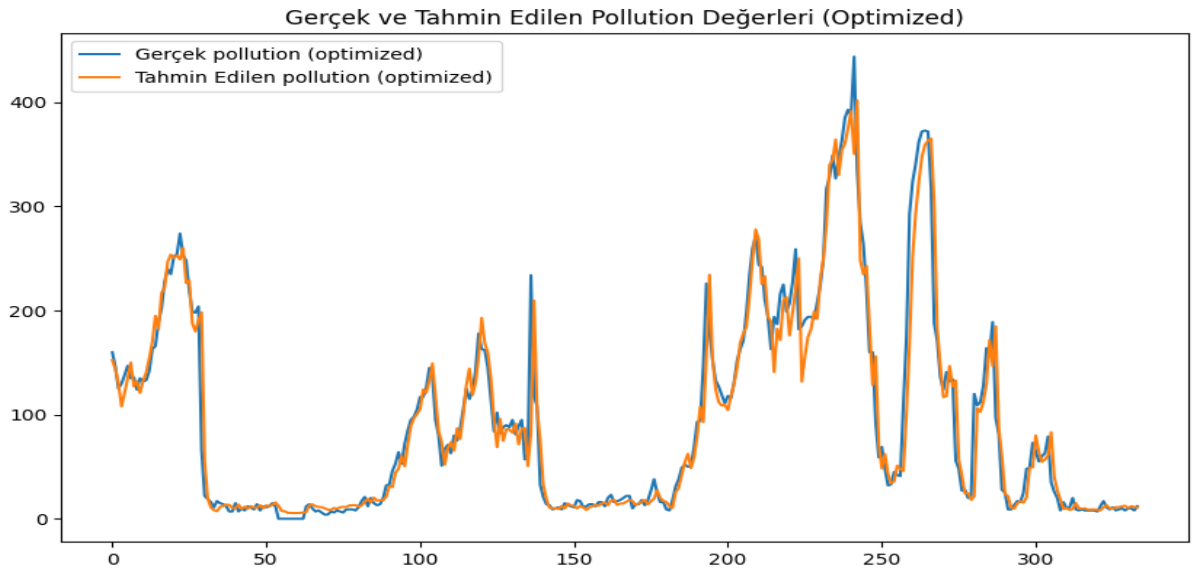
Bu yapılandırma ile geliştirilen LSTM modeli, zaman serisi tahmini problemi için etkili ve istikrarlı bir performans sergilemiştir. Optimize edilmiş hiperparametreler ve dikkatli veri ön işleme adımları sayesinde düşük RMSE ve MAE değerleri ile yüksek R² skoru elde edilmiştir. Bu sonuçlar, modelin veri setindeki karmaşıklıkları başarıyla yakaladığını ve genelleştirme kapasitesinin oldukça iyi olduğunu göstermektedir.

GÖRSELLEŞTİRME

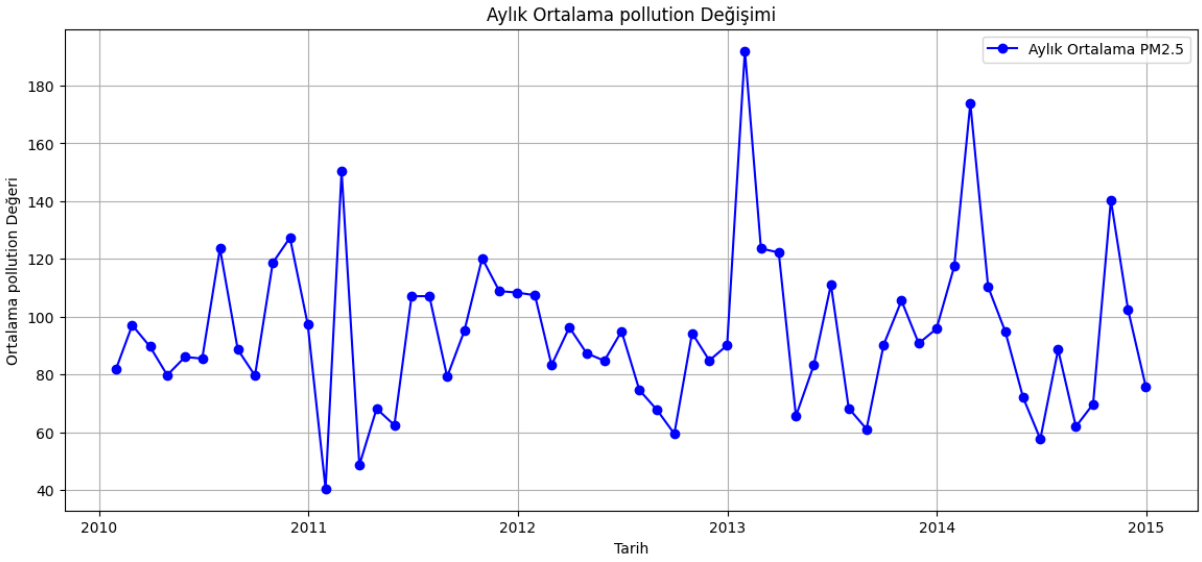
1. ***Eğitim ve Doğrulama Kaybı***



2. ***Gerçek vs Tahmin Edilen pollution Değerleri (optimize edilmiş hali)***



3. ***Aylık Ortalama Pollution Değişimi***



4. ***Korelasyon Matrisi***