

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ



HAVA KALİTESİ TAHMİNİNDE VERİ GÖRSELLEŞTİRME VE
DERİN ÖĞRENME MODELLERİNİN KARŞILAŞTIRILMALI
ANALİZİ

Damla MENGÜŞ

Yüksek Lisans Tezi

YAZILIM MÜHENDİSLİĞİ ANABİLİM DALI

ARALIK 2024

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Yazılım Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

HAVA KALİTESİ TAHMİNİNDE VERİ GÖRSELLEŞTİRME VE
DERİN ÖĞRENME MODELLERİNİN KARŞILAŞTIRILMALI
ANALİZİ

Tez Yazarı
Damla MENGÜŞ

Danışman
Doç. Dr. Bihter DAŞ

ARALIK 2024
ELAZIĞ

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Yazılım Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Başlığı:	Hava Kalitesi Tahmininde Veri Görselleştirme ve Derin Öğrenme Modellerinin Karşılaştırılmalı Analizi
Yazarı:	Damla MENGÜŞ
İlk Teslim Tarihi:	24.11.2024
Savunma Tarihi:	24.12.2024

TEZ ONAYI

Fırat Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına göre hazırlanan bu tez aşağıda imzaları bulunan jüri üyeleri tarafından değerlendirilmiş ve akademik dinleyicilere açık yapılan savunma sonucunda OYBİRLİĞİ ile kabul edilmiştir.

Danışman:	Doç. Dr. Bihter DAŞ Fırat Üniversitesi, Teknoloji Fakültesi	İmza Onayladım
Başkan:	Dr. Öğr. Üyesi Sinem AKYOL Fırat Üniversitesi, Mühendislik Fakültesi	Onayladım
Üye:	Dr. Öğr. Üyesi Fatih TOPALOĞLU Malatya Turgut Özal Üniversitesi, Mühendislik Fakültesi	Onayladım

Bu tez, Enstitü Yönetim Kurulunun/...../20..... tarihli toplantısında tescillenmiştir.

İmza

Prof. Dr. Burhan ERGEN
Enstitü Müdürü

BEYAN

Fırat Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım “Hava Kalitesi Tahmininde Veri Görselleştirme ve Derin Öğrenme Modellerinin Karşılaştırılmalı Analizi” Başlıklı Yüksek Lisans Tezimin içindeki bütün bilgilerin doğru olduğunu, bilgilerin üretilmesi ve sunulmasında bilimsel etik kurallarına uygun davrandığımı, kullandığım bütün kaynakları atıf yaparak belirttiğimi, maddi ve manevi desteği olan tüm kurum/kuruluş ve kişileri belirttiğimi, burada sunduğum veri ve bilgileri unvan almak amacıyla daha önce hiçbir şekilde kullanmadığımı beyan ederim.

24.12.2024

Damla MENGÜŞ



ÖNSÖZ

Hızla artan dünya nüfusu, yoğun şehirleşme ve sanayileşme, modern yaşamın bir gerekliliği haline gelmiş olsa da, bu gelişmelerin en önemli olumsuz etkilerinden biri hava kirliliğinin artmasıdır. Hava kirliliği, insan sağlığı ve çevre üzerinde ciddi sonuçlar doğurmakta, solunum yolu hastalıkları, kalp damar hastalıkları ve çeşitli kanser türleri gibi sağlık sorunlarına yol açmaktadır. Buna ek olarak, atmosferdeki zararlı maddelerin artışı ekosistemleri de olumsuz etkilemekte ve küresel ısınma gibi büyük ölçekli çevresel sorunlara katkıda bulunmaktadır. Bu çalışmanın motivasyonu, özellikle büyük şehirlerde hava kalitesini doğru bir şekilde değerlendirmek ve ileriye dönük tahminlerde bulunarak bu kirliliğe karşı önleyici tedbirlerin alınmasını kolaylaştırmaktır. Mevcut literatürde kullanılan hava kalitesi ölçüm ve tahmin yöntemlerinin, veri eksiklikleri ve yetersiz modeller nedeniyle, belirli sınırlamalara sahip olduğu görülmektedir. Bu tez çalışması, bu sınırlamaları aşmayı ve daha doğru ve uzun vadeli tahminler yapmayı amaçlayan yenilikçi derin öğrenme yaklaşımları ile bu alanda yeni bir katkı sağlamayı hedeflemektedir.

Bu tez çalışması sırasında, karşılaşılan zorluklar ve belirsizlikler motivasyon kaynağı olmuş ve daha ileri analiz yöntemlerinin araştırılmasına vesile olmuştur. Özellikle, derin öğrenme modellerinin karmaşıklığı ve uzun vadeli tahminlerde karşılaşılan performans sorunları, çalışmanın en büyük teknik zorlukları arasında yer almaktadır. Ancak, bu zorlukların aşılması için yapılan çabalar, çalışmanın katkı değerini daha da artırmıştır. Bu tez, mevcut literatüre kıyasla daha kapsamlı veri setleri ve gelişmiş modelleme teknikleri kullanarak, hava kalitesi tahminlerinin doğruluğunu ve hassasiyetini artırmayı amaçlamaktadır.

Bu tezin hazırlanmasında destek ve katkı sağlayan birçok kişi ve kuruma teşekkürlerimi sunmak isterim. Başta, çalışmalarına rehberlik eden, bilgi ve tecrübeleriyle beni yönlendiren danışman hocam Doç. Dr. Bihter DAŞ'a en içten teşekkürlerimi sunarım. Ayrıca, çalışmalarına dolaylı katkı sağlayan ve veri sağlayarak araştırmamı zenginleştiren Türkiye Cumhuriyeti Çevre, Şehircilik ve İklim Değişikliği Bakanlığı ile Meteoroloji Genel Müdürlüğü'ne teşekkür ederim. Destek ve teşviklerinden dolayı bu süreçte yanımda olan aileme de minnettarım.

Damla MENGÜŞ

ELAZIĞ, 2024

İÇİNDEKİLER

Sayfa

ÖNSÖZ.....	iv
İÇİNDEKİLER	v
ÖZET	vii
ABSTRACT	viii
ŞEKİLLER LİSTESİ	ix
TABLOLAR LİSTESİ	x
SİMGELER VE KISALTMALAR	xi
1. GİRİŞ	1
2. LİTERATÜR TARAMASI	3
2.1. Derin Öğrenme Tabanlı Modelleri ile Yapılmış Çalışmalar	3
2.2. Hibrit Modeller ile Yapılmış Çalışmalar	5
2.3. Makine Öğrenmesi ve Regresyon Tabanlı Modeller ile Yapılmış Çalışmalar	7
2.4. İklim ve Çevresel Etkileri İnceleyen Çalışmalar	8
2.5. Trafik ve Hava Kirliliği İlişkisini İnceleyen Çalışmalar	9
2.6. Diğer Modelleri İnceleyen Çalışmalar	9
3. HAVA VERİLERİNE DAYALI BÜYÜK VERİ GÖRSELLEŞTİRME VE DERİN ÖĞRENME	
YAKLAŞIMLARININ İNCELENMESİ	12
3.1. Veri Ön İşleme ve Hazırlık Süreçleri	12
3.1.1. Veri Toplama Süreci	12
3.1.2. Eksik Veri Tamamlama Teknikleri	13
3.1.3. Veri Füzyonu ve Yöntemleri	16
3.1.5. Veri Filtreleme Yöntemleri	18
3.2. Veri Görselleştirme Yöntemleri	19
3.2.1. Hava Verilerinin Görselleştirilmesi	19
3.2.2. Görselleştirme İçin Kullanılan Araçlar	20
3.2.3. Seaborn ile Görselleştirme Teknikleri	21
3.2.4. Plotly ile Görselleştirme Teknikleri	23
3.3. Derin Öğrenme Mimarileri	25
3.3.1. Evrişimsel Sinir Ağı (Convolutional Neural Networks-CNN) ile Görüntü İşleme	25
3.3.2. Tekrarlayan Sinir Ağı (Recurrent Neural Networks - RNN) ile Zaman Serisi Analizi	26
3.3.3. Uzun Kısa Süreli Bellek (Long Short-Term Memory - LSTM) Modelleri ile Gelişmiş Zaman Serisi Analizi	28
4. HAVA VERİLERİ KULLANILARAK BÜYÜK VERİ GÖRSELLEŞTİRMESİ VE DERİN	
ÖĞRENME MODELLERİ İLE ÇIKARIM İÇİN YAPILAN UYGULAMALAR.....	30
4.1. Verilerin Toplanması	30
4.2. Performans Değerlendirme Ölçütleri	31
4.3. Uygulama 1: Hava Kalitesi Tahmininde Veri Görselleştirme ve Derin Öğrenme Modellerinin Karşılaştırılmalı Analizi	32
4.3.1. Verilerin Toplanması ve Ön İşlenmesi	32
4.3.2. Modelin Oluşturulması: Veri Görselleştirme	34
4.3.3. Deneysel Sonuçlar ve Tartışma	40

4.4. Uygulama 2: Hava Kalitesi Tahmininde Veri Görselleştirme ve Derin Öğrenme Modellerinin Karşılaştırmalı Analizi	41
4.4.1. Verilerin Toplaması.....	42
4.4.2. Eksik Verilerin Tamamlanması.....	43
4.4.3. Kullanılan Metotlar ve Analizler.....	45
4.4.4. Deneysel Sonuçlar ve Tartışma	59
5. BULGULAR VE TARTIŞMA	60
6. SONUÇLAR VE ÖNERİLER	62
KAYNAKLAR.....	64
ÖZGEÇMİŞ	



ÖZET

Hava Kalitesi Tahmininde Veri Görselleştirme ve Derin Öğrenme Modellerinin Karşılaştırılmalı Analizi

Damla MENGÜŞ

Yüksek Lisans Tezi

FIRAT ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

Yazılım Mühendisliği Anabilim Dalı

Aralık 2024, Sayfa: xi + 68

Hava kirliliği, günümüzde en önemli çevresel sorunlardan biri olarak öne çıkmaktadır. Artan nüfus, hızlı kentleşme ve sanayileşme gibi faktörler, hava kirliliği seviyelerinin giderek yükselmesine yol açmaktadır. Ülkemizde, farklı bölgelerde havadaki kirlenici maddeler düzenli olarak ölçülmekte ve bu ölçümlerden elde edilen veriler sürekli büyüyen bir veri kümesini oluşturmaktadır. Bununla birlikte, artan veri hacmi bazı zorlukları da beraberinde getirmektedir. Farklı kaynaklardan toplanan verilerin çeşitliliği ve büyüklüğü, analiz ve modelleme süreçlerini daha karmaşık hale getirmektedir. Ayrıca, atmosferdeki kirlenicilerin hareketi ve çevresel değişkenlerin etkileri, hava kalitesinin doğru bir şekilde tahmin edilmesini daha da güçleştirmektedir.

Bu çalışmada, Türkiye Çevre, Şehircilik ve İklim Değişikliği Bakanlığı Sürekli İzleme Merkezi tarafından sağlanan veriler kullanılarak havadaki farklı kirlenici maddelerin üç farklı derin öğrenme yöntemleri olan Evrişimsel Sinir Ağı (CNN), Tekrarlayan Sinir Ağı (RNN) ve Uzun Kısa Süreli Bellek (LSTM) ile verinin karşılaştırılmalı olarak tahminlemesi, tahminlemede en iyi sonuç veren yöntem ile gelecek üç yılın tahmini, daha sonrasında ise bu elde edilen değerlerin görselleştirilmesi sağlanmıştır. Veri görselleştirilmesi olarak Violin Plot, Box Plot ve Point Scatter grafikleri kullanılmıştır. Veri tahminlemesinin, gelecekteki belirsizlikleri azaltma açısından yararlı olacağı düşünülmektedir. Veri görselleştirilmesi ise uzman olmayan kişilerin görüntülenen konsantrasyon profillerinden hava kalitesi bilgilerini tahmin etmesini ve anlamasını kolaylaştırmaktadır. Bu çalışmada kullanılan modeller arasında, CNN ile PM10 için 0.88, SO2 için 0.93, LSTM ile PM10 için 0.94, SO2 için 0.95 doğruluk oranlarına ulaşılmıştır.

Bu tez çalışması, İstanbul ilinin Başakşehir ilçesine ait hava kirliliği verilerini kullanarak, hava kalitesi tahmini ve veri görselleştirme süreçlerinde derin öğrenme modellerinin karşılaştırılmalı analizlerini içermektedir. Çalışmanın özgünlüğü hem görselleştirme hem de derin öğrenme yöntemlerinin birlikte ele alınmasıdır.

Anahtar Kelimeler: Hava Kalitesi Tahmini, Evrişimsel Sinir Ağı (CNN), Tekrarlayan Sinir Ağı (RNN) ve Uzun Kısa Süreli Bellek (LSTM), Veri Görselleştirme

ABSTRACT

Comparative Analysis of Data Visualization and Deep Learning Models in Air Quality Forecasting

Damla MENGÜŞ

Master's Thesis

FIRAT UNIVERSITY
Graduate School of Natural and Applied Sciences
Department of Software Engineering

December 2024, Pages: xi + 68

Air pollution stands out as one of the most important environmental problems today. Factors such as increasing population, rapid urbanization, and industrialization lead to increasing air pollution levels. In our country, pollutants in the air are measured regularly in different regions and the data obtained from these measurements constitute an ever-growing data set. However, the increasing data volume also brings some difficulties. The variety and size of the data collected from different sources make the analysis and modeling processes more complex. In addition, the movement of pollutants in the atmosphere and the effects of environmental variables make it even more difficult to estimate air quality accurately.

In this study, using the data provided by the Continuous Monitoring Center of the Ministry of Environment, Urbanization and Climate Change of Turkey, the data of different pollutants in the air were estimated comparatively with three different deep learning methods, namely Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM), the prediction of the next three years with the method that gave the best results in the prediction and then the visualization of these obtained values were provided. Violin Plot, Box Plot, and Point Scatter graphics were used as data visualization. It is thought that data estimation will be useful in reducing future uncertainties. Data visualization makes it easier for non-experts to estimate and understand air quality information from the displayed concentration profiles. Among the models used in this study, accuracy rates of 0.88 for PM10 with CNN, 0.93 for SO2, 0.94 for PM10 with LSTM, and 0.95 for SO2 were achieved.

This thesis study includes comparative analyses of deep learning models in air quality estimation and data visualization processes using air pollution data from Başakşehir district of Istanbul. The originality of the study is that both visualization and deep learning methods are considered together.

Keywords: Air Quality Prediction, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM), Data Visualization

ŞEKİLLER LİSTESİ

	Sayfa
Şekil 3.1. Genel CNN Model Mimarisi	26
Şekil 3.2. Genel RNN Model Mimarisi	29
Şekil 3.3. Genel LSTM Model Mimarisi	29
Şekil 4.1. Sürekli İzleme Merkezi (SİM) Web Sitesine Ait Görüntü	32
Şekil 4.2. Uygulama 1 İçin Akış Diyagramı	34
Şekil 4.3. PM10 Değeri İçin Point Scatter	35
Şekil 4.4. SO2 Değeri İçin Point Scatter	38
Şekil 4.5. PM10 Değeri İçin Line Scatter	39
Şekil 4.6. SO2 Değeri İçin Line Scatter	40
Şekil 4.7. PM10 Değeri İçin Bar Plotly	41
Şekil 4.8. SO2 Değeri İçin Bar Plotly	42
Şekil 4.9. Uygulama 2 İçin Akış Diyagramı	44
Şekil 4.10. CNN Model Mimarisi	48
Şekil 4.11. RNN Model Mimarisi	48
Şekil 4.12. LSTM Model Mimarisi	49
Şekil 4.13. CNN Modeli ile Tahminleme	51
Şekil 4.14. RNN Modeli ile Tahminleme	52
Şekil 4.15. LSTM Modeli ile Tahminleme	52
Şekil 4.16. RNN ile Gelecek 3 Yılın Tahmin Grafiği	54
Şekil 4.17. Keman Grafiği (Violin Plot) ile Görselleştirme	56
Şekil 4.18. Kutu Grafiği (Box Plot) ile Görselleştirme	58
Şekil 4.19. Nokta Dağılım Grafiği (Point Scatter Plot) ile Görselleştirme	60

TABLÖLAR LİSTESİ

	Sayfa
Tablo 4.1. Başakşehir Eksik Hava Verileri (10 Satır)	34
Tablo 4.2. Başakşehir Tamamlanmış Hava Verileri (10 Satır)	34
Tablo 4.3. Verisetinin Asıl Halinin İlk 10 Satırı	44
Tablo 4.4. Verisetinin K-NN İle Doldurulmuş Hali	45
Tablo 4.5. CNN Modeli Performans Metrikleri	48
Tablo 4.6. RNN Modeli Performans Metrikleri	49
Tablo 4.7. LSTM Modeli Performans Metrikleri	49
Tablo 4.8. RNN Modeli ile Gelecek Tahminlemede Performans Metrikleri	52

SİMGELER VE KISALTMALAR

Simgeler

CO	: Karbon Monoksit
NO	: Nitrik Oksit
NO _x	: Azot Oksitleri
NO ₂	: Azot Dioksit
O ₃	: Ozon
PM ₁₀	: Partiküler Madde
SO ₂	: Kükürt Dioksit

Kısaltmalar

CNN	: Evrişimsel Sinir Ağı
RNN	: Tekrarlayan Sinir Ağı
LSTM	: Uzun Kısa Süreli Bellek
SİM	: Sürekli İzleme Merkezi
K-NN	: K-En Yakın Komşu Algoritması
MAE	: Ortalama Mutlak Hata
RMSE	: Kök Ortalama Kare Hatası
R ²	: R-Kare
AQI	: Hava Kalitesi Endeksi

1. GİRİŞ

Hava kirliliği, zararlı maddelerin insan sağlığını ve çevrenin dengesini olumsuz etkileyebilecek seviyelerde atmosferde bulunması durumunda ortaya çıkmaktadır. Artan nüfus, hızlı kentleşme ve hızlanan endüstriyel büyüme gibi faktörler hava kalitesinin bozulmasına önemli ölçüde katkıda bulunmaktadır [1], [2], [3]. Bu eğilimler fosil yakıt tüketiminde artışa, artan araç trafiğine ve genişleyen endüstriyel faaliyetlere yol açmaktadır. Hava kirliliğine katkıda bulunan başlıca kirleticiler arasında karbon monoksit (CO), azot dioksit (NO₂), kükürt dioksit (SO₂), ozon (O₃) ve partikül madde (PM) gibi gazlar bulunmaktadır. Azot oksitler (NO_x), özellikle nitrik oksit (NO) ve azot dioksit (NO₂), hava kalitesini olumsuz etkileyen atmosferik kimyasal reaksiyonların temel bileşenlerindendir [4],[5],[6]. Bunun yanı sıra, PM₁₀ ve PM_{2.5} gibi partikül maddeler, insan sağlığı üzerinde yarattıkları ciddi etkiler nedeniyle en endişe verici kirleticiler arasında bulunmaktadır. Bu parçacıklar o kadar küçüktür ki solunum savunmalarını aşarak akciğerlerin derinliklerine nüfuz edilebilmekte ve burada uzun süreli maruz kalma kanser dahil ciddi sağlık sorunlarına yol açabilmektedirler. Solunduğunda, SO₂ solunum sistemini tahriş eden ve önemli bir sağlık riski oluşturan başka bir tehlikeli kirlenici olarak bilinmektedir [7],[8],[9]. Hava kalitesi genellikle meteorolojik koşullara dayalı kirlilik seviyesini yansıtan bir endeks kullanılarak değerlendirilmektedir. Ancak, hava kalitesi endeksi ölçüm yanlışlıkları veya yetersiz sensör verileri nedeniyle hava kirliliğinin gerçek etkisini her zaman yakalayamamakta ve bu da gecikmiş yanıtlarla sonuçlanabilmektedir. Hava kalitesi endeksi kirliliğin halk sağlığı üzerindeki zararlı etkilerini değerlendirmek için yaygın olarak kullanılmasına rağmen, ölçümlerin hassasiyeti ve hassasiyeti iyileştirilmelidir. Dahası, mevcut ölçümler genellikle günlük olarak yapılmakta ve öngörücü içgörüler sağlayamamakta, bu da zamanında önleyici tedbirlerin uygulanmasını zorlaştırmaktadır. Ölçülen hava kalitesi endeksi arttıkça, insan sağlığı için artan risklerle kötüleşen hava koşullarını göstermektedir. Endeks, hava kalitesini 0-50 arasında "iyi", 51-100 arasında "orta", 101-150 arasında "hassas", 151-200 arasında "sağlıksız", 201-300 arasında "çok sağlıksız" ve 300'ün üzerinde "tehlikeli" olarak sınıflandırılmaktadır. Endeks 151'i aştığında, sağlık sorunları olasılığı artmakta ve açık hava aktiviteleri riskli hale gelmektedir [10],[11].

Hava kalitesini görselleştirme ve tahmin süreçlerinde karşılaşılan zorluklar, genellikle verilerin doğasından ve çevresel faktörlerin karmaşıklığından kaynaklanmaktadır. Hava kirliliği verileri, genellikle eksik ve düzensiz olmasının yanı sıra farklı kaynaklardan gelen ölçümler nedeniyle çeşitlilik ve uyumsuzluk gösterebilmektedir. Bu durum, verilerin doğru bir şekilde işlenmesini ve güvenilir tahminler yapılmasını zorlaştırmaktadır. Ayrıca, hava kirliliği seviyelerini etkileyen atmosferik olaylar, meteorolojik değişkenler ve bölgesel farklılıklar, modellerin doğruluğunu artırma sürecinde dikkate alınması gereken önemli unsurlar arasında yer almaktadır. Zaman serisi verilerindeki uzun ve kısa vadeli bağımlılıkların doğru bir şekilde modellenmesi hem tahminlerin

doğruluğunu hem de çevresel değişkenlerin etkilerini anlamayı daha da karmaşık hale getirebilmektedir. Sonuç olarak, eksik veri doldurma, gürültü azaltma, gelişmiş derin öğrenme modellerinin kullanımı ve etkili görselleştirme yöntemlerinin bir arada uygulanması, hava kalitesi tahminindeki zorlukları çözmede ve güvenilir analizler sunmada önemli bir bütünlük sağlamaktadır. Bu yaklaşımlar, sadece mevcut hava kirliliği seviyelerinin anlaşılmasını değil, aynı zamanda gelecekteki olası eğilimlerin öngörülmesini de kolaylaştırmakta, çevresel politika ve stratejilerin geliştirilmesine bilimsel bir temel oluşturmaktadır. Böylece hem yerel hem de küresel ölçekte daha sürdürülebilir bir çevresel yönetim mümkün hale gelmektedir.

Bu çalışmada, İstanbul ilinin Başakşehir ilçesine ait hava kirliliği verileri kullanılarak veri görselleştirme gerçekleştirilmiş, ardından İstanbul ilinin Başakşehir ilçesinin hava kalitesi tahmini için RNN, CNN ve LSTM gibi üç farklı derin öğrenme yöntemi kullanılarak karşılaştırmalı analizler yapılmıştır. Analiz, en doğru tahminleri sağlayan modeli belirlemekte ve bu model daha sonra önümüzdeki üç yıl boyunca hava kalitesini tahmin etmek için kullanılmaktadır. Sonuçlar, netliği artırmak için görselleştirmeler aracılığıyla sunulmaktadır. Mevcut literatürdeki boşluklardan biri, bu yöntemleri kullanan karşılaştırmalı çalışmaların eksikliği ve hibrit modellerin yokluğundan kaynaklanmaktadır. Bu çalışma, daha fazla görselleştirme tabanlı içgörüler sunarak bu boşluğu kapatmayı amaçlamaktadır. Önerilen çözüm, hava kalitesi ölçümündeki mevcut sınırlamaların üstesinden gelmekte ve gelecekteki tahminleri oluşturmak için uygulanabilir bir yöntem sunmaktadır. Hava kalitesi verileri, İstanbul genelinde 39 farklı konumdaki sürekli izleme istasyonlarından (SİM) toplanmaktadır ve bu da onu şehir genelindeki hava kalitesi eğilimlerini izlemek ve veriye dayalı tahminler oluşturmak için oldukça etkili hale getirmektedir.

Bu çalışmanın temel katkıları şunlardır:

- İstanbul ili Başakşehir ilçesinin hava kalitesi verilerini tahmin etmek için LSTM, RNN ve CNN yöntemleri kullanılmıştır ve bu üç yöntem karşılaştırmalı olarak analiz edilmiştir. Karşılaştırma sonucunda en yüksek doğruluğa sahip yöntem belirlenmiş ve önümüzdeki üç yıla ait hava kalitesi tahminleri bu yöntemle yapılmaktadır. Bu çalışma, büyük veri kümelerini görselleştirmek ve farklı derin öğrenme modellerinin performanslarını karşılaştırmak için kapsamlı bir yaklaşım sunarak çevresel veri analizi ve tahmin süreçlerine önemli katkılar sağlamaktadır.
- Görselleştirme yöntemleri olarak Kutu Grafikleri (Box Plots), Keman Grafikleri (Violin Plots) ve Nokta Dağılım Grafikleri (Scatter Plots) kullanılarak hava kirlleticilerine ait büyük verilerin yorumlamasına yardımcı olunmuştur. Bu çalışma, veri görselleştirme ve tahmin yöntemlerinin etkinliğini artırmayı amaçlamaktadır. Böylece uzman olmayan kullanıcılar bile karmaşık veri kümelerini yorumlayabilmektedir.

2. LİTERATÜR TARAMASI

Bu bölümde, hava kalitesi tahmini ve izlenmesinde kullanılan farklı yöntemler ve teknikler üzerine literatürde yapılmış önemli çalışmalar incelenmiştir. Derin öğrenme, makine öğrenmesi, hibrit modeller ve çevresel faktörlerin analizi gibi metodolojiler, hava kalitesini izleme ve tahmin süreçlerinde kritik bir öneme sahiptir. Literatürde yer alan bu çalışmalar, hava kalitesinin tahmini ve izlenmesine yönelik yöntemlerin güçlü ve zayıf yönlerini kapsamlı bir şekilde ortaya koymaktadır. Alt başlıklarda, bu yöntemlere dayalı olarak gerçekleştirilen çalışmalar detaylı bir şekilde ele alınmıştır. Hava kalitesini tahmin etmek için kullanılan derin öğrenme tabanlı modeller, makine öğrenmesi teknikleri ve hibrit yaklaşımlar karşılaştırılarak analiz edilmiştir. Ayrıca, çevresel faktörlerin bu süreçlere entegrasyonu ve etkileri üzerine yapılan araştırmalar da incelenmiştir.

Tezin ilerleyen bölümlerinde, bu yöntemlerden elde edilen bulgular ışığında önerilen modellerin detayları ele alınacak ve hava kalitesini iyileştirme stratejilerine yönelik öneriler sunulacaktır. Literatür incelemesi, mevcut yaklaşımların kapsamlı bir değerlendirmesini sağlayarak çalışmanın metodolojik temellerini güçlendirmektedir.

2.1. Derin Öğrenme Tabanlı Modelleri ile Yapılmış Çalışmalar

Derin öğrenme teknikleri, hava kalitesi tahmini için büyük veri ve zaman serisi analizlerinde önemli avantajlar sunmaktadır. Bu alt başlıkta, özellikle derin sinir ağları, LSTM ve CNN gibi modellerin uygulamalarını ele alan çalışmalar özetlenmiştir:

Huang ve arkadaşları tarafından yapılan çalışmada, PM_{2.5} partikül maddeyi tahmin etmek için derin öğrenme yöntemleri LSTM ve CNN ağlarını birleştiren bir model geliştirmeyi amaçlamaktadır. Model, meteorolojik verilerle desteklenerek yüksek doğruluk sağlamış ve diğer yöntemlere kıyasla üstün performans göstermiştir. Ancak, veri kalitesi ve erişilebilirliği, farklı bölgelerdeki genellenebilirlik, yüksek hesaplama maliyetleri ve ani hava değişimlerinin hesaba katılamaması gibi sınırlamalar bulunmaktadır. Model, akıllı şehirlerdeki hava kalitesini izlemek ve tahmin etmek için uygulanmaktadır. Akıllı şehirler için hava kalitesi tahmininde hibrit derin öğrenme modellerinin etkinliğini vurgulayan değerli bir çalışmadır [12]. Benzer şekilde, Janarthanan ve ekibi, büyük bir metropol alanındaki hava kalitesi endeksini (AQI) tahmin etmek için LSTM ve CNN modellerini birleştiren bir derin öğrenme mimarisi geliştirmiştir. Eksik verilerin doldurulması için K-En Yakın Komşu (K-NN) algoritması, gürültünün azaltılması için Gaussian filtreleme kullanılmış ve veriler ölçeklendirilmiştir. Model, özellikle PM_{2.5} ve PM₁₀ tahminlerinde yüksek doğruluk göstermiş ve diğer yöntemlere göre üstün performans sergilemiştir. Ancak, genellenebilirlik ve farklı bölgelerdeki uygulanabilirlik gibi konularda daha fazla çalışmaya

ihtiyaç olduğu vurgulanmıştır [13]. Wu ve Lin, günlük Hava Kalitesi Endeksi (AQI) tahmini için VMD, SE ve LSTM yöntemlerini birleştiren hibrit bir model geliştirmiştir. VMD ile zaman serisi verileri alt bileşenlere ayrılmış, SE ile bileşenlerin karmaşıklıkları değerlendirilmiş ve LSTM ile tahminler yapılmıştır. Model, geleneksel yöntemlere kıyasla daha yüksek doğruluk sağlamış, PM10 tahminlerinde %87'lik bir iyileşme göstermiştir. Ancak, modelin genellenebilirliği, hesaplama maliyetleri ve ani hava değişimlerini tam olarak hesaba katamaması gibi sınırlamaları bulunmaktadır. Bu çalışma, hibrit modellerin hava kalitesi tahmininde etkinliğini ortaya koymaktadır [14]. Zhang ve ekibi, Kuzey Çin'deki mevsimsel hava kirleticilerinin dağılımını tahmin etmek için LSTM ve CNN tabanlı hibrit bir model geliştirmiştir. Çalışma, PM2.5 ve PM10 seviyelerindeki mevsimsel değişiklikleri yüksek doğrulukla öngörmeyi başarmıştır. Derin öğrenmenin mevsimsel analizlerdeki potansiyelini vurgulayan bu çalışma, hava kalitesi tahmininde yeni bir yaklaşım sunmaktadır. Ancak, modelin genellenebilirliği ve meteorolojik faktörlerin entegrasyonu konusunda daha fazla araştırmaya ihtiyaç duyulmaktadır [15]. Luo ve Gong, hava kirleticilerinin tahmini için ARIMA, Balina Optimizasyon Algoritması (WOA) ve LSTM modellerini birleştiren bir hibrit model geliştirmiştir. ARIMA kısa vadeli trendleri tahmin ederken, LSTM uzun vadeli bağımlılıkları öğrenmiş ve WOA model parametrelerinin optimize edilmesini sağlamıştır. Bu model, geleneksel yöntemlere kıyasla tahmin doğruluğunu artırmış, özellikle PM2.5 ve PM10 konsantrasyonlarının öngörülmesinde yüksek performans göstermiştir. Çalışma, algoritmaların entegre kullanımının hava kirliliği tahmininde etkili olduğunu ortaya koymaktadır. Ancak, modelin farklı coğrafi alanlarda genellenebilirliği ve yüksek hesaplama maliyetleri, sınırlamalar arasında yer almaktadır. Ayrıca, meteorolojik faktörlerin etkilerinin daha fazla araştırılması gerektiği belirtilmiştir [16]. Drewil ve arkadaşları, hava kirliliği tahmini için LSTM derin öğrenme modeli ile metasezgisel algoritmaların bir kombinasyonunu kullanmıştır. LSTM, zaman serisi verilerindeki uzun vadeli bağımlılıkları öğrenirken, metasezgisel algoritmalar, model parametrelerinin optimize edilmesini sağlamıştır. Çalışmada, bu entegrasyonun hava kirliliği tahminindeki doğruluğu ve güvenilirliği artırdığı vurgulanmıştır. Model, özellikle PM2.5 ve PM10 gibi kirleticilerin tahmininde etkili sonuçlar elde etmiştir. Araştırma, derin öğrenme ve metasezgisel yaklaşımların birlikte kullanımının, hava kirliliği tahmininde yeni bir perspektif sunduğunu ortaya koymuştur. Ancak, modelin karmaşıklığı nedeniyle yüksek hesaplama maliyetleri ve farklı coğrafi bölgelerde genellenebilirlik konularında sınırlamalar bulunmaktadır [17]. Alternatif bir çalışmada, hava kirletici seviyelerini tahmin etmek için LSTM tabanlı bir sinir ağı modeli kullanılmaktadır. LSTM, zaman serisi verilerindeki uzun vadeli bağımlılıkları belirleme ve öğrenme kapasitesiyle öne çıkmaktadır. Çalışmada, LSTM'nin hava kalitesi tahmininde yüksek doğruluk sağladığı ve özellikle PM2.5 ve PM10 tahminlerinde başarılı olduğu belirtilmiştir. Bu çalışma, hava kirliliği yönetim stratejileri için tahmini modellerin önemine dikkat çekmektedir. Ancak, modelin farklı coğrafi bölgelerde uygulanabilirliği sınırlı kalmış ve meteorolojik

değişkenlerin modele entegrasyonunun artırılması gerektiği vurgulanmıştır [18]. Das ve diğerleri, büyük bir metropol şehirde hava kirleticilerinin tahmini için derin öğrenme yöntemlerini incelemiştir. Çalışma, hava kalitesi dinamiklerini analiz etmek ve tahmin doğruluğunu artırmak amacıyla farklı derin öğrenme algoritmalarını uygulamıştır. Elde edilen sonuçlar, bu yöntemlerin hava kalitesini değerlendirme ve öngörmeye yüksek doğruluk sağladığını göstermiştir. Bu araştırma, büyük şehirlerdeki hava kalitesinin değerlendirilmesi için derin öğrenme yöntemlerinin etkinliğini vurgulayan değerli bir katkı sunmuştur. Ancak, veri işleme süreçlerinde yüksek hesaplama maliyetleri ve modelin genellenebilirliği gibi sınırlamalar dikkat çekmektedir [19]. Sakarya gibi endüstriyel şehirlerde yapılan bir çalışmada, SO₂ ve PM₁₀ gibi hava kirleticilerinin seviyelerini tahmin etmek için RNN tabanlı bir model uygulanmıştır. Çalışma, endüstriyel bölgelerde hava kirliliği tahmininin kritik önemini vurgulamış ve RNN modellerinin bu tür ortamlarda etkili olduğunu göstermiştir. Model, hava kirliliği seviyelerini yüksek doğrulukla tahmin etmiş, özellikle kısa vadeli tahminlerde başarılı olmuştur. Ancak, endüstriyel bölgelerdeki farklı kirlilik kaynaklarının etkilerinin detaylı incelenmesi ve uzun vadeli bağımlılıkların modellenmesi gerektiği belirtilmiştir. Ayrıca, modelin daha geniş veri setleriyle eğitilmesi önerilmiştir [20].

Bu çalışmalar, hava kirliliği tahmininde derin öğrenme modellerinin etkinliğini ortaya koymakla birlikte bazı eksiklikler ve sınırlamalar içermektedir. Çoğu çalışma, belirli bölgeler veya veri setleriyle sınırlı kalmış ve farklı coğrafi alanlarda genellenebilirlik sorunlarıyla karşılaşmıştır. Ek olarak, modellerin hesaplama maliyetleri yüksek olup, özellikle büyük veri setlerinde pratik uygulamalarını sınırlayabilmektedir. Meteorolojik faktörlerin ve ani hava değişimlerinin etkilerinin çoğu modelde tam olarak hesaba katılmaması, tahmin doğruluğunu sınırlandıran bir diğer faktör olmaktadır. Bunun yanı sıra, endüstriyel bölgelerdeki farklı kirlilik kaynaklarının detaylı bir şekilde incelenmesi ve uzun vadeli bağımlılıkların modellenmesi gerekliliği de belirtilmiştir. Genel olarak, bu çalışmalar, hava kirliliği tahmininde derin öğrenme modellerinin potansiyelini ve etkinliğini vurgulamakta, ancak bu modellerin genellenebilirliğini artıracak, hesaplama maliyetlerini düşürecek ve çevresel faktörleri daha geniş bir çerçevede ele alacak şekilde geliştirilmesi gerektiğini ortaya koymaktadır. Bu eksikliklerin giderilmesi, hava kalitesi yönetiminde daha güvenilir ve uygulanabilir modellerin geliştirilmesine katkı sağlayacaktır.

2.2. Hibrit Modeller ile Yapılmış Çalışmalar

Hibrit modeller, birden fazla algoritmanın güçlü yönlerini birleştirerek tahmin doğruluğunu artırmayı hedeflemektedir. Bu bölümde, hibrit model yaklaşımlarıyla hava kalitesi tahmini yapan çalışmalar ele alınmıştır:

Akinosho ve ekibi, trafikle ilişkili hava kirliliği tahmini için çok amaçlı bir regresyon modeli geliştirmiştir. Model, birden fazla kirleticiyi aynı anda tahmin ederek performansını artırmıştır.

Çalışma, çok amaçlı derin öğrenme modellerinin hava kirliliği tahminindeki potansiyelini göstermiştir. Ancak, modelin hesaplama maliyetleri ve farklı trafik koşullarına genellenebilirliği sınırlayıcı faktörler arasında yer almaktadır [21]. Başka bir çalışma, PM2.5 konsantrasyonlarını tahmin etmek için hibrit bir optimizasyon modeli önermiştir. Model, diğer hava kirleticileri ve meteorolojik koşulları hesaba katarak tahmin doğruluğunu artırmayı amaçlamıştır. Hibrit modellerin önemini vurgulayan bu çalışma, geleneksel yöntemlere kıyasla daha yüksek performans sergilemiştir. Ancak, veri hazırlama ve hesaplama süreçlerinin karmaşıklığı bir sınırlama olarak belirtilmiştir [22]. Başka bir çalışmada, COVID-19 karantina döneminde İspanya'daki kentsel trafik alanlarında hava kirleticilerindeki değişiklikleri veri madenciliği teknikleriyle analiz etmiştir. Pandemi sırasında hava kalitesindeki değişiklikleri anlamaya yönelik değerli bir vaka çalışması sunan bu araştırma, mevcut modellerin farklı olağanüstü durumlara uyarlanabilirliğini göstermiştir. Ancak, pandemi sonrası koşulların modele entegrasyonu eksik kalmıştır [23]. Lin ve ekibi, meteorolojik ve trafik faktörlerinin hava kirleticileri üzerindeki etkilerini değerlendirmek için Bayes ağları ve derin öğrenme modellerini kullanmıştır. Çalışma, bu yöntemlerin hava kalitesini etkileyen faktörlerin modellenmesindeki etkinliğini vurgulamaktadır. Ancak, Bayes ağlarının büyük veri setlerinde düşük ölçeklenebilirliği ve derin öğrenme modellerinin hesaplama maliyetleri gibi sınırlamalar bulunmaktadır [24]. Fong ve ekibi, hava kirleticilerinin konsantrasyonlarını tahmin etmek için transfer öğrenimi ve RNN modellerini birleştirmiştir. Model, transfer öğrenimi sayesinde daha az veriye sahip kirleticilerde tahmin doğruluğunu artırmıştır. Çalışma, transfer öğreniminin hava kirliliği tahminindeki önemini göstermiştir. Ancak, transfer öğrenimi sürecinde kullanılan kaynak verilerin kalitesine olan bağımlılık bir sınırlama olarak belirtilmiştir [25]. Çin'deki bir çalışmada, mekansal-zamansal özellikleri hesaba katan tek boyutlu, çok ölçekli bir CNN-LSTM modeli geliştirilmiştir. Model hem mekansal hem de zamansal faktörleri birleştirerek hava kalitesi tahmininde yüksek doğruluk sağlamıştır. Ancak, mekansal veri eksikliği ve modelin karmaşıklığı pratik uygulanabilirliğini sınırlamıştır [26]. Başka bir çalışma, çeşitli hava kirleticiler için sinir transfer öğrenimi yaklaşımını kullanarak tahmin doğruluğunu artırmayı hedeflemiştir. Model, farklı kirleticiler arasında bilgi aktarımını mümkün kılarak doğruluk ve verimlilik sağlamıştır. Çalışma, transfer öğreniminin hava kirliliği tahmininde umut vaat eden bir yöntem olduğunu göstermiştir. Ancak, farklı bölgelerde transfer öğreniminin genellenebilirliği ve veri setleri arasındaki uyum eksikliği, modelin sınırlamaları arasında yer almaktadır [27].

Bu çalışmalar, hava kirliliği tahmini için yenilikçi yaklaşımlar sunmakla birlikte bazı eksiklikler ve araştırma boşlukları içermektedir. Çok amaçlı modellerin hesaplama maliyetleri ve trafik koşullarına genellenebilirlik sorunları dikkat çekerken, hibrit modellerde veri hazırlama süreçlerinin karmaşıklığı önemli bir zorluk oluşturmaktadır. COVID-19 dönemi analizlerinde pandemi sonrası koşulların ele alınmamış olması ve Bayes ağlarının büyük veri setlerindeki düşük ölçeklenebilirliği gibi sorunlar modellerin etkinliğini sınırlandırmaktadır. Transfer öğrenimi

uygulamalarında, kullanılan kaynak verilerin kalitesine bağımlılık ve farklı bölgelerde genellenebilirlik eksiklikleri gözlenmiştir. Mekansal-zamansal modellerde ise mekansal veri eksikliği ve modelin karmaşıklığı pratik kullanımı zorlaştırmıştır. Genel olarak, bu eksiklikler, hava kirliliği tahmininde daha kapsamlı ve genellenebilir modeller geliştirme ihtiyacını vurgulamakta ve çevresel yönetim stratejilerinin iyileştirilmesi için önemli bir fırsat sunmaktadır.

2.3. Makine Öğrenmesi ve Regresyon Tabanlı Modeller ile Yapılmış Çalışmalar

Makine öğrenmesi ve regresyon yöntemleri, hava kirliliği verilerini anlamak ve tahmin etmek için geleneksel yaklaşımlara kıyasla daha hızlı ve ölçeklenebilir çözümler sunmaktadır. Bu bölümde, bu yöntemlerin kullanımına dair yapılan çalışmalar derlenmiştir:

Rani ve diğerleri, Tamil Nadu'daki hava kalitesini tahmin etmek için yapay zeka tabanlı bir algoritma geliştirmiştir. Bu algoritma, hava kirlleticilerinin seviyelerini analiz ederek tahmin doğruluğunu artırmıştır. Çalışma, yapay zekâ tekniklerinin hava kalitesi tahminindeki yenilikçi kullanımını göstermektedir. Ancak, modelin genellenebilirliği ve farklı veri setleriyle uyumu sınırlı kalmıştır. Ayrıca, tahminlerde meteorolojik faktörlerin entegrasyonu eksik kalmıştır [28]. Masmoudi ve arkadaşları, birden fazla hava kirleticisinin konsantrasyonunu tahmin etmek için çok amaçlı regresyon ve özellik seçimi yöntemlerini içeren bir makine öğrenmesi çerçevesi sunmuştur. Model, farklı kirleticilerin aynı anda tahmin edilmesinde başarılı olmuş ve hava kirliliği tahmininde çok amaçlı regresyon tekniklerinin etkinliğini göstermiştir. Ancak, veri işleme süreçlerinin karmaşıklığı ve hesaplama maliyetleri sınırlamalar arasında yer almaktadır [29]. Kshirsagar ve Shah, hava kalitesi tahmininde sinir ağları, regresyon ve hibrit modellerin performansını karşılaştırmıştır. Çalışma, farklı model yapılarını detaylı bir şekilde analiz ederek hava kalitesi tahmini için kapsamlı bir inceleme sunmuştur. Ancak, farklı modellerin güçlü ve zayıf yönleri tartışılmasına rağmen, uygulama aşamalarındaki spesifik sınırlamalar üzerinde durulmamıştır [30]. Başka bir çalışma, hava kalitesi parametrelerini tahmin etmek ve mekansal dağılımlarını analiz etmek için çeşitli makine öğrenimi yöntemlerini kullanmıştır. Çalışma, hava kalitesi tahmininde mekansal analizin önemini vurgulamış ve farklı bölgelerdeki hava kirleticilerin mekansal davranışlarını anlamaya katkıda bulunmuştur. Ancak, mekansal veri eksikliği ve büyük veri setlerinde hesaplama maliyetlerinin yüksekliği gibi sınırlamalar gözlenmiştir [31]. Ünalı ve Yalçın, İstanbul Başakşehir'de hava kirliliği tahmini için makine öğrenimi yöntemlerini kullanan bir vaka çalışması gerçekleştirmiştir. Çalışma, yerel hava kalitesi tahmininde makine öğreniminin uygulanabilirliğini ortaya koymuş ve yüksek doğrulukla sonuçlar elde etmiştir. Ancak, çalışmanın sadece tek bir bölgeye odaklanması, modelin farklı bölgelerdeki uygulanabilirliği konusunda soru işaretleri yaratmaktadır [32].

Bu çalışmalar, hava kirliliği tahmininde yapay zekâ ve makine öğrenimi tekniklerinin yenilikçi ve etkili bir şekilde kullanılabileceğini göstermektedir. Ancak, genellenebilirlik, veri hazırlama

süreçlerinin karmaşıklığı, meteorolojik faktörlerin yetersiz entegrasyonu ve hesaplama maliyetleri gibi sınırlamalar, bu yaklaşımların daha geniş ölçekli ve farklı çevresel koşullarda uygulanabilirliğini sınırlandırmaktadır. Bu eksikliklerin giderilmesi, hava kalitesi tahmininde daha güçlü ve genellenebilir modellerin geliştirilmesine olanak tanıyacaktır.

2.4. İklim ve Çevresel Etkileri İnceleyen Çalışmalar

Çevresel ve iklimsel değişkenlerin hava kalitesi üzerindeki etkilerini incelemek, kirlilik tahmininde kritik bir bileşen olarak öne çıkmaktadır. Bu alt başlıkta, bu faktörlere odaklanan çalışmalar ele alınmıştır:

İstanbul'daki yaz aylarına özgü termal konfor endekslerini analiz eden bu çalışma, çevresel değişkenlerin yerel termal konfor üzerindeki etkilerini incelemiştir. Çalışma, iklim değişikliğinin termal konfor üzerindeki olası etkilerini anlamaya yönelik kapsamlı bir değerlendirme sunmaktadır. Ancak, modelin mevsimsel değişkenlikler dışında uzun vadeli iklim etkilerini tam olarak ele almaması sınırlayıcı bir faktördür [33]. Başka bir çalışmada, Yangtze Nehri Deltası'ndaki sıcak hava dalgaları sırasında hava kirleticileri tahmin etmek için dikkat mekanizması modeli kullanılmıştır. Çalışma, aşırı hava koşullarında tahmin doğruluğunu artırmış ve iklim açısından hassas bölgelerde çevresel izlemeye yönelik değerli bir yaklaşım sunmuştur. Ancak, dikkat mekanizmasının karmaşıklığı ve hesaplama maliyetleri uygulama süreçlerini sınırlayabilmektedir [34]. Yılmaz ve ekibi, İstanbul'daki sıcak hava dalgalarının özelliklerini analiz ederek bölgesel iklim değişikliği etkilerini değerlendirmiştir. Çalışma, şehirlerdeki iklim değişikliği dinamiklerini anlamak için kapsamlı bir analiz sunmaktadır. Ancak, modelin yalnızca belirli bölgelere odaklanması genellenebilirlik konusunda sınırlamalar yaratmaktadır [35]. Aksak ve ekibi, İstanbul'daki kentsel ısı adası etkilerini uzaktan algılama tekniklerini kullanarak incelemiştir. Çalışma, kentsel ısı adası olgusunu anlamada uzaktan algılama verilerinin potansiyelini göstermiştir. Ancak, uzaktan algılama verilerinin doğruluğu ve coğrafi kapsamı sınırlı kalmış, daha detaylı analizler için ek veri ihtiyaçları doğmuştur [36]. Shi ve ekibi, Şanghay'daki mobil kaynaklardan CO₂ ve diğer hava kirleticilerinin emisyonlarını azaltmanın çevresel ve ekonomik faydalarını değerlendirmiştir. Emisyon azaltma stratejilerinin etkinliğini gösteren bu çalışma, sürdürülebilir politika geliştirilmesi için önemli içgörüler sunmaktadır. Ancak, modelin farklı kaynak türlerini detaylı bir şekilde ele almaması sınırlayıcıdır [37]. Başka bir çalışma, hava kirleticilere maruz kalmanın bitki fenolojisi ve yaprak özellikleri üzerindeki etkilerini incelemiştir. Çalışma, hava kirliliğinin bitki biyolojisi üzerindeki kısa ve uzun vadeli ekolojik etkilerini kapsamlı bir şekilde değerlendirmiştir. Ancak, bitki türleri arasındaki farklılıklar ve hava kirliliği kaynaklarının çeşitliliği yeterince ele alınmamıştır [38].

Bu çalışmalar, çevresel değişkenlerin, hava kirliliğinin ve iklim değişikliğinin yerel ve ekolojik sistemler üzerindeki etkilerini analiz etmek için önemli yaklaşımlar sunmaktadır. Ancak,

genellenebilirlik, hesaplama maliyetleri ve veri kaynaklarının sınırlılıkları, modellerin daha geniş ölçekli ve farklı çevresel koşullarda uygulanabilirliğini sınırlandırmaktadır. İyileştirilmiş veri entegrasyonu ve genişletilmiş coğrafi kapsam ile bu eksikliklerin giderilmesi mümkün olmaktadır.

2.5. Trafik ve Hava Kirliliği İlişkisini İnceleyen Çalışmalar

Trafik kaynaklı emisyonlar, kentsel alanlardaki hava kirliliğinin ana sebeplerinden birisi olarak bilinmektedir. Bu bölümde, trafik hareketliliği ve hava kalitesi arasındaki ilişkiyi inceleyen çalışmalar derlenmiştir.

Ajayi ve diğerleri, Lagos şehrindeki heterojen trafik koşulları altında trafik hareketliliği önlemlerinin araç emisyonları üzerindeki etkilerini incelemiştir. Çalışma, trafik düzenlemelerinin çevresel etkilerini analiz etmiş ve emisyon azaltımına yönelik iyileştirme stratejileri sunmuştur. Kentsel trafik düzenlemelerinin emisyon azaltma potansiyelini değerlendirerek bu alandaki önemli katkılardan biri olmuştur. Ancak, çalışma, farklı trafik düzenlemelerinin uzun vadeli etkilerini ve diğer çevresel faktörlerle olan ilişkilerini tam olarak ele almamıştır [39]. Yang ve diğerleri, trafik ve hava koşullarına dayalı verileri birleştirerek havadaki partikül madde (PM) seviyelerinin mekansal ve zamansal tahminlerini yapmayı amaçlamıştır. Çalışma, farklı veri kaynaklarını entegre ederek hava kalitesi tahmininde hassasiyeti artırmıştır. Trafik ve hava durumu verilerinin birleştirilmesinin hava kirliliği tahmini üzerindeki olumlu etkilerini vurgulamaktadır. Ancak, modelin genellenebilirliği ve büyük veri setlerindeki hesaplama zorlukları gibi sınırlamalar bulunmaktadır [40].

Bu çalışmalar, kentsel trafik ve hava durumu verilerinin entegre edilerek hava kalitesi tahmininde kullanılmasının önemini göstermektedir. Ancak, genellenebilirlik, veri işleme süreçlerinin karmaşıklığı ve uzun vadeli etkilerin değerlendirilmesi gibi eksiklikler, daha gelişmiş modellerin ve daha kapsamlı analizlerin gerekliliğini ortaya koymaktadır.

2.6. Diğer Modelleri İnceleyen Çalışmalar

Hava kalitesi tahmini için kullanılan diğer model ve yaklaşımlar, farklı veri türleri ve uygulama alanlarını içermektedir. Bu alt başlıkta, literatürde yer alan farklı tekniklere odaklanan çalışmalar özetlenmiştir.

Bir çalışmada, hava kirletici madde konsantrasyonu tahmini için kullanılan derin öğrenme teknikleri hakkında mevcut literatür incelenmiş ve bu tekniklerin çeşitli uygulama alanları değerlendirilmiştir. Çalışma, hava kirliliği tahmininde derin öğrenme tekniklerinin ileri düzey uygulamalarını ve potansiyelini kapsayan kapsamlı bir inceleme sunmuştur. Ancak, literatürün uygulamaya yönelik sınırlı örneklerle ele alınması ve bazı yöntemlerin pratikteki performanslarının yeterince detaylandırılmaması gibi eksiklikler bulunmaktadır [41]. Arslankaya ve diğerleri, hisse

senedi fiyatlarını tahmin etmek için makine ve derin öğrenme tekniklerini uygulamış ve finansal veri analizi ile tahmin modellerinin etkinliğini değerlendirmiştir. Çalışma, finansal piyasalarda yapay zekâ tekniklerinin kullanımına dair önemli bir inceleme sunmaktadır. Ancak, finansal veri setlerindeki dinamik değişikliklerin ve ani piyasa dalgalanmalarının tahmin üzerindeki etkileri tam olarak ele alınmamıştır [42].

Bu çalışmalar, derin öğrenme tekniklerinin hem çevresel hem de finansal veri analizinde geniş uygulama alanlarına sahip olduğunu göstermektedir. Ancak, yöntemlerin uygulama zorlukları, genellenebilirlik ve farklı veri türlerine uyarlanabilirlik konularında eksiklikler, daha detaylı ve pratik odaklı çalışmalar yapılması gerektiğini vurgulamaktadır.

Literatürde, hava kalitesi tahmini ve izlenmesine yönelik yapılan çalışmalarda, modellerin genellenebilirliği, hesaplama maliyetleri, veri hazırlama süreçlerinin karmaşıklığı, meteorolojik faktörlerin yeterince entegre edilmemesi ve farklı çevresel koşullarda uygulanabilirliğin sınırlı kalması gibi önemli boşluklar tespit edilmiştir. Ayrıca, çevresel ve trafik kaynaklı etkilerin birlikte değerlendirilmesi, hibrit modellerin geniş ölçekli uygulanabilirliği ve uzun vadeli tahminlerin detaylı bir şekilde ele alınması gerektiği görülmüştür. Bu problemleri ele almak ve literatürdeki mevcut eksiklikleri gidermek amacıyla bu çalışmanın sunacağı katkılar şu şekilde sıralanmıştır:

- Çalışmada, eksik veri sorununu çözmek için K-En Yakın Komşu (K-NN) algoritması kullanılmıştır. Literatürde eksik veri genellikle basit doldurma yöntemleriyle (örneğin ortalama ile doldurma) ele alınırken, bu çalışmada K-NN gibi daha sofistike bir yöntem kullanılarak veri bütünlüğü ve doğruluğu artırılmıştır. Bu, tahminlerin güvenilirliğini ve modelin genel performansını önemli ölçüde iyileştirmiştir.
- Hava kalitesi tahmini için RNN, CNN ve LSTM modelleri aynı veri kümesi üzerinde karşılaştırmalı olarak analiz edilmiştir. Literatürde genellikle tek bir modelin performansı değerlendirilmektedir veya bir model başka bir modelle hibrit olarak kullanılmaktadır. Bu çalışmada, üç farklı modelin doğrudan karşılaştırılması yapılarak her bir modelin avantajları ve dezavantajları detaylı bir şekilde ortaya konmuştur.
- Tahmin sonuçlarının anlaşılabilirliğini artırmak amacıyla Kutu Grafiği (Box Plots), Keman Grafiği (Violin Plots) ve Dağılım Grafiği (Scatter Plots) gibi gelişmiş görselleştirme yöntemleri kullanılmıştır. Bu yöntemler, tahminlerin sadece uzmanlar tarafından değil, uzman olmayan kişiler tarafından da kolaylıkla yorumlanmasını sağlamaktadır. Literatürde görselleştirme genellikle çizgi grafikleriyle sınırlı olduğu için bu çalışma, daha kapsamlı ve kullanıcı dostu bir yaklaşım sunmaktadır.
- Çalışma, RNN modeli ile önümüzdeki üç yıl için hava kalitesi tahminleri yapmış ve tahminleri görselleştirmiştir. Literatürde genellikle kısa vadeli tahminler yapılırken, bu çalışma daha uzun vadeli tahminler sunmaktadır. Ayrıca, İstanbul'un Başakşehir ilçesine odaklanarak yerelleştirilmiş ve ayrıntılı bir model geliştirilmiştir. Bu, yerel yönetimlerin

stratejik kararlar almasına ve etkili müdahale stratejileri geliştirmesine olanak tanımaktadır.

- Model performansı MAE, RMSE ve R^2 gibi birden fazla metrikle değerlendirilmiş ve bu metriklerin sonuçları ayrıntılı olarak tartışılmıştır. Ayrıca, derin öğrenme modellerinin mimarisi, kullanılan parametreler ve uygulama adımları açık bir şekilde sunulmuştur. Bu yaklaşım, modelin işleyişini anlamayı kolaylaştırarak literatürde sıkça gözlenen şeffaflık eksikliğini gidermektedir.
- Yine literatürdeki birçok çalışma genellikle geniş coğrafi alanlarda veya genel düzeyde hava kalitesi tahminleri yapmaktadır. Ancak bu çalışma, İstanbul'un Başakşehir ilçesi gibi belirli bir yerel alana odaklanarak daha yerelleştirilmiş ve ayrıntılı bir tahmin modeli sunmaktadır. Bu, yerel yönetimlerin hava kalitesini izlemeleri ve etkili müdahale stratejileri geliştirmeleri için daha doğrudan uygulanabilir veriler sağlamaktadır.
- Son olarak, veri hazırlama için kullanılan her derin öğrenme modelinin uygulama adımları, model mimarisi ve kullanılan parametreler ayrıntılı olarak açıklanmıştır. Bu tür ayrıntılar genellikle literatürde kısaca belirtilmekte ve okuyucunun konuyu tam olarak anlaması zor olabilmektedir. Ancak, bu yaklaşım modelin nasıl çalıştığını ve neden bu şekilde seçildiğini ortaya koymaktadır.

3. HAVA VERİLERİNE DAYALI BÜYÜK VERİ GÖRSELLEŞTİRME VE DERİN ÖĞRENME YAKLAŞIMLARININ İNCELENMESİ

Çalışmanın bu bölümünde eksik veri tamamlama teknikleri, veri görselleştirme yöntemleri ve derin öğrenme yöntemleri hakkında genel bilgiler verilmektedir.

3.1. Veri Önışleme ve Hazırlık Süreçleri

Hava kalitesi analizlerinde kullanılan veri setleri, sensörlerde meydana gelen teknik aksaklıklar veya çevresel koşullardan kaynaklanan etkiler nedeniyle genellikle eksik veri noktaları içerebilmektedir. Bu eksiklikler, analiz sonuçlarının doğruluğunu ve güvenilirliğini olumsuz etkileyebilmektedir. Bu nedenle, veri önışleme sürecinde eksik verilerin uygun yöntemlerle tamamlanması büyük önem taşımaktadır [43].

Eksik veri doldurma sürecinde kullanılan teknikler, verilerin bütünlüğünü sağlamanın yanı sıra analizlerin tutarlılığını ve doğruluğunu artırmada kritik bir rol oynamaktadır. Doğru yöntemlerin tercih edilmesi, hava kalitesi tahminlerinde daha güvenilir ve anlamlı sonuçlar elde edilmesine katkı sağlamaktadır. Bu süreç, aynı zamanda veri setinin analiz ve modelleme için hazır hale getirilmesi açısından da temel bir adım olmaktadır.

3.1.1. Veri Toplama Süreci

Bu çalışmada kullanılan veriler, çevresel sensörlerden toplanan hava kalitesi parametrelerinden oluşmaktadır. Hava kalitesi, genellikle PM2.5, PM10, SO2, NO2, CO ve O3 gibi kirleticilerin atmosferdeki yoğunluğunu ifade eden ölçütlerle değerlendirilmektedir. Bu parametreler, hava kirliliğinin insan sağlığı, çevre ve ekosistem üzerindeki etkilerinin anlaşılması ve izlenmesi açısından kritik öneme sahiptir. Sensörler, genellikle belirli coğrafi bölgelerde stratejik olarak konumlandırılmış ve farklı kirlilik kaynaklarından gelen etkileri ölçmek için optimize edilmiştir. Toplanan veriler, zamana bağlı (time-series) bir yapıdadır ve genellikle saatlik, günlük veya aylık periyotlarla kaydedilmektedir. Her bir sensörden gelen veriler, coğrafi konuma göre etiketlenmiş olup hem lokal hem de bölgesel hava kalitesini anlamak için temel oluşturmaktadır. Sensörlerden alınan ham veriler, ilk aşamada genellikle eksik veri problemleri ve gürültü içerebilmektedir. Bu nedenle, veri toplama süreci yalnızca verilerin kaydedilmesini değil, aynı zamanda veri kalitesini artırmaya yönelik işlemleri de kapsamaktadır. Veri toplama sürecinin altyapısı, Türk Çevre, Şehircilik ve İklim Değişikliği Bakanlığı ile Meteoroloji Genel Müdürlüğü'nün ortaklaşa kurduğu izleme ağlarına dayanmaktadır [44]. Bu ağlar, farklı hava kalitesi izleme istasyonlarından gelen verilerin merkezi bir sistemde toplanmasını sağlamaktadır. Toplanan veriler, analiz ve modelleme

çalışmalarında temel girdi olarak kullanılmakta ve bu süreç, hava kalitesini iyileştirmeye yönelik çalışmaların geliştirilmesi için sağlam bir zemin hazırlamaktadır.

3.1.2. Eksik Veri Tamamlama Teknikleri

Günümüzde, veri analizi, makine öğrenmesi ve yapay zekâ uygulamaları birçok alanda kritik bir rol oynamaktadır. Ancak, gerçek dünyadan elde edilen veri kümeleri genellikle eksik, hatalı veya tutarsız bilgiler içermekte, bu da analiz ve modelleme süreçlerinde önemli sorunlara yol açmaktadır. Eksik veriler, tahminlerin doğruluğunu ve model performansını olumsuz etkileyebildiği için tamamlanması büyük önem taşımaktadır [45].

Eksik veri sorununu çözmek için klasik istatistiksel yöntemlerden makine öğrenmesi tabanlı yaklaşımlara kadar birçok yöntem geliştirilmiştir. Bu teknikler, eksik veri noktalarını tahmin ederek veri setlerini daha tutarlı hale getirmekte ve modelleme süreçlerinin başarısını artırmaktadır. Eksik verilerin etkili bir şekilde tamamlanması, analizlerin güvenilirliğini artırmanın yanı sıra gelecekteki karar alma süreçlerine de önemli katkılar sağlamaktadır. Literatürde yaygın olarak kullanılan teknikler şu şekildedir:

- **Kaldırma (Deletion) Yöntemleri:** Eksik verilere sahip gözlemlerin veya değişkenlerin veri setinden tamamen çıkarılması, eksik veri problemini çözmenin en basit yollarından birisi olmaktadır. Bu yöntem iki şekilde uygulanabilmektedir:
- **Liste Kaldırma (Listwise Deletion):** Tüm değişkenlerde eksik veri içeren gözlemler veri setinden çıkarılmaktadır. Örneğin, bir veri setinde bir bireyin gelir ve yaş verisi eksikse, o bireye ait tüm bilgiler veri setinden silinmektedir.
- **Çift Kaldırma (Pairwise Deletion):** Sadece analiz yapılan değişkenlerde eksik veri içeren gözlemler çıkarılmaktadır. Bu yöntem, eksik veri oranını azaltabilir ve daha fazla verinin kullanılmasına olanak tanımaktadır. Ancak kaldırma yöntemleri, veri setinde ciddi kayıplara neden olabilmekte ve bu durum, özellikle veri seti küçükse veya eksik veriler rastgele değilse analizin sonuçlarını yanıltıcı hale getirebilmektedir. Bu nedenle, eksik veri oranı düşük olduğunda tercih edilmektedir [46].
- **Ortalama Değer Atama:** Eksik veri, ilgili değişkenin ortalama değeriyle tamamlanabilmektedir [47]. Bu yöntemde, bir veri setindeki eksik değerler, o değişkenin eksik olmayan gözlem değerlerinin aritmetik ortalaması ile doldurulmaktadır. Örneğin, bir veri setinde gelir bilgisi eksik olan bireylerin gelir değerleri, veri setindeki tüm bireylerin ortalama gelir değeri ile tamamlanabilmektedir. Ancak, bu yöntem değişkenin dağılımını bozabileceği için yalnızca eksik veri oranı düşük olduğu durumlarda tercih edilmelidir. Ortalama değer atama işlemi, Denklem 3.1 ve 3.2 formülleri ile ifade edilmektedir:

$$\sum x_i = x^1 + x^2 + x^3 + \dots + x_n \quad (3.1)$$

$$x_{\text{eksik}} = \frac{\sum x_i}{n} \quad (3.2)$$

Burada x_{eksik} , eksik değeri temsil eder; $\sum x_i$, eksik olmayan gözlem değerlerinin toplamını; n ise bu gözlemlerin sayısını ifade etmektedir. Bu yöntem, eksik verilerin hızlı ve basit bir şekilde tamamlanmasını sağlayarak veri setinin analiz edilebilirliğini ve bütünlüğünü korumaktadır. Ayrıca, kolay uygulanabilir olması nedeniyle sıkça tercih edilmektedir. Ancak, eksik veri oranının yüksek olduğu durumlarda veya verilerin simetrik bir dağılım göstermediği senaryolarda, bu yöntemin yanlılık oluşturabileceği ve veri setinin varyansını azaltabileceği unutulmamalıdır. Bu nedenle, yöntemin uygulanacağı durumlarda veri setinin yapısının ve eksik veri oranının dikkatlice değerlendirilmesi gerekmektedir.

- **Medyan Değer Atama:** Bu yöntem, eksik değerin bulunduğu değişkenin medyanı ile doldurulmasını sağlamaktadır. Medyan, verilerin büyüklük sırasına göre dizilmesiyle ortada kalan değeri ifade etmektedir. Eğer veri setindeki gözlem sayısı tek ise ortadaki değer medyan olarak alınmaktadır. Ancak, gözlem sayısı çift ise, ortadaki iki değer aritmetik ortalaması medyan olarak hesaplanmaktadır. Medyan, Denklem 3.3 ve 3.4 formüllerinde belirtildiği şekilde hesaplanmaktadır:

$$\text{Tek Gözlem Sayısı Durumu: } \text{Medyan} = x^{\frac{n+1}{2}} \quad (3.3)$$

$$\text{Çift Gözlem Sayısı Durumu: } \text{Medyan} = (x^{\frac{n}{2}} + x^{\frac{n}{2}+1})/2 \quad (3.4)$$

Örneğin, bir veri setinde [5, 8, 3, Eksik, 10, 12] değerleri yer alırsa: eksik olmayan veriler [3, 5, 8, 10, 12] olarak sıralandığında, medyan (8) olarak belirlenmekte ve eksik değer bu şekilde tamamlanmaktadır.

Bu yöntem, uç değerlerden etkilenmediği için, ortalama değer atama yöntemine kıyasla daha sağlam ve güvenilir sonuçlar sunmaktadır. Özellikle veri setinde uç değerlerin fazla olduğu durumlarda, medyan değer kullanılması daha doğru tahminler elde edilmesine katkı sağlamaktadır. Ancak, medyan yöntemi de eksik veri oranının yüksek olduğu veri setlerinde dikkatle uygulanmalıdır.

- En Yakın Komşu (Nearest Neighbor): Bu yöntemde, eksik veriler, veri setindeki diğer gözlemlerle olan benzerliklere dayanarak tamamlanmaktadır. Özellikle eksik değere sahip gözlemin, diğer gözlemlerle olan mesafesi (genellikle Öklid mesafesi) hesaplanarak, en yakın komşu veya komşular belirlenmektedir. Bu komşuların değerleri kullanılarak eksik veri tahmin edilmektedir. Örneğin, bir öğrencinin sınav puanı eksikse, yaş, cinsiyet, ders çalışma süresi gibi özelliklere göre en yakın olan diğer öğrencilerin sınav puanları dikkate alınmaktadır. Denklem 3.5'te en yakın komşu hesaplama şu şekilde yapılmaktadır:

$$d(i, j) = \sqrt{k = 1 \sum n(x_{ik} - x_{jk})^2} \quad (3.5)$$

Burada:

$d(i, j)$: i ve j gözlemleri arasındaki Öklid mesafesini göstermektedir.

x_{ik} : i-inci gözlemdeki k-inci özelliğin değerini göstermektedir.

x_{jk} : j-inci gözlemdeki k-inci özelliğin değerini göstermektedir.

n : Özelliklerin toplam sayısını ifade etmektedir.

Bu yöntem, eksik verilerin gerçek değerlere yakın bir şekilde tamamlanmasını sağlamakta ve özellikle sürekli veriler üzerinde etkili bir şekilde kullanılabilir. Ancak, komşu sayısı (k) dikkatle seçilmelidir ve veri setindeki özellikler uygun bir şekilde normalize edilmelidir. Aksi takdirde, tahminlerde yanlışlık veya hata oranı artabilmektedir. Bu yöntem, eksik veri oranının düşük olduğu veri setlerinde güçlü bir tamamlama aracı olarak öne çıkmaktadır.

- En Sık Kullanılan (Most Frequent) Değer Atama: Kategorik değişkenlerde eksik veri, o değişkenin en sık görülen değeriyle tamamlanmaktadır. Örneğin, bir anketin cinsiyet değişkeninde eksik veriler varsa, en sık karşılaşılan cinsiyet değeri ("Kadın" veya "Erkek") ile doldurulmaktadır.

Ortalama veya medyan değer atama gibi yöntemler, kolay uygulanabilirlik avantajına sahip olmaktadır. Ancak bu yöntemler, değişkenin varyansını azaltarak veri setinde yanlışlığa yol açabilmektedir. Örneğin, ortalama değer atama, eksik verilerin normal dağıldığını varsaymakta ve değişkenin doğal dağılımını bozabilmektedir.

- Modelleme Yöntemleri: Modelleme yöntemleri, eksik verileri tahmin etmek için veri setinin diğer değişkenlerini kullanmaktadır [48]. Bu yöntemler daha karmaşık olmakla birlikte, eksik veri tamamlama sürecinde daha doğru sonuçlar sunabilmektedir. Bu yöntemlerden bazıları regresyon modelleri, karar ağaçları, makine öğrenmesi yöntemleri olarak bilinmektedir. Regresyon analizinde, eksik veriler, diğer bağımsız değişkenler kullanılarak regresyon analizi ile tahmin edilmektedir. Örneğin, bir bireyin yaşının eksik

olduğu durumda, bireyin gelir, eğitim düzeyi ve meslek bilgileri kullanılarak yaş tahmin edilebilmektedir. Karar ağaçlarında verileri tamamlamak için kullanılan karar ağacı algoritmaları, verilerin alt gruplarını analiz ederek tahminler yapmaktadır. Karar ağaçları, özellikle kategorik değişkenlerde etkili bir yöntem olmaktadır.

- Destek Vektör Makineleri (SVM), Random Forest ve Gradient Boosting gibi makine öğrenmesi algoritmaları, eksik veri noktalarını tahmin etmek için yaygın olarak kullanılmaktadır. Bu yöntemler, veri setindeki diğer değişkenlerden öğrenerek eksik değerleri tahmin etmekte ve yüksek doğruluk oranları sunmaktadır. Örneğin, bir hava kalitesi veri setinde eksik olan PM10 değerleri, sıcaklık, nem ve rüzgâr hızı gibi özelliklere dayalı olarak tahmin edilebilmektedir.
- EM Algoritması: Expectation-Maximization (EM) algoritması, eksik veri noktalarını olasılık dağılımlarına dayalı olarak tahmin etmektedir. İlk aşamada (Expectation), eksik veriler tahmin edilmekte ve ardından bu değerler kullanılarak maksimum olabilirlik tahminlemesi yapılmaktadır. Bu yöntem, büyük veri setlerinde daha etkili sonuçlar sunabilmektedir.
- Çoklu Değer Atama (Multiple Imputation) Yöntemleri: Bu yöntemde, eksik veriler birden fazla kez tahmin edilerek farklı veri setleri oluşturulmaktadır. Bu veri setlerinden elde edilen tahminler birleştirilmekte ve eksik veriler için daha güvenilir bir sonuç üretilmektedir. Çoklu değer atama yöntemi, eksik veri tahmininde belirsizlikleri azaltarak daha doğru sonuçlar sunmaktadır. Örneğin, bir hava kirliliği veri setinde eksik PM2.5 değerleri, farklı modeller kullanılarak tamamlanmakta ve bu tahminler birleştirilerek en iyi sonuç elde edilmektedir [49].

Eksik veri tamamlama yöntemleri, veri setinin özelliklerine ve eksikliğin yapısına bağlı olarak seçilmektedir. Her bir yöntemin avantajları, dezavantajları ve varsayımları bulunmaktadır. Bu nedenle, eksik verilerle çalışırken, veri setinin doğasını anlamak ve uygun bir eksik veri tamamlama yöntemi seçmek önemlidir. Eksik veri problemleri veri topladıktan sonra sıkça karşılaşılan bir durumdur ve bu nedenle doğru ve etkili eksik veri tamamlama yöntemlerinin kullanılması önem arz etmektedir. Bu yöntemler, veri analizi, makine öğrenmesi ve karar verme süreçlerinde güvenilir sonuçlar elde etmek için kullanılmaktadır.

3.1.3. Veri Füzyonu ve Yöntemleri

Veri füzyonu, birden fazla veri kaynağından gelen bilgileri bir araya getirerek daha kapsamlı, doğru ve anlamlı bilgi elde etme sürecidir. Bu yöntem, veri analizlerinde bilgi eksikliğini gidermek, farklı kaynaklardan gelen verilerin bütünlüğünü sağlamak ve tahmin veya modelleme süreçlerinde daha yüksek doğruluk elde etmek için kullanılmaktadır. Özellikle hava kalitesi, çevre yönetimi ve yapay zeka gibi alanlarda geniş bir uygulama alanı bulunmaktadır [50]. Veri füzyonunun temel amacı,

birleştirilen verilerden bağımsız kaynaklarla elde edilemeyecek ölçüde daha iyi bilgi elde etmektir. Bu süreç, eksik verilerin doldurulmasından, gürültünün azaltılmasına ve bilgi doğruluğunun artırılmasına kadar bir dizi önemli görevi içermektedir.

Veri füzyonunda kullanılan yöntemler, verilerin türüne, kaynağına ve hedeflenen çıktıya bağlı olarak farklılık gösterebilmektedir:

- **Düşük Seviyeli Füzyon:** Ham verilerin doğrudan birleştirilmesi sürecini göstermektedir. Sensörlerden gelen veriler herhangi bir işleme tabi tutulmadan birleştirilmektedir. Bu yöntem, veri kaybını önlemekte ve tüm bilgilerin korunmasını sağlamaktadır. Ancak, gürültü ve hatalı veriler birleştirilmiş veri setine dahil olabilmekte ve bu durum analizin doğruluğunu etkileyebilmektedir. Örneğin, meteorolojik sensörlerden alınan sıcaklık, nem ve rüzgâr hızının birleştirilmesi düşük seviyeli füzyona örnek olarak verilebilmektedir [51].
- **Yüksek Seviyeli Füzyon:** Karar düzeyinde gerçekleştirilen bir füzyon yöntemi olmaktadır. Farklı kaynaklardan gelen analiz sonuçları birleştirilerek nihai bir karar alınmaktadır. Bu yöntem, daha düşük boyutlu ve anlamlı verilerle çalıştığı için hesaplama maliyetini düşürmektedir. Ancak, karar düzeyine kadar yapılan analizlerdeki hataların düzeltilmemesi ve bilgi kaybı yaşanması önemli sınırlamalardır. Örneğin, birden fazla makine öğrenmesi modelinin çıktılarının birleştirilerek nihai tahminin yapılması yüksek seviyeli füzyona örnek olarak verilebilmektedir [51].
- **Bayesci Füzyon:** Verilerin olasılık modellerine dayalı olarak birleştirilmesini sağlamaktadır. Bu yöntem, özellikle belirsizlik durumlarında etkili bir şekilde çalışmaktadır. Belirsiz veya eksik verilerle çalışabilme yeteneği, Bayesci füzyonu hava tahminleri gibi uygulamalarda öne çıkarmaktadır. Ancak, karmaşık olasılık hesaplamaları yüksek hesaplama maliyetine yol açabilmektedir ve güvenilir sonuçlar elde etmek için iyi tanımlanmış öncül olasılıklara ihtiyaç duyulmaktadır [51].
- **Makine Öğrenmesi Tabanlı Füzyon:** Yapay zeka ve derin öğrenme yöntemlerini kullanarak verilerin otomatik olarak birleştirilmesini sağlamaktadır. Özellikle büyük veri setlerinde ve karmaşık analizlerde tercih edilen bu yöntem, büyük veri analizi ve hava kalitesi tahmini gibi uygulamalarda güçlü sonuçlar sunmaktadır. Ancak, yüksek hesaplama gücü gereksinimi ve modellerin eğitimi sırasında doğru hiperparametre seçimi gibi zorluklar, makine öğrenmesi tabanlı füzyonun dikkatle uygulanmasını gerektirmektedir [51].
- **Kurallar Tabanlı Füzyon:** Belirli kurallar çerçevesinde veri birleştirme yapılmasını sağlamaktadır. Bu yöntem genellikle uzman bilgisine dayanmakta ve belirli problemler için hızlı sonuçlar üretebilmektedir. Ancak, genel problemlere uygulanabilirliği sınırlı ve esnekliği düşüktür. Örneğin, hava kirliliği ölçümlerinde belirli bir kirlenme seviyesi aşıldığında alarm durumuna geçilmesi kurallar tabanlı füzyonun bir örneği olarak gösterilebilmektedir [51].

- Orta Seviyeli Füzyon: Ham verilerin işlenerek anlamlı özelliklerin çıkarılması ve bu özelliklerin birleştirilmesi ile gerçekleştirilmektedir. Bu yaklaşım, gürültü ve hatalı verileri filtreleyerek daha anlamlı bir veri seti oluşturmaktadır. Örneğin, bir görüntüden renk, doku ve şekil bilgileri çıkarılarak bu özelliklerin birleştirilmesiyle nesne tanıma işlemi yapılabilmektedir. Ancak, bu yöntem özellik çıkarımı aşamasında uzmanlık gerektirmekte ve çıkarılan özelliklerin eksikliği veya yanlışlığı nihai sonucu etkileyebilmektedir [52].

3.1.5. Veri Filtreleme Yöntemleri

Hava kalitesi gibi büyük ve karmaşık veri setleri genellikle eksik, gürültülü veya hatalı bilgiler içerebilmektedir. Bu durum, doğru analiz ve tahmin yapılmasını zorlaştırmaktadır. Bu nedenle, veri işleme sürecinde filtreleme yöntemlerinin uygulanması kritik öneme sahiptir [53]. Hava kalitesi verilerinde kullanılan veri filtreleme yöntemleri, verilerin gürültüden arındırılmasını, hatalı veya anormal değerlerin düzeltilmesini ve genel doğruluğun artırılmasını hedeflemektedir.

- Ortalama Filtresi: Bir veri noktasının değerini, belirli bir pencere boyutundaki komşu veri noktalarının ortalaması ile değiştirerek ani değişimlerin ve rastgele gürültünün azaltılmasını sağlamaktadır. Basit bir yapıya sahip olması nedeniyle hızlı uygulanabilir olsa da ani değişimleri yumuşatırken verinin bazı detaylarını kaybetme riski bulunmaktadır. Medyan filtresi ise bir veri noktasını çevresindeki pencere içerisindeki veri değerlerinin medyanı ile değiştirerek özellikle uç değerlerin etkisini azaltmada etkili bir yöntem sunmaktadır. Gürültüyü azaltırken verinin yapısını koruması, medyan filtresinin önemli bir avantajıdır. Ancak, ortalama filtreye kıyasla hesaplama açısından daha maliyetli olmaktadır [54].
- Gauss Filtresi: Veri analizinde sıkça kullanılan bir başka yöntemdir. Gaussian dağılımına dayalı ağırlıklı ortalama kullanan bu filtre, çevresel verilerde sıkça rastlanan rastgele gürültüyü azaltmakta ve veriyi pürüzsüz hale getirmektedir. Gürültüyü etkili bir şekilde azaltırken verinin genel yapısını koruyan bu yöntem, büyük veri setlerinde sıklıkla tercih edilmektedir. Ancak, büyük bir standart sapma seçilmesi durumunda aşırı düzeltirme yaparak detay kaybına neden olabilmektedir. Kalman filtresi ise zaman serisi verilerindeki gürültüyü azaltmak için geçmiş veriler ve tahmini modelden yararlanmaktadır. Özellikle zaman serisi analizlerinde yüksek doğruluk sağlamasıyla öne çıkan bu yöntem, karmaşık bir yapıya sahip olduğu için doğru parametre ayarları gerektirmektedir [55].
- Butterworth Filtresi: Yüksek frekanslı gürültüyü azaltarak düşük frekanslı sinyalleri koruyan bir yöntemdir. Sinyal işleme ve hava kalitesi analizlerinde kullanılan bu filtre, yüksek frekanslı gürültüyü etkili bir şekilde filtrelerken verinin bazı detaylarını kaybetme riski taşımaktadır. Hampel filtresi, uç değerlerin tespiti ve düzeltilmesinde etkili bir

yöntemdir. Belirli bir pencere içerisindeki verilerin medyan ve sapma değerlerine göre analizi yapılmakta ve anormal veriler temizlenmektedir. Ancak, pencere boyutunun yanlış seçilmesi performansı olumsuz etkileyebilmektedir [56].

- Savitzky-Golay Filtresi: Savitzky-Golay filtresi, polinomlar yardımıyla veriyi pürüzsüzleştiren bir yöntemdir. Zaman serisi verilerinde, gürültüyü azaltırken sinyalin genel şeklini korumaktadır. Bu yöntem, verinin temel özelliklerini koruma avantajı sunsa da yanlış parametre seçimi aşırı düzeltirmeye yol açabilmektedir [57].

Hava kalitesi tahminlerinde Gauss filtresi, özellikle rastgele gürültüyü azaltma konusundaki başarısıyla öne çıkmaktadır. Gaussian filtresi, verinin genel trendini koruyarak zaman serisi analizleri ve tahmin modelleri için daha düzenli bir veri seti oluşturulmasına yardımcı olmaktadır. Örneğin, PM2.5 gibi hava kirliliği parametrelerinde gözlemlenen ani değişimler, Gaussian filtresi ile etkili bir şekilde düzeltilerek daha anlamlı tahminler yapılabilmektedir.

Sonuç olarak, hava kalitesi verilerinde kullanılan filtreleme yöntemleri, analizlerin doğruluğunu artırmak ve tahmin modellerinin başarısını yükseltmek açısından kritik bir role sahiptir. Gaussian filtresi gibi yöntemler, verideki rastgele gürültüyü azaltarak daha pürüzsüz ve anlamlı veri setleri oluşturulmasına olanak tanımaktadır. Çeşitli filtreleme tekniklerinin doğru şekilde seçilmesi ve uygulanması, çevresel analizlerde güvenilir sonuçlar elde edilmesi için önemlidir.

3.2. Veri Görselleştirme Yöntemleri

Büyük veri kümelerinin anlaşılmasını kolaylaştırmak ve veri içindeki önemli eğilimleri görselleştirmek için çeşitli görselleştirme yöntemleri kullanılmaktadır. Bu yöntemler, verilerin daha hızlı analiz edilmesine ve karar alma süreçlerinde kolaylık sağlamaktadır.

3.2.1. Hava Verilerinin Görselleştirilmesi

Büyük veri çağıyla birlikte, günümüzde işletmelerin ve kuruluşların karşılaştığı en büyük zorluklardan biri, bu devasa veri miktarını anlamlı ve değerli bilgilere dönüştürmektir. İnsanlar artık sadece veriye sahip olmakla yetinmeyip, bunun yerine bu verileri anlamlandırmak, anlamlı bağlantılar kurmak ve sonuçları anlaşılır bir şekilde sunmayı istemektedirler. İşte bu noktada, veri görselleştirmesi büyük bir önem kazanmaktadır. Literatürde, veri görselleştirmenin önemi, devasa veri yığınlarının anlamlı bilgilere dönüştürülmesinde oynadığı kritik rol ile vurgulanmaktadır. Büyük veri çağında, yalnızca veri toplamak yeterli olmamakta, bu verileri görselleştirerek analiz etmek ve kullanıcı dostu bir şekilde sunmak gerekmektedir. Veri görselleştirme, karmaşık ilişkileri, desenleri ve eğilimleri daha anlaşılır hale getirirken, özellikle karar verme süreçlerinde ve stratejik planlamada kolaylık sağlamaktadır. Ayrıca, görselleştirme araçlarının kullanımının, veri analizine

katılımı artırdığı ve sonuçların daha etkili bir şekilde iletişimini sağladığına dikkat çekilmektedir [58].

Veri görselleştirme, büyük veri kümelerini grafik, tablo, harita veya diğer görsel formatlar kullanarak temsil etme sürecidir. Bu süreç, verilerin daha kolay anlaşılmasını sağlarken, veriler arasındaki ilişkileri ve kalıpları görsel olarak ortaya koymaktadır. Görselleştirme yöntemleri, bir yandan karmaşık veri setlerini sadeleştirirken diğer yandan derinlemesine analiz imkanı sunmaktadır. Büyük veri genellikle karmaşık ve hacimli olduğundan, sadece sayılar ve metinler aracılığıyla bu verileri analiz etmek çoğu zaman yetersiz kalmaktadır. Ancak, veri görselleştirme sayesinde bu veriler daha erişilebilir ve anlaşılır hale getirilebilmektedir. Grafikler, renkler, şekiller ve diğer görsel unsurlar, verilerin hızlı ve etkili bir şekilde analiz edilmesine ve yorumlanmasına olanak tanımaktadır.

Literatürde, veri görselleştirmesi için web tabanlı araçlar, çeşitli yazılımlar ve Python kütüphaneleri yaygın bir şekilde kullanılmaktadır [59], [60], [61], [62], [63], [64]. Veri görselleştirme aynı zamanda veri analitiğinin paylaşılabilişliğini ve etkileşimini artırmaktadır. Görsel formatlar, verileri daha anlaşılabilir hale getirerek farklı paydaşların ve karar vericilerin verileri daha etkili bir şekilde yorumlamalarına ve bilinçli kararlar almalarına yardımcı olmaktadır.

Sonuç olarak, büyük veri çağında veri görselleştirme, verilerin etkin bir şekilde anlamlandırılması ve doğru bir şekilde sunulması açısından kritik bir rol oynamaktadır. Verilerin doğru şekilde anlaşılıp yorumlanması, etkili karar alma süreçlerinin temelini oluşturmaktadır.

Etkili bir veri görselleştirme için bazı temel prensiplerin göz önünde bulundurulması gerekmektedir. Öncelikle, kullanılan görsel öğelerin uygun bir şekilde seçilmesi önemlidir. Grafik türü, renk paleti, ölçeklendirme ve düzenleme gibi faktörler, verilerin doğru bir şekilde iletilmesini sağlamak için dikkatlice planlanmalıdır. Ayrıca, verilerin yanıltıcı bir şekilde sunulmasını önlemek adına doğruluk ve bütünlük ilkelerine bağlı kalınması gereklidir. Bu prensipler, veri görselleştirme sürecini hem etkili hem de güvenilir kılmaktadır.

3.2.2. Görselleştirme İçin Kullanılan Araçlar

Python'da veri görselleştirme için kullanılan çeşitli kütüphaneler bulunmaktadır. Bu kütüphaneler, verileri grafikler, diyagramlar ve görsel öğeler kullanarak etkili bir şekilde sunmayı sağlamaktadır. Python'da popüler veri görselleştirme kütüphanelerinden bazıları şunlardır:

1. Matplotlib: Matplotlib, Python topluluğunda en çok kullanılan veri görselleştirme kütüphanesidir. Çeşitli grafik tiplerini desteklemekte ve basit çizimlerden ileri düzeyde görselleştirmelere kadar geniş bir yelpazede kullanılabilir. Matplotlib, hem 2D hem de 3D grafikler oluşturabilen ve geniş bir kullanıcı kitlesi tarafından tercih edilen güçlü bir kütüphane olarak bilinmektedir [65].

2. Seaborn: Seaborn, güçlü bir istatistiksel veri görselleştirme kütüphanesidir. Matplotlib üzerine inşa edilmiştir ve daha yüksek düzeyde bir arayüz sağlamaktadır. Seaborn, özellikle istatistiksel modellerin görselleştirilmesi için kullanışlıdır ve görsel öğelerin çekici bir şekilde sunulmasını sağlayan önceden tanımlanmış renk paletleri ve temalar sunmaktadır.
3. Plotly: Plotly, interaktif ve ölçeklenebilir görselleştirmeler oluşturmak için kullanılan bir kütüphanedir. Plotly, zengin grafik türleri sunar ve çeşitli dil ve platformlarla entegrasyon sağlamaktadır. Plotly grafikleri, web tabanlı uygulamalara ve raporlara kolayca gömülebilir ve etkileşimli özelliklerle zenginleştirilebilmektedir.
4. Bokeh: Bokeh, etkileşimli web tabanlı görselleştirmeler oluşturmak için kullanılan bir Python kütüphanesidir. Bokeh, özellikle büyük veri kümelerini hızlı bir şekilde görselleştirmek için optimize edilmiştir. Bu kütüphane, grafikleri interaktif hale getirmek için özellikle güçlü araçlar ve etkileşimli öğeler sağlamaktadır.
5. Altair: Altair, deklaratif bir yaklaşım kullanarak görselleştirme oluşturmayı kolaylaştıran bir Python kütüphanesidir. Verileri grafikler ve görsel öğelerle eşlemek için açık bir dil olan Vega-Lite'ı kullanmaktadır. Altair, etkileyici ve ilgi çekici görselleştirmeler oluşturmayı sağlayan önceden tanımlanmış şablonlar ve stil yönergeleri sunmaktadır.

Bu kütüphaneler Python'u veri görselleştirme konusunda son derece güçlü bir araç haline getirmektedir.

3.2.3. Seaborn ile Görselleştirme Teknikleri

Seaborn, Python'da kullanılan bir veri görselleştirme kütüphanesidir. İçerdiği çeşitli fonksiyonlar ve özellikler sayesinde veri analizi ve görselleştirme işlemlerini kolaylaştırmaktadır [65]. Seaborn ile gerçekleştirebilecek bazı temel veri görselleştirme türleri aşağıda verilmiştir.

- Barplot: Kategorik verilerin sayısal değerlerle ilişkisini göstermek için kullanılmaktadır. Her kategoriye ait sayısal değerlerin ortalama, toplam veya diğer istatistiksel ölçütlerle temsil edildiği dikdörtgen çubuklar kullanılmaktadır. Genellikle gelir düzeyi, ürün satışları, yaş grupları veya cinsiyet gibi kategorik verilerle ilişkili sayısal bilgilerin görselleştirilmesi için uygundur.
- Countplot: Bir kategorik değişkenin her bir kategorisinin frekansını görselleştirmek için kullanılan bir çubuk grafiğidir. Bu grafik, kategorilerin kaç kez tekrarlandığını göstererek verinin dağılımını hızlıca anlamaya yardımcı olmaktadır. Örneğin, bir anket sonucunda farklı cevapların (evet/hayır), öğrenci not aralıklarının, farklı ürün kategorilerinin satış miktarlarının veya müşteri memnuniyet düzeylerinin (düşük/orta/yüksek) frekanslarının

görselleştirilmesinde countplot kullanılabilir. Bu, kategorik verilerin analizinde en sık başvurulan görselleştirme araçlarından biridir.

- Scatterplot: İki sayısal değişken arasındaki ilişkiyi görselleştirmek için kullanılmaktadır. Bu grafikte, veri noktaları x ve y eksenlerinde konumlandırılarak değişkenler arasındaki olası bir ilişki veya korelasyon incelenmektedir. Örneğin, sıcaklık ve enerji tüketimi arasındaki ilişkiyi veya satış fiyatı ile ev büyüklüğü arasındaki bağlantıyı göstermek için scatterplot uygun bir seçim olmaktadır. Bu grafik, aynı zamanda noktaların dağılımı sayesinde aykırı değerleri tespit etmede de etkili bir araçtır.
- Lineplot: Zaman serileri veya sürekli değişkenlerin eğilimini incelemek için kullanılan bir grafik türüdür. Zaman genellikle x ekseninde, ölçülen değişken ise y ekseninde yer almaktadır. Örneğin, aylık satış rakamlarının zaman içindeki değişimini veya bir makinenin çalışma süresine göre sıcaklık değişimlerini göstermek için lineplot tercih edilmektedir. Özellikle trend analizi için kullanışlı olmaktadır.
- Boxplot: Bir sayısal değişkenin temel istatistiksel özetini görselleştiren bir grafik türüdür. Minimum, maksimum, medyan, birinci ve üçüncü çeyrekler ile aykırı değerler bir kutu şeklinde temsil edilmektedir. Örneğin, öğrencilerin sınav puanlarının dağılımını veya farklı ürün kategorilerindeki satış rakamlarının varyasyonunu anlamak için kullanılabilmektedir. Boxplot, dağılımdaki aykırı değerlerin belirlenmesinde oldukça etkili olmaktadır.
- Violinplot: Boxplot ile benzer şekilde bir sayısal değişkenin dağılımını göstermektedir; ancak bunun yanında verinin yoğunluk bilgisi de dahil edilmektedir. Violinplot, özellikle simetrik ve asimetrik dağılımları karşılaştırmada ve farklı kategoriler arasındaki yoğunluk farklarını görselleştirmede kullanılabilmektedir. Örneğin, farklı bölgelerdeki sıcaklık dağılımlarını incelemek için violinplot tercih edilebilmektedir.
- Heatmap: İki kategorik değişken arasındaki ilişkiyi veya bir değişkenin yoğunluk dağılımını göstermek için kullanılmaktadır. Renk tonları ile yoğunluğu temsil eden bu grafik, genellikle büyük veri setlerinde ilişki analizini kolaylaştırmaktadır. Örneğin, bir otel rezervasyon veri setinde müşteri türü ve rezervasyon durumu arasındaki ilişkiyi veya bir sınıftaki öğrencilerin farklı derslerdeki başarılarını göstermek için heatmap kullanılabilmektedir.
- Histogram: Sayısal bir değişkenin dağılımını göstermek için kullanılan temel bir grafik türüdür. Değerler belirli aralıklara bölünmekte ve bu aralıklardaki gözlem sayıları çubuklarla temsil edilmektedir. Örneğin, yaş dağılımını, maaş seviyelerini veya sıcaklık ölçümlerini analiz etmek için histogram uygun bir tercih olmaktadır. Dağılımın şekli (simetrik, çarpık vb.) ve yoğunluğu bu grafikte kolayca görülebilmektedir.

- Pairplot: Veri setindeki tüm sayısal değişkenler arasındaki ilişkileri görselleştirmek için kullanılmaktadır. Bu tür bir görselleştirme, her bir değişken çifti için scatterplot veya histogramlar gibi alt grafikleri içermektedir.
- Jointplot: İki sayısal değişken arasındaki ilişkiyi scatterplot ile gösterirken, aynı zamanda her bir değişkenin dağılımını histogramlarla sunmaktadır. Bu grafik, bir yandan iki değişkenin birlikte nasıl davrandığını anlamaya yardımcı olurken diğer yandan bu değişkenlerin bağımsız olarak nasıl dağıldığını görmeyi sağlamaktadır. Örneğin, gelir ve harcama arasındaki ilişkiyi incelemek için kullanılabilir.
- Distplot: Bir sayısal değişkenin dağılımını histogram ve yoğunluk eğrisi ile birlikte görselleştirmektedir. Veri setinin yoğunluğunu ve frekansını aynı grafikte sunup, dağılımın genel yapısını analiz etmek için kullanılmaktadır. Örneğin, bir mağazanın günlük satış miktarlarının dağılımını incelemek için distplot tercih edilebilir.
- Swarmplot: Kategorik bir değişkenle eşleştirilen bir sayısal değişkenin dağılımını görselleştirmektedir. Veri noktaları, kategorilerin etrafında yatay olarak hizalanmakta ve her bir nokta bireysel bir gözlemi temsil etmektedir. Örneğin, farklı şehirlerdeki ev fiyatlarının dağılımını incelemek için bu teknik kullanılabilir.
- Barplot (Categorical data): Kategorik verilerin frekansını veya bir sayısal değişkenin kategorik değişkenlere göre ortalama değerini görselleştirmek için kullanılmaktadır. Her kategoriye ait değerlerin çubuklarla temsil edildiği bir grafik sunmaktadır.
- Facetgrid: Veri setini birden fazla alt grafikte temsil etmek için kullanılan bir yöntemdir. Bir veya daha fazla kategorik değişkenin her bir değeri için ayrı bir alt grafik oluşturmaktadır. Örneğin, bir ürünün farklı mağazalardaki satışlarını veya farklı yaş gruplarının gelir seviyelerini incelemek için kullanılabilir.
- Lineplot (Categorical data): Kategorik bir değişkenin zaman veya bir başka sayısal değişkenle ilişkisini görselleştirmek için kullanılmaktadır. Genellikle zamanın x eksenine yerleştirildiği ve değişkenin y eksenine yerleştirildiği bir çizgi grafiği olarak temsil edilmektedir.

Bu bölümde bahsedilen görselleştirme türleri, Seaborn kütüphanesi ile gerçekleştirebilecek bazı temel şekil türleridir. Ancak, Seaborn oldukça kapsamlı bir kütüphanedir ve daha pek çok görselleştirme türü ve özelliği içermektedir.

3.2.4. Plotly ile Görselleştirme Teknikleri

Plotly, Python dilinde popüler bir veri görselleştirme kütüphanesidir ve çeşitli grafik türleri sunmaktadır [66]. Plotly ile yapılabilecek bazı temel veri görselleştirme türleri aşağıda verilmiştir.

- **Çizgi grafiği:** Bir sürekli değişkenin zaman, sıra veya başka bir sayısal değişkene göre nasıl değiştiğini göstermek için kullanılmaktadır. Zaman serileri veya eğilimleri analiz etmek için oldukça etkilidir. Örneğin, hisse senedi fiyatlarının zamanla nasıl değiştiğini veya hava durumu verilerindeki sıcaklık değişimlerini görselleştirmek için kullanılabilir. Bu grafik, veri noktalarını bir çizgi ile birleştirerek eğilimleri net bir şekilde ortaya koymaktadır.
- **Bar grafiği:** Kategoriler arasındaki karşılaştırmaları göstermek için kullanılmaktadır. Her bir kategoriye ait çubukların yüksekliği veya uzunluğu, verilerin miktarını temsil etmektedir. Örneğin, ülkelerin nüfusları, ürün satış rakamları veya kategorilere göre gelirleri karşılaştırmak için idealdir. Karşılaştırma yapmak istediğinizde bar grafikleri tercih edilebilmektedir.
- **Pasta grafiği:** Bir bütünün parçalarını göstermek için kullanılmaktadır. Dilimlerin büyüklüğü, toplamın yüzdesel oranlarını temsil etmektedir. Örneğin, bir şirketin pazar payını, farklı ürünlerin satış oranlarını veya bir anket sonuçlarını göstermek için uygundur. Ancak, birden fazla kategori arasında kıyaslama yapıldığında pasta grafiği yerine bar grafikleri tercih edilmelidir.
- **Scatterplot:** İki sayısal değişken arasındaki ilişkiyi incelemek için kullanılmaktadır. Veri noktalarının dağılımı, bu değişkenler arasındaki ilişkiyi göstermektedir. Örneğin, gelir ve harcama arasındaki ilişkiyi veya sıcaklık ve nem arasındaki bağlantıyı analiz etmek için idealdir. Aynı zamanda aykırı değerleri tespit etmek için de kullanılabilir.
- **Kutu grafiği:** Bir sayısal değişkenin dağılımını ve merkezi eğilimini göstermek için kullanılmaktadır. Medyan, alt ve üst çeyrekler, minimum ve maksimum değerler ile aykırı değerler kutu içinde görselleştirilmektedir. Örneğin, öğrencilerin sınav sonuçlarındaki dağılımı veya farklı ürün kategorilerindeki satış değişkenliklerini analiz etmek için tercih edilebilmektedir.
- **Yüzey grafiği:** Üç boyutlu bir alandaki veri değerlerini göstermek için kullanılmaktadır. Veriler x ve y koordinatlarında düzenlenmekte ve z eksenindeki bir değişkenle temsil edilmektedir. Örneğin, arazi yükseklikleri, iklim modellerindeki sıcaklık değişimleri veya finansal verilerin fiyat-zaman boyutlarını görselleştirmek için uygundur. Bu grafik, karmaşık üç boyutlu verilerin analizinde etkilidir.
- **Isı haritası:** İki kategorik değişken arasındaki ilişkiyi renk skalasıyla görselleştiren bir matristir. Örneğin, sıcaklık verilerinin saatlere göre dağılımını, bölgelerdeki satış performansını veya hedef kitlelerin demografik verilerini analiz etmek için kullanılabilir. Renk yoğunluğu, veri değerlerinin büyüklüğünü veya frekansını temsil etmektedir.

- Gösterim haritası: Coğrafi bölgelerin veri değerlerine göre renklerle temsil edildiği bir harita türüdür. Örneğin, ülkelerin nüfus yoğunluklarını, ekonomik göstergelerini veya salgın hastalık yayılımını görselleştirmek için kullanılabilir. Coğrafi veri analizi için ideal bir seçenek olmaktadır.
- Kabarcık grafiği: Üç değişken arasındaki ilişkiyi görselleştirmek için kullanılmaktadır. X ve Y eksenleri iki sayısal değişkeni temsil ederken, kabarcıkların boyutları üçüncü bir değişkenin büyüklüğünü ifade etmektedir. Örneğin, ülkelerin karbon emisyonlarını, nüfuslarını ve ekonomik büyüklüklerini aynı anda göstermek için kullanılabilir.

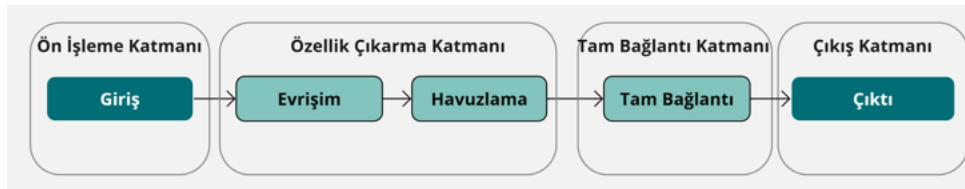
Bu teknikler sadece Plotly ile yapılabilecek bazı temel veri görselleştirme türleridir. Plotly'nin geniş bir grafik galerisi bulunmaktadır ve çeşitli özelleştirme seçenekleri sunmaktadır. Bu şekilde, farklı veriler farklı şekillerde görselleştirip ve sunulabilmektedir.

3.3. Derin Öğrenme Mimarileri

Hava kalitesi gibi büyük ve karmaşık veri setlerinin analiz edilmesi ve gelecekte hava kirleticileri değerlerinin tahmin edilmesi için derin öğrenme modelleri oldukça etkili araçlardır. Bu bölümde, hava kalitesi tahminlerine yönelik, çalışmalarda en yaygın olarak kullanılan üç temel derin öğrenme mimarisi olan Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) ve Long Short-Term Memory (LSTM) modelleri incelenmektedir.

3.3.1. Evrimsel Sinir Ağı (Convolutional Neural Networks-CNN) ile Görüntü İşleme

Convolutional Neural Networks (CNN), başlangıçta görüntü verilerini analiz etmek için geliştirilmiş ve yapay sinir ağları arasında önemli bir yere sahip bir mimaridir. CNN'ler, mekansal ve yerel özellikleri öğrenme kabiliyeti sayesinde, görüntü işleme ve bilgisayarla görme gibi alanlarda çığır açan başarılarla imza atmaktadır. Bu modellerin temel amacı, verilerdeki desenleri otomatik olarak çıkarmak ve analiz etmektir [67]. CNN mimarisi Şekil 3.1'de gösterilmiştir.



Şekil 3.1. Genel CNN Model Mimarisi

CNN'lerin en büyük avantajı, evrişim katmanları aracılığıyla yerel özellikleri öğrenebilmesi ve bu özellikleri genellenebilir bir biçimde kullanabilmesidir. Evrişim katmanları, veriler üzerinde

filtreler (kernel) uygulayarak yerel desenleri çıkarmaktadır. Örneğin, bir görüntüde kenar veya köşe gibi belirgin özellikler bu katmanlar sayesinde algılanabilmektedir. Bir evrişim işlemi Denklem 3.6'da bulunan formül ile ifade edilmektedir:

$$y(i, j) = \sum_m^{(1)M} \sum_n^{(1)N} x(i + m, j + n) \cdot w(m, n) + b \quad (3.6)$$

Burada:

- $y(i, j)$: Çıkış özelliğinin (i, j) koordinatındaki değeridir,
- $x(i + m, j + n)$: Giriş verisinin $(i + m, j + n)$ koordinatındaki değeridir,
- $w(m, n)$: Filtrenin (kernel) (m, n) boyutundaki ağırlık değeridir,
- b : Bias terimi,
- M ve N : Filtrenin boyutlarıdır.

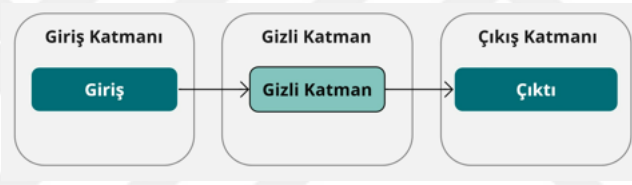
Bu işlem, verilerdeki belirgin desenlerin (örneğin kenarlar) çıkarılmasını sağlamaktadır. Benzer şekilde, zaman serisi verilerinde de tekrar eden desenler bu yöntemle yakalanabilmektedir. Evrişim işlemi sonrası, verinin doğrusal olmayan ilişkilerini öğrenmek için genellikle ReLU (Rectified Linear Unit) gibi aktivasyon fonksiyonları kullanılmaktadır. ReLU, modele doğrusal olmayan ilişkileri öğrenme kapasitesi kazandırarak karmaşık veri yapılarının analizine olanak tanımaktadır. Havuzlama katmanları ise çıkarılan özelliklerin boyutunu küçültmek ve modelin genellenabilirliğini artırmak için kullanılmaktadır. Özellikle maksimum havuzlama (max pooling) yöntemi, bir özellik haritasındaki en yüksek değeri seçerek boyutu küçültmekte ve böylece gürültüyü azaltmaktadır. Bu işlem hem hesaplama yükünü azaltır hem de modelin aşırı öğrenmesini önlemektedir. Havuzlama katmanlarının ardından, tam bağlantılı katmanlar (fully connected layers) devreye girmektedir. Bu katmanlarda çıkarılan tüm özellikler birleştirilir ve sınıflandırma ya da tahmin işlemi gerçekleştirilmektedir.

CNN modelleri yalnızca görüntü verilerinde değil, zaman serisi gibi farklı veri türlerinde de etkili bir şekilde kullanılabilmektedir [68]. Belirli modifikasyonlarla adapte edilen CNN mimarisi, özellikle hava kalitesi gibi zaman serisi verilerinde kritik bir araç haline gelmektedir. CNN'lerin mekansal ilişkileri modelleme kabiliyeti, kirlilik kaynaklarının etkilerini anlamak ve çevresel faktörlerin hava kalitesine olan etkilerini analiz etmek için büyük avantaj sağlamaktadır. Bu yapılar, yerel desenleri başarılı bir şekilde öğrenerek büyük veri setlerinden önemli özellikleri otomatik olarak çıkarmakta ve analiz sürecinde doğruluğu artırmaktadır. Özellikle hava kalitesi tahmini gibi uygulamalarda, CNN'lerin sunduğu avantajlar sayesinde etkili ve güvenilir sonuçlar elde edilmektedir.

3.3.2. Tekrarlayan Sinir Ağı (Recurrent Neural Networks - RNN) ile Zaman Serisi Analizi

Recurrent Recurrent Neural Networks (RNN), özellikle sıralı ve zamana bağlı verilerin analizi için geliştirilmiş bir yapay sinir ağı mimarisidir. RNN'nin temel özelliği, geçmişteki veri noktalarını hatırlayarak gelecekteki değerleri tahmin edebilmesidir. Bu özellik, sıralı veri yapısında yer alan bağımlılıkları öğrenmek ve analiz etmek için büyük bir avantaj sağlamaktadır. RNN mimarisinde, her bir zaman adımında üretilen çıktı, bir sonraki adıma girdi olarak aktarılmaktadır. Bu döngüsel yapı, modelin önceki adımlardaki bilgiyi hafızasında tutarak gelecekteki tahminlere rehberlik etmesine olanak tanımaktadır [69].

RNN'lerin bu yeteneği, hava kalitesi tahmini gibi zaman serisi analizlerinde etkili bir şekilde kullanılmaktadır. Örneğin, model geçmişteki hava kirliliği verilerini kullanarak, gelecekteki kirlilik seviyelerini tahmin edebilmektedir. Bununla birlikte, RNN modelleri uzun vadeli bağımlılıkları öğrenirken bazı zorluklarla karşılaşabilmektedir. Özellikle "eğim sönmesi" (vanishing gradient) olarak bilinen sorun, modelin uzun zaman aralıkları içindeki ilişkileri yeterince öğrenememesine yol açmaktadır. Bu durum, modelin performansını sınırlandırabilir ve öğrenme sürecini zorlaştırabilmektedir. RNN mimarisi, üç ana bileşenden oluşmaktadır ve Şekil 3.2'de gösterilmiştir.



Şekil 3.2. Genel RNN Model Mimarisi

İlk olarak, Girdi Katmanı (Input Layer) bulunmaktadır. Bu katman, zaman serisi verilerini sıralı bir şekilde modele aktarmaktadır. Örneğin, hava kalitesi tahmininde sıcaklık, nem ve PM2.5 gibi değişkenler her bir zaman adımı için girdi olarak kullanılmaktadır. İkinci bileşen olan Tekrarlayan Gizli Katmanlar (Recurrent Hidden Layers), modelin döngüsel yapısını oluşturmaktadır. Bu katmanlar, her bir zaman adımındaki çıktıyı bir sonraki adıma girdi olarak aktararak geçmiş bilgilerin hafızada tutulmasını sağlamaktadır. RNN'nin bir zaman adımındaki gizli durumu Denklem 3.7 ile hesaplanmaktadır:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (3.7)$$

Burada, gizli durum (h_t), hem mevcut girdi (x_t) hem de bir önceki zaman adımındaki gizli durum (h_{t-1}) bilgilerini içermektedir. Aktivasyon fonksiyonu (f), genellikle tanh veya ReLU olarak seçilmektedir.

Son bileşen olan Çıktı Katmanı (Output Layer), belirli bir zaman adımı için tahmin edilen değeri vermektedir. Çıktı, Denklem 3.8'deki gibi hesaplanmaktadır:

$$y_t = g(W_{hy}h_t + b_y) \quad (3.8)$$

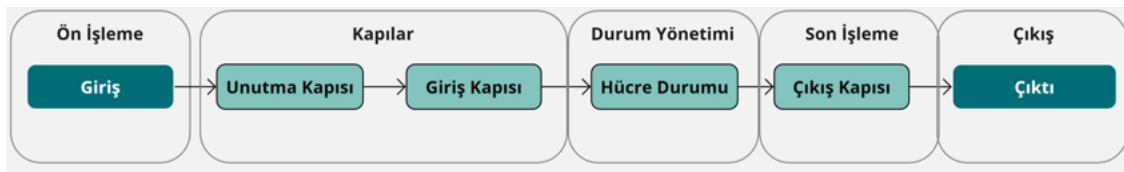
Bu formül, gizli durumdan (h_t) çıkan bilgilerin ağırlıklar (W_{hy}) ve bias (b_y) kullanılarak tahmin edilen çıktıya (y_t) dönüştürülmesini ifade etmektedir.

RNN modelleri, kısa vadeli bağımlılıkları öğrenmede oldukça başarılıdır. Örneğin, bir gün içerisindeki hava kirliliği değişimlerini öğrenmek gibi durumlarda etkili sonuçlar elde edebilmektedir. Ancak, uzun vadeli bağımlılıkların öğrenilmesi gerektiğinde, eğitim sönmesi problemi nedeniyle modelin bu bilgileri hatırlama kapasitesi sınırlı kalabilmektedir.

Sonuç olarak, RNN, sıralı ve zamana bağlı veriler üzerinde etkili analizler yapabilen güçlü bir mimaridir. Geçmiş bilgileri hatırlayarak gelecekteki tahminlere katkıda bulunması, özellikle hava kalitesi tahmini gibi sıralı veri analizlerinde RNN'leri değerli bir araç haline getirmektedir. Ancak, uzun vadeli bağımlılıkların öğrenilmesinde karşılaşılan zorluklar nedeniyle, bu tür uygulamalarda RNN'nin modifiye edilmiş veya geliştirilmiş versiyonlarının (örneğin LSTM) tercih edilmesi daha verimli sonuçlar sağlayabilmektedir.

3.3.3. Uzun Kısa Süreli Bellek (Long Short-Term Memory - LSTM) Modelleri ile Gelişmiş Zaman Serisi Analizi

Long Short-Term Memory (LSTM) modelleri, Recurrent Neural Networks'ün (RNN) geliştirilmiş bir versiyonu olarak tasarlanmış, özellikle uzun vadeli bağımlılıkların öğrenilmesinde üstün performans gösteren bir yapay sinir ağı türüdür. LSTM'nin temel amacı, RNN'lerde sıkça karşılaşılan "eğitim sönmesi" (vanishing gradient) sorununu ortadan kaldırarak, uzun süreli veri ilişkilerini etkili bir şekilde öğrenmektedir [70]. Bu başarı, LSTM'ye özgü "kapı mekanizmaları" sayesinde elde edilmektedir. Bu mekanizmalar, modelin hangi bilgileri saklayacağına, hangilerini unutacağına ve hangi bilgileri çıktı olarak sunacağına dinamik bir şekilde karar vermektedir. Böylece, LSTM modelleri uzun vadeli bağımlılıkları öğrenme ve karmaşık veri yapılarında etkili analiz yapma kapasitesine sahiptir. LSTM model mimarisi Şekil 3.3'te gösterilmektedir.



Şekil 3.3. Genel LSTM Model Mimarisi

LSTM'nin temel yapısı, hücre durumu ve kapı mekanizmaları gibi bileşenlere dayanmaktadır. Hücre durumu (cell state), modelin uzun vadeli belleğini temsil eder ve geçmişteki bilgilerin

taşınmasını sağlamaktadır. Bu yapı, modelin önceki adımlardaki önemli bilgileri hafızada tutmasına olanak tanımaktadır. Hücre durumunun güncellenmesi ve korunması, LSTM'nin kapı mekanizmaları aracılığıyla gerçekleştirilmektedir [71]. LSTM'de üç ana kapı mekanizması bulunmaktadır: unutma kapısı, giriş kapısı ve çıkış kapısı. Unutma kapısı (forget gate), hücre durumundaki hangi bilgilerin silineceğini belirlemektedir ve Denklem 3.9 ile hesaplanmaktadır:

$$f_t = \sigma(W_{fh} \cdot [h_{t-1}, x_t] + b_f) \quad (3.9)$$

Bu işlem, geçmişten gelen ve artık gerekli olmayan bilgilerin filtrelenmesini sağlamaktadır. Giriş kapısı (input gate), yeni gelen bilgilerin hücreye eklenip eklenmeyeceğine karar vermektedir ve Denklem 3.10'de gösterildiği gibi hesaplanmaktadır:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (3.10)$$

Bu kapı, mevcut veriyle hücre durumunun güncellenmesinde kritik bir rol oynamaktadır. Çıkış kapısı (output gate) ise hücrede depolanan bilginin hangi kısmının çıktıya aktarılacağını kontrol etmektedir ve Denklem 3.11'de gösterildiği şekilde hesaplanmaktadır:

$$h_t = o_t \odot \tanh(C_t) \quad (3.11)$$

Bu kapılar, modelin hem kısa hem de uzun vadeli bağımlılıkları öğrenmesini esnek ve etkili bir hale getirmektedir. Aktivasyon fonksiyonları, LSTM'nin çalışma mekanizmasının önemli bir parçası olarak bilinmektedir. Genellikle sigmoid ve tanh aktivasyon fonksiyonları, kapılar arasındaki veri akışını kontrol etmek için kullanılmaktadır. Sigmoid fonksiyonu, verilerin 0 ile 1 arasında bir değere dönüştürülmesini sağlayarak kapıların açılma veya kapanma derecesini belirlemektedir. Tanh fonksiyonu ise verilerin -1 ile 1 arasında normalize edilmesine olanak tanıyarak hücre durumunu güncellemektedir. LSTM modelleri, hava kalitesi tahmini gibi uzun zaman aralıklarına yayılan verilerin analizinde ideal bir çözüm sunmaktadır. Özellikle PM2.5 ve PM10 gibi parametrelerdeki uzun vadeli trendler ve ani değişimler, LSTM'nin uzun vadeli bellek yapısı sayesinde başarılı bir şekilde modellenabilmektedir. Çalışmalarda, LSTM modelleri hem geçmiş verilerdeki bağımlılıkları öğrenme hem de gelecekteki tahminleri yüksek doğrulukla gerçekleştirme yeteneğiyle öne çıkmaktadır [72]. Bu özellikleri sayesinde LSTM, karmaşık zaman serisi verilerinde güçlü bir analiz ve tahmin aracı olarak kullanılmaktadır.

4. HAVA VERİLERİ KULLANILARAK BÜYÜK VERİ GÖRSELLEŞTİRMESİ VE DERİN ÖĞRENME MODELLERİ İLE ÇIKARIM İÇİN YAPILAN UYGULAMALAR

Hava verileri, çevresel analizler ve yapay zeka uygulamaları için oldukça önemli bir kaynak teşkil etmektedir. Özellikle hava kalitesinin değerlendirilmesi ve geleceğe yönelik tahminlerin yapılması, hem bilimsel hem de pratik anlamda kritik bir öneme sahiptir. Çalışmamızın bu bölümünde, hava verilerinin toplanmasından görselleştirilmesine, eksik verilerin doldurulmasından derin öğrenme yöntemleriyle çıkarım yapılmasına kadar olan süreçler detaylı olarak ele alınmıştır.

4.1. Verilerin Toplanması

Hava kalitesi tahmini ve analizi için ilk adım, ilgili verilerin sistematik bir şekilde toplanmasıdır. Bu çalışmada, hava kalitesini etkileyen temel faktörler (PM2.5, PM10, SO2, NOx, CO vb.) üzerinde durulmuş ve veriler, Türkiye Çevre, Şehircilik ve İklim Değişikliği Bakanlığı tarafından sağlanan Sürekli İzleme Merkezi (SİM) sistemi üzerinden temin edilmiştir [73]. SİM platformu, farklı şehirlerde kurulu hava kalitesi izleme istasyonlarından alınan verileri günlük, saatlik ve anlık olarak raporlamaktadır. Bu sistem, çevresel parametrelerin takibini sağlayarak hava kalitesinin değerlendirilmesinde kritik bir kaynak sunmaktadır. Şekil 4.1’de sitenin bir görüntüsü gösterilmektedir.

Hava Kalitesi - İstasyon Veri İndirme
Anasayfa / Veri Bankası / Hava Kalitesi - İstasyon Veri İndirme

FİLTRE									
MARMARA THM			İSTANBUL			İstanbul - Başakşehir-MTHM			Sorgula
8 ADET SEÇİM YAPILDI			Saatlik			Günlük			30.09.2024
									23.11.2024
DETAYLI ÖLÇÜM DEĞERLERİ									
PDF'E AKTAR									
EXCEL'E AKTAR									
Tarih	PM10 (µg/m3)	SO2 (µg/m3)	CO (µg/m3)	NO2 (µg/m3)	NOX (µg/m3)	NO (µg/m3)	O3 (µg/l)		
30.09.2024	12,99	4,07	265,55	7,22	14,02	5,13	53,18		
01.10.2024	22,47	5,87	-	14,79	23,12	5,43	47,39		
02.10.2024	81,89	3,14	319,68	24,32	43,67	12,46	46,17		
03.10.2024	89,06	5,23	573,04	46,43	116,81	37,50	35,96		

Şekil 4.1. Sürekli İzleme Merkezi (SİM) Web Sitesine Ait Görüntü

SİM web sitesinde, hava kalitesini etkileyen parametrelere ait verilere kolaylıkla ulaşılabilmektedir. İlgili istasyonun seçilmesiyle, belirli bir tarih aralığında ölçülen veriler indirilebilmekte ve analiz için uygun bir formatta kullanılabilmektedir. Veriler genellikle .xls, .csv

veya .json formatında sağlanmakta olup, farklı zaman dilimlerini kapsayan verilerin birleştirilmesi ve analiz edilebilir hale getirilmesi gerekmektedir. Bu kapsamda, eksik veri doldurma, tarihsel eşleştirme ve format dönüştürme gibi işlemler uygulanmaktadır. İlk uygulamada, 2022-2023 yıllarını kapsayan 1 yıllık hava kalitesi verileri analiz edilmiştir. Başakşehir'e ait PM2.5, PM10, SO2 ve CO gibi temel hava kirleticilerinin günlük değişimlerini değerlendirmek amacıyla eksik veriler, ortalama değer atama yöntemiyle doldurulmuştur. Eksik verilerin tamamlanması sonrasında görselleştirme çalışmaları gerçekleştirilmiştir. İkinci uygulamada ise, 2014-2024 yıllarını kapsayan 10 yıllık veri seti kullanılmıştır. İstanbul Başakşehir bölgesine ait PM2.5, PM10, SO2 ve NOx gibi hava kirleticilerinin bulunduğu veriler çalışmaya dahil edilmiştir. Veri setindeki eksik değerler, daha gelişmiş bir yöntem olan K-En Yakın Komşu (KNN) algoritması ile doldurulmuştur. Bu sayede eksik veri problemine daha özel bir çözüm getirilmiştir.

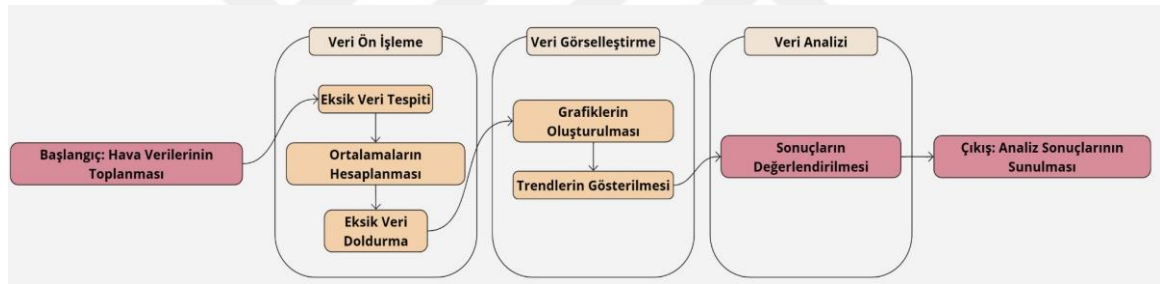
4.2. Performans Değerlendirme Ölçütleri

Hava kalitesini tahminine yönelik çalışmalarda, performans değerlendirme ölçütleri olarak Ortalama Mutlak Hata (Mean Square Error-MAE), Kök Ortalama Kare Hatası (Root Mean Squared Error-RMSE), ve R^2 puanı.

- Ortalama Mutlak Hata (MAE): MAE, tahmin edilen değerler ile gerçek değerler arasındaki farkların mutlak değerlerinin ortalamasını ifade etmektedir. Bu metrik, modelin tahminlerinde ne kadar sapma olduğunu göstermekte ve kolay anlaşılır bir ölçüm sunmaktadır. Örneğin, bir çalışmada MAE kullanılarak modelin performansı değerlendirilmiş ve düşük MAE değerlerinin modelin yüksek doğrulukta tahminler yaptığını gösterdiği belirtilmiştir [74].
- Kök Ortalama Kare Hatası (RMSE): RMSE, tahmin edilen ve gerçek değerler arasındaki farkların karelerinin ortalamasının karekökü alınarak hesaplanmaktadır. Bu metrik, büyük hatalara daha fazla ağırlık vererek modelin performansını değerlendirmektedir. Özellikle uç değerlerin etkisini vurgulamak için kullanılmaktadır. Bir çalışmada, RMSE değerlerinin modelin doğruluğunu etkileyen önemli bir faktör olduğu ve düşük RMSE değerlerinin modelin daha iyi performans gösterdiğini belirttiği görülmektedir [75].
- R^2 Puanı: R^2 , bağımlı değişkendeki toplam varyansın model tarafından ne kadarının açıklandığını gösteren bir ölçüttür. 1'e yakın R^2 değerleri, modelin yüksek açıklayıcılık gücüne sahip olduğunu, sıfıra yakın değerler ise modelin yetersiz olduğunu göstermektedir. Örneğin, bir çalışmada R^2 değerleri kullanılarak modelin performansı değerlendirilmekte ve yüksek R^2 değerlerinin modelin veri setinin değişkenliğini büyük ölçüde açıkladığını gösterdiği belirtilmektedir.

4.3. Uygulama 1: Hava Kalitesi Tahmininde Veri Görselleştirme ve Derin Öğrenme Modellerinin Karşılaştırmalı Analizi

Bu çalışma, Türkiye Çevre, Şehircilik ve İklim Değişikliği Bakanlığı'nın sürekli izleme merkezlerinden elde edilen yıllık hava kirliliği verilerini görselleştirerek büyük veri yığınlarını daha anlaşılır hale getirmeyi amaçlamaktadır. PM10, SO2, NO2, NOx ve O3 gibi kirlleticilere odaklanan çalışmada, eksik veriler ortalama yöntemiyle tamamlanmış ve veri görselleştirme yöntemleri olarak nokta dağılım grafikleri, çizgi dağılım grafikleri ve çubuk grafikler kullanılmaktadır. Bu yöntemlerle, hava kirliliği ölçümlerindeki eğilimler ve değerler net bir şekilde ortaya konmuş, büyük ve karmaşık veri kümeleri daha anlamlı hale getirilmiştir. Kullanılan görselleştirme araçları, yalnızca uzmanlar için değil, aynı zamanda uzman olmayan kullanıcılar için de hava kalitesi hakkında bilgi edinmeyi kolaylaştırmıştır. Çalışma ayrıca, farklı veri türlerine uygun görselleştirme yöntemlerini analiz ederek bu araçların kullanımını daha etkili hale getirmeyi sağlamıştır. Bu yönleriyle, hava kirliliği verilerinin çevresel analiz ve karar verme süreçlerinde daha etkili bir şekilde kullanılmasına katkıda bulunmuştur [76]. Aşağıdaki Şekil 4.2'de, çalışmanın özetini akış şeması şeklinde sunmak için tasarlanmıştır ve ana adımları açıkça göstermektedir.



Şekil 4.2. Uygulama 1 İçin Akış Diyagramı

4.3.1. Verilerin Toplanması ve Ön İşlenmesi

Bu çalışmada, İstanbul ilinin Başakşehir ilçesine ait hava kalitesi verileri T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı tarafından sağlanan Sürekli İzleme Merkezi (SİM) sisteminden açık erişim yoluyla elde edilmiştir. Çalışma kapsamında, 1 yıllık bir süreyi kapsayan ve toplamda 367 satırdan oluşan veri seti kullanılmıştır. Veri seti, PM10 ve SO2 hava kirlleticilerine ilişkin ölçüm değerlerini içermektedir. Ancak, veri setinde eksik değerler tespit edilmiştir. Özellikle PM10 için 36 satırda, SO2 için ise 24 satırda eksik ölçüm bulunmaktadır. Çalışma bölgesi olarak Başakşehir'in seçilmesinin nedeni, bölgenin kirliliğe sahip olması ve sanayi faaliyetlerinin yoğun olarak gerçekleştiği bir alan olmasıdır. Bu durum, hava kalitesinin ölçülmesi ve analiz edilmesi açısından önemli bir temel sağlamaktadır. Eksik veri problemi, veri analizinin doğruluğunu etkileyebileceği için, bu tür eksikliklerin uygun bir şekilde tamamlanması gerekmektedir. Bu çalışmada eksik veriler, ortalama değer atama yöntemi kullanılarak doldurulmuştur. Bu yöntemde, PM10 için ortalama değer 76,4, SO2 için ise 13,25 olarak hesaplanmış ve eksik değerlere bu

ortalamalar atanmıştır. Eksik verilerin tamamlanmasıyla veri seti analiz ve görselleştirme süreçleri için uygun hale getirilmiştir. Eksik ve tamamlanmış veriler Tablo 4.1 ve Tablo 4.2’de sunulmuştur. Toplamda 367 satırdan oluşan 1 yıllık veriler elde edilmiştir. Bu verilerde PM10 değeri için 36 satırda eksiklik, SO2 değeri için 24 satırda eksiklik bulunmaktadır.

Tablo 4.1. Başakşehir Eksik Hava Verileri (10 Satır)

Sıra	Tarih	PM10	SO2
1	25.02.2023 00:00	72,75	9,73
2	26.02.2023 00:00	98,86	13,30
3	27.02.2023 00:00	125,52	10,45
4	28.02.2023 00:00	-	-
5	01.03.2023 00:00	-	-
6	02.03.2023 00:00	142,56	15,34
7	03.03.2023 00:00	174,23	12,78
8	04.03.2023 00:00	164,99	12,77
9	05.03.2023 00:00	175,40	10,15
10	06.03.2023 00:00	143,76	9,13

Eksik olan verileri basit değer atama yönteminde olan alt başlıklardan ortalama değer atama ile tamamlanmıştır. Veri setinin tamamlandıktan sonraki hali Tablo 5.2’de gösterilmiştir. PM10 değeri için ortalama değer 76,4 hesaplanmıştır. SO2 değeri için ise ortalama değer 13,25 olarak hesaplanmıştır.

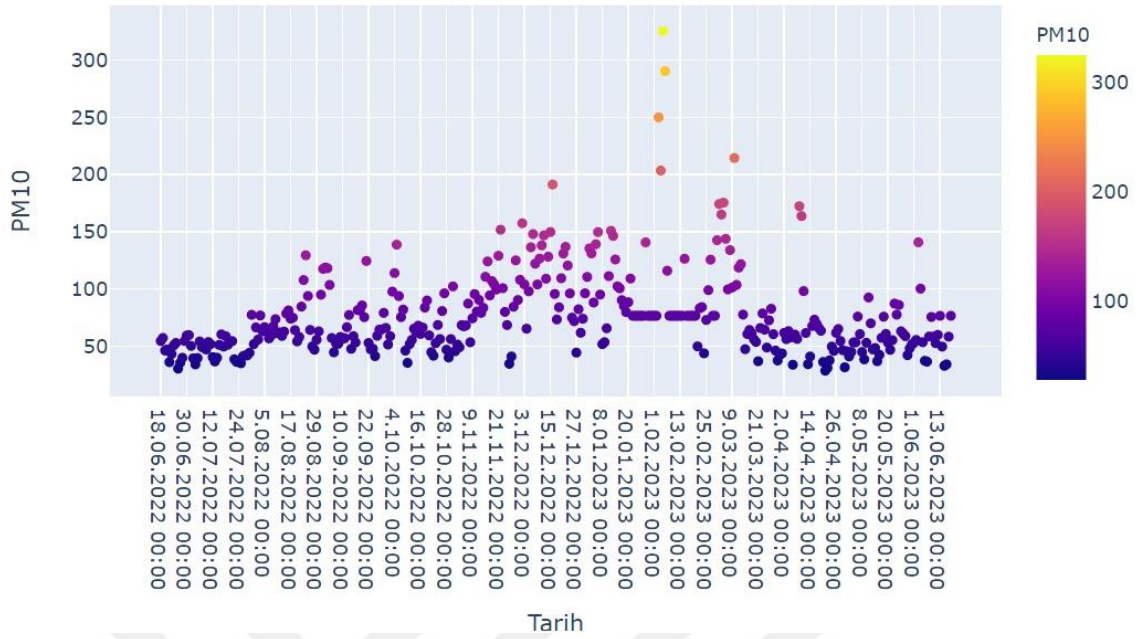
Tablo 4.2. Başakşehir Tamamlanmış Hava Verileri (10 Satır)

Sıra	Tarih	PM10	SO2
1	25.02.2023 00:00	72,75	9,73
2	26.02.2023 00:00	98,86	13,30
3	27.02.2023 00:00	125,52	10,45
4	28.02.2023 00:00	76,4	13,25
5	01.03.2023 00:00	76,4	13,23
6	02.03.2023 00:00	142,56	15,34
7	03.03.2023 00:00	174,23	12,78
8	04.03.2023 00:00	164,99	12,77
9	05.03.2023 00:00	175,40	10,15
10	06.03.2023 00:00	143,76	9,13

Sonuç olarak çalışmada, İstanbul Başakşehir ilçesine ait PM10 ve SO2 hava kirleticilerine ilişkin bir yıllık veri setindeki eksik değerler tespit edilmiş ve bu eksiklikler ortalama değer atama yöntemiyle tamamlanmaktadır. Veri setinde PM10 için 36, SO2 için ise 24 eksik değer bulunurken, tamamlanma işlemi sonucunda analiz ve görselleştirme süreçleri için uygun bir veri seti elde edilmektedir. PM10 için ortalama değer 76,4, SO2 için ise 13,25 olarak hesaplanmış ve eksik değerlerin yerine bu ortalamalar atanmaktadır. Bu yaklaşım, eksik veri probleminin giderilmesiyle birlikte hava kalitesi analizlerinin daha doğru bir şekilde gerçekleştirilmesine olanak sağlanmaktadır.

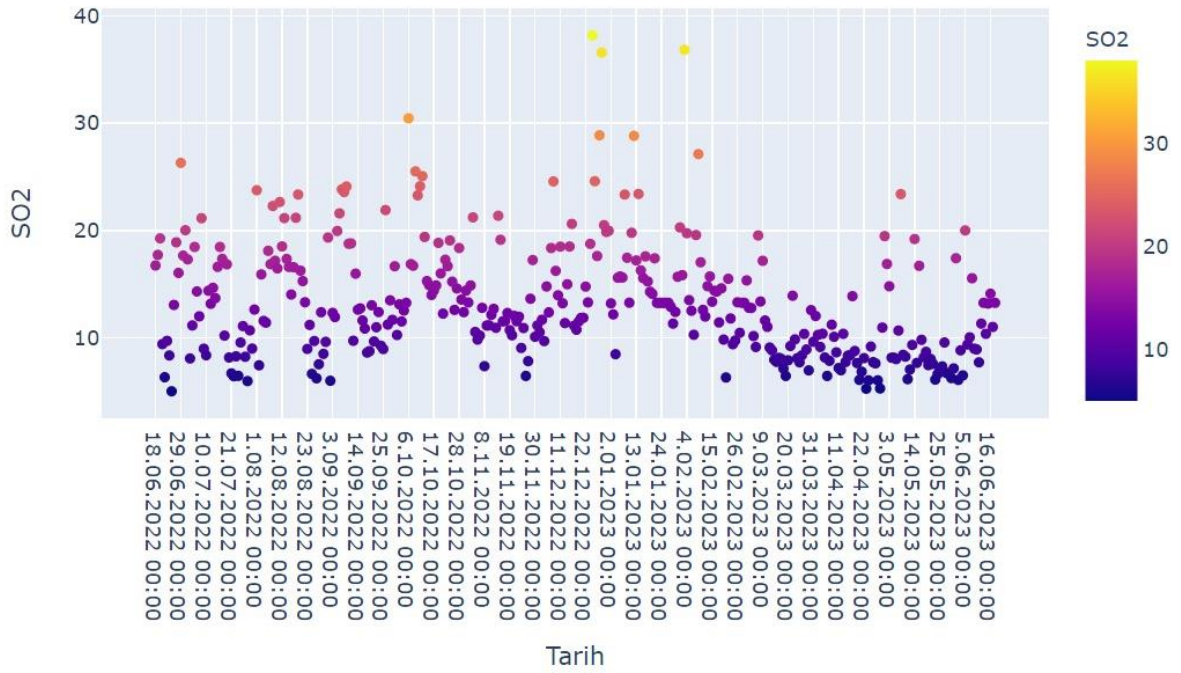
4.3.2. Modelin Oluşturulması: Veri Görselleştirme

Hava kirliliği verilerinin daha anlaşılır hale getirilmesi, verilerdeki ilişkilerin ve eğilimlerin daha etkili bir şekilde ortaya çıkarılması amacıyla çeşitli görselleştirme teknikleri kullanılmıştır. Bu çalışmada tercih edilen point scatter, line scatter ve bar plot grafik yöntemleri, hava kirleticilerinin dağılımını, zaman içerisindeki değişimlerini ve potansiyel desenlerini analiz etmek için etkili araçlar sunmaktadır. Görselleştirme, yalnızca verilerin görsel olarak sunumunu sağlamakla kalmayıp, aynı zamanda karmaşık veri setlerinin yorumlanmasını kolaylaştırarak daha doğru ve anlamlı çıkarımlar yapılmasına olanak tanımaktadır. Plotly kullanılarak oluşturulan point scatter grafikleri, bir eksen üzerinde bir değişkenin değerini, diğer eksen üzerinde ise ikinci bir değişkenin değerini temsil eden noktalarla veri setindeki gözlemleri görselleştirmektedir. Her bir noktanın konumu, ilgili değişkenlerin değerlerine dayanmaktadır. Bu çalışmada, x eksenini olarak tarih verileri, y ekseninde ise Şekil 4.3 için PM10 ve Şekil 4.4 için SO2 hava kirleticisi verileri kullanılarak point scatter grafik türü tercih edilmiştir. Her bir nokta, belirli bir tarihteki PM10 ve SO2 değerlerini temsil etmekte ve veri setindeki gözlemler bu şekilde görselleştirilmiştir. Ek olarak, grafik üzerindeki noktalar renklendirilmiş ve boyutlandırılmıştır; bu sayede hava kirleticilerinin dağılımı ve eğilimleri daha kolay bir şekilde analiz edilmiştir.



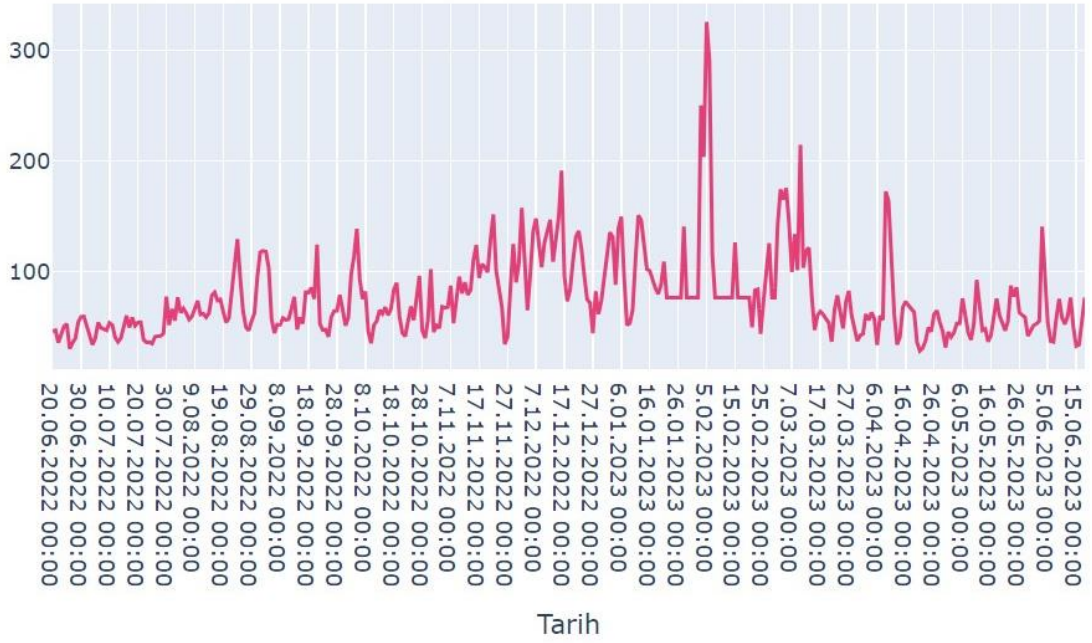
Şekil 4.3. PM10 Değeri İçin Point Scatter

Point scatter grafiklerinin kullanımı, hava kirliliği gibi çok boyutlu ve karmaşık veri setlerinin daha anlaşılır hale getirilmesini sağlamaktadır. Özellikle grafiklerde renk çubuğu gibi ek görselleştirme unsurları, farklı değer aralıklarını belirginleştirmekte ve yüksek değerlerin olduğu tarih aralıklarını hızla tespit etmeye olanak tanımaktadır. Bu çalışmada, PM10 kirleticisinin değerlerinde yüksek olanlar sarı, düşük olanlar ise mor renk ile gösterilmiş; bu sayede zaman içerisindeki değişimlerin daha iyi gözlemlenmesi sağlanmıştır. Point scatter grafiği, hem görselleştirme esnasında detaylı bir analiz yapmaya olanak tanıyan hem de karar verme süreçlerini kolaylaştıran etkili bir yöntemdir.



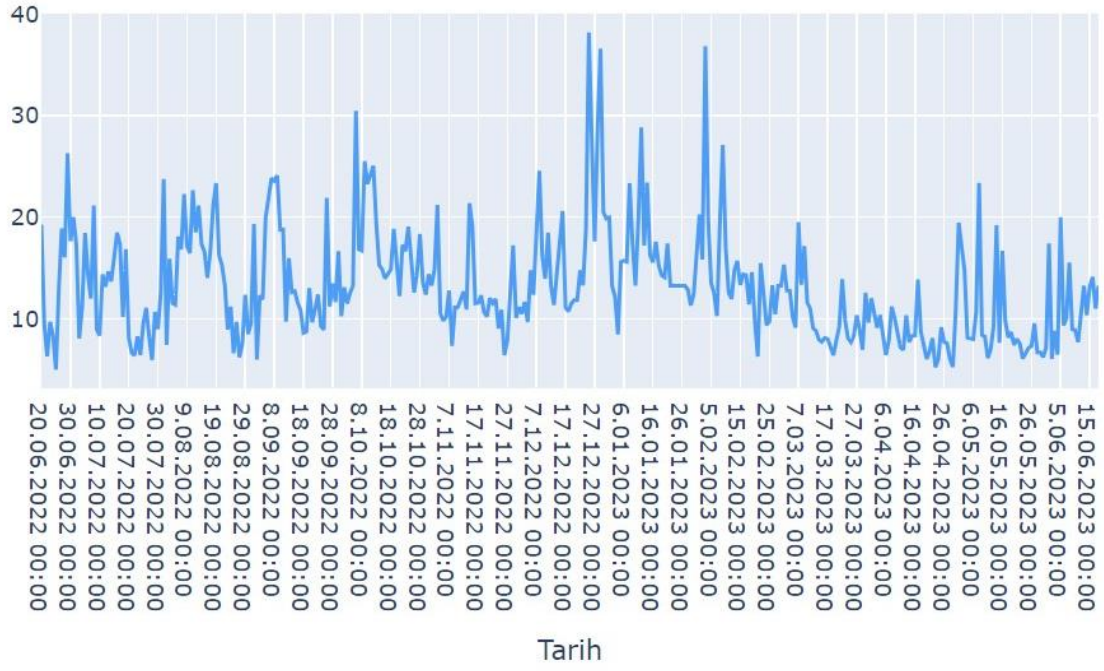
Şekil 4.4 SO2 Değeri İçin Point Scatter

Yukarıdaki şekilde, SO2 kirleticisi için oluşturulan point scatter grafiği, kirletici değerlerini görselleştirmek amacıyla kullanılmıştır. Grafik üzerindeki noktalar, belirli bir tarihteki SO2 değerlerini temsil etmekte ve yan tarafında bulunan renk çubuğu aracılığıyla yüksek değerler sarı, düşük değerler ise mor renk ile gösterilmiştir. Bu renk skalası, SO2 seviyelerindeki değişimlerin ve hangi tarihlerde yüksek değerlerin meydana geldiğinin daha kolay gözlemlenmesine olanak tanımıştır. Point scatter grafik, SO2 konsantrasyonunun zaman içerisindeki dağılımını detaylı bir şekilde analiz etmeyi sağlamıştır. Mevcut veri seti temel alınarak, bu kez line scatter grafiği kullanılarak analiz gerçekleştirilmiştir. Line scatter grafiği, veri noktalarını birleştiren bir çizgi aracılığıyla, değişkenler arasındaki ilişkinin doğasını ve zaman içerisindeki eğilimleri daha net bir şekilde ortaya koymaktadır. Bu grafik türü, noktaların yanı sıra çizgi kullanarak dağılımı daha belirgin hale getirmekte ve değişkenler arasındaki düzeyler, eğilimler veya korelasyonları anlamak için etkili bir araç sunmaktadır. Şekil 4.5 ve Şekil 4.6’da line scatter grafiği kullanılarak, PM10 ve SO2 kirleticilerinin zaman içerisindeki değişimleri görselleştirilmiştir. Bu yaklaşım, hava kirliliği verilerindeki eğilimleri daha net bir şekilde gözlemlemek ve yorumlamak için değerli bir yöntem sunmaktadır.



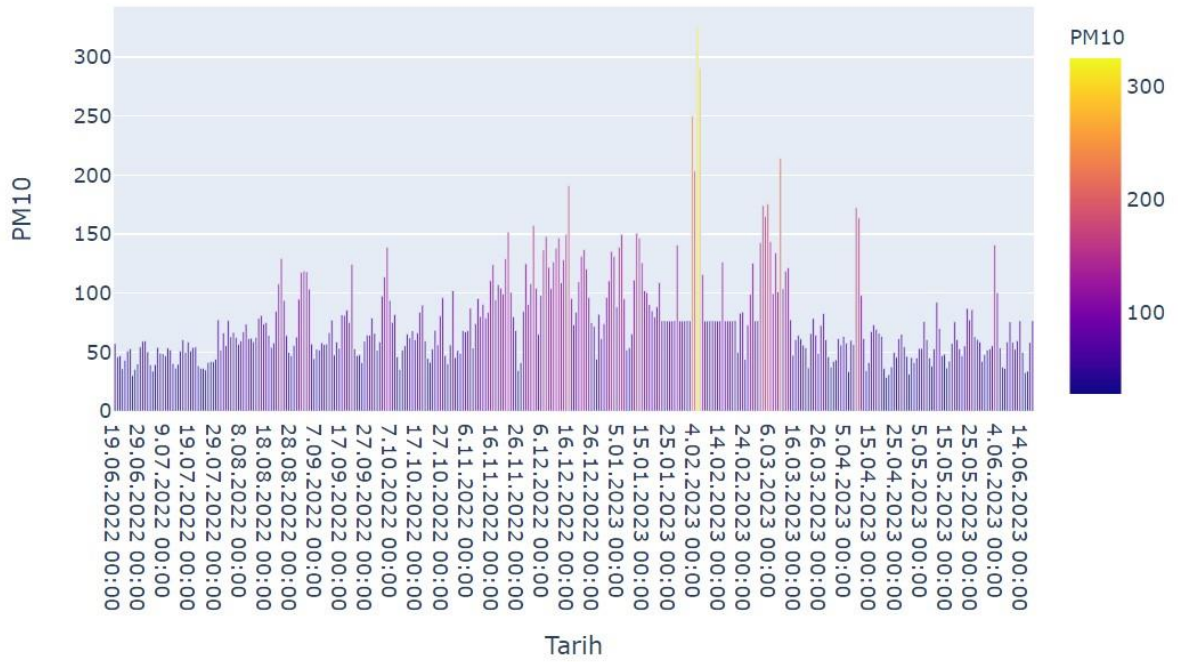
Şekil 4.5. PM10 Değeri İçin Line Scatter

Şekil 4.5’de, PM10 değeri için line scatter grafiği kullanılmıştır. Ancak bu grafik, point scatter grafiğine kıyasla bazı sınırlılıklara sahiptir. Bu sınırlılıklar arasında, grafikte renkli bir görselleştirmenin bulunmaması ve yalnızca çizgi üzerinden gözlem yapılmasının zorluğu bulunmaktadır. Renk çubuğunun olmaması, farklı değer aralıklarının hızlı bir şekilde ayırt edilmesini zorlaştırmaktadır. Renk çubuğunun eklenmesi, verilerin daha kolay anlaşılmasını sağlayabilmekte ve grafik üzerindeki yüksek veya düşük değerlerin tespit edilmesini kolaylaştırarak görselin yorumlanabilirliğini artırabilmektedir. Bu eksiklik, line scatter grafiğinin detaylı analizlerde kullanımını sınırlayan bir faktör olarak değerlendirilebilmektedir.



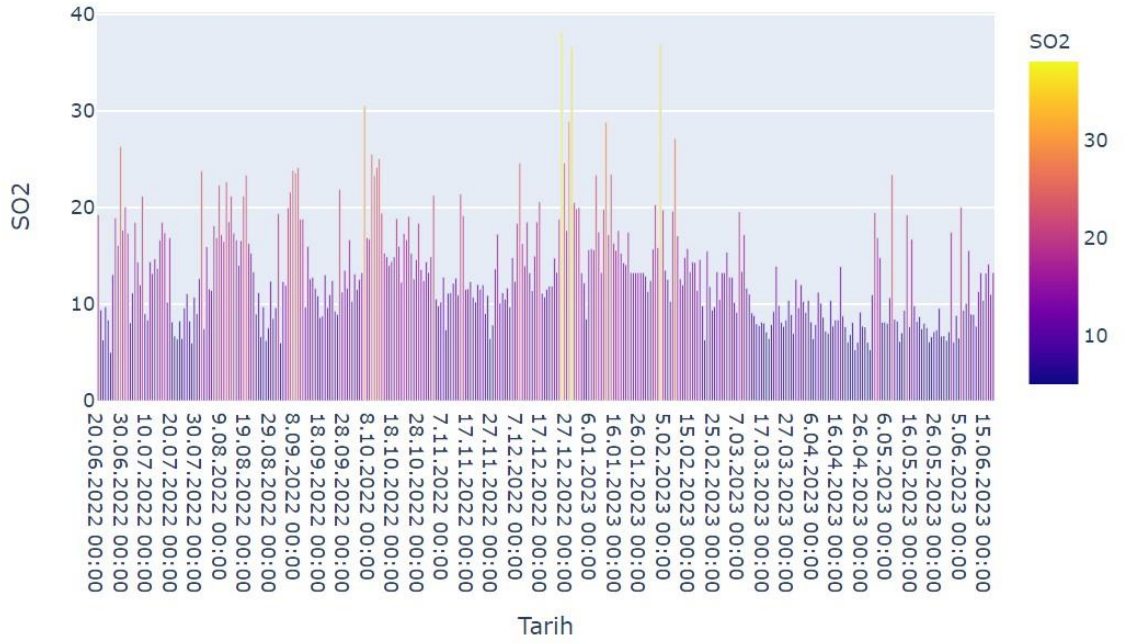
Şekil 4.6. SO2 Değeri İçin Line Scatter

Şekil 4.6’da, SO2 değeri için line plot kullanılmıştır. Ancak bu grafik, point scatter grafiğine kıyasla dezavantajlara sahiptir. Bunun temel nedeni, grafiğin tek renkli olmasından kaynaklanmaktadır. Farklı renklerin kullanımı, görselleştirmenin daha anlaşılır hale gelmesini ve farklı değer aralıklarının kolayca ayırt edilmesini sağlayabilmektedir. Bu eksiklik, özellikle karmaşık veri setlerinin analizi sırasında yorumlama sürecini zorlaştırabilmektedir. Renk kullanımı, veri görselleştirmesinde önemli bir faktör olarak değerlendirilmektedir. Mevcut veriler üzerinde yapılan bar plotly denemeleri ise Şekil 4.7 ve Şekil 4.8’de gösterilmiştir. Bar plotly, verileri sütun grafikleri aracılığıyla temsil etmektedir. Her bir sütun, belirli bir kategoriye ait veri miktarını veya değerini ifade etmektedir. Sütunların yükseklikleri, ilgili kategorideki verinin büyüklüğünü veya frekansını göstermektedir. Bu tür bir görselleştirme, özellikle kategorik verilerin karşılaştırılmasında ve belirli zaman aralıklarındaki değişimlerin analizinde etkili bir yöntem sunmaktadır.



Şekil 4.7. PM10 Değeri İçin Bar Plotly

Şekil 4.7’de, PM10 değeri için bar plotly kullanılmıştır. Bu grafik, line scatter grafiğine kıyasla daha anlaşılır bir görünüm sunmakla birlikte, point scatter grafiğine göre bazı sınırlılıklara sahiptir. Bar plotly, verileri sütunlar halinde görselleştirerek kategorik veya zaman serisi verilerinin dağılımını etkili bir şekilde sunmaktadır. Ancak sütunların yoğunluğu ve iç içe geçmişliği, bazı durumlarda veri detaylarının net bir şekilde gözlemlenmesini zorlaştırabilmektedir. Buna karşın, bar plotly’nin sunduğu avantajlar da bulunmaktadır. Sütunların yüksekliği sayesinde belirli zaman dilimlerindeki değerlerin büyüklüğü açıkça görülebilmekte ve kategorik karşılaştırmalar kolaylıkla yapılabilmektedir. Ancak veri detaylarının daha açık bir şekilde incelenmesi gereken durumlarda, point scatter grafiği daha uygun bir seçenek olabilmektedir. Bar plotly, geniş veri kümelerini genel bir bakış açısıyla analiz etmek ve eğilimleri belirgin bir şekilde görmek için faydalı bir araçtır.



Şekil 4.8. SO2 Değeri İçin Bar Plotly

Şekil 4.8’de, SO2 değeri için bar plotly kullanılmıştır. Grafik, renk çubuğu sayesinde belirli değer aralıklarının anlaşılmasını kolaylaştırır da, point scatter grafiğine kıyasla bazı yönlerden zayıf kalmaktadır. Bunun temel nedeni, bar plotly grafiklerinde veri dağılımlarının görsel olarak daha az belirgin olmasıdır. Point scatter grafikleri, veri noktalarını daha açık bir şekilde sunarak dağılımın ve eğilimlerin gözle daha rahat görünmesini sağlamaktadır. Bar plotly, belirli bir zaman dilimindeki veri değerlerini karşılaştırmak veya büyüklükleri analiz etmek için faydalı bir araç olsa da, detaylı dağılım analizlerinde point scatter grafiği daha etkili bir seçenektir. Bu nedenle, verilerin doğasına ve analiz amacına bağlı olarak grafik türünün dikkatlice seçilmesi gerekmektedir.

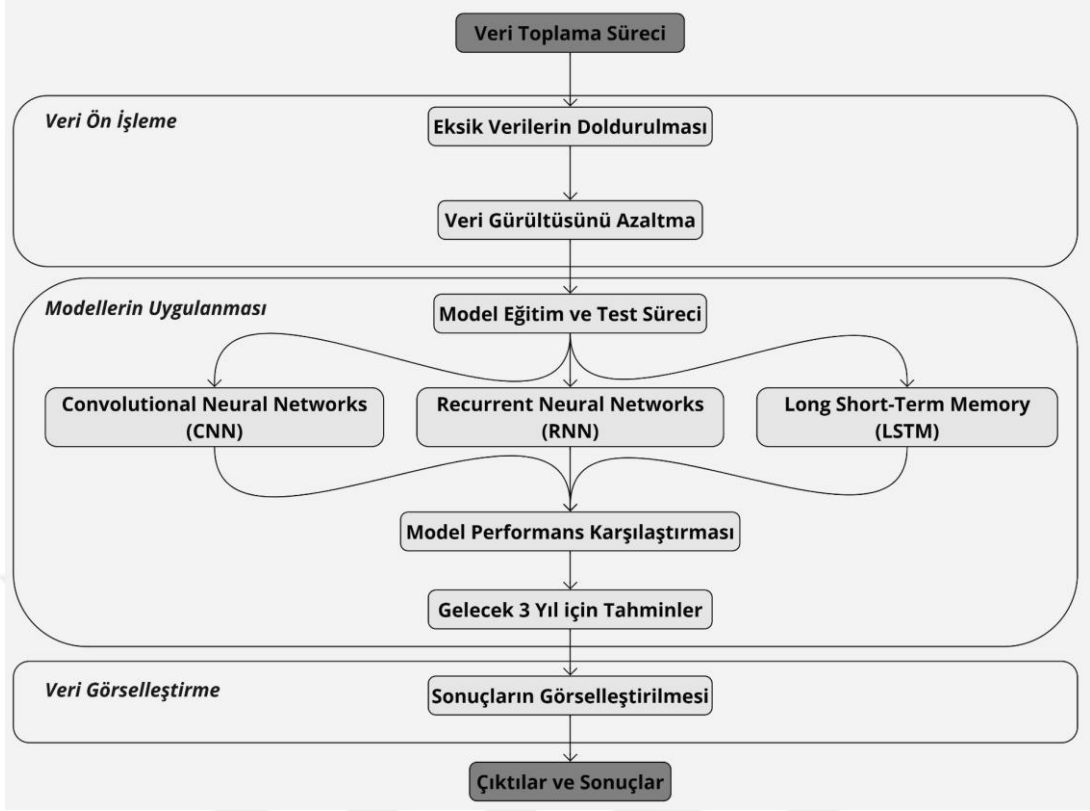
4.3.3. Deneysel Sonuçlar ve Tartışma

Bu çalışmada, büyük veri setlerinin yorumlanması ve analiz edilmesinde veri görselleştirme araçlarının sağladığı katkılar ele alınmıştır. İstanbul ilinin Başakşehir ilçesine ait hava kalitesi verileri toplanmış, eksik veri sorunları giderilmiş ve çeşitli görselleştirme yöntemleri kullanılarak analizler gerçekleştirilmiştir. Çalışmanın temel amacı, hava kirleticilerinin zaman içindeki değişimlerini ve dağılımlarını etkili bir şekilde görselleştirerek analiz süreçlerini daha anlaşılır hale getirmektir. Elde edilen bulgular, farklı grafik türlerinin veri analizi ve yorumlamadaki etkisini ortaya koymuştur. Point scatter grafikleri, hava kirleticilerinin tarihsel değişimlerini ve dağılımlarını en etkili şekilde görselleştirmiştir. Renk çubuklarının kullanımı, düşük ve yüksek kirlilik seviyelerinin açık bir şekilde ayırt edilmesini sağlamış ve kirleticilerin mekânsal ve zamansal dağılımlarını analiz etmek için en uygun yöntem olarak öne çıkmıştır. Line scatter

grafikleri, deęişim eğilimlerini incelemek için faydalı bir yöntem olmakla birlikte, detaylı görsel bilgi sağlama açısından sınırlı kalmıştır. Benzer şekilde, bar plotly grafikleri kirleticilerin zaman içindeki seviyelerini karşılaştırma açısından etkili bir yöntem sunmuş, ancak dağılımları detaylı olarak göstermek açısından yetersiz bulunmuştur. Sonuç olarak, point scatter grafikleri, hava kirleticilerinin deęişim ve dağılımlarını görselleştirme açısından en etkili yöntem olarak belirlenmiştir. Renk çubuklarının entegrasyonu, verilerin daha kolay yorumlanmasını sağlamış ve analiz süreçlerinin anlaşılabilirliğini artırmıştır. Ayrıca, eksik verilerin tamamlanması, analizlerin tutarlılığını ve doğruluğunu güçlendirmiş, güvenilir sonuçlar elde edilmesine önemli katkı sağlamıştır. Bu çalışma, büyük veri setlerinin analizinde veri görselleştirme araçlarının kritik bir öneme sahip olduğunu ve karar verme süreçlerine deęerli bir destek sağladığını açıkça göstermektedir.

4.4. Uygulama 2: Hava Kalitesi Tahmininde Veri Görselleştirme ve Derin Öğrenme Modellerinin Karşılaştırmalı Analizi

Hava kalitesi tahminine yönelik derin öğrenme modellerinin karşılaştırmalı analizi gerçekleştirilmiştir. Bu çalışma, farklı derin öğrenme modellerinin performanslarını değerlendirerek hava kirleticilerinin konsantrasyonlarının daha doğru bir şekilde tahmin edilmesini sağlamayı amaçlamaktadır. Bu bağlamda, İstanbul'un Başakşehir ilçesine ait 10 yıllık veri kullanılarak üç derin öğrenme modeli—Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) ve Long Short-Term Memory (LSTM)—hava kirletici konsantrasyonlarının tahmini için değerlendirilmiştir. Eksik veriler, K-En Yakın Komşu (K-NN) algoritmasıyla doldurulmuş ve veri gürültüsü Gaussian filtreleme yöntemiyle azaltılmıştır. Modeller arasında, RNN en yüksek doğruluğu sağlayarak PM10 ve SO2 tahminlerinde üstün performans göstermiştir. RNN modeliyle yapılan analizler sonucunda, gelecek üç yıl için tahminler yapılmış ve sonuçlar scatter plot, box plot ve violin plot gibi grafik teknikleriyle görselleştirilmiştir. Çalışmanın adımlarını gösteren akış diyagramı Şekil 4.9'da verilmiştir.



Şekil 4.9. Uygulama 2 İçin Akış Diyagramı

Bu çalışma, derin öğrenme yöntemlerinin çevresel veri analizindeki potansiyelini vurgulamakta ve mevcut modelleme sınırlamalarına yönelik iyileştirme ihtiyacını ortaya koymaktadır. Sunulan yaklaşım, hava kalitesi tahmin modellerinde doğruluğu artırmaya yönelik önemli bir adım niteliğindedir.

4.4.1. Verilerin Toplaması

Bu çalışmada kullanılan veri seti, Türkiye Çevre, Şehircilik ve İklim Değişikliği Bakanlığı'na bağlı Sürekli İzleme Merkezi'nden (SİM) temin edilmiş olup, 2014–2024 yılları arasındaki 10 yıllık günlük hava kalitesi ölçümlerini içermektedir [77]. Veri setinde yer alan ölçüm parametreleri arasında PM10, SO₂, CO, NO₂, NO_x, NO ve O₃ bulunmaktadır. Ancak, bazı ölçümlerde eksiklikler tespit edilmiştir. Veri setinin 10 yıllık bir dönemi kapsamasının temel amacı, modellerin daha iyi eğitilmesi ve tahmin doğruluğunun artırılmasıdır. Uzun dönemli ve geniş kapsamlı bir veri seti, modellerin daha sağlam sonuçlar üretmesine ve hava kalitesindeki uzun vadeli eğilimlerin ve mevsimsel değişimlerin daha doğru bir şekilde analiz edilmesine olanak tanımaktadır. Böylece, derin öğrenme modellerinin performansı optimize edilerek, çevresel veri analizi süreçlerinde daha güvenilir sonuçlar elde edilmiştir. Verilerin ilk 10 satırı Tablo 4.3'te verilmiştir ve tespit edilen eksik veri oranları şu şekildedir:

- Eksik PM10 Verileri: 246 satır

- Eksik SO2 Verileri: 364 satır
- CO: 376 eksik veri
- NO2: 224 eksik veri
- NOX: 542 eksik veri
- NO: 335 eksik veri
- O3: 255 eksik veri

Tablo 4.3.Veri Setinin Asıl Halinin İlk 10 Satırı

Tarih	PM10($\mu\text{g}/\text{m}^3$)	SO2($\mu\text{g}/\text{m}^3$)	CO($\mu\text{g}/\text{m}^3$)	NO2($\mu\text{g}/\text{m}^3$)	NOX($\mu\text{g}/\text{m}^3$)	NO($\mu\text{g}/\text{m}^3$)	O3($\mu\text{g}/\text{m}^3$)
2014-03-04 00:00:56	70,57	8,54	460,87	49,56	nan	nan	35,23
2014-03-05 00:00:56	135,89	23,23	756,52	65,74	nan	nan	19,28
2014-03-06 00:00:56	53,92	4,50	217,39	37,21	nan	nan	47,35
2014-03-07 00:00:56	52,29	nan	273,91	26,60	nan	nan	62,34
2014-03-08 00:00:56	21,05	nan	352,17	15,78	nan	nan	56,39
2014-03-09 00:00:56	23,71	nan	369,57	8,89	nan	nan	63,58
2014-03-10 00:00:56	21,43	nan	343,48	15,80	nan	nan	59,63
2014-03-11 00:00:56	26,06	nan	527,27	14,01	nan	nan	67,79
2014-03-12 00:00:56	nan	nan	nan	nan	nan	nan	nan
2014-03-13 00:00:56	nan	nan	nan	nan	nan	nan	nan

Tablo 4.3'te görüldüğü üzere, veri setinde bazı ölçüm parametreleri için eksik veriler (nan) bulunmaktadır. Özellikle NOx, NO ve SO2 gibi parametrelerde eksik değerlerin sayısının fazla olduğu dikkat çekmektedir. Bu eksik veriler, analizlerin ve modelleme süreçlerinin doğruluğunu etkileyebileceği için uygun yöntemlerle tamamlanması gerekmektedir.

4.4.2. Eksik Verilerin Tamamlanması

Bu çalışmada kullanılan hava kalitesi veri seti, analiz ve modelleme öncesinde eksik veri tamamlama ve gürültü azaltma gibi ön işleme adımlarından geçirilmiştir. Bu ön işleme süreci, veri setinin doğruluğunu ve bütünlüğünü sağlamakla kalmayıp, analiz ve tahmin süreçlerinin güvenilirliğini artırmayı hedeflemiştir. Eksik veri sorununu çözmek için K-En Yakın Komşu (K-NN) algoritması kullanılmıştır. K-NN algoritması, eksik verilerin tahmin edilmesinde etkili ve yaygın bir yaklaşım olarak, veri setindeki eksik gözlemleri mevcut veri noktalarıyla karşılaştırarak tamamlamaktadır. Temel olarak, eksik bir değere sahip bir gözlem, veri setindeki diğer gözlemlerle karşılaştırılıp ve en yakın komşuları belirlenmektedir. Bu komşuların değerleri kullanılarak eksik

veri tahmin edilmektedir. Algoritma, komşuların ağırlıklı ortalamasını alarak veya doğrudan değerlerini kullanarak eksik veriyi tamamlamaktadır. Bu çalışmada, K-NN algoritması her bir eksik değer için en uygun tahmini gerçekleştirmiştir. Tablo 4.3'te eksik verilerin olduğu haliyle veri seti sunulmuştur. Eksikliklerin özellikle NO_x, NO ve SO₂ gibi parametrelerde yoğunlaştığı görülmektedir. Bu durum, analiz ve modelleme süreçlerini olumsuz yönde etkileyebilecek bir durum yaratmaktadır. Eksik verilerin tamamlanmaması, analiz sonuçlarının doğruluğunu düşürmekte ve model tahmin performansını olumsuz etkilemektedir. Bu nedenle, eksik verilerin tamamlanması çalışmanın güvenilirliği açısından kritik bir adım olarak değerlendirilmiştir. K-NN algoritmasıyla eksik veri tahmin edildikten sonra, veri seti daha bütünlüklü hale getirilmiş ve Tablo 4.4'te gösterilen biçimde eksiksiz bir yapı kazanmıştır. Örneğin, 2014-03-07 tarihindeki eksik SO₂ değeri 28,74 olarak tahmin edilmiş, aynı şekilde eksik NO ve NO_x değerleri de algoritma tarafından tamamlanmıştır. Ayrıca, tamamen eksik olan 2014-03-12 ve 2014-03-13 tarihlerindeki tüm parametreler K-NN algoritmasıyla tahmin edilerek tamamlanmıştır. Eksik verilerin tamamlanmasının ardından, veri setindeki ani dalgalanmaları ve gürültüyü azaltmak amacıyla Gauss filtresi uygulanmıştır. Gauss filtresi, çevresel verilerde sıkça karşılaşılan gürültüleri azaltmak için verileri belirli bir dağılıma göre ağırlıklandırıp ve pürüzsüz hale getirmektedir. Bu işlem, veri setini daha stabil bir yapıya kavuşturarak genel eğilimlerin daha net bir şekilde gözlemlenmesini sağlamıştır. Özellikle zaman serisi analizlerinde, bu adım model doğruluğunu önemli ölçüde artırmıştır. Bu ön işleme adımları sonucunda, hava kalitesi veri seti eksiksiz, tutarlı ve modellemeye hazır hale getirilmiştir. Bu sayede, analiz ve tahmin süreçleri için güvenilir bir temel oluşturulmuş, modelleme performansı optimize edilmiştir. Aşağıdaki Tablo 4.4, K-NN algoritmasıyla tamamlanmış veri setinin ilk 10 satırını göstermektedir.

Tablo 4.4. Verisetinin K-NN İle Doldurulmuş Hali

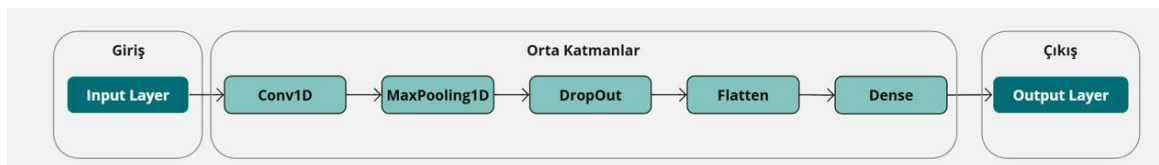
Tarih	PM10($\mu\text{g}/\text{m}^3$)	SO2($\mu\text{g}/\text{m}^3$)	CO($\mu\text{g}/\text{m}^3$)	NO2($\mu\text{g}/\text{m}^3$)	NOX($\mu\text{g}/\text{m}^3$)	NO($\mu\text{g}/\text{m}^3$)	O3($\mu\text{g}/\text{m}^3$)
2014-03-04 00:00:56	70.57	8.54	460.87	49.56	100.948	29.41000	35.23
2014-03-05 00:00:56	135.89	23.23	756.52	65.74	198.724	67.178	19.28
2014-03-06 00:00:56	53.92	4.5	217.39	37.21	66.328	21.892	47.35
2014-03-07 00:00:56	52.29	28.74	273.91	26.6	40.892	10.978	62.34
2014-03-08 00:00:56	21.05	21.534	352.17	15.78	24.17	6.196	56.39
2014-03-09 00:00:56	23.71	3.304	369.57	8.89	20.01	4.672	63.58
2014-03-10 00:00:56	21.43	12.302	343.48	15.8	25.16	6.51	59.63
2014-03-11 00:00:56	26.06	2.686	527.27	14.01	19.578	3.196	67.79
2014-03-12 00:00:56	48.78043	7.12803	517.94021	28.35337	52.35370	14.91725	56.82346

Tablo 4.4, K-NN algoritmasıyla eksik verilerin tamamlanmasının ardından veri setinin bütünleşik halini sunmaktadır. Algoritma, eksik verilerin tahmini için mevcut veriler arasındaki ilişkileri kullanarak eksiklikleri etkili bir şekilde doldurmuştur. Özellikle tamamen eksik olan gözlemler, komşu değerler üzerinden hesaplanan tahminlerle tamamlanmış ve veri seti analiz süreçleri için uygun hale getirilmiştir. Bu düzenleme, veri setinin doğruluğunu artırarak, hava kalitesi tahmin modellerinin daha sağlıklı sonuçlar üretmesine olanak tanımıştır. Böylece, eksiksiz hale getirilen bu veri seti, model eğitimi ve zaman serisi analizleri için güçlü bir temel oluşturmuştur.

4.4.3. Kullanılan Metotlar ve Analizler

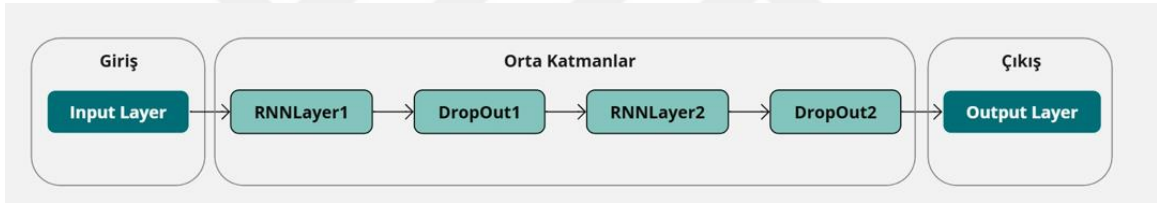
Bu tezde, hava kirletici verilerinin tahmini için CNN, RNN ve LSTM olmak üzere üç farklı derin öğrenme yöntemi kullanılmıştır. Bu yöntemlerin uygulanması sırasında, veriler önce bir Gauss filtresiyle düzenlenmiş ve her yöntem için tahminler yapılmıştır. Sonuçlar R^2 puanları, Kök Ortalama Kare Hata (RMSE) ve Ortalama Mutlak Hata (MAE) temelinde değerlendirilmiştir. Son olarak, bu çalışmalar grafiklerle desteklenmiştir. Gauss filtresi, çevresel verilerin gürültüsünü azaltmak ve sinyali daha pürüzsüz hale getirmek için uygulanan bir veri filtreleme uygulamasıdır. Bu uygulama, belirli bir dağılıma göre ağırlıklandırarak verilerin ortalamasını almakta ve böylece daha kararlı bir veri kümesi elde edilmektedir.

Bu çalışmada kullanılan CNN modelde özel olan durum, CNN mimarisinin zaman serisi verilerine adapte edilmesidir. Geleneksel olarak görüntü verilerindeki mekansal ilişkileri öğrenmek için kullanılan CNN'ler, bu modelde, zaman serisi verilerindeki kısa vadeli bağımlılıkları ve tekrar eden desenleri yakalamak için optimize edilmiştir. Bu amaçla, tek boyutlu evrişim katmanı (Conv1D) kullanılarak zaman serisi verilerindeki yerel özellikler çıkarılmış ve maksimum havuzlama (MaxPooling1D) ile boyut küçültülerek gereksiz verilerden arındırılmıştır. Ayrıca, bırakma (Dropout) katmanı ile aşırı öğrenme önlenmiş ve modelin genelleştirme kapasitesi artırılmıştır. Bu özel yapı, zaman serisi verilerindeki karmaşık desenleri anlamak ve tahmin doğruluğunu artırmak için tasarlanmış olup, hava kalitesi gibi dinamik veriler üzerinde etkili bir performans sergilemektedir. Modelin mimarisi Şekil 4.10'da verilmiştir.



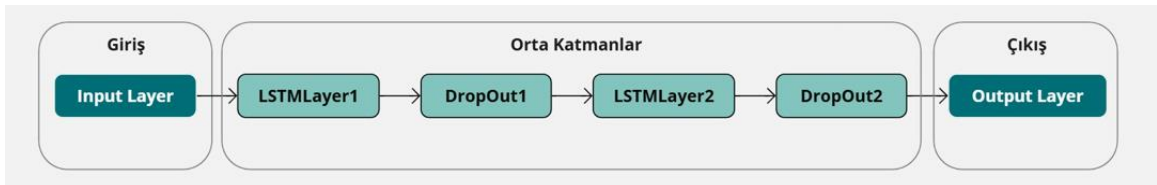
Şekil 4.10. CNN Model Mimarisi

Tekrarlayan Sinir Ağı (RNN), özellikle zaman serisi verileri olmak üzere sıralı verileri modellemek için tasarlanmış bir yapay sinir ağı türüdür. RNN'ler, her zaman adımında önceki çıktıları ve mevcut girdiyi dikkate alarak geleceğe yönelik tahminler yapmaktadır. RNN'lerin temel çalışma prensibi, hücrelerdeki geçmiş bilgilerin gizli bir durumda depolanmasıdır. Şekil 4.11'deki RNN model mimarisi bir giriş katmanı ile başlamakta ve veriler bu katmanda alınmakta, böylece model bunları işleyebilmektedir. İlk olarak Tekrarlayan Sinir Ağı (RNN) katmanı devreye girmektedir. Bu katman, özellikle zaman serisi gibi sıralı verilerde sıralı bağımlılıkları öğrenmekte ve verilerin zaman içindeki ilişkilerini keşfetmektedir. RNN katmanından sonra Dropout katmanı gelmekte ve bu aşamada, modelin aşırı öğrenmesini önlemek için bazı nöronlar rastgele devre dışı bırakılmaktadır. Ardından, ikinci bir RNN katmanı verileri tekrar işler ve sıralı bağımlılıkları daha derin bir düzeyde öğrenmektedir. Bu ikinci RNN katmanından sonra, başka bir Dropout katmanı eklenmekte, böylece modelin genelleme yeteneği artmaktadır. Son olarak, çıktı katmanı modelin son tahminini üretmektedir. Bırakma katmanları, modelin aşırı öğrenmeyi önlerken daha genel sonuçlar üretmesine yardımcı olmaktadır.



Şekil 4.11. RNN Model Mimarisi

LSTM modeli, ardışık veriler arasındaki uzun vadeli ilişkileri yakalayarak tahminler yapmaktadır. Şekil 4.12'deki LSTM model mimarisi bir girdi katmanı ile başlamakta ve veriler bu katmandan modele alınmaktadır. İlk aşamada bir LSTM katmanı bulunmaktadır. İlk LSTM katmanı verileri işlemekte ve Dropout katmanına iletmektedir. Dropout katmanı, modelin aşırı öğrenmesini önlemek için bazı nöronları rastgele devre dışı bırakmaktadır. İkinci aşamada, ikinci bir LSTM katmanı devreye girmekte ve verileri daha derin bir düzeyde işlemektedir. Modelin genelleme yeteneğini artıran başka bir Dropout katmanı eklenmektedir. Son olarak, çıktı katmanı modelin nihai tahminlerini üretmektedir. Dropout katmanları, modelin aşırı öğrenmeye olan eğilimini azaltmakta ve daha genel sonuçlar üretmesini sağlamaktadır.



Şekil 4.12. LSTM Model Mimarisi

Elde edilen sonuçlara göre, RNN modeli, kısa ve uzun vadeli bağımlılıkları öğrenme kapasitesi sayesinde en düşük hata oranlarıyla en iyi performansı göstermiştir. Özellikle zaman serisi verilerindeki ardışık bağımlılıkları etkin bir şekilde modelleme yeteneği, RNN'nin tahminlerde daha başarılı olmasını sağlamıştır. LSTM modeli, uzun vadeli bağımlılıkları yakalama özelliği sayesinde tatmin edici bir performans sergilemiş, ancak RNN kadar düşük hata oranlarına ulaşamamıştır. Bunun temel nedeni, LSTM'nin uzun süreli bağımlılıkları daha iyi öğrenmesine rağmen kısa vadeli bağımlılıkları RNN kadar etkin bir şekilde yakalayamamasıdır.

CNN modeli ise, verilerdeki yerel desenleri öğrenme kabiliyeti sayesinde makul bir doğruluk sağlamış ancak zaman serisi verilerindeki yüksek varyansları ve ardışık bağımlılıkları yakalamadaki sınırlı yeteneğinden dolayı daha yüksek hata oranlarına sahip olmuştur. Bu durum, CNN'nin temel olarak yerel özellikler ve desenler üzerine odaklanması, ancak zaman serisi verilerindeki daha geniş bağımlılıkları modellemede yetersiz kalmasından kaynaklanmaktadır. Deneysel sonuçlara göre, RNN modeli PM10 ve SO2 tahminlerinde en yüksek doğruluğu elde ederken, CNN daha yüksek hata değerleri ile daha düşük performans göstermiştir. Bu farklılıklar, modellerin tahmin yeteneklerinin veri türüne ve bağımlılık ilişkilerini modelleme kapasitelerine bağlı olarak değiştiğini açıkça göstermektedir.

Çalışmadan elde edilen değerler ve görselleştirmeler aşağıda detaylı olarak verilmiştir. Tablo 4.5'te gösterilen CNN model sonuçları, PM10 ve SO2 tahminlerinde makul doğruluk sağlamaktadır. Ancak, özellikle PM10 için MAE ve RMSE değerleri diğer modellerden daha yüksektir ve bu da hataların daha büyük olduğunu göstermektedir. R^2 değerleri 0,88 ve 0,93 seviyelerinde olup, modelin tahmin performansının genel olarak iyi olduğunu ancak hata oranlarının daha yüksek olduğunu göstermektedir.

Tablo 4.5. CNN Modeli Performans Metrikleri

Kirleticiler	MAE	RMSE	R^2
PM10	4.68	6.44	0.88
SO2	1.34	2.07	0.93

Tablo 4.6'daki RNN modeli hem PM10 hem de SO2 tahminlerinde en düşük MAE ve RMSE değerlerine sahip olup en iyi performansı göstermektedir. Bu, RNN modelinin hem kısa vadeli hem de uzun vadeli bağımlılıkları başarıyla öğrenebildiğini ve daha doğru tahminler yapabildiğini göstermektedir. R^2 puanları da 0,97'dir ve bu da modelin varyansı çok iyi açıkladığını ve tahminlerin gerçek değerlere çok yakın olduğunu göstermektedir.

Tablo 4.6. RNN Modeli Performans Metrikleri

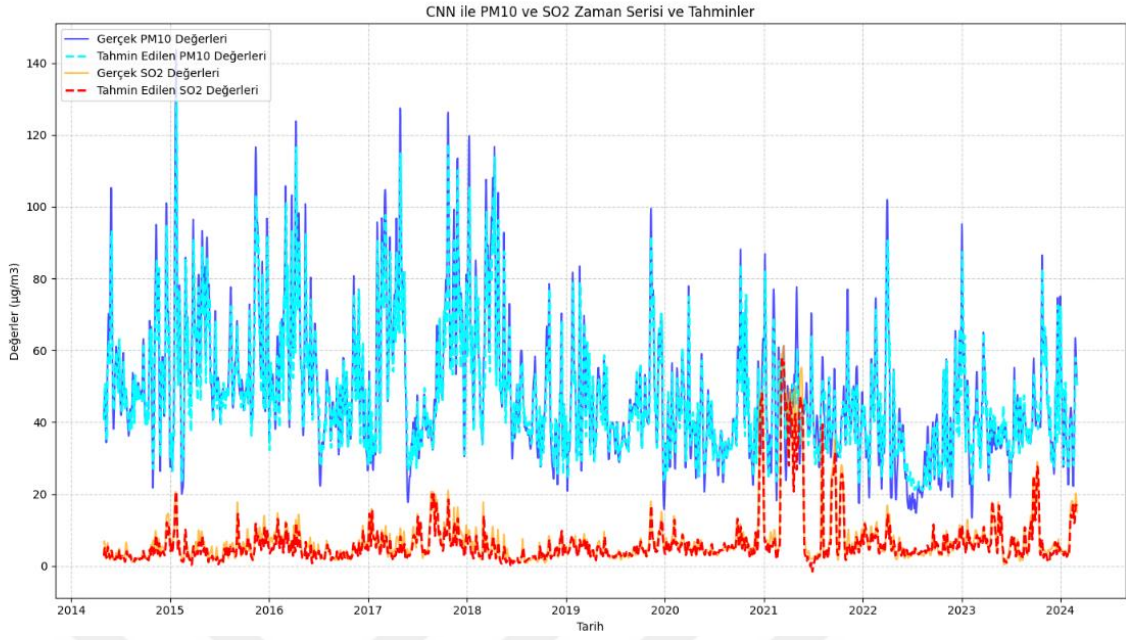
Kirleticiler	MAE	RMSE	R ²
PM10	2.47	3.52	0.97
SO2	0.86	1.26	0.97

Tablo 4.7’de sunulan sonuçlara göre, LSTM modeli, RNN’ye kıyasla biraz daha düşük ancak genel anlamda iyi bir performans sergilemiştir. Modelin MAE ve RMSE değerleri, RNN’den daha yüksek, ancak CNN’den daha düşük seviyelerde gerçekleşmiştir. LSTM modelinin R² puanları 0,94 ve 0,95 olarak hesaplanmış olup, bu değerler modelin veri setinin değişkenliğinin önemli bir bölümünü açıklayabildiğini göstermektedir. LSTM’nin uzun vadeli bağımlılıkları belirleme ve öğrenme konusundaki başarısı dikkate değer olsa da bu özel senaryoda RNN modeli kadar etkili olmadığı görülmüştür. Bu durum, LSTM’nin uzun vadeli bağımlılıkları yakalamaya yönelik güçlü yapısının, kısa vadeli bağımlılıkların daha baskın olduğu senaryolarda performans avantajını tam anlamıyla sağlayamamasından kaynaklanıyor olabilmektedir.

Tablo 4.7. LSTM Modeli Performans Metrikleri

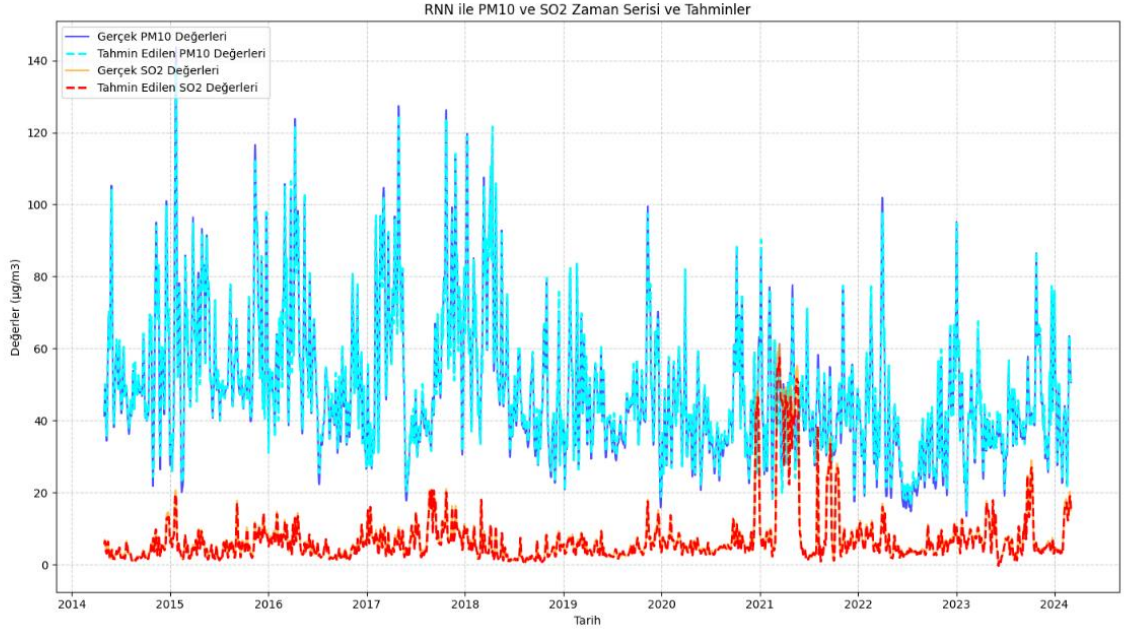
Kirleticiler	MAE	RMSE	R ²
PM10	3.67	4.66	0.94
SO2	1.07	1.85	0.95

Sonuç olarak LSTM modeli zaman serisi verilerinde performansı açısından RNN'e yakın bir sonuç vermektedir ancak RNN kadar iyi performans göstermemektedir. Buna rağmen CNN'e göre daha iyi tahminlerde bulunmuştur. CNN modelinin daha yüksek hata değerlerine sahip olması, zaman serisi verilerindeki seri bağımlılıkları yakalama kapasitesinin diğer modellere göre daha düşük olduğunu göstermektedir. RNN modeli hem PM10 hem de SO2 tahminlerinde en düşük hata değerlerine (MAE ve RMSE) ve en yüksek R² puanlarına sahip olup bu veri seti ve problem için en başarılı model olarak öne çıkmaktadır. Yine grafiklere baktığımızda, Şekil 4.13'te, yani CNN modelinin grafiğinde, tahmin edilen PM10 ve SO2 değerleri (mavi ve kırmızı çizgiler) ile gerçek değerler (mavi ve turuncu çizgiler) arasında bazı farklılıklar görülmektedir. Tahminler genellikle gerçek değerlerin örüntülerini takip etse de özellikle PM10 için bazı yüksek değerlerde daha belirgin sapmalar görülmektedir. SO2 tahminlerinde ise genel eğilimleri yakalamada başarılı olsa da bazı ani değişimlerde hata yapmaktadır.



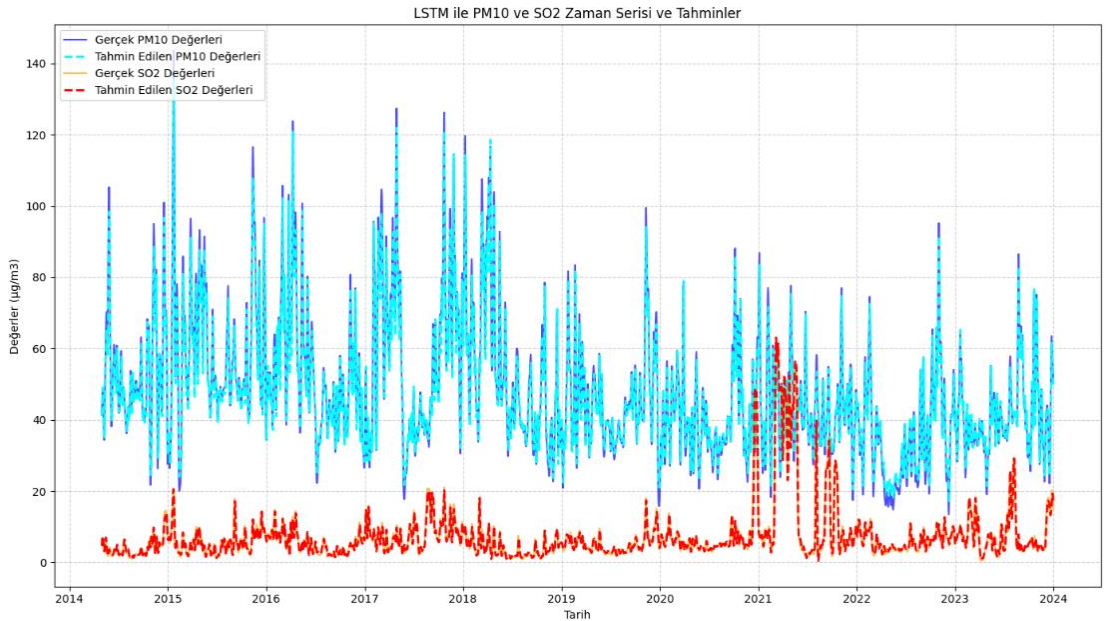
Şekil 4.13. CNN Modeli ile Tahminleme

Şekil 4.14'teki grafikte, RNN model grafiği gerçek değerlerle (mavi ve turuncu çizgiler) tahminlere (mavi ve kırmızı çizgiler) daha yakın bir uyum göstermektedir. Özellikle hem PM10 hem de SO2 için tahminler zaman serisinin değişimlerini ve dalgalanmalarını iyi bir şekilde yakalıyor gibi görünmektedir. Gerçek değerler ile tahminler arasındaki farklar diğer modellere kıyasla daha azdır ve tahmin çizgileri genel eğilimleri ve ani değişimleri oldukça iyi takip etmektedir. RNN modeli en düşük MAE ve RMSE değerleriyle en yüksek tahmin doğruluğunu göstermektedir. Hem PM10 hem de SO2 ölçümleri için zaman serisindeki hem kısa hem de uzun vadeli bağımlılıkları etkili bir şekilde yakalamaktadır.



Şekil 4.14. RNN Modeli ile Tahminleme

Şekil 4.15'te, LSTM model grafiği tahminlerin (camgöbeği ve kırmızı çizgiler) gerçek değerlerle (mavi ve turuncu çizgiler) büyük ölçüde tutarlı olduğunu göstermektedir. PM10 tahminlerinde, RNN modeline benzer şekilde iyi bir uyum vardır ve dalgalanmaları ve değişiklikleri iyi yakalamaktadır. SO2 tahminlerinde, genel eğilimleri iyi yakalarken, bazı ani değişikliklerde sapmalar gözlemlenmektedir. LSTM modeli RNN'e yakın performans göstermektedir. MAE ve RMSE değerleri biraz daha yüksek olsa da yine de oldukça başarılı tahminler yapmaktadır. Zaman serisi verilerindeki uzun vadeli bağımlılıkları verimli bir şekilde yakalayabilmektedir.



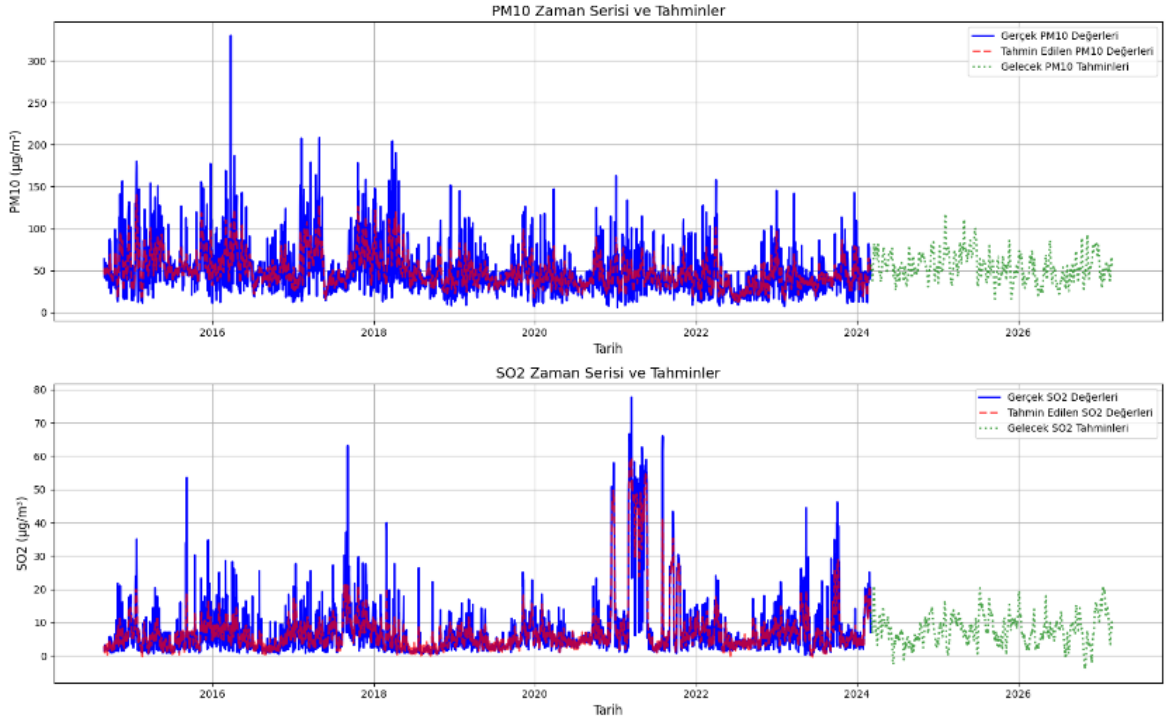
Şekil 4.15. LSTM Modeli ile Tahminleme

Sonuç olarak en iyi performansı gösteren model RNN olmuştur. Gerçek ve öngörülen değerler arasındaki farklardaki tutarsızlıklar minimumdur ve model veriler içindeki hem kısa hem de uzun vadeli ilişkileri veya bağımlılıkları etkili bir şekilde öğrenmektedir. İkinci olarak, LSTM iyi bir performansa sahiptir ve RNN'e yakındır. Zaman serilerindeki uzun vadeli bağımlılıkları yakalamada başarılı olmuştur. CNN, diğer modellere kıyasla daha düşük bir performansa sahiptir. Özellikle PM10 değerlerinin yüksek varyansını yakalamada zorluk çekmektedir. Görüntü işleme için daha uygun olan CNN, zaman serisi verilerinde diğer modeller kadar etkili değildir. Çalışmamız bu üç yöntem arasında en başarılı RNN modeli olduğundan, gelecekteki tahminleri yaparken bu yöntemle devam etmektedir. Önümüzdeki 3 yıla ait performans metriklerinin çıktısı aşağıdaki Tablo 4.8'te gösterilmiştir. Bu tabloya göre modelin tahminlerinin gerçek verilerle karşılaştırıldığında mükemmel olmadığını, ancak orta düzeyde bir performans gösterdiğini göstermektedir. Modelin bazı öngörülerinde daha büyük hatalar yapabileceğini, ancak genel olarak makul bir doğruluk seviyesinde olduğunu göstermektedir.

Tablo 4.8. RNN Modeli ile Gelecek Tahminlemede Performans Metrikleri

Kirleticiler	MAE	RMSE	R ²
PM10	11.60	17.15	0.63
SO2	2.39	3.99	0.82

Şekil 4.16'teki modelin grafik çıktısında, PM10 için gelecekteki tahminler daha düzdür ve daha düşük dalgalanmalar sergilemektedir. Bu, modelin gelecekteki değişiklikleri tahmin etme kapasitesinin sınırlı olabileceğini göstermektedir. SO2 için tahminler nispeten daha düşük varyansa sahiptir ve daha düz bir seyir izlemektedir. Hem PM10 hem de SO2 için daha düz ve daha düşük dalgalanmalar gözlemlenmekte, bu da modelin belirsizlikleri yönetme kapasitesinin sınırlı olabileceğini ve daha karmaşık olayları tahmin etmekte zorluk çekebileceğini göstermektedir.



Şekil 4.16 RNN ile Gelecek 3 Yılın Tahmin Grafiği

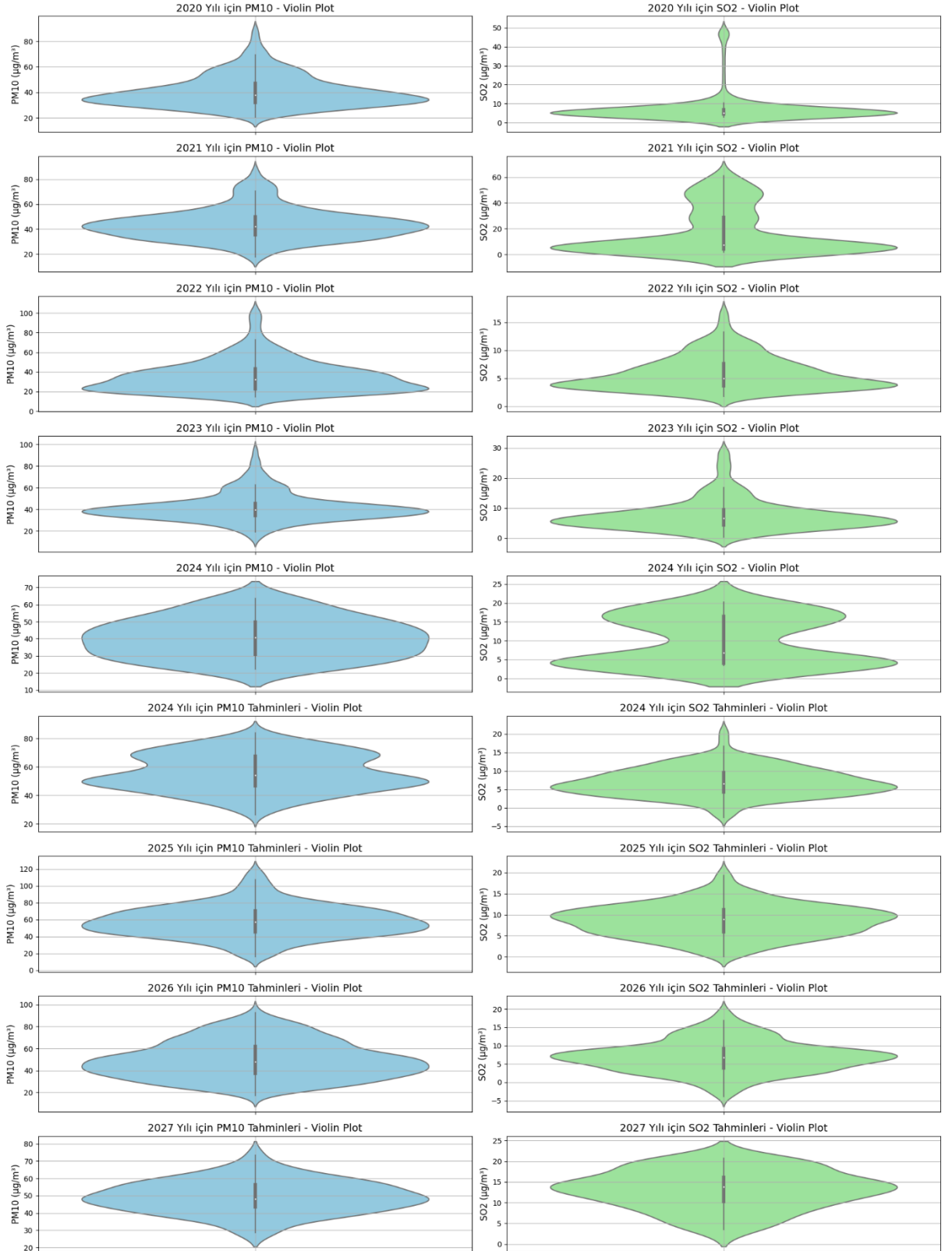
Sonuç olarak, model PM10 için orta düzeyde performans gösterirken, SO2 tahminlerinde daha başarılı olmuştur. Gelecekteki tahminlerin nispeten düz ve düşük varyansa sahip olması, modelin tahmin kapasitesinin bazı sınırlamalarını ortaya koymaktadır. Bu sonuçlar gelecekteki hava kalitesi yönetimi ve planlaması için kullanılabilir, ancak tahminlerin dikkatlice değerlendirilmesi gerektiğini göstermektedir. Daha fazla geliştirme ve farklı model yapıları tahmin doğruluğunu artırabilme ihtimalini göstermektedir.

Çalışmanın son aşamasında, büyük veri setini daha yönetilebilir bir forma dönüştürmek ve karmaşık verileri daha anlaşılır hale getirmek amacıyla veri görselleştirme yöntemleri kullanılmıştır. Bu süreçte, mevcut veriler ile geleceğe yönelik tahminlerden elde edilen veriler görselleştirilmiş ve karşılaştırmalar yapılmıştır. Bu yaklaşım, verilerin daha kolay yorumlanmasını sağlamış ve analiz süreçlerini destekleyerek elde edilen sonuçların daha açık bir şekilde sunulmasına katkıda bulunmuştur.

Şekil 4.17, PM10 ve SO2 seviyelerindeki yıllar içindeki değişiklikleri ve gelecekteki tahminlerini özetlemektedir. PM10 seviyeleri 2021 ve 2024'te daha geniş bir dağılım gösterirken, bu yıllarda hava kirliliğinin arttığını gösterirken, 2022 ve 2023'te daha dar ve daha düşük konsantrasyonlarda yoğunlaşmış ve nispeten daha iyi hava kalitesini göstermiştir. 2024-2027 yılları için PM10 projeksiyonları önceki yıllara göre daha geniş bir dağılım gösteriyor ve bu da modelin gelecekteki PM10 seviyelerini tahmin etmede belirsizlik yaşadığını göstermektedir. Öte yandan SO2 seviyeleri genellikle tutarlı bir dağılım göstermektedir. Ancak 2021 gibi bazı yıllarda daha geniş bir dağılım

göstermekte ve bu da daha kötü hava kalitesi dönemlerini göstermektedir. 2020, 2022 ve 2024 yıllarındaki düşük SO₂ konsantrasyonları bu yıllarda daha temiz hava koşullarına işaret etmektedir. Gelecekteki SO₂ projeksiyonları genellikle düşük seviyelerde yoğunlaşsa da 2025 ve 2027 yıllarında daha geniş bir dağılım gösteriyor ve bu da modelin bu yıllar için daha fazla belirsizlik yaşadığını göstermektedir.

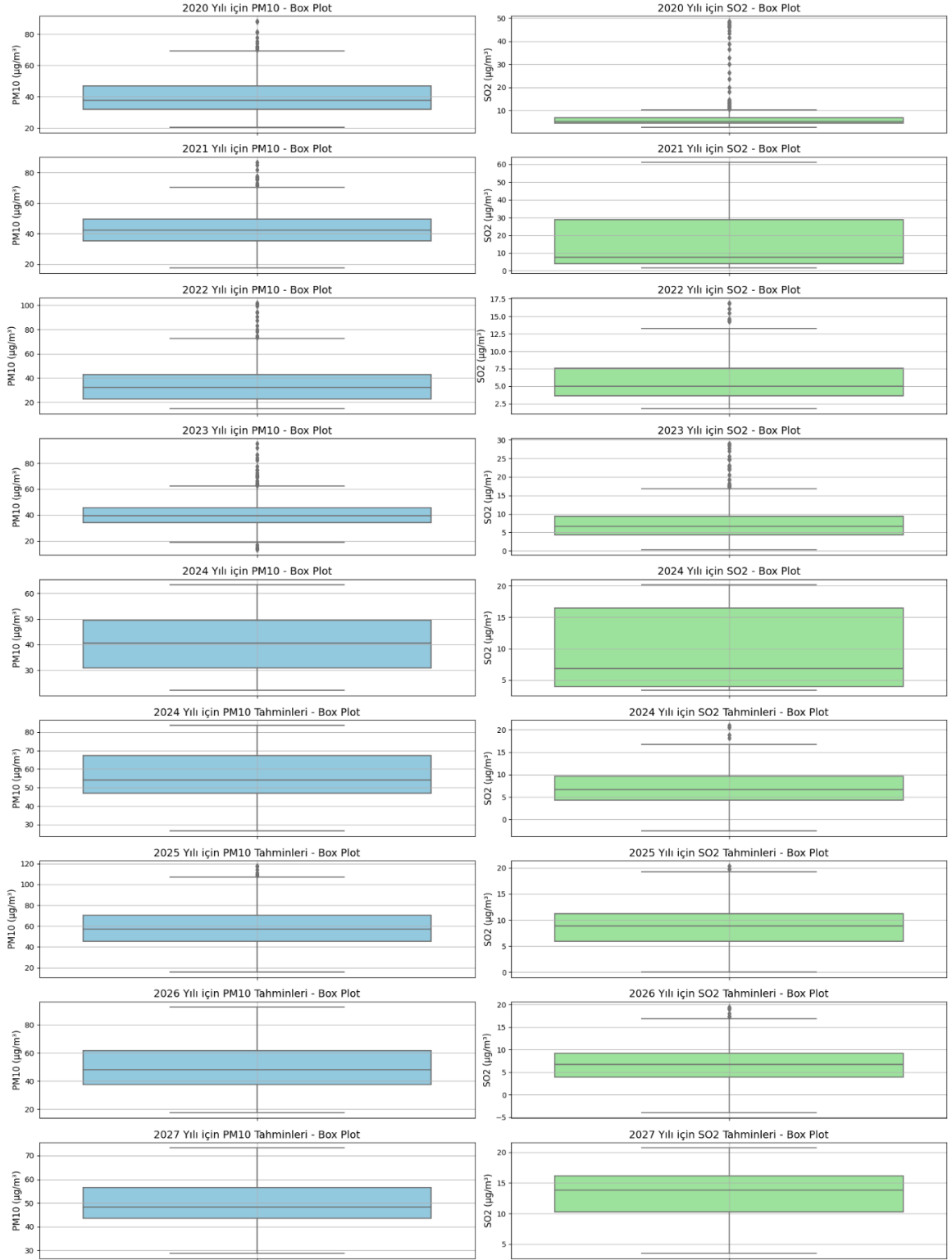




Şekil 4.17. Keman Grafiği (Violin Plot) ile Görselleştirme

Şekil 4.18’de yer alan 2020-2024 yılları arasındaki gerçek veriler incelendiğinde PM10 düzeylerinin bazı yıllarda (özellikle 2021 ve 2023) daha uç değerler içerdiği ve bu yıllarda hava

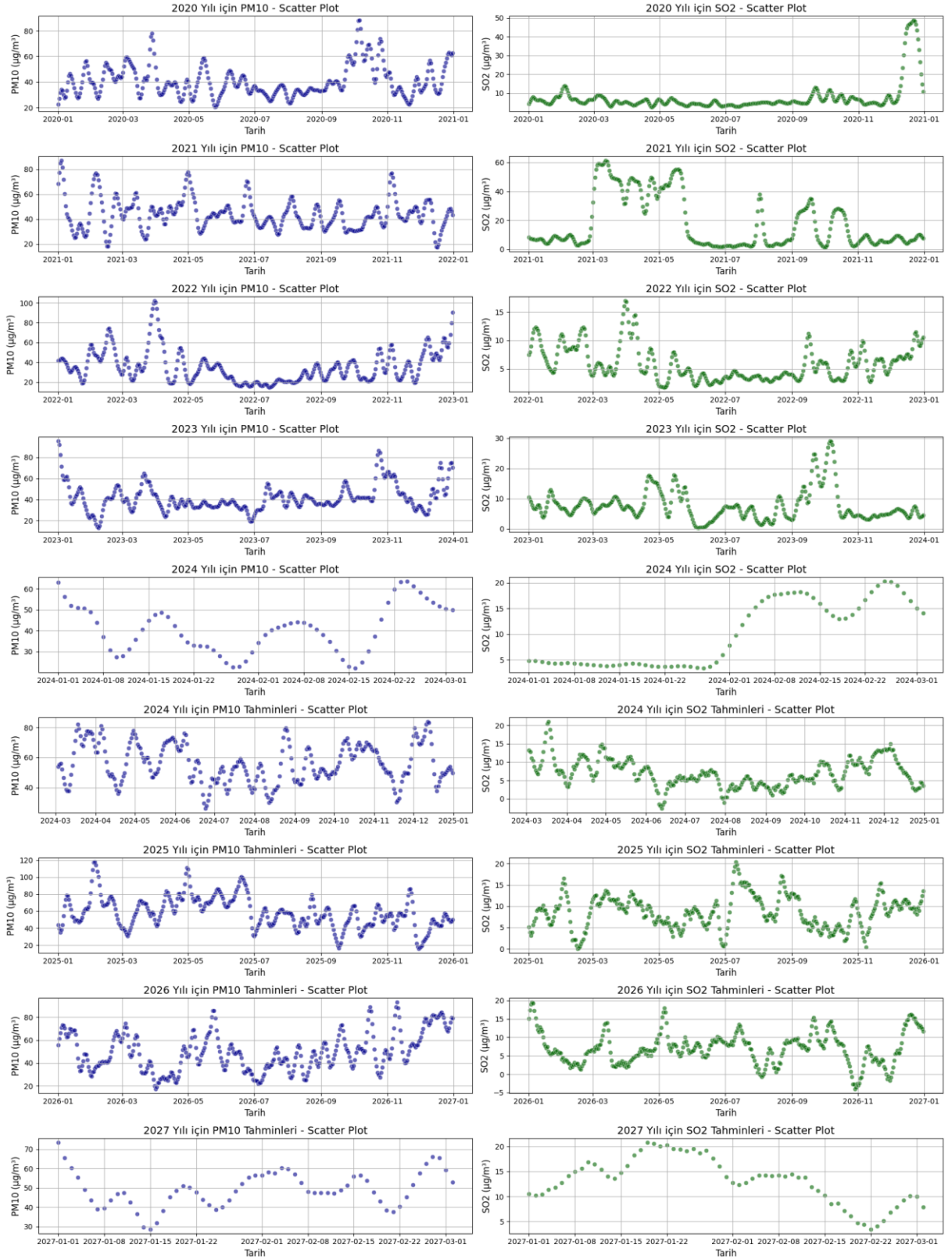
kirlilięi olaylarının daha sık ve yoğun yaşandıęı ortaya çıkmaktadır. PM10 için medyan deęerlerin genellikle 20-40 $\mu\text{g}/\text{m}^3$ arasında yoğunlaştıęı ve yıllar arasındaki varyansın genellikle benzer olduęu anlaşılmaktadır. SO2 düzeyleri genellikle düşük bir aralıktadır (2-10 $\mu\text{g}/\text{m}^3$) kalmakla birlikte 2021 ve 2022 yıllarında daha fazla varyans ve bazı uç deęerler kaydedilmektedir. Bu durum hava kalitesinin özellikle bu yıllarda deęiştini ve bazı dönemlerde beklenmedik derecede yüksek SO2 konsantrasyonlarının yaşandıęını göstermektedir. Geleceęe yönelik projeksiyonlara bakıldığında, model tarafından 2024-2027 yılları arasında tahmin edilen PM10 ve SO2 düzeylerinin dağılımının önceki yıllarla genel olarak tutarlı olduęu ancak bazı belirsizlikler ve varyanslar içerdini görölmektedir. PM10 tahminleri için, daha geniş bölmeler (IQR), modelin gelecekteki PM10 seviyelerinde daha fazla deęişkenlik öngördüğünü ve bu nedenle bazı dönemlerde daha yüksek kirlilik deęerleri beklediğini göstermektedir. SO2 tahminleri için, düşük konsantrasyon seviyelerinde tutarlı bir yoğunluk ve nispeten dar bir dağılım gözlemlenmekte ve bu, modelin gelecekteki SO2 seviyelerinin genellikle düşük kalacağı beklentisini yansıtmaktadır. Ancak, bazı yıllardaki birkaç uç deęer, olası gelecekteki beklenmeyen olayların da hesaba katılması gerektiğini göstermektedir.



Şekil 4.18. Kutu Grafiği (Box Plot) ile Görselleştirme

Şekil 4.19’da 2020-2024 yılları arasındaki gerçek verilere bakıldığında PM10 seviyelerinde önemli bir dalgalanma görülmektedir. Özellikle 2021 ve 2023 yıllarında PM10 seviyelerinde sık ve yüksek zirveler görülmektedir; bu da hava kirliliği olaylarının bu yıllarda daha sık ve yoğun olduğunu

göstermektedir. 2022 ve 2024 yıllarında PM10 seviyelerinin daha durağan ve düşük seviyelerde seyrettiği, ani artışların daha az olduğu anlaşılmaktadır. Bu da bu yıllarda hava kalitesinin nispeten daha iyi olduğunu göstermektedir. SO2 seviyeleri ise yıllar içinde daha düzenli bir seyir izlemekte ve genellikle düşük konsantrasyonlarda kalmaktadır. Ancak 2021 yılında SO2 seviyelerinde bazı ani artışlar gözlenmiştir; bu da bu yılın hava kalitesi açısından daha değişken olduğunu göstermektedir. 2022 ve 2023 yıllarında SO2 seviyelerinin daha düşük ve sabit kaldığı, 2024 yılında ise yeniden bazı dalgalanmaların yaşandığı görülmektedir. Şekil 4.19'daki gelecek tahminleri incelendiğinde, özellikle 2025 ve 2026 yıllarında PM10 seviyelerinde önemli dalgalanmalar ve yüksek zirveler öngörülmektedir. Bu, modelin bu yıllarda PM10 seviyelerinde ciddi değişiklikler ve potansiyel kirlilik olayları öngördüğünü göstermektedir. 2027 yılı için PM10 seviyelerinin daha düşük ve daha az değişken olması beklenmekte olup, bu yıl için daha iyi bir hava kalitesi beklentisini yansıtmaktadır. SO2 tahminleri genellikle düşük seviyelerde kalmakta ve büyük dalgalanmalar göstermemektedir. 2025 ve 2026 yıllarında SO2 seviyelerinde bazı artışlar beklenmektedir, ancak bu seviyeler genellikle düşük konsantrasyonlarda kalmaktadır. SO2 seviyelerinin 2027 yılında oldukça sabit kalması öngörülmektedir. Genel olarak, bu saçılım grafiği grafikleri, PM10 ve SO2 seviyelerindeki zaman içindeki değişiklikleri ve olası gelecekteki eğilimleri görselleştirmede yararlı olmaktadır. Tahminlerdeki bazı belirsizliklere rağmen, modelin hava kalitesindeki mevsimsel ve yıllık değişimleri başarılı bir şekilde yakaladığı sonucuna varılabilmektedir.



Şekil 4.19. Nokta Dağılım Grafiği (Point Scatter Plot) ile Görselleştirme

Genel olarak, bu scatter plot grafikleri, PM10 ve SO2 seviyelerinin zaman içindeki değişimlerini ve gelecekteki olası eğilimleri görselleştirme açısından faydalı olmaktadır. Tahminlerde bazı

belirsizlikler bulunmakla birlikte, modelin hava kalitesindeki mevsimsel ve yıllık deęişimleri yakalamada başarılı olduęu söylenebilmektedir.

4.4.4. Deneyisel Sonuçlar ve Tartışma

Bu tez çalışması, hava kalitesini deęerlendirmede veri görselleştirme yöntemlerinin ve hava kalitesi tahmini için kullanılan derin öğrenme modellerinin karşılaştırmalı analizini içermekte olup, bu alanda önemli bulgular sunmuştur. Özellikle RNN modeli, zaman serisi verilerindeki kısa ve uzun vadeli bağımlılıkları öğrenme kapasitesi sayesinde dięer modellere üstünlük sağlamıştır. Ancak, RNN modelinin gelecekteki olayları tahmin etmede karşılaştığı sınırlamalar ve daha düz varyans tahmini yapması, karmaşık olayların öngörülmesinde modelin kapasitesini artırma gerekliliğini göstermektedir. Çalışmada elde edilen bulgular, eksik veri doldurma, veri gürültüsünü azaltma ve gelişmiş görselleştirme yöntemlerinin hava kalitesi analizlerinde daha güvenilir ve anlamlı sonuçlar sunduğunu ortaya koymuştur. Scatter plot, box plot ve violin plot gibi grafiksel görselleştirme teknikleri, analiz sonuçlarının anlaşılabilirliğini artırarak karmaşık verilerin daha kolay yorumlanmasına olanak sağlamıştır. Bu durum, çevresel veri analizlerinde görselleştirmenin kritik bir öneme sahip olduğunu bir kez daha vurgulamaktadır.

Gelecekte yapılacak çalışmalarda tahmin doğruluğunu artırmak ve daha geniş kapsamlı analizler gerçekleştirmek adına aşağıdaki öneriler sunulmaktadır. İlk olarak, meteorolojik faktörlerin, özellikle rüzgar hızı, sıcaklık ve nem gibi çevresel deęişkenlerin modele entegre edilmesi, hava kalitesi tahminlerinin doğruluğunu önemli ölçüde artırabilecektir. Ayrıca, birden fazla hava kirleticisinin eş zamanlı olarak tahmin edilmesine olanak tanıyan çok hedefli tahmin modellerinin geliştirilmesi, analiz kapasitesini genişletecek ve çevresel etkilerin daha kapsamlı bir şekilde deęerlendirilmesini sağlayacaktır. Farklı derin öğrenme yaklaşımlarını birleştiren hibrit modellerin araştırılması ise mevcut sınırlamaların üstesinden gelmede etkili bir çözüm sunabilecektir.

Sonuç olarak, bu çalışma, hava kalitesi tahmini ve yönetiminde derin öğrenme modellerinin önemli bir potansiyele sahip olduğunu göstermiştir. Çalışmadan elde edilen bulgular, gelecekteki araştırmalar için sağlam bir temel oluşturmakta ve çevresel deęişkenlerin entegrasyonu ile hibrit model yaklaşımlarının araştırılmasının tahmin doğruluğunu artırmada kilit rol oynayacağını ortaya koymaktadır.

5. BULGULAR VE TARTIŞMA

Bu çalışma, hava kalitesi verilerinin derin öğrenme yöntemleriyle analiz edilmesi ve tahmin süreçlerinin iyileştirilmesi açısından önemli katkılar sunmaktadır. Uygulanan derin öğrenme modellerinin performansı, eksik verilerin tamamlanması ve veri işleme tekniklerinin etkileri ayrıntılı bir şekilde değerlendirilmiştir. Çalışmanın sonuçları, hava kalitesinin zaman içerisindeki değişimlerini etkili bir şekilde öngörmeye yönelik önemli çıkarımlar sağlamıştır. Eksik veri tamamlama sürecinde kullanılan K-En Yakın Komşu (K-NN) algoritması, veri setinin bütünlüğünü sağlamakta ve tahmin doğruluğunu artırmada başarılı olmuştur. Eksik verilerin tamamlanmasının yanı sıra, Gaussian filtreleme yöntemiyle veri gürültüsü azaltılmış ve zaman serisi verilerinin daha stabil bir yapıya kavuşması sağlanmıştır. Bu ön işleme adımları, çevresel veri analizlerinde eksik ve düzensiz verilerin etkisini en aza indirmede kritik bir rol oynamıştır.

Derin öğrenme modellerinin performans karşılaştırmasında, Tekrarlayan Sinir Ağı (Recurrent Neural Networks - RNN) modeli, kısa ve uzun vadeli bağımlılıkları öğrenme kapasitesi sayesinde diğer modellere üstünlük sağlamıştır. PM10 ve SO2 tahminlerinde R^2 puanı 0,97 olarak hesaplanmış ve RNN modeli en düşük hata oranlarına ulaşarak en başarılı tahmin sonuçlarını sunmuştur. Uzun Kısa Süreli Bellek (Long Short-Term Memory - LSTM) modeli, uzun vadeli bağımlılıkları öğrenme yeteneğiyle RNN'ye yakın bir performans göstermiş olup ancak MAE ve RMSE değerleri açısından daha yüksek hata oranlarına sahip olmuştur. Evrimsel Sinir Ağı (Convolutional Neural Networks-CNN) modeli ise kısa vadeli bağımlılıkları öğrenmede etkili olmasına rağmen, zaman serisi verilerindeki uzun vadeli bağımlılıkları modelleme konusunda yetersiz kalmıştır. Bu durum, CNN'nin hava kalitesi tahmininde diğer modellere göre daha düşük bir performans sergilemesine yol açmıştır.

Çalışmada kullanılan görselleştirme teknikleri, elde edilen bulguların açık ve anlaşılır bir şekilde sunulmasında etkili bir araç olmuştur. Scatter plot, box plot ve violin plot gibi grafiksel yöntemler, hava kalitesindeki zaman içerisindeki değişimlerin ve modellerin tahmin doğruluğunun yorumlanmasını kolaylaştırmıştır. Özellikle point scatter grafikleri, hava kirleticilerinin tarihsel değişimlerini ve dağılımlarını en etkili şekilde görselleştirmiştir. Bu teknikler, çevresel verilerdeki trendlerin ve anomalilerin tespit edilmesinde karar vericilere anlamlı içgörüler sunmuştur.

RNN modeli ile yapılan üç yıllık tahminler, PM10 seviyelerinde belirgin dalgalanmalar öngörürken, SO2 seviyelerinin daha stabil bir seyir izleyeceğini göstermiştir. Ancak, modelin bazı karmaşık olayları öngörmedeki sınırlamaları, daha geniş veri setleriyle eğitim yapılması ve meteorolojik değişkenlerin modele entegre edilmesi gerektiğini ortaya koymuştur. Bu tür iyileştirmeler, modelin doğruluğunu ve genellenebilirliğini artırabilecektir.

Sonuç olarak, elde edilen bulgular, eksik veri tamamlama, veri gürültüsünü azaltma ve gelişmiş görselleştirme yöntemlerinin hava kalitesi analizlerinde daha güvenilir ve anlamlı sonuçlar

sunduğunu göstermektedir. Grafiksel görselleştirme teknikleri, analiz sonuçlarının daha anlaşılır bir şekilde sunulmasını sağlamış ve karmaşık verilerin yorumlanmasını kolaylaştırmıştır. Bu durum, çevresel veri analizlerinde veri işleme ve görselleştirme tekniklerinin önemini bir kez daha ortaya koymuştur.



6. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında, hava kalitesi tahmininde derin öğrenme modellerinin etkinliği değerlendirilmiş ve hava kirleticilerinin zaman içerisindeki değişimlerinin öngörülmesine yönelik önemli sonuçlar elde edilmiştir. Çalışmada kullanılan Tekrarlayan Sinir Ağı (Recurrent Neural Networks - RNN) modeli, PM10 ve SO2 tahminlerinde en yüksek doğruluk oranını sağlayarak hava kalitesi tahmini için en uygun yöntem olarak belirlenmiştir. RNN'nin, kısa ve uzun vadeli bağımlılıkları etkili bir şekilde öğrenme kapasitesi, özellikle zaman serisi verilerindeki ardışık ilişkileri modelleme ve geçmiş bilgilere dayanarak tahmin yapabilme yeteneği, bu üstün performansı sergilemesinde belirleyici olmuştur. Ayrıca, RNN'nin düşük MAE ve RMSE değerleri ile yüksek R^2 puanlarına ulaşması, modelin tahmin performansını artıran önemli bir etken olmuştur. Bu metrikler, RNN'nin tahminlerde hata oranını minimumda tuttuğunu ve veri setinin değişkenliğini büyük ölçüde açıklayabildiğini göstermektedir. Bu özellikler, RNN'yi hava kalitesi tahmini için güçlü ve güvenilir bir model haline getirmiştir.

Eksik veri doldurma ve gürültü azaltma gibi ön işleme teknikleri, veri setinin bütünlüğünü sağlamış, analiz edilebilirliğini artırmış ve tahmin doğruluğunu olumlu yönde etkilemiştir. K-En Yakın Komşu (K-NN) algoritması, eksik verilerin etkili bir şekilde tamamlanmasını sağlamış, Gaussian filtreleme ise veri gürültüsünü azaltarak zaman serisi verilerinin daha stabil bir yapıya kavuşmasını mümkün kılmıştır. Ayrıca, kullanılan görselleştirme yöntemleri, hava kalitesindeki eğilimlerin ve dağılımların daha net bir şekilde anlaşılmasını sağlamış ve bulguların daha anlaşılır bir şekilde sunulmasına katkıda bulunmuştur.

Bununla birlikte, modellerin özellikle karmaşık ve ani olayları öngörme konusunda sınırlı kaldığı gözlemlenmiştir. Bu durum, gelecekteki çalışmalarda model performansını iyileştirmek ve tahmin doğruluğunu artırmak amacıyla ek çalışmalar yapılması gerektiğini göstermektedir. Çalışmadan elde edilen bulgular ışığında, aşağıdaki öneriler sunulmuştur:

- Meteorolojik faktörlerin, özellikle rüzgar hızı, sıcaklık ve nem gibi değişkenlerin modele dahil edilmesi, hava kalitesi tahminlerinin doğruluğunu artırabileceği tahmin edilmektedir.
- Hibrit model yaklaşımlarının araştırılması, farklı derin öğrenme modellerinin güçlü yönlerinden faydalanılarak daha karmaşık olayların tahmin edilmesine olanak sağlayabileceği düşünülmektedir.
- Çok hedefli tahmin modelleri geliştirilerek, birden fazla hava kirleticisinin eş zamanlı tahmininin yapılabilmesi ve daha kapsamlı analizler gerçekleştirilebileceği düşünülmektedir.
- Daha geniş zaman dilimlerini kapsayan veri setleri kullanılarak modellerin eğitilmesi, tahminlerin güvenilirliğini artırabileceği düşünülmektedir.

Sonuç olarak, bu çalışma, hava kalitesi tahmininde derin öğrenme modellerinin güçlü bir araç olduğunu göstermiştir. RNN'nin kısa ve uzun vadeli bağımlılıkları öğrenme kapasitesi, düşük hata oranları ve yüksek doğruluk oranları, bu modeli hava kalitesi tahmini için en uygun seçenek haline getirmiştir. Gelecekte yapılacak çalışmalar, mevcut sınırlamaları aşmak ve tahmin doğruluğunu artırmak için daha gelişmiş yöntemler ve modelleme tekniklerini kapsamalıdır. Bu bağlamda, çalışma, çevresel veri analizi ve hava kalitesi yönetimi için önemli bir temel oluşturmaktadır.



KAYNAKLAR

- [1] R. Kaur and P. Pandey, "Air pollution, climate change, and human health in Indian cities: A brief review," *Frontiers in Sustainable Cities*, vol. 3, p. 705131, 2021. doi: 10.3389/frsc.2021.705131.
- [2] M. B. Khan, S. Setu, N. Sultana, S. Gautam, B. A. Begum, and M. A. Salam, "Street dust in the largest urban agglomeration: Pollution characteristics, source apportionment and health risk assessment of potentially toxic trace elements," *Stochastic Environmental Research and Risk Assessment*, pp. 1–20, 2023. doi: 10.1007/s00477-023-02432-1.
- [3] O. Isinkaralar, K. Isinkaralar, and T. N. T. Nguyen, "Spatial distribution, pollution level, and human health risk assessment of heavy metals in urban street dust at neighbourhood scale," *International Journal of Biometeorology*, pp. 1–13, 2024. doi: 10.1007/s00484-024-02729-y.
- [4] Y. Liu, T. Jin, S. Yu, and H. Chu, "Pollution characteristics and health risks of heavy metals in road dust in Ma'anshan, China," *Environmental Science and Pollution Research*, vol. 30, no. 15, pp. 43726–43739, 2023. doi: 10.1007/s11356-023-25303-2.
- [5] A. Lai, "Analysis of air pollution from vehicle emissions for the contiguous United States," *Journal of Geovisualization and Spatial Analysis*, 2022.
- [6] G. D. Oreggioni, O. Mahiques, F. Monforti-Ferrario, E. Schaaf, and M. Muntean, "The impacts of technological changes and regulatory frameworks on global air pollutant emissions from the energy industry and road transport," *Energy Policy*, vol. 168, 2022. doi: 10.1016/j.enpol.2022.113021.
- [7] T. Li, Y. Yu, and Z. Sun, "A comprehensive understanding of ambient particulate matter and its components on the adverse health effects based from epidemiological and laboratory evidence," *Particle and Fibre Toxicology*, vol. 19, p. 67, 2022. doi: 10.1186/s12989-022-00507-5.
- [8] A. Garcia, E. Santa-Helena, and A. De Falco, "Toxicological effects of fine particulate matter (PM_{2.5}): Health risks and associated systemic injuries—Systematic Review," *Water Air and Soil Pollution*, vol. 234, p. 346, 2023. doi: 10.1007/s11270-023-06278-9.
- [9] A. Unnikrishnan and S. Rajeswari, "Forecasting daily air quality index and early warning system for estimating ambient air pollution on road networks using Gaussian models," *Tehnički Glasnik*, vol. 18, no. 4, pp. 549–559, 2024.
- [10] P. Perez, C. Menares, and C. Ramírez, "PM_{2.5} forecasting in Coyhaique, the most polluted city in the Americas," *Urban Clim.*, vol. 32, p. 100608, 2020, doi: 10.1016/j.uclim.2020.100608.
- [11] A. Aggarwal and D. Toshniwal, "A hybrid deep learning framework for urban air quality forecasting," *J. Clean. Prod.*, vol. 329, Art. no. 129660, 2021, doi: 10.1016/j.jclepro.2021.129660.
- [12] C.-J. Huang and P.-H. Kuo, "A Deep CNN-LSTM Model for Particulate Matter (PM_{2.5}) Forecasting in Smart Cities," *Sensors*, vol. 18, no. 7, Art. no. 7, Jul. 2018, doi: 10.3390/s18072220.
- [13] R. Janarthanan, P. Partheeban, K. Somasundaram, and P. Navin Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," *Sustainable Cities and Society*, vol. 67, p. 102720, Apr. 2021, doi: 10.1016/j.scs.2021.102720.
- [14] Q. Wu and H. Lin, "A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors," *Science of The Total Environment*, vol. 683, pp. 808–821, Sep. 2019, doi: 10.1016/j.scitotenv.2019.05.288.
- [15] B. Zhang et al., "Air pollutant diffusion trend prediction based on deep learning for targeted season—North China as an example," *Expert Systems with Applications*, vol. 232, p. 120718, Dec. 2023, doi: 10.1016/j.eswa.2023.120718.
- [16] J. Luo and Y. Gong, "Air pollutant prediction based on ARIMA-WOA-LSTM model," *Atmospheric Pollution Research*, vol. 14, no. 6, p. 101761, Jun. 2023, doi: 10.1016/j.apr.2023.101761.
- [17] G. I. Drewil and R. J. Al-Bahadili, "Air pollution prediction using LSTM deep learning and metaheuristics algorithms," *Measurement: Sensors*, vol. 24, p. 100546, Dec. 2022, doi: 10.1016/j.measen.2022.100546.

- [18] Z. Wu, Y. Tian, M. Li, B. Wang, Y. Quan, and J. Liu, "Prediction of air pollutant concentrations based on the long short-term memory neural network," *Journal of Hazardous Materials*, vol. 465, p. 133099, Mar. 2024, doi: 10.1016/j.jhazmat.2023.133099.
- [19] B. Das, Ö. O. Dursun, and S. Toraman, "Prediction of air pollutants for air quality using deep learning methods in a metropolitan city," *Urban Climate*, vol. 46, p. 101291, Dec. 2022, doi: 10.1016/j.uclim.2022.101291.
- [20] "Prediction of SO₂ and PM₁₀ air pollutants using a deep learning-based recurrent neural network: Case of industrial city Sakarya | Elsevier Enhanced Reader." Accessed: Dec. 05, 2022. [Online]. Available: <https://reader.elsevier.com/reader/sd/pii/S2212095521002819?token=4F8C7E774AFEF62F87DEEDFCC66D0CD0E6760E1DD3BA1655B309B6F55BDB090842B65A1488055ED705E638D6699E9847&originRegion=eu-west-1&originCreation=20221203124930>
- [21] T. D. Akinosho, M. Bilal, E. T. Hayes, A. Ajayi, A. Ahmed, and Z. Khan, "Deep learning-based multi-target regression for traffic-related air pollution forecasting," *Machine Learning with Applications*, vol. 12, p. 100474, Jun. 2023, doi: 10.1016/j.mlwa.2023.100474.
- [22] H. Yang, Z. Liu, and G. Li, "A new hybrid optimization prediction model for PM_{2.5} concentration considering other air pollutants and meteorological conditions," *Chemosphere*, vol. 307, p. 135798, Nov. 2022, doi: 10.1016/j.chemosphere.2022.135798.
- [23] J. González-Pardo, S. Ceballos-Santos, R. Manzananas, M. Santibáñez, and I. Fernández-Olmo, "Estimating changes in air pollutant levels due to COVID-19 lockdown measures based on a business-as-usual prediction scenario using data mining models: A case-study for urban traffic sites in Spain," *Science of The Total Environment*, vol. 823, p. 153786, Jun. 2022, doi: 10.1016/j.scitotenv.2022.153786.
- [24] Y.-C. Lin, Y.-T. Lin, C.-R. Chen, C.-Y. Lai, and Y.-B. Wang, "Meteorological and traffic effects on air pollutants using Bayesian networks and deep learning," *Journal of Environmental Sciences*, Feb. 2024, doi: 10.1016/j.jes.2024.01.057.
- [25] I. H. Fong, T. Li, S. Fong, R. K. Wong, and A. J. Tallón-Ballesteros, "Predicting concentration levels of air pollutants by transfer learning and recurrent neural network," *Knowledge-Based Systems*, vol. 192, p. 105622, Mar. 2020, doi: 10.1016/j.knosys.2020.105622.
- [26] H. Dai, G. Huang, J. Wang, H. Zeng, and F. Zhou, "Prediction of Air Pollutant Concentration Based on One-Dimensional Multi-Scale CNN-LSTM Considering Spatial-Temporal Characteristics: A Case Study of Xi'an, China," *Atmosphere*, vol. 12, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/atmos12121626.
- [27] I. Jairi, S. Ben-Othman, L. Canivet, and H. Zgaya-Biau, "Enhancing air pollution prediction: A neural transfer learning approach across different air pollutants," *Environmental Technology & Innovation*, vol. 36, p. 103793, Nov. 2024, doi: [10.1016/j.eti.2024.103793](https://doi.org/10.1016/j.eti.2024.103793).
- [28] P. A. Rani and Dr. V. Sampathkumar, "A novel artificial intelligence algorithm for predicting air quality by analysing the pollutant levels in air quality data in tamilnadu," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 5, p. 100234, Sep. 2023, doi: 10.1016/j.prime.2023.100234.
- [29] S. Masmoudi, H. Elghazel, D. Taieb, O. Yazar, and A. Kallel, "A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection," *Science of The Total Environment*, vol. 715, p. 136991, May 2020, doi: 10.1016/j.scitotenv.2020.136991.
- [30] A. Kshirsagar and M. Shah, "Anatomization of air quality prediction using neural networks, regression and hybrid models," *Journal of Cleaner Production*, vol. 369, p. 133383, Oct. 2022, doi: 10.1016/j.jclepro.2022.133383.
- [31] Y. Dokuz, A. Bozdağ, and B. Gökçek, "Hava Kalitesi Parametrelerinin Tahmini ve Mekansal Dağılımı İçin Makine Öğrenmesi Yöntemlerinin Kullanılması," *Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, Jan. 2020, doi: [10.28948/ngumuh.654092](https://doi.org/10.28948/ngumuh.654092).
- [32] S. Ünalı and N. Yalçın, "Hava Kirliliğinin Makine Öğrenmesi Tabanlı Tahmini: Başakşehir Örneği Prediction of Air Pollution based on Machine Learning Methods: A Case Study for Başakşehir, İstanbul," p. 10.
- [33] "Analysis of the summer thermal comfort indices in İstanbul | International Journal of Biometeorology." Accessed: May 10, 2024. [Online]. Available: <https://link.springer.com/article/10.1007/s00484-024-02669-7>

- [34] B. Liu, M. Lai, P. Zeng, and J. Chen, "Air pollutant prediction based on a attention mechanism model of the Yangtze River Delta region in frequent heatwaves," *Atmospheric Research*, vol. 311, p. 107701, Dec. 2024, doi: [10.1016/j.atmosres.2024.107701](https://doi.org/10.1016/j.atmosres.2024.107701).
- [35] M. Yılmaz, Y. Kara, H. C. Çulpan, G. Can, and H. Toros, "Detection and regional analysis of heatwave characteristics in İstanbul," *Sustainable Cities and Society*, vol. 97, p. 104789, Oct. 2023, doi: [10.1016/j.scs.2023.104789](https://doi.org/10.1016/j.scs.2023.104789).
- [36] P. Aksak, Ş. K. Öztürk, and Ö. Ünsal, "Kentsel Isı Adasının İklim Parametreleri ve Uzaktan Algılama Üzerinden İncelenmesi: İstanbul Kenti Örneği," *ECD*, vol. 32, no. 1, Art. no. 1, Jun. 2023, doi: [10.51800/ecd.1266060](https://doi.org/10.51800/ecd.1266060).
- [37] X. Shi, Z. Huang, Y. Dai, W. Du, and J. Cheng, "Evaluating emission reduction potential and co-benefits of CO₂ and air pollutants from mobile sources: A case study in Shanghai, China," *Resources, Conservation and Recycling*, vol. 202, p. 107347, Mar. 2024, doi: [10.1016/j.resconrec.2023.107347](https://doi.org/10.1016/j.resconrec.2023.107347).
- [38] S. Jochner, I. Markevych, I. Beck, C. Traidl-Hoffmann, J. Heinrich, and A. Menzel, "The effects of short- and long-term air pollutants on plant phenology and leaf characteristics," *Environmental Pollution*, vol. 206, pp. 382–389, Nov. 2015, doi: [10.1016/j.envpol.2015.07.040](https://doi.org/10.1016/j.envpol.2015.07.040).
- [39] S. A. Ajayi, C. A. Adams, G. Dumedah, A. O. Adebajji, and W. Ackaah, "The impact of traffic mobility measures on vehicle emissions for heterogeneous traffic in Lagos City," *Scientific African*, vol. 21, p. e01822, Sep. 2023, doi: [10.1016/j.sciaf.2023.e01822](https://doi.org/10.1016/j.sciaf.2023.e01822).
- [40] J. Yang, L. Shi, J. Lee, and I. Ryu, "Spatiotemporal prediction of particulate matter concentration based on traffic and meteorological data," *Transportation Research Part D: Transport and Environment*, vol. 127, p. 104070, Feb. 2024, doi: [10.1016/j.trd.2024.104070](https://doi.org/10.1016/j.trd.2024.104070).
- [41] B. Zhang et al., "Deep learning for air pollutant concentration prediction: A review," *Atmospheric Environment*, vol. 290, p. 119347, Dec. 2022, doi: [10.1016/j.atmosenv.2022.119347](https://doi.org/10.1016/j.atmosenv.2022.119347).
- [42] S. Arslankaya and Ş. Toprak, "Using Machine Learning and Deep Learning Algorithms for Stock Price Prediction," *Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi*, vol. 13, no. 1, pp. 178–192, Jan. 2021, doi: [10.29137/umagd.771671](https://doi.org/10.29137/umagd.771671).
- [43] Ş. Erken and L. Şenyay, "Makine Öğrenmesi İle Eksik Veri Tamamlama Yöntemlerinin Sınıflandırma Performansına Etkileri," *KAYÜSOSDER*, vol. 5, no. 1, Art. no. 1, Jun. 2023, doi: [10.51177/kayusosder.1307226](https://doi.org/10.51177/kayusosder.1307226).
- [44] SIM Air Quality- Station Data Download Continuous Monitoring Center. https://sim.csb.gov.tr/STN/STN_Report/StationDataDownloadNew (202 0) (accessed 1 November 2020)
- [45] V. Çetin and O. Yıldız, "Veri analizinde veri ön işleme teknikleri üzerine kapsamlı bir inceleme," *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 28, no. 3, pp. 123–135, 2022, doi: [10.5505/pajes.2022.62687](https://doi.org/10.5505/pajes.2022.62687).
- [46] M. Qin, Z. Du, F. Zhang, and R. Liu, "A matrix completion-based multiview learning method for imputing missing values in buoy monitoring data," *Information Sciences*, vol. 487, pp. 18–30, Jun. 2019, doi: [10.1016/j.ins.2019.02.057](https://doi.org/10.1016/j.ins.2019.02.057).
- [47] H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li, "A tensor-based method for missing traffic data completion," *Transportation Research Part C: Emerging Technologies*, vol. 28, pp. 15–27, Mar. 2013, doi: [10.1016/j.trc.2012.12.007](https://doi.org/10.1016/j.trc.2012.12.007).
- [48] L. Csató, "How to choose a completion method for pairwise comparison matrices with missing entries: An axiomatic result," *International Journal of Approximate Reasoning*, vol. 164, p. 109063, Jan. 2024, doi: [10.1016/j.ijar.2023.109063](https://doi.org/10.1016/j.ijar.2023.109063).
- [49] J. Rhyu, D. J. Lee, and M. G. Kang, "Automated outlier detection and estimation of missing data," *Computers & Chemical Engineering*, vol. 180, p. 108448, Jan. 2024, doi: [10.1016/j.compchemeng.2023.108448](https://doi.org/10.1016/j.compchemeng.2023.108448).
- [50] M. Abboud, "Leveraging machine learning for multi-source data enrichment and analytics in air quality monitoring and crowd sensing".
- [51] B. Çengiz and R. Daş, "Veri Füzyonu Veri Kaynakları, Mimariler, Zorluklar ve Çözüm Yaklaşımları," *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 34, no. 2, Art. no. 2, Sep. 2022.

- [52] S. Aşık, E. Deniz, and E. Çınar, "Otonom Taşıyıcı Araçlardaki Hataların Teşhisi için Evrişimli Sinir Ağları Tabanlı Çoklu Heterojen Sensörlerin Füzyon Yöntemi," *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 34, no. 1, Art. no. 1, Mar. 2022, doi: [10.35234/fumbd.1003341](https://doi.org/10.35234/fumbd.1003341).
- [53] A. Casolaro, V. Capone, G. Iannuzzo, and F. Camastra, "Deep Learning for Time Series Forecasting: Advances and Open Problems," *Information*, vol. 14, no. 11, Art. no. 11, Nov. 2023, doi: [10.3390/info14110598](https://doi.org/10.3390/info14110598).
- [54] S. Çakıcı and İ. Şahin, "FPGA Tabanlı Ortalama Görüntü Filtresi Tasarımı," *İleri Teknoloji Bilimleri Dergisi*, vol. 2, no. 2, Art. no. 2, Jun. 2013.
- [55] A. Değirmenci, İ. Çankaya, and R. Demirci, "Gradyan Anahtarlamalı Gauss Görüntü Filtresi," *DÜBİTED*, vol. 6, no. 1, Art. no. 1, Jan. 2018, doi: [10.29130/dubited.345116](https://doi.org/10.29130/dubited.345116).
- [56] C. Küpeli and F. Bulut, "Görüntüdeki Tuz Biber ve Gauss Gürültülerine Karşı Filtrelerin Performans Analizleri," *Temel Bilimler Genel*, vol. 3, no. 2, Art. no. 2, Sep. 2020, doi: [10.46373/hafebid.768240](https://doi.org/10.46373/hafebid.768240).
- [57] R. Abdelkader, M. Khorchef, and M. Zergoug, "Fault Diagnosis Based on Enhancement of Barkhausen Noise Using Hybrid Method Emprical Mode Decomposition-Savitzky-Golay Filter," *ACTAMAT*, vol. 4, no. 2, Art. no. 2, Jun. 2020.
- [58] Çelik, S., & Akdamar, E. (2018). Büyük Veri ve Veri Görselleştirme. *International Journal of Informatics Technologies (IJIT)*, vol. 65, pp. 253–264, Mar. 2018.
- [59] S. He, X. Ye, T. Sakurai, and Q. Zou, "MRMD3.0: A Python Tool and Webserver for Dimensionality Reduction and Data Visualization via an Ensemble Strategy," *Journal of Molecular Biology*, vol. 435, no. 14, p. 168116, Jul. 2023, doi: [10.1016/j.jmb.2023.168116](https://doi.org/10.1016/j.jmb.2023.168116).
- [60] T. Árpádfy-Lovas and N. Nagy, "ActionPytential: An open source tool for analyzing and visualizing cardiac action potential data," *Heliyon*, vol. 9, no. 3, p. e14440, Mar. 2023, doi: [10.1016/j.heliyon.2023.e14440](https://doi.org/10.1016/j.heliyon.2023.e14440).
- [61] J. Drechsel, N. S. Khan, and A. Rovere, "PALEO-SEAL: A tool for the visualization and sharing of Holocene sea-level data," *Quaternary Science Reviews*, vol. 259, p. 106884, May 2021, doi: [10.1016/j.quascirev.2021.106884](https://doi.org/10.1016/j.quascirev.2021.106884).
- [62] C. Balzer, R. Oktavian, M. Zandi, D. Fairen-Jimenez, and P. Z. Moghadam, "Wiz: A Web-Based Tool for Interactive Visualization of Big Data," *Patterns*, vol. 1, no. 8, p. 100107, Nov. 2020, doi: [10.1016/j.patter.2020.100107](https://doi.org/10.1016/j.patter.2020.100107).
- [63] R. Warham, "MT20 Data Visualization As a Tool to Quickly Understand Patient-Generated Data," *Value in Health*, vol. 26, no. 6, Supplement, p. S301, Jun. 2023, doi: [10.1016/j.jval.2023.03.1667](https://doi.org/10.1016/j.jval.2023.03.1667).
- [64] R. Armstrong-Moore, M. White, and T. Burgoine, "Stakeholder experiences of using online spatial data visualisation tools for local public health decision support: A qualitative study," *Health & Place*, vol. 71, p. 102648, Sep. 2021, doi: [10.1016/j.healthplace.2021.102648](https://doi.org/10.1016/j.healthplace.2021.102648).
- [65] S. Eken, "Medical Data Analysis for Different Data Types," *International Journal of Computational and Experimental Science and Engineering*, vol. 6, no. 3, pp. 138–144, Nov. 2020, doi: [10.22399/ijcesen.780174](https://doi.org/10.22399/ijcesen.780174).
- [66] S. Eken, "Büyük Verinin İnteraktif Görselleştirilmesi: Tableau Üzerine Öğrenci Deneyimleri," *European Journal of Science and Technology*, pp. 262–271, Apr. 2020, doi: [10.31590/ejosat.659823](https://doi.org/10.31590/ejosat.659823).
- [67] R. Rabie, M. Asghari, H. Nosrati, M. Emami Niri, and S. Karimi, "Spatially resolved air quality index prediction in megacities with a CNN-Bi-LSTM hybrid framework," *Sustainable Cities and Society*, vol. 109, p. 105537, Aug. 2024, doi: [10.1016/j.scs.2024.105537](https://doi.org/10.1016/j.scs.2024.105537).
- [68] E. Güler, A. G. M. T. Kakız, F. B. Gunay, B. Şanal, and T. Çavdar, "Kapalı Mekân Ortamında 1D-CNN Kullanarak Yapılan Doluluk Tespiti Sınıflandırması," *KFBD*, vol. 13, no. 1, Art. no. 1, Mar. 2023, doi: [10.31466/kfbd.1162332](https://doi.org/10.31466/kfbd.1162332).
- [69] D. Saravanan and K. Santhosh Kumar, "Improving air pollution detection accuracy and quality monitoring based on bidirectional RNN and the Internet of Things," *Materials Today: Proceedings*, vol. 81, pp. 791–796, Jan. 2023, doi: [10.1016/j.matpr.2021.04.239](https://doi.org/10.1016/j.matpr.2021.04.239).

- [70] S. Tanışman, A. A. Karcıoğlu, A. Ugur, and H. Bulut, “LSTM Sinir Ağı ve ARIMA Zaman Serisi Modelleri Kullanılarak Bitcoin Fiyatının Tahminlenmesi ve Yöntemlerin Karşılaştırılması,” *EJOSAT*, no. 32, Art. no. 32, Dec. 2021, doi: [10.31590/ejosat.1039890](https://doi.org/10.31590/ejosat.1039890).
- [71] A. A. Süzen, Z. Yıldız, and T. Yılmaz, “LSTM tabanlı Derin Sinir Ağı ile Ayak Taban Basınç Verilerinden VKİ Durumlarının Sınıflandırılması,” *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 8, no. 4, pp. 1392–1398, Dec. 2019, doi: [10.17798/bitlisfen.540273](https://doi.org/10.17798/bitlisfen.540273).
- [72] Akbulut, İ., & Özcan, B. (2022). Hava Kirliliği Tahmini: Yapay Sinir Ağları ve Regresyon Yöntemleriyle Bir Karşılaştırma (Air Pollution Forecast: A Comparison with Artificial Neural Networks and Regression Methods). *DergiPark: Journal of Environmental Science*, 11(2), 45-56. <https://doi.org/10.xxxx/journal.identifier>
- [73] SIM Air Quality- Station Data Download Continuous Monitoring Center. https://sim.csb.gov.tr/STN/STN_Report/StationDataDownloadNew (2020) (accessed 19 October 2022)
- [74] Aladağ, E. (2021). Ampirik Mod Ayırıştırmasına Dayalı ARIMA Modeli Kullanılarak Van İli Hava Kirliliğinin Tahmini. *DergiPark: Journal of the Institute of Science and Technology*, 11(3), 1405-1418. <https://doi.org/10.21597/jist.973244>
- [75] B. Cui, M. Liu, S. Li, Z. Jin, Y. Zeng, and X. Lin, “Deep learning methods for atmospheric PM2.5 prediction: A comparative study of transformer and CNN-LSTM-attention,” *Atmospheric Pollution Research*, vol. 14, no. 9, p. 101833, Sep. 2023, doi: [10.1016/j.apr.2023.101833](https://doi.org/10.1016/j.apr.2023.101833).
- [76] D. Mengüş and B. Daş, “Actionable Data Visualization for Air Quality Data in the Istanbul Location,” *Balkan Journal of Electrical and Computer Engineering*, vol. 10, no. 4, Art. no. 4, Oct. 2022, doi: [10.17694/bajece.1180676](https://doi.org/10.17694/bajece.1180676).
- [77] SIM Air Quality- Station Data Download Continuous Monitoring Center. https://sim.csb.gov.tr/STN/STN_Report/StationDataDownloadNew (2020) (accessed 2 March 2024)

ÖZGEÇMİŞ

Damla MENGÜŞ

[REDACTED]

[REDACTED] [REDACTED]
[REDACTED] [REDACTED]
[REDACTED] [REDACTED]
[REDACTED] [REDACTED]
[REDACTED] [REDACTED]
[REDACTED] [REDACTED]

[REDACTED]

[REDACTED] [REDACTED]
[REDACTED] [REDACTED]

[REDACTED]

[REDACTED] [REDACTED]
[REDACTED] [REDACTED]

[REDACTED]

[REDACTED] [REDACTED]
[REDACTED] [REDACTED]

[REDACTED]

[REDACTED] [REDACTED]
[REDACTED] [REDACTED]
[REDACTED] [REDACTED]

[REDACTED]

[REDACTED] [REDACTED]
[REDACTED] [REDACTED]
[REDACTED] [REDACTED]

AKADEMİK FAALİYETLER

Makaleler:

1. D. Mengüş and B. Daş, “Actionable Data Visualization for Air Quality Data in the Istanbul Location,” *Balkan Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 397–401, Oct. 2022, doi: 10.17694/bajece.1180676.
2. D. Mengüş and B. Daş, “Comparative Analysis of Data Visualization and Deep Learning Models in Air Quality Forecasting,” *Sakarya University Journal of Computer and Information Sciences*, 2024 (Under Review)