



**MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE HAVA KİRLİLİĞİ TAHMİNİ:
MANİSA VE ZONGULDAK ÖRNEĞİ**

Büşra DUYGU ÇELİK

**YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

HAZİRAN 2022

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Büşra DUYGU ÇELİK

28/06/2022

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE HAVA KİRLİLİĞİ TAHMİNİ: MANİSA VE ZONGULDAK ÖRNEĞİ

(Yüksek Lisans Tezi)

Büşra DUYGU ÇELİK

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Haziran 2022

ÖZET

Dünyadaki artan nüfusla beraber meydana gelen enerji kullanımındaki artış, şehirleşme ve ısınma aktiviteleri sonucunda ortaya çıkan hava kirliliği her geçen gün artarak büyük bir çevre sorunu haline gelmiştir. Hava kirliliği problemi, atmosferde bulunan hava bileşiklerinin değişmesine, asit yağmurlarına, ozon tabakasının incelmesine, doğada bulunan tüm canlılar üzerindeki olumsuz etkilere sebep olarak ekosisteme ciddi zarar vermektedir. Hava kirliliğinin önceden bir model oluşturularak değerlendirilmesi, oluşan kirliliğe karşı önlemler almakta ve hava kalitesinin iyileştirme çözüm süreçlerini kolaylaştırmakta büyük önem taşımaktadır. Hava kirliliğinin birçok parametreden etkilenmesi, her parametrenin ölçülememesi ve modellemeler sonunda elde edilen yüksek doğruluk bu konuda makine öğrenmesi tekniklerinin kullanılmasının temel nedenidir. Bu tez çalışmasında Zonguldak ve Manisa iline ait 2016-2021 yılları arasında ölçülen PM_{10} , SO_2 , CO , NO_2 , NO_x , O_3 kirlleticileri ile Sıcaklık, Nem, Yağış, Rüzgâr Hızı ve Basınç meteoroloji parametrelerinden oluşan bir veri seti kullanılmıştır. Veri seti literatürde iyi bilinen Destek Vektör Makinaları, Karar Ağaçları, Naif Bayes, Rastgele Orman, K-En Yakın Komşu ve Uzun Kısa Süreli Bellek yöntemleriyle modellenmiş ve Kök Ortalama Kare Hatası (KOKH), Ortalama Mutlak Hata (OMH), ve Açıklayıcılık Katsayısı (R^2) ile modellerin başarısı karşılaştırılmıştır. İki istasyon için LSTM %87 değeri ile diğer yöntemlerden daha iyi sonuç verdiği görülmüştür. En düşük performansı ise %36 değer ile Naif Bayes olmuştur.

Bilim Kodu : 92432
Anahtar Kelimeler : Makine öğrenmesi, hava kirliliği, hava kirliliği tahmini
Sayfa Adedi : 81
Danışman : Prof. Dr. Nursal ARICI

AIR POLLUTION PREDICTION WITH MACHINE LEARNING METHODS: MANISA AND ZONGULDAK EXAMPLE

(M. Sc. Thesis)

Büşra DUYGU ÇELİK

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

June 2022

ABSTRACT

With the increase in energy use, urbanization, and heating activities that occur with the increasing world population, air pollution has become a major environmental problem, increasing day by day. The problem of air pollution causes serious damage to the ecosystem by causing changes in air compounds in the atmosphere, acid rain, depletion of the ozone layer, and negative effects on all living creatures. Estimating air pollution using a model is crucial in taking measures for possible pollution and facilitating air quality improvement processes. The fact that air pollution is dependent on many parameters, the inability to measure all parameters and the high accuracy obtained at the end of the modelling are the main reasons for using machine learning techniques in this regard. In this thesis, a data set consisting of PM₁₀, SO₂, CO, NO₂, NO_x, O₃ pollutants and meteorological parameters of Temperature, Humidity, Precipitation, Wind Speed and Pressure measured between 2016-2021 in Zonguldak and Manisa provinces are used. The data set was modelled with Support Vector Machines, Decision Trees, Naive Bayes, Random Forest, K-Nearest Neighbor and Long Short-Term Memory methods, which are well known in the literature. Model performances measured and compared using accuracy, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Explanatory Coefficient (R^2) metrics. The LSTM, gives better results than other methods with a 0.87 accuracy value for both stations. The lowest performance is Naive Bayes with an accuracy value of 0.36.

Science Code : 92432

Key Words : Machine learning, air pollution, air pollution prediction

Page Number : 81

Supervisor : Prof. Dr. Nursal ARICI

TEŞEKKÜR

Yüksek lisans eğitimi süresince iletişimi hiç koparmadan huzurlu bir çalışma ortamı sağlayan, desteğini esirgemeyen, ilerlemekte zorlandığım anlarda yönlendirmeleriyle ufkumu açan, değerli bilgi ve tecrübeleriyle akademik gelişimime ve çalışmama katkı sağlayan danışman hocam (Akademik Annem) Prof. Dr. Nursal ARICI' ya en içten teşekkürlerimi sunarım.

Eğitim hayatım boyunca maddi manevi desteğini sürekli arkamda hissettiğim canım aileme ve özellikle annem Muazzez ÇELİK'e, yoğun çalışmalarım boyunca sabır ve anlayış göstererek her daim bana moral, motivasyon ve ilham kaynağı olan sevgili eşim Muhammed Çağrı ÇELİK'e ve bu süreçte bana eşlik eden, iç motivasyonumda büyük katkısı olan "Canım oğluma" teşekkür ederim.

Çalışmada kullandığım kirletici ve meteorolojik verileri sağlayan Meteoroloji Genel Müdürlüğü ve Çevre Şehircilik Bakanlığı'na teşekkür ederim.

Son olarak, bu tez çalışmasını 2210-A Genel Yurt İçi Yüksek Lisans Burs Programı kapsamında destekleyen Türkiye Bilimsel ve Teknolojik Araştırma Kurumuna (TÜBİTAK) teşekkür ederim.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
ŞEKİLLERİN LİSTESİ.....	x
SİMGELER VE KISALTMALAR.....	xiii
1. GİRİŞ.....	1
2. KAVRAMSAL VE KURAMAL ÇERÇEVE.....	5
2.1. Hava Kirliliği	5
2.2. Kirleticiler	5
2.2.1. PM ₁₀	6
2.2.2. SO ₂	6
2.2.3. CO	7
2.2.4. Azot oksitler (NO _x , NO ₂).....	7
2.2.5. O ₃	7
2.3. Meteorolojik Veriler.....	8
2.3.1. Meteorolojik veriler ve kirleticiler arasındaki ilişki.....	8
2.4. Kuramsal Temeller	9
2.4.1. Makine öğrenmesi yöntemleri.....	9
2.4.2. Literatürde yapılan benzer çalışmalar	23
3. MATERYAL VE YÖNTEM	31
3.1. Veri Seti ve Özellikleri	31

	Sayfa
3.1.1. Çalışma alanı ve istasyon konumu.....	31
3.1.2. Çalışma alanına ait iklimsel özellikler	31
3.1.3. Veri seti ve temini.....	32
3.1.4. Veri setinin hazırlanması ve normalleştirilmesi	38
3.1.5. Performans değerlendirme ölçütleri	41
4. MAKİNE ÖĞRENMESİ YÖNTEMLERİ VE BULGULARI	43
4.1. Manisa-Soma İstasyon Sonuçları	43
4.1.1. RO makine öğrenmesi sonuçları	44
4.1.2. DVM makine öğrenmesi sonuçları	46
4.1.3. NB makine öğrenmesi sonuçları.....	48
4.1.4. KEYK makine öğrenmesi sonuçları.....	50
4.1.5. KA makine öğrenmesi sonuçları.....	52
4.1.6. LSTM makine öğrenmesi sonuçları	54
4.2. Zonguldak-Kuzyaka İstasyon Sonuçları.....	57
4.2.1. Çalışma alanı ve istasyon konumu.....	58
4.2.2. DVM makine öğrenmesi sonuçları	60
4.2.3. NB makine öğrenmesi sonuçları	62
4.2.4. KEYK makine öğrenmesi sonuçları.....	64
4.2.5. KA makine öğrenmesi sonuçları.....	66
4.2.6. LSTM makine öğrenmesi sonuçları	68
4. SONUÇ VE ÖNERİLER	73
KAYNAKLAR	75
ÖZGEÇMİŞ	81

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1. Ulusal hava kalite indeksi kesme noktaları.....	6
Çizelge 2.2. Hava kirliliği tahmini için makine öğrenmesi kullanan çalışmaların kıyaslanması.....	29
Çizelge 3.1. İstasyondan alınan verilerin örneği	32
Çizelge 3.2. Manisa kirletici parametrelere ait tanımlayıcı istatistikler.....	36
Çizelge 3.3. Zonguldak kirletici parametrelere ait tanımlayıcı istatistikler	37
Çizelge 3.4. Performans metrikleri	41
Çizelge 4.1. Manisa-Soma istasyonuna ait PM ₁₀ tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları	43
Çizelge 4.2. Manisa-Soma istasyonuna ait PM ₁₀ tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları	43
Çizelge 4.3. Uygulanan LSTM tahmin modeline ait parametrelerin değerleri.....	55
Çizelge 4.4. Zonguldak-Kuzyaka istasyonuna ait PM ₁₀ tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları	56
Çizelge 4.5. Zonguldak-Kuzyaka istasyonuna ait SO ₂ tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları	57
Çizelge 4.6. Uygulanan LSTM tahmin modeline ait parametrelerin değerleri.....	58

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 2.1. Rastgele Orman algoritması yapısı	11
Şekil 2.2. K-EYK algoritma yapısı	14
Şekil 2.3. Karar ağacı yapısı	15
Şekil 2.4. Derin öğrenmenin makine öğrenmesi ve yapay zeka ile ilişkisi	17
Şekil 2.5. Yapay nöron hücresinin yapısı	17
Şekil 2.6. RNN yapısı	18
Şekil 2.7. RNN yapısının açık gösterimi	19
Şekil 2.8. LSTM mimarisinin yapısı.....	20
Şekil 2.9. Unutma kapısı.....	20
Şekil 2.10. Girdi kapısı	21
Şekil 2.11. Çıktı kapısı.....	22
Şekil 2.12. Hücre durumu	23
Şekil 3.1. a, b) Zonguldak ve Manisa ili hava kalitesi verilerinin temin edildiği istasyonlar	31
Şekil 3.2. Manisa iline ait verilerin birleştirilmesi.....	33
Şekil 3.3. Manisa iline ait korelasyon matrisi.....	34
Şekil 3.4. Zonguldak iline ait korelasyon matrisi	35
Şekil 3.5. Manisa iline ait meteorolojik ve kirletici parametrelerin grafiksel gösterimi	37
Şekil 3.6. Zonguldak iline ait meteorolojik ve kirletici parametrelerin grafiksel gösterimi	38
Şekil 3.7. Zonguldak iline ait özniteliklerin formatı.....	38
Şekil 3.8. a, b, c.) Zonguldak iline ait veri setinden eksik veriler çıkarılmadan öncesi ve sonrası	39
Şekil 3.9. Zonguldak ili veri setine ait PM10, SO2, basınç ve sıcaklık parametrelerine ait kutu grafikleri.....	40

Şekil	Sayfa
Şekil 3.10. Normalize edilen verilerin örnek satır görünümü.....	41
Şekil 4.1. a) Gerçek ve tahmin PM10 değerlerin karşılaştırma grafiği b) RO kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	45
Şekil 4.2. a) Gerçek ve Tahmin SO2 Değerlerin karşılaştırma grafiği b) RO kullanılarak tahmin edilen SO2 değerlerinin dağılım grafiği	46
Şekil 4.3. a) Gerçek ve tahmin PM10 değerlerin karşılaştırma grafiği b) DVM kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	47
Şekil 4.4. a) Gerçek ve tahmin SO2 değerlerin karşılaştırma grafiği b) DVM kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	48
Şekil 4.5. a) Gerçek ve tahmin PM10 değerlerin karşılaştırma grafiği b) NB kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	79
Şekil 4.6. a) Gerçek ve tahmin SO2 değerlerin karşılaştırma grafiği b) NB kullanılarak tahmin edilen SO2 değerlerinin dağılım grafiği	50
Şekil 4.7. a) Gerçek ve Tahmin PM10 değerlerin karşılaştırma grafiği b) KEYK kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	51
Şekil 4.8. a) Gerçek ve tahmin SO2 değerlerin karşılaştırma grafiği b) KEYK kullanılarak tahmin edilen SO2 değerlerinin dağılım grafiği	52
Şekil 4.9. a) Gerçek ve Tahmin PM10 değerlerin karşılaştırma grafiği b) KA kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	53
Şekil 4.10. a) Gerçek ve tahmin SO2 değerlerin karşılaştırma grafiği b) KA kullanılarak tahmin edilen SO2 değerlerinin dağılım grafiği	54
Şekil 4.11. a) Gerçek ve tahmin PM10 değerlerin karşılaştırma grafiği b) LSTM kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	56
Şekil 4.12. a) Gerçek ve tahmin SO2 değerlerin karşılaştırma grafiği b) LSTM kullanılarak tahmin edilen SO2 değerlerinin dağılım grafiği	57
Şekil 4.13. a) Gerçek ve tahmin PM10 değerlerin karşılaştırma grafiği b) RO kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	59
Şekil 4.14. a) Gerçek ve Tahmin SO2 Değerlerin karşılaştırma grafiği b) RO kullanılarak tahmin edilen SO2 değerlerinin dağılım grafiği	60
Şekil 4.15. a) Gerçek ve tahmin PM10 değerlerin karşılaştırma grafiği b) DVM kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği.....	61

Şekil	Sayfa
Şekil 4.16. a) Gerçek ve tahmin SO ₂ değerlerin karşılaştırma grafiği b) DVM kullanılarak tahmin edilen SO ₂ değerlerinin dağılım grafiği	62
Şekil 4.17. a) Gerçek ve Tahmin PM ₁₀ Değerlerin karşılaştırma grafiği b) NB kullanılarak tahmin edilen PM ₁₀ değerlerinin dağılım grafiği.....	63
Şekil 4.18. a) Gerçek ve tahmin SO ₂ değerlerin karşılaştırma grafiği b) NB kullanılarak tahmin edilen SO ₂ değerlerinin dağılım grafiği	64
Şekil 4.19. a) Gerçek ve tahmin PM ₁₀ değerlerin karşılaştırma grafiği b) KEYK kullanılarak tahmin edilen PM ₁₀ değerlerinin dağılım grafiği.....	65
Şekil 4.20. a) Gerçek ve tahmin SO ₂ değerlerin karşılaştırma grafiği b) KEYK kullanılarak tahmin edilen SO ₂ değerlerinin dağılım grafiği	66
Şekil 4.21. a) Gerçek ve tahmin PM ₁₀ değerlerin karşılaştırma grafiği b) KA kullanılarak tahmin edilen PM ₁₀ değerlerinin dağılım grafiği.....	67
Şekil 4.22. a) Gerçek ve tahmin SO ₂ değerlerin karşılaştırma grafiği b) KA kullanılarak tahmin edilen SO ₂ değerlerinin dağılım grafiği	68
Şekil 4.23. a) Gerçek ve tahmin PM ₁₀ Değerlerin karşılaştırma grafiği b) LSTM kullanılarak tahmin edilen PM ₁₀ değerlerinin dağılım grafiği.....	70
Şekil 4.24. a) Gerçek ve tahmin SO ₂ Değerlerin karşılaştırma grafiği b) LSTM kullanılarak tahmin edilen SO ₂ değerlerinin dağılım grafiği	71

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

μg

kg

m^2

m^3

mbar

$^{\circ}\text{C}$

α

Açıklamalar

Mikrogram

Kilogram

Metrekare

Metreküp

Milibar

Santigrat

Learning Rate (Öğrenme Oranı)

Kısaltmalar

CBS

C

CNN

C_xH_y

ÇDR

ÇSB

DSÖ

DVM

DAE

GA

GRU

GSR

HKİ

IoT

K-EYK

KA

Açıklamalar

Coğrafi Bilgi Sistemi

Karbon

Convolution Neural Network

Hidrokarbonlar

Çoklu Doğrusal Regresyon

Çevre Şehircilik Bakanlığı

Dünya Sağlık Örgütü

Destek Vektör Makinesi

Deep Auto Encoder

Genetik Algoritma

Gate Recurrent Unit

Gauss Süreç Regresyonu

Hava Kalite İndeksi

Internet of Things

K-En Yakın Komşu

Karar Ağacı

Kısaltmalar**Açıklamalar**

KOKH	Kök Ortalama Kare Hatası
LSTM	Long Short Term Memory
LR	Lineer Regresyon
MAE	Mean Absolute Error
MGM	Meteoroloji Genel Müdürlüğü
MLP	Multi Layer Perceptron
NB	Naive Bayes
N	Azot
NO₂	Azot dioksit
NO_x	Nitrik oksit
O₃	Ozon
OMH	Ortalama Mutlak Hata
PM	Partikül Madde
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RO	Rastgele Orman
S	Kükürt
SO₂	Kükürt dioksit
TSA	Tekrarlayan Sinir Ağları
UKSB	Uzun Kısa Süreli Bellek
YSA	Yapay Sinir Ağları

1. GİRİŞ

Dünya üzerinde artan nüfus ile meydana gelen enerji kullanımındaki artış, sanayileşme, endüstriyel gelişim, şehirleşme ve ısınma ihtiyacında kullanılan yakıtlar hava kirliliğine sebep olarak dünyamız için büyük tehdit oluşturmaktadır. Hava kirliliği; yirmi birinci yüzyılın en yaygın sorunu olmakla birlikte atmosferde bulunan gaz, su buharı, toz ve duman şeklindeki kirleticilerin, dünya üzerinde yaşayan insanlara, hayvanlara ve bitkilere zarar meydana getirecek seviyeye yükselmesidir [1]. Hava kirliliği canlıların sağlığı üzerinde ciddi bir risk oluşturmaktadır. Hava kalitesinin artırılması amacıyla kirleticilerin sebep olduğu hava kirliliğinin değerlendirilmesi, kontrol altında tutulması ve ileriye yönelik tahminde bulunulması büyük önem taşımaktadır [2]. Havada bulunan kirleticiler özellikle çocuk ve yaşlılarda astım, bronşit, zatürre ve kalp krizi gibi kronik hastalıklara neden olur veya kötüleştirir [3]. Partikül madde (PM), kükürt dioksit (SO_2), ozon (O_3), ve azot dioksit (NO_2) toplum sağlığı üzerinde en çok zarara sebep olan kirleticilerdir. Sanayinin gelişmiş olduğu bölgelerde gerçekleşen endüstriyel faaliyetler ve petrol, kömür gibi fosil yakıtların yanması sonucunda oluşan PM_{10} ve SO_2 , özellikle solunum yolu üzerinde olumsuz etkilere sebep olmaktadır. Tüm bu etkiler göz önünde bulundurulduğunda doğadaki canlıların hayat kalitesini önemli ölçüde etkileyen hava kalitesinin ve kirletici konsantrasyonlarının önceden tahmini ve bu tahmine dayalı meydana gelebilecek zararların önlenmesi büyük önem taşımaktadır.

Bu çerçevede yapılan çalışmalarda günümüze gelene dek daha çok geleneksel istatistiksel ve matematiksel yöntemlerle modellemeler ve analizler yapılmaktadır. Ancak her geçen gün artan veri miktarı ve çeşitliliği büyük veri kavramı çerçevesinde yeni modelleme yöntemlerine başvurulmasına da yol açmaktadır. Bu açıdan bakıldığında hava kalitesi değerlendirme çalışmaları çevre bilimleri kadar bilgisayar bilimlerinin de veri modellemesi ve analizi çerçevesinde ilgi alanına girmektedir. Hava kirliliği ve kalitesi ile ilgili alan yazın çalışmaları incelendiğinde bilgisayar bilimlerinin konusu olan yapay zeka temelli yöntemlerle daha hassas sonuçlar elde edildiğini göstermektedir [4]. Termik santrallerin yer aldığı sanayi alanlarını içeren Zonguldak-Kuzyaka ve Manisa-Soma, PM_{10} ve SO_2 kirliliğinin yoğun yaşandığı bölgelerdir. Bu tez çalışmasının amacı, yapay zekanın bir alt disiplini olan makine öğrenmesi yöntemleri ile hava kirliliği tahminine yönelik analiz çalışmalarının etkinliğini araştırmaktır. Bu amaç doğrultusunda Manisa ve Zonguldak

illerinde PM₁₀ ve SO₂ kirletici parametreleri ile makine öğrenmesi yöntemleri kullanılarak modelleme ve analiz çalışmaları gerçekleştirilmiştir. Bu suretle, yapay zeka temelli yöntemlerin etkinliği ve başarımı ortaya konulmuştur. Tez çalışması kapsamında Manisa ve Zonguldak illerine ait meteorolojik veriler (rüzgar, nem, yağış, sıcaklık, basınç) ve PM₁₀, SO₂, CO, NO₂, NO_x, O₃ kirletici parametreleri dahil edilerek hava kirliliği modelleme ve tahmin çalışmaları gerçekleştirilmiştir.

Çalışmada kullanılan kirletici verileri Çevre Şehircilik Bakanlığı (ÇSB)'ndan, meteorolojik veriler Meteoroloji Genel Müdürlüğü (MGM)'nden alınmıştır [5,6]. Zonguldak-Kuzyaka ve Manisa-Soma Hava Kalitesi İzleme istasyonu termik santrallerin yer aldığı sanayi bölgelerinde bulunmaktadır. Aralık 2016 -Nisan 2021 tarihleri arasında PM₁₀ ve SO₂ kirleticilerine ait konsantrasyon verileri ve meteorolojik faktörler (sıcaklık, basınç, nem, rüzgar hızı, yağış miktarı) alınarak iki ayrı veri seti oluşturulmuştur. Yapılan çalışmalarda elde edilen sonuç ve bulgular bakımından solunum yolu hastalıklarını ciddi oranda etkileyen PM₁₀ ve SO₂ kirleticilerinin ele alınması, çalışma veri setinin Covid-19 pandemi dönemini de kapsamı bakımından büyük önem taşımaktadır [7].

Oluşturulan iki ayrı veri setine yaygın olarak kullanılan Destek Vektör Makinesi (DVM), K-En Yakın Komşu (K-EYK), Karar Ağacı (KA), Rastgele Orman (RO), ve Naive Bayes (NB), Uzun Kısa Süreli Bellek (UKSB) makine öğrenme algoritmaları uygulanmış ve tahmin başarıları kıyaslanmıştır. Çalışmada kullanılan verileri bölme işleminde %80 eğitim ve %20 test verisi alınarak model denenmiştir. DVM, KEYK, KA, RO, NB modellerinin kod geliştirme aşamasında Python programlama dili kütüphanelerden Scipy, Numpy, Pandas, Sklearn, Seaborn ve LSTM modeli geliştirilirken Tensorflow ve Keras kütüphanelerinden faydalanılmıştır. Alınan sonuçların görselleştirme aşamasında ise Matplotlib kütüphanesi kullanılmıştır. Her bir yöntemin tahmin başarı performansı açıklayıcılık katsayısı (R^2), Kök Ortalama Kare Hatası (KOKH- Root Mean Square Error-RMSE) ve Ortalama Mutlak Hata (OMH- Mean Absolute Error-MAE) metriklerine göre kıyaslanmıştır.

Tez çalışmasının sonraki bölümleri şu şekilde yapılandırılmıştır: İkinci bölümünde, konuyla ilgili kavramsal ve kuramsal temeller ele alınmış olup öncelikle hava kirliliği, kirleticiler ve meteorolojik veriler hakkında kavramsal bilgiler verilmiştir. Ardından meteorolojik verilerin hava kirliliği üzerindeki etkileri açıklanmış ve tahminleme modelinde kullanılan makine öğrenmesi yöntemleri hakkında kuramsal bilgiler verilmiştir. Son olarak bu bölümde

literatürdeki meteorolojik verilerle hava kirliliği arasındaki ilişkiyi ve hava kirliliği tahmini için kullanılan yapay zeka yöntemlerini konu alan alan yazın çalışmaları incelenmiştir. Üçüncü bölümde çalışmada kullanılan verilerin temini, veri ön işleme aşamaları ve veri kaynağı olan istasyonlar hakkında bilgi verilmiştir. Ayrıca çalışma sonuçlarının kıyaslanmasında kullanılacak olan performans metrikleri açıklanmıştır. Dördüncü bölümde uygulanan modeller performans metrikleri baz alınarak kıyaslanmış ve son bölümde genel bir değerlendirme yapılmıştır.

2. KAVRAMSAL VE KURAMSAL ÇERÇEVE

Bu bölümde tez kapsamında değerlendirmeye alınan hava kirliliği, kirlетici maddeler, meteorolojik veriler, kirlетici maddeler ile meteorolojik veriler arasındaki ilişki açıklanmaktadır.

2.1. Hava Kirliliği

Hava kirliliği, atmosferdeki su buharı, toz, duman, gaz halindeki kirlетicilerin seviyesinin artarak doğadaki tüm canlılara zarar verecek sınırlara ulaşmasıdır. Hava kirliliğinde önemli rol oynayan maddeler beş başlık altında toplanabilir. Hidrokarbonlar (C_xH_y), kükürt(S)-karbon(C)-azot(N) emisyonları ve partikül maddelerdir (PM). Partikül maddeler kendi içinde boyutlarına göre PM_{10} ve $PM_{2,5}$ olmak üzere ikiye ayrılmaktadır. Bu beş kirlетiciden solunum sistemi üzerinde en çok tahribata sebep olan PM_{10} ve SO_2 'tir.

Artan nüfus ve hızla büyümeye devam eden sanayileşme, özellikle insan sağlığını kötü etkileyecek önemli bir boyuta gelmiştir ve her yıl havadaki kirlilik sebebiyle meydana gelen ölümler 7 milyon civarındadır [8]. Hava kirliliğinin başlıca sebep olduğu hastalıklar arasında astım, kalp krizi ve rahatsızlıkları, bronşit, zatürre bulunmaktadır. Ayrıca üst solunum yolu rahatsızlıklarını tetikleyip daha da kötüleşmesine sebep olmaktadır.

Çevremizdeki hava kalitesinin sağlığımız üzerindeki etkisinin ne olduğunu anlamamıza yarayan ölçek ise Hava kalite indeksi (HKİ)'dir. HKİ, havada bulunan kirlетici konsantrasyonlarına bakılarak iyi, orta, hassas, sağlıksız, kötü ve tehlikeli olacak şekilde beş seviye ile gösterilmektedir. Aynı zamanda HKİ, hava kirliliği kaynaklı hastalıkları bulunan insanlar için sağlık durumlarını korumak amacıyla tedbir almalarına yardımcı olmaktadır. Yani çevresel etkenlerin insan sağlığını hangi derecede etkilediğini gösteren HKİ, sağlık açısından tehdit olabilecek durumların önüne geçmek için kullanılmaktadır. Çizelge 2.1'de Ulusal hava kalite indeksine ait kesme noktaları verilmiştir.

Çizelge 2.1. Ulusal hava kalite indeksi kesme noktaları [9]

İndeks	H Kİ	SO ₂ [µg/m ³]	NO ₂ [µg/m ³]	CO[µg/m ³]	O ₃ [µg/m ³]	PM ₁₀ [µg/m ³]
		1 Sa. Ort.	1 Sa. Ort.	8 Sa. Ort.	8 Sa. Ort.	24 Sa. Ort.
İyi	0- 50	0-100	0-100	0-5500	0-120	0-50
Orta	51- 100	101-250	101-200	5501-10000	121-160	51-100
Hassas	101-150	251-500	201-500	10001-16000	161-180	101-260
Sağlıksız	151-200	501-850	501-1000	16001-24000	181-240	261-400
Kötü	201-300	851-1100	1001-2000	24001-32000	241-700	401-520
Tehlikeli	301-500	>1101	>2001	>32001	>701	>521

2.2. Kirleticiler

Hava kalitesini etkileyen temel kirleticilerden PM₁₀, SO₂, CO, Azot oksitler (NO₂, NO_x), O₃ kirleticisi maddeleri açıklanmıştır.

2.2.1. PM₁₀

Partikül maddeler (PM), havadaki asılı kalan katı ve sıvı parçacıklardan meydana gelen kirleticilerdir. Boyutlarına göre ikiye ayrılmakta olup (PM₁₀ ve PM_{2,5}) sağlık üzerinde ciddi negatif etkilere sahiptir. Özellikle boyut olarak küçüldükçe solunum yolu hastalıklarına etkisi daha fazla olmaktadır. Kirliliğin olduğu bölgelerde, partikül maddelere uzun süre maruz kalınması halinde akciğerlerde biriken kirleticisi oranı artmaktadır ve bunun sonucu olarak hem solunum hem de kalp rahatsızlıkları oldukça sık gözlemlenmektedir [10]. Sanayi, madencilik, trafik ve ısınma faaliyetleri nedeniyle yakılan fosil yakıtlar, volkan patlamaları, yangın ve toz fırtınaları partikül maddelerin oluşmasındaki temel kaynaklardır.

2.2.2. SO₂

Kükürt içerikli fosil yakıtların yanmasıyla ortaya çıkmaktadır. Genellikle sanayi, trafik ve ısınma faaliyetleri sonucunda oluşan bir kirleticidir. Doğada bulunan başka kirleticisi maddelerle birleşip katı veya sıvı partiküller şeklinde taşınabilmektedir [11]. Renksiz, keskin bir kokuya sahip olan kükürt dioksit (SO₂) gazı kısa süre maruz kalma durumunda bile

solunum sisteminde ciddi tahribata yol açmaktadır. Solunum sistemindeki tahribatın dışında gözlerde hassasiyet, astım krizi ve kalp rahatsızlıkları gibi sağlık problemlerine de sebep olmaktadır [12].

2.2.3. CO

Yakıtların yetersiz yanması sebebiyle ortaya çıkan Karbon monoksit (CO) gazı, renksiz, tatsız ve kokusuzdur. Orman yangınları, yoğun trafik ve egzoz gazları atmosfere yüksek oranda CO yayılmasına sebep olmaktadır. Bu yoğun CO gazı, doğada bulunan tüm canlıların sağlığını tehdit etmektedir. Özellikle insanlardaki kan hücrelerinde bulunan ve oksijen taşıyan hemoglobine bağlanarak damarlara taşınan oksijen miktarını düşürmektedir [13]. Bu azalma sonucu yeterli oksijen alamayan sinir dokuları, kalp ve beyin gibi hayati organlar düzgün çalışmayabilir. Aynı zamanda yüksek miktarda CO gazına maruz kalan bireylerde algılama ve görmede zayıflık oluşabilmektedir [14].

2.2.4. Azot oksitler (NO_x , NO_2)

Yüksek sıcaklıklarda meydana gelen renksiz, kokusuz ve yüksek reaktif olan Azot oksitler (NO_x), havada bulunan en önemli kirleticilerdendir. Atmosferde genellikle azot oksit (NO), diazot monoksit (N_2O), azot dioksit (NO_2) türleri bulunmaktadır. NO_x 'ler, genellikle fosil yakıtlar ve endüstriyel atıkların yanması sonucu oluşmaktadır. NO_x 'ler havada bulunan su ile istenmeyen tepkimelere girerek asit yağmurlarının oluşmasına neden olmaktadır. Ayrıca ekosistemde bulunan canlı hayatını olumsuz yönde etkilemektedir. Özellikle NO_2 gazı akciğerlerde ciddi hasarlara yol açarak solunum rahatsızlıklarına sebebiyet vermektedir [15].

2.2.5. O_3

Ozon(O_3), kararsız, açık mavi renkte, keskin bir kokuya sahip oldukça zararlı bir gazdır. İnsan sağlığı üzerinde ciddi bir tehdit oluşturan O_3 gazı, yüksek oranlarda maruz kalındığında göz ve akciğer hasarlarına sebep olmaktadır. Ayrıca, orta seviyede O_3 miktarlarına uzun ve tekrarlı maruz kalınması durumunda kronik etkileri olduğu gözlemlenmiştir. Ozon gazı genellikle gündüz güneşli saatlerde meydana gelir. Dolayısıyla ozon seviyeleri yaz aylarında en yüksek seviyelere ulaşmaktadır [16].

2.3. Meteorolojik Veriler

Sıcaklık, basınç, yağış, nem ve rüzgâr hava kirliliğini etkileyen başlıca meteorolojik faktörlerdir. Özellikle sonbahar ve kış aylarında hava kirliliğinin artmasında etkisinin olduğu düşünülmektedir [17].

Sıcaklık: Soğuk havalarda artan ısınma faaliyetleri sonucunda oluşan hava kirliliği sıcaklık etkenine bağlıdır. Hava sıcaklıklarının yüksek olduğu dönemlerde hava kirliliği çok fazla hissedilmemektedir.

Basınç: Soğuk havalarda oluşan yüksek basıncın etkisi ile fosil yakıtların yanması sonucu oluşan kirleticilerin etkisi daha net hissedilmektedir. Bu durumun tam tersi alçak basıncın olduğu bölgelerde ise kirlilik aynı yoğunlukta hissedilmemektedir.

Rüzgar: Rüzgarın yönü, hızı, frekansı hava kirliliğini en fazla etkileyen faktördür. Özellikle hakim rüzgar yönü incelenen bölgenin yerleşimiyle uyumlu olmaması halinde kirlilik daha fazla hissedilmektedir. Kış aylarında rüzgarın etkisi kirliliğin hissedilmesinde önemli rol oynamaktadır.

Yağış ve Nem: Havada bulunan nem ve su buharının kirleticilerle bir araya gelerek asit yağmurlarını meydana getirdiği bilinmektedir [18].

2.3.1. Meteorolojik veriler ve kirleticiler arasındaki ilişki

Hava kirliliği üzerine etki eden faktörlerden önde gelen meteorolojik faktörler temel olarak sıcaklık, basınç, rüzgar, yağış ve nemdir. Hava kirliliği alanında yapılan bazı çalışmalarda özellikle rüzgar hızı, nem ve sıcaklığın hava kalitesi üzerinde önemli etkilere sahip olduğuna dair sonuçlar elde edilmiştir [19]. Bu çalışmalardan Kara, Konya iline ait SO₂ ve PM₁₀ kirleticilerinin günlük ve aylık dağılımlarını incelemiş ve meteorolojik faktörlerle arasındaki ilişkiyi incelemiştir [20]. Özellikle kış mevsiminde yüksek basınç ve düşük rüzgar hızının PM₁₀ ve SO₂ kirleticilerini etkilediği gözlemlenmiştir. Kış mevsimindeki PM₁₀ ve SO₂ kirletici emisyon miktarlarının %53 SO₂ ve %48 PM₁₀ oranında meteorolojik faktörlere bağlı olduğu görülmüştür.

Sari ve Esen, çalışmalarında, Türkiye’deki 103 ölçüm istasyonundan aldıkları veriler ile PM_{10} ve SO_2 kirletici seviyelerini analiz etmişler ve meteorolojik faktörler ile ilişkisini araştırmışlardır. İncelemeleri sonucunda kirletici konsantrasyon değerleri üzerinde özellikle sıcaklık ve rüzgar hızı parametrelerinin daha fazla etkisi olduğu anlaşılmıştır [21].

Tepe ve Doğan, Türkiye’nin güneyinde bulunan dört ilin (Adana, Antalya, Hatay ve Mersin) hava kalitesini araştırmış ve kış döneminde kirliliğin özellikle PM_{10} yoğunluğunda bir artış hatta limit değerini aştığı sonucunu elde etmişlerdir [22].

Öztürk ve Bayram, yaptıkları çalışmada Van ilindeki kirletici konsantrasyonları ile sıcaklık arasındaki ilişkiyi incelemişler ve hava kirliliğinin kış mevsiminde arttığını fakat yine de hava kalitesinin iyi seviyede olduğunu gözlemlemişlerdir. 2015-2016 yıllarına ait SO_2 ocak ayı ortalaması $47 \mu g/m^3$ iken ekimde bu ortalama $4 \mu g/m^3$ olarak hesaplanmıştır. Özellikle SO_2 kirleticisi ile sıcaklık arasında ters bir ilişkinin söz konusu olduğu sonucunu çıkarmışlardır [23].

2.4. Kuramsal Temeller

Hava kirliliğinin ileriye yönelik tahmin edilmesi amacıyla yapılan çalışmaların sayısı gün geçtikçe artmaktadır. Bilindiği üzere yapay zeka birçok alanda uygulanmakta ve yüksek başarılar elde edilmektedir. Hava kirliliği tahmininde de özellikle geleneksel istatistik yöntemlere göre daha yüksek başarılı sonuçlar veren yapay zeka yöntemlerine başvurulmaktadır. Tez çalışması kapsamında uygulanan makine öğrenmesi yöntemlerine ilişkin kuramsal temeller aşağıda açıklanmaktadır.

2.4.1. Makine öğrenmesi yöntemleri

Makine öğrenmesi, yapay zekanın alt dalı olup [24] temelde bilgisayarlara bir veri topluluğu içerisindeki karmaşık örüntüleri tespit ettirmeyi ve akılcı kararlar verdirebilmeyi amaçlamaktadır [25].

1980’lerde ortaya çıkan makine öğrenmesi zamanla çağın teknolojisinde önemli bir yer edinmiştir. Ayrıca, günlük yaşamdaki birçok problemin çözümünde ve ortaya faydalı bir ürün çıkarılmasında rol oynamaktadır. Makine öğrenmesi yöntemleri geliştirilirken

mühendislik, biyoloji ve istatistik gibi bilim dallarından yararlanılmaktadır. Temelde tahmin, kümeleme ve sınıflandırma amaçlı algoritma geliştirmekte kullanılmaktadır. Genellikle makine öğrenmesi algoritmalarının işleyiş adımları;

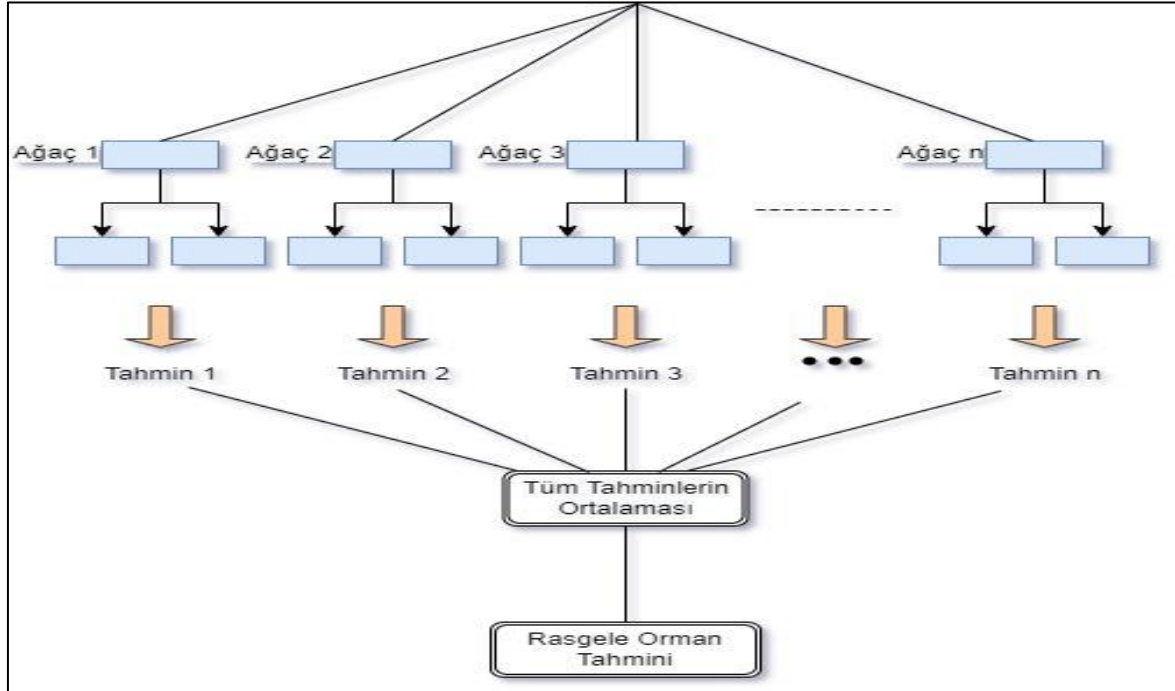
- Örnek veri setinin oluşturulması,
- Özelliklerin belirlenmesi ve uygulanacak olan algoritmanın seçimi,
- Eğitim veri seti ile modelin eğitilmesi
- Son olarak eğitim veri setinden bağımsız bir veri seti oluşturularak modelin test edilmesidir [26].

Hava kalitesi tahmin sistemlerinde kullanılan makine öğrenmesi algoritmasının seçimi çok önem arz etmektedir. Literatürde yapılan inceleme sonucu özellikle 2017'den itibaren, atmosferdeki hava kirliliğinin tahmininde makine öğrenmesi algoritmalarının uygulanmasında ciddi bir artış olduğu gözlemlenmiştir [27]. Masih tarafından yapılan sistematik literatür incelemesinde 2013-2018 yılları arasında yapılan 32 çalışma ele alınmıştır. Bu sistematik literatür incelemesi, 2017'den beri atmosferik kirliliği tahmin etmek için makine öğrenimi tekniklerinin uygulanmasının önemli ölçüde arttığını ortaya koymaktadır. Masih'in yaptığı çalışmaya göre, hava kirliliğinin birden fazla değişkenden etkilenmesi, problemin lineer modelleme ile çözülememesine yol açmaktadır. Ayrıca makine öğrenmesi algoritmalarıyla elde edilen sonuçların yüksek doğruluk değerlerine sahip olması, bu algoritmaların geleneksel yaklaşımlara göre daha uygun olduğunu ve neden tercih edilmesi gerektiğini açıklamaktadır. Tez çalışmasında da özellikle başarı oranlarının yüksek olması sebebiyle tercih edilen makine öğrenmesi teknikleri aşağıda açıklanmıştır.

Rastgele orman (RO)

Makine öğrenmesi algoritmalarından rastgele orman algoritması (Random Forest Algorithms) hem sınıflandırma hem de regresyon maksatlı uygulanabilmektedir. Çoğunlukla sınıflandırma amacıyla kullanılmaktadır. Temelde birden fazla karar ağacının bir araya gelerek oluşturduğu rastgele orman algoritması oluşturulan veri setinde rastgele bir orman oluşturmaktadır ve oluşan bu yeni ormanların her biri için tahmin yapar. Son adımda tahmin sonuçlarından en iyi olanı çözüm olarak seçer [28]. Sonuçların ortalamasını alarak aşırı

öğrenmenin önüne geçen rastgele orman algoritması, birbirinden bağımsız değişkenlerle farklı karar ağaçları üzerinden değerlendirdiği için başarısı oldukça yüksektir [29].



Şekil 2.1. Rastgele Orman algoritması yapısı [30]

- Rastgele orman algoritmasının temel adımları kısaca açıklanmıştır:
- İlk olarak kullanılan veri setinden rastgele örnekler seçilir.
- Seçilen her bir örnek için yeni bir karar ağacı oluşturulur ve meydana gelen her bir karar ağacına ait bir tahmin sonucu elde edilir.
- Ardından elde edilen sonuçlar üzerinde eğer problem sınıflandırma ise mod alma, şayet problem regresyon ise ortalama alma kullanılacaktır.
- Son adımda ise en çok oylanan tahmin sonucu seçilir [31].

Algoritmanın performansı, 3.2.8’de açıklanan ölçütlere göre değerlendirilmektedir.

Destek vektör makinesi (DVM)

Destek vektör makinesi (Support Vector Machine) algoritması, iki veya daha fazla boyuta ait verileri doğru ya da hiper düzlem aracılığıyla birbirinden ayırıp sınıflandırma ve regresyon analizi yapmaktadır. Bu algoritmanın en temel amacı verilen veri noktalarını

mümkün mertebede en iyi halleriyle sınıflandırmaktır. Temelde veri setinin doğrusal ve doğrusal olmamasına göre ikiye ayrılır [32].

Doğrusal destek vektör makinesi algoritmasıyla sınıflandırmada, veri setini ikiye ayıran sayısız karar doğrusu çizilir ve içlerinden en uygun olan karar doğrusu seçilir. Veri setine eklenebilecek yeni veriye karşı karar doğrusunun dayanıklı olabilmesi için iki sınıfa ait sınır çizgisine de yakın mesafede olması gerekir. Bu çizgilere en yakın olduğu noktalara ise destek noktaları denmektedir [33].

Doğrusal olmayan destek vektör makinesi algoritmasında ise veri setinde doğrusal bir hiper düzlem çizilemediğinden çekirdek numarası kullanılır. Çekirdek yöntemi (kernel trick) olarak bilinen yöntemde doğrusal olmayan ya da iki boyut ile açıklanamayan değişimleri boyut arttırıp çözüyormuş gibi hile yapar. Eğer modelde aşırı öğrenme ile karşılaşılırsa derecesi düşürülür. Tam tersi durumda ise derece yükseltilir. En çok tercih edilen Polinomal çekirdek ve Gaussian çekirdek yöntemleridir.

Denetimli öğrenme yöntemlerinden biri olan DVM, modelin eğitilmesinde kullanılacak olan eğitim verisine ve eğitilen modelin test etmek amacıyla gerekli olan test verisine ihtiyaç duymaktadır. Eğitim veri setinin boyutunun az olduğu zamanlarda ve çok sınıflı alanlarda bile çekirdek yöntemiyle çok iyi genelleme yapabilir. Ayrıca en büyük avantajı hızlı olmasıdır. Özellikle hava kirleticiler konsantrasyonlarının tahmininde önemli rol oynamaktadır [34].

Naive bayes (NB)

Naif Bayes (Naive Bayes) algoritması (NB) temelde olasılık teoremine dayanan makine öğrenmesi yöntemlerinden biridir. Tahmin yapmak için her sınıfın olasılığını hesaplayıp kullanan basit ve güçlü bir algoritmadır. Yani modelin oluşturulması basit ancak belirli özelliklerin arasındaki bağımsızlık varsayımından dolayı zorlukla karşılaşılabılır. NB büyük veri kümeleri içinde kullanışlı olmakla beraber dengesiz veri setlerinde de çalışabilmektedir. Algoritma, bir örnek için her bir olasılığı hesaplayıp elde ettiği en yüksek olasılık değerine göre örneği sınıflandırmaktadır. NB yönteminde, test veri kümesinde var olan ancak eğitim veri kümesinde olmayan veriler için tahmin gerçekleştirememektedir. Karşılaşılan bu

probleme “Sıfır frekans” denir [35]. Literatürdeki çalışmalarda bu problemi çözmek için Laplace tahmini benzeri düzenleme teknikleri kullanılmaktadır.

Temel çalışma mantığı, bir sonraki duruma daha önceki durumlara bakılarak karar verilmesi üzerinedir. Kısacası bu algoritmadaki mantık önceki deney temelinde çalışmaktadır. Naive Bayes algoritmasını gerçek hayattaki problemlere uygularken ilk veri seti tanımlanır. Veri setindeki sınıflar açıkça görülebilmektedir. Bu sebeple sınıf soyutlaması basitçe yapılabilmektedir. Gerçekleşen bir durumu üreten bir etkeni izleme olasılığı, $P(L|S)$ ile adlandırılan ana olabilirlik kavramıdır. $P(L|S)$ 'nin hesaplanmasında kullanılan denklem aşağıda verilmiştir:

$$P(L|S) = \frac{P(L) \times P(S|L)}{P(S)} \quad (2.1)$$

$P(S)$: Gerçekleşen S olayının her şeyden bağımsız olarak gözlemlenme olasılığıdır.

$P(L)$: Gerçekleşen olayın diğer tüm etkenlerden bağımsız tutularak L faktörünün gözlemlenme olasılığıdır.

$P(S|L)$: Sonsal olasılıktır ve L olayı gerçekleştiğinde S'nin koşullu gözlenmesi olasılığıdır.

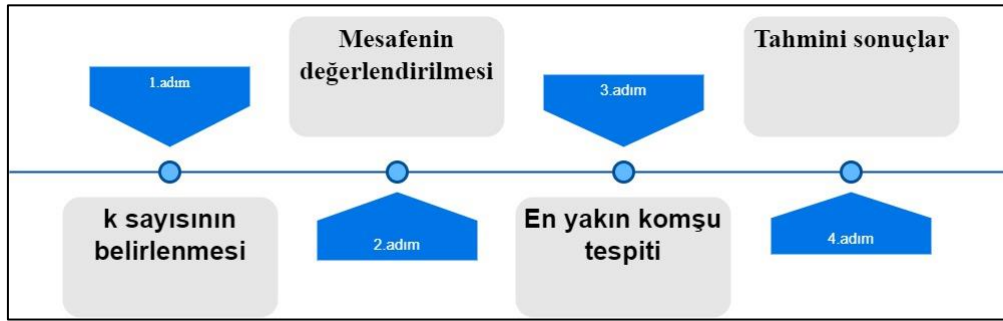
NB yönteminin çalışma prensibi; oluşturulan sisteme belirli oranda eğitim verisi alınmasıyla başlamaktadır. Sisteme alınan bu eğitim verilerinin her biri bir sınıfa ait olmalıdır. Bu eğitim verileri üzerinden ortaya konan olasılık yöntemleriyle sisteme alınan test verileri işlenmektedir. Gerçekleştirilen bu işlemde önceden elde edilen olasılık değerleri kullanılmaktadır. Ardından alınan test verilerinin sınıfı tespit edilir ve eğitim verisinin sayısı arttıkça test verilerinin gerçek sınıfının doğruluğu tespit etmek kolaylaşır.

K en yakın komşu (K-EYK)

K En Yakın Komşu (K-nearest neighbors algorithm) algoritması, uygulanması basit, yaygın, regresyon ve sınıflandırma problem çözümlerinde sıklıkla tercih edilen bir makine öğrenmesi yöntemidir. K-EYK yöntemi ilk adımında veri setinde bulunan tüm veriler etiketlenmektedir. Daha sonra etiketlenmiş olan bu veriler kullanılarak eğitim veri seti

oluşturulmaktadır. k sayıda sınıf merkezi belirlenir ve uzaklık fonksiyonu seçilir. Yöntemin uygulanma aşamasında önceden karşılaşılmamış bir veri ile karşılaşıldığında seçilen uzaklık algoritması devreye sokularak kullanılarak yüz yüze gelinen verinin eğitim veri setindeki verilerle olan mesafesi ölçülür. Ardından bu veriler arasından uzaklığı en kısa olan k sayıdaki veri eğitim veri setinden seçilerek sınıflama kümesi oluşturulur. Son adımda, seçilen yeni verinin sınıfı, sınıflama kümesinin en sık içerdiği sınıf olarak tanımlanır ve oluşturulan model sona erdirilir [36].

K-EYK algoritmasında kabul edilmiş olan komşuların sayısını gösteren k sayısının belirlenmesi çok önemlidir. k sayısı pozitif tam sayı seçilmelidir. Seçilen bir veri seti için, ön işleme aşaması başlangıçta gerçekleştirilmelidir. Ardından K-EYK süreci dört ana adımda ilerler:



Şekil 2.2. K-EYK algoritma yapısı

- Öncelikli olarak rastgele k sayısı belirlenir.
- k sayıda komşunun mesafe ölçümü yöntemleri ile uzaklıkları hesaplanır. Mesafenin hesaplanmasında kullanılan en yaygın ölçümler Öklid mesafesi, Manhattan mesafesi ve Hamming mesafesidir. Özellikle Öklid mesafesi sıklıkla tercih edilmektedir. İki nokta arasındaki Öklid mesafesi denklem (2.2) kullanılarak hesaplanabilir.

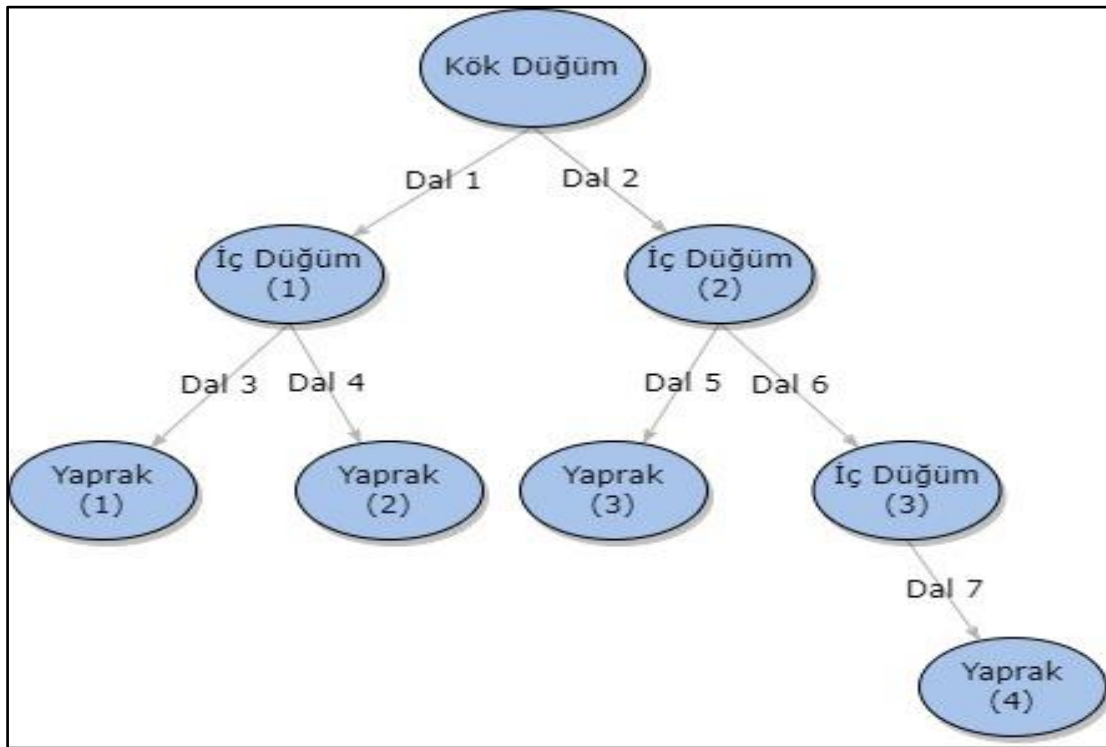
$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2} \quad (2.2)$$

Burada d , test verilerinin ögesi x ile her bir eğitim ögesi x' arasındaki mesafeyi temsil etmektedir.

- Ölçülen mesafeler sıralandıktan sonra en küçük uzaklığa bağlı olarak en yakın komşular tespit edilir.
- Son olarak k en yakın komşunun ortalaması hesaplanır [37].

Karar ağacı (KA)

Karar Ağacı (*Decision tree*) algoritması, örneklerin entropisine dayanan bir makine öğrenmesi yaklaşımıdır. Oldukça popüler ve basit bir makine öğrenmesi yöntemidir. KA yöntemi, önceden eğitimde kullandığı verilerden elde edilen karar kurallarını öğrenerek en son çıktı değişkenine ait sınıfı tahmin etmeyi hedefleyen bir model gerçekleştirmektedir. KA, bir ya da birden çok düğüm ve düğümlere bağlı dallardan meydana gelir. Ağaçta bulunan her düğüm bir özelliği gösterir. Dalların her biri bir kararı temsil eder [38].



Şekil 2.3. Karar ağacı yapısı

KA, örneklerin entropisine dayanan bir makine öğrenme prosedürüdür. Denetimli bir algoritma olan karar ağacı algoritması, bir sınıflandırma yöntemi veya problemle ilgili bir regresyon belirleyicisi olarak hizmet edebilir. Bir kök düğümden başlar ve entropiye dayanarak, örnekler bir yaprak düğüme ulaşana kadar bölünür.

Karar ağaçları, kullanımı ve anlaşılması basit tahmine dayalı öğrenme algoritmalarıdır. Akış diyagramı benzeri ağaç yapısı ile her sonuçla ilişkili tüm olası yolları temsil ederler. Görselleştirilebilmesi ve kolay yorumlanabilmesi avantajlarındanır. Ayrıca elde edilen sonuçların güvenilirliği ve yüksek performansa sahip olması da tercih sebeplerindendir [39].

KA algoritmalarının genellikle karşılaştığı belli başlı sorunlar vardır. Oluşan ağaç, gürültü ve eksik veri gibi anormalliklerden etkilenmektedir. Aykırı değerler ve oluşabilecek anormallikler aşırı öğrenmeye sebep olabilir. İstatistiksel önlemler kullanılarak budama işlemiyle ağaçtan en düşük güvenilirliğe sahip dalları çıkarılır ve aşırı öğrenme bu şekilde çözülebilir. Bilgilendirici nitelikleri doğru seçmek için entropi ve gini indeksi kullanılabilir [40]. Ayrıca veri setinin eğitim ve test olarak bölünmesi de bir sorundur. Küçük boyutlu eğitim veri setlerinde sınıflandırma doğru bir şekilde gerçekleşmeyebilir. Tam tersi büyük boyutlu eğitim veri seti ayarlandığında ise aşırı öğrenme oluşabilir.

Derin öğrenme ve ilişkili diğer yöntemler

Son zamanlarda gerçekleştirilen hava kalitesinin tahmini ve hava kirliliği problemlerinin çözülmesine yönelik çalışmalarda makine öğrenmesi yöntemleri oldukça sık tercih edilmeye başlanmıştır. Özellikle zaman serilerine uygulanabilen ve yüksek başarıya sahip olan derin öğrenme algoritmalarının popülerliği gün geçtikçe artmaktadır [41].

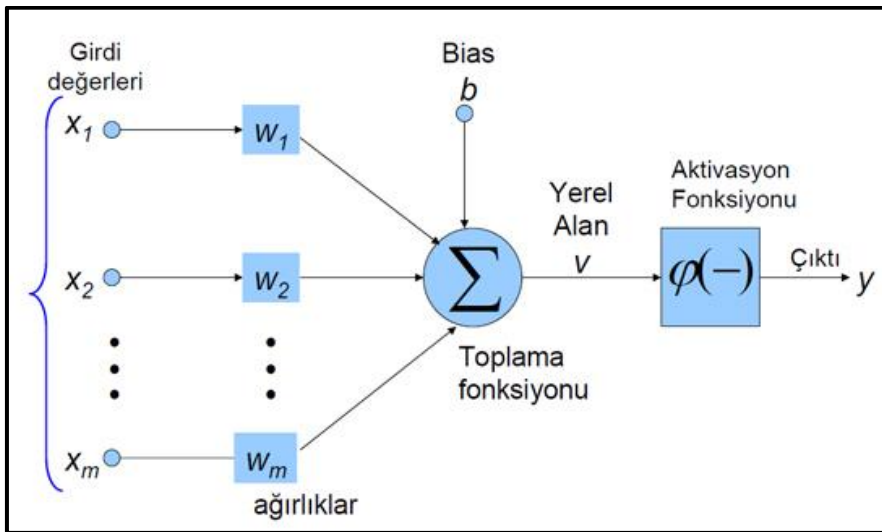
Derin öğrenme hem makine öğrenmesi hem de yapay zeka çalışma alanı içinde yer almaktadır. Bu kapsamda derin öğrenmeyle ilişkili çalışma alanları Şekil 2.4' te gösterilmektedir. Birçok doğrusal olmayan katman kullanarak özellik çıkarma ve dönüştürme gerçekleştirdiği için başarısı yüksektir. Kullandığı bu katmanlar birbiri ile ardışık olup bir önceki katmandaki bilgiyi bir sonraki katman girdi olarak kullanır [42].



Şekil 2.4. Derin öğrenmenin makine öğrenmesi ve yapay zeka ile ilişkisi [25-43]

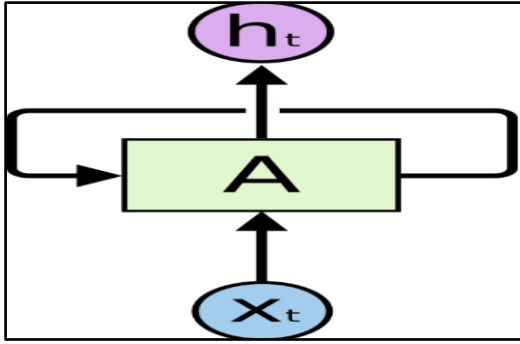
Hava kirliliği tahmininde kullanılan derin öğrenme algoritmalarından Uzun Kısa Süreli Bellek (Long Short Term-LSTM) algoritmasını açıklayabilmek için öncelikle yapay sinir ağları ve tekrarlayan sinir ağlarına yer verilmiştir.

İnsan beyнинin taklit edilmesi ile ortaya çıkan yapay sinir ağları, hava kirliliğinin modellenmesinde oldukça sık tercih edilen bir makine öğrenmesi algoritmasıdır [44,45]. Temelde beyindeki sinir hücrelerinin çalışma mantığından yola çıkılarak karmaşık giriş verilerini alarak çözümleme yapmaya çalışır. En temel yapı taşı biyolojik olarak beyin hücresine benzetilen yapay nöronlarda işlemleri gerçekleştirir. Bu yapay nöron hücrelerinde Şekil 2.5'te görüldüğü gibi girdi değerleri, ağırlıklar, toplama fonksiyonu, aktivasyon fonksiyonu ve çıkış olmak üzere toplamda beş temel elemandan oluşmaktadır.



Şekil 2.5. Yapay nöron hücresinin yapısı [42]

Yapay sinir ağlarında her nöron hücresi aldığı bilgiyi işleyerek bir sonraki yapay nöron hücresine aktarmaktadır. Bazı yapay sinir ağlarında ise bu öğrenme yöntemi geri beslemeli olarak modellenir. Birimleri arasında döngüsel bağlantı bulunan tekrarlayan sinir ağları (Recurrent Neural Network-RNN), diğer yapay sinir ağlarından farklı olarak bir hafızaya sahiptir. Derin öğrenme algoritmaları arasındaki en ayırıcı özelliği bu hafızasının varlığı ve kendi içinde döngüsel yapısıdır [46]. Geri beslemeli olarak öğrenen modelde daha önceki katmanlardaki hatalara bakılır ve bir sonraki katmanda bu hatalar, tüm noktalar ağırlıklandırılarak en düşük seviyeye indirgenmeye çalışılır. Temel bir RNN hücresinin yapısı Şekil 2.6'da gösterilmektedir.



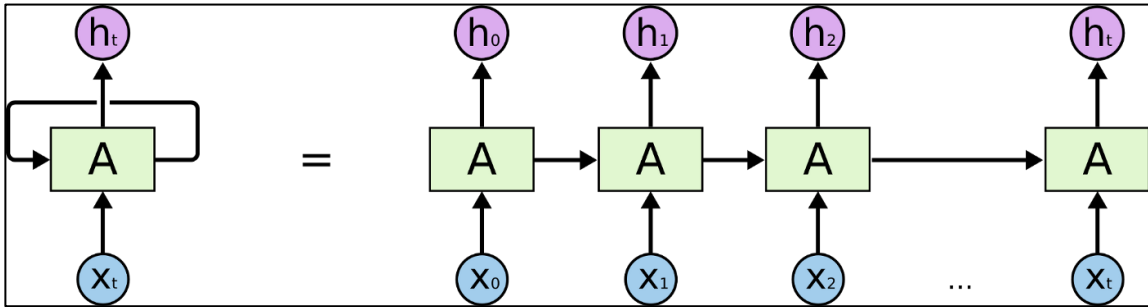
Şekil 2.6. RNN yapısı [47]

A: Sinir ağını temsil eder.

x_t = Giriş değerlerini temsil eder.

h_t = Çıkış değerlerini temsil eder.

RNN'ler birden çok sinir ağı içermektedir ve her bir sinir ağı içerdiği bilgiyi bir sonraki girişe aktarmaktadır. Her çıkış şekil 2.7.' de görüldüğü üzere kendi girişi ve bir önceki sinir ağının çıkışına bağlıdır.



Şekil 2.7. RNN yapısının açık gösterimi [47]

$$h_t = f_w(h_{t-1}, x_t) \quad (2.3)$$

h_t : Şuanki h değerini temsil eder.

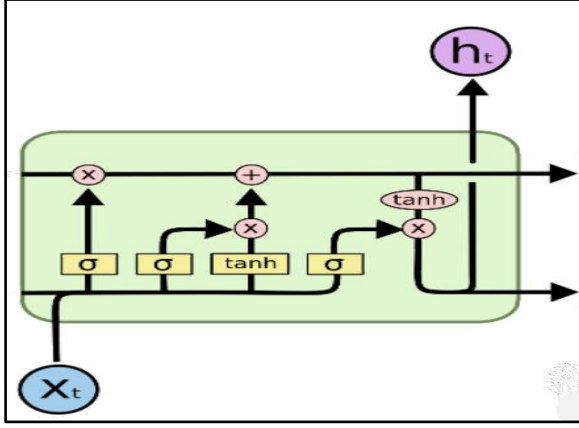
h_{t-1} : Bir önceki h değerini temsil eder.

x_t : Şuanki x vektörünü temsil eder.

f_w = Ağırlıkları temsil eder.

Belirli bir girdinin gizli katman üzerindeki ve dolayısıyla ağ çıktısı üzerindeki etkisi, bir RNN boyunca yayılırken ya azalır ya da üstel olarak büyür. Pratikte, standart RNN'lerin erişebileceği bağlamsal bilgi aralığı, ilgili girdi ve hedef olaylar arasındaki yaklaşık on zaman adımıyla sınırlıdır. Gradyanlar, sinir ağlarının ağırlıklarını güncellemek için kullanılan değerlerdir [46]. RNN'ler, zamanla çok tekrar gerçekleştirdiklerinde veriyle alakalı bazı parametreleri veriyi daha az değiştirdiği sebebiyle sistem dışı bırakabilir. Bu sistemden atılan parametreler, çok başta sinir ağına eklenmişse sistemi kullanan bu bilgilere erişmekte zorluk yaşayabilmektedir. Meydana gelen kaybolan gradyan sorunu, gradyanın zaman içinde geri yayılırken küçülmesidir. Gradyan değeri aşırı derecede küçük olursa, öğrenmeye çok fazla katkıda bulunmaz. Unutulmakta olan ya da sistemden atılan bu bilgileri daha gelişmiş bir bellekte tutan mimariler geliştirilmiştir ve bu mimarilerden biri olan LSTM, rastgele zamanlarda hatırlama gerçekleştirerek RNN'lerde oluşan uzun süreli unutma problemlerini ortadan kaldırmıştır.

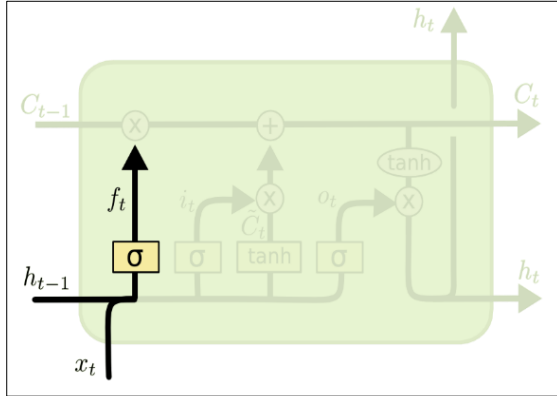
Uzun kısa süreli bellek ağları, kaybolan gradyan problemini çözmek için kullanılan bir RNN yöntemidir [49]. Hava kirliliğini tahmin etmek için en çok kullanılan yöntemlerden biridir. LSTM, hücrenin içine ve dışına bilgi akışını düzenleyen giriş kapısı, çıkış kapısı ve unutma kapısı olmak üzere 3 ana bölümden oluşur. Bu parçalar, standart RNN ile gizli katmandaki nöronlar yerine bellek bloklarında bulunur. Bu kapılar ve hücre sayesinde hangi bilginin hatırlanacağı, hangisinin unutulacağı belirlenir. Yapısında bulunan hücre, değerleri rastgele zamanlarda hatırlar. Zaman serilerinde oldukça sık tercih edilmektedir. Şekil 2.8.'de LSTM Mimarisi gösterilmektedir.



Şekil 2.8. LSTM Mimarisinin yapısı [45,47]

LSTM hücresinin içinde bulunan dört kapı ve işlevleri kısaca açıklanmıştır:

Unutma kapısı (forget gate)



Şekil 2.9. Unutma kapısı [45,47]

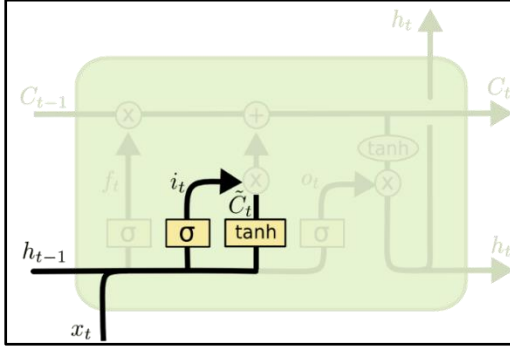
Hangi bilgilerin unutulacağı veya tutulacağı kararına varan kapıdır. Önceki hücreden gelen bilgi(h_t) ve mevcut bilgi(x_t) sigmoid aktivasyon fonksiyonuna sokulur ve sonucuna göre karar verilir. 0 olan bilgiler unutulur ve 1 olan bilgiler Hücre durumu ile taşınmaya devam eder [49].

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (2.4)$$

Kapılar, sigmoid aktivasyonları içerir. Bir sigmoid aktivasyonu, tanh aktivasyonuna benzemektedir. Yani -1 ile 1 arasındaki değerleri ezmek yerine, 0 ile 1 arasındaki değerleri

ezer. Ağ, hangi verilerin önemli olup olmadığını öğrenebilir, önem derecesine göre unutulabilir veya hafızada tutabilir.

Girdi kapısı (input gate)



Şekil 2.10. Girdi kapısı [45,47]

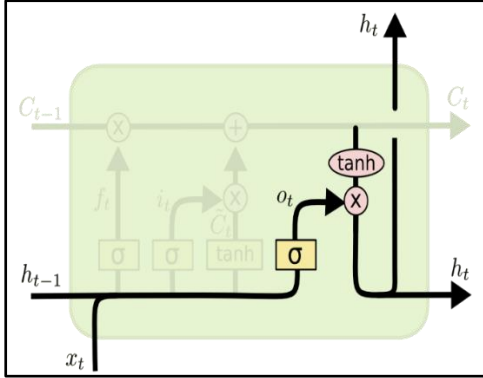
Hücre durumu güncellemesi yapar. Önceki ve mevcut bilginin sigmoid işlemi sonucuna göre güncelleme yapıp yapılmayacağı sonucuna varılır. 0 olan bilgi önemsiz ve 1 olan bilgi önemli olarak kabul edilir. Ayrıca bu kapıda ağı düzenleme işlemi için veriyi -1 ve 1 aralığına sıkıştıran tanh aktivasyon fonksiyonu da kullanılır. Daha sonrasında sigmoid ve tanh fonksiyonu çıktıları çarpılır ve hangi bilginin güncelleneceği kararına varılır [49].

Tanh aktivasyonu, ağ üzerinden akan değerleri düzenlemeye yardımcı olmak için kullanılır. Tanh işlevi, değerleri her zaman -1 ile 1 arasında olacak şekilde sıkıştırır.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2.5)$$

$$\tilde{C} = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (2.6)$$

Çıktı kapısı (output gate)



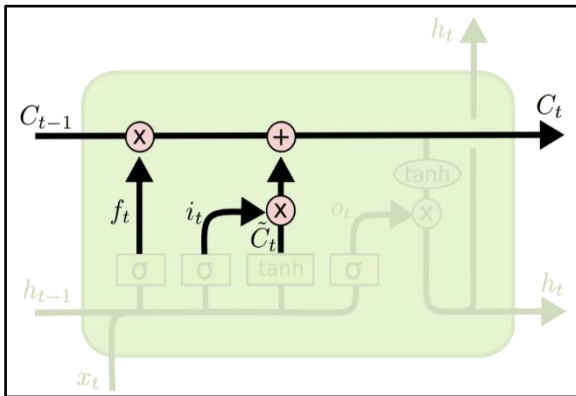
Şekil 2.11. Çıktı kapısı [47]

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (2.7)$$

$$h_t = o_t * \tanh(C_t) \quad (2.8)$$

Çıkış kapısı ise bir sonraki hücrenin girişini(h_{t+1}) belirler. Ayrıca tahmin yapmak için de kullanılır. Öncelikle önceki hücreye ait bilgiyi ve mevcut girişin bilgisini sigmoid fonksiyondan geçirilir. Daha sonra Hücre durumu üzerinde var olan bilgiyi tanh fonksiyonundan geçirilir. Son olarak iki sonucu çarparak hangi bilgilerin bir sonraki hücre için giriş(h_{t+1}) olacağına karar verilir. Mevcut hücre için kapı işlemleri tamamlandığında bir sonraki hücreye gidecek olan Hücre durumu ve hücrenin giriş bilgisi olarak tanımladığı giriş(h_t) bilgilerine karar verilmiş olur [49].

Hücre durumu (cell state)



Şekil 2.12. Hücre durumu [47]

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (2.9)$$

Tahmin yapmak için anlamlı bilgileri hücreler boyunca taşıyan bir iletişim hattı ve ağın hafızası olarak açıklanabilir. Bu sayede kısa süreli bellek problemi çözülerek eski veriler ağ zinciri boyunca taşınabilir. Hücre durumunun bu yolculuğu boyunca taşınması gereken bilgiler ise kapılar aracılığı ile belirlenir. Bu kapılar hangi bilginin gerekli veya gereksiz olduğunu belirleyebilir. Kapılar ise verileri 0 ile 1 aralığına sıkıştıran sigmoid aktivasyon fonksiyonunu kullanır. Sigmoid aktivasyonu sonucunda 0 olan bilgiler unutulur ve 1 olan bilgiler ise hücre durumu ile ilerlemeye devam eder.

2.4.2. Literatürde yapılan benzer çalışmalar

Hava kirliliği problemlerini çözmek ve bu süreci yönetmek zaman gerektirmektedir. Hava kirliliğinin veya kalitesinin tahmini, bu problemlerin meydana getirdiği zararların önlenmesine ya da en aza indirilmesine yardımcı olmaktadır. Bu tahminlerin zamanında ve önceden gerçekleştirilmesi, oluşabilecek ciddi hava kirliliği olaylarının meydana gelmeden engellenebilmesinde ve koruyucu tedbirler alınmasında büyük önem taşımaktadır.

Literatürdeki çalışmalar incelendiğinde geleneksel istatistiksel yöntemler, maliyet ve tahmin performansı yönünden sınırlılıklar içermektedir [50]. Son zamanlarda birçok alanda uygulanan ve popülerlik kazanan makine öğrenmesi yöntemleri, hava kirliliği ve tahmini için de uygulanabilmektedir. Hava kirliliği ve hava kalitesinin tahmini üzerine farklı ülke ve şehirlere ait kirletici konsantrasyon verileri ele alınarak çeşitli çalışmalar gerçekleştirilmiştir.

Deleawe ve arkadaşları, akıllı ortamlarda hava kalitesinin bir göstergesi olarak CO₂ seviyelerini tahmin etmek için makine öğrenimi teknolojilerinin kullanımını araştırmışlardır. Akıllı ortamlarda sensör bilgilerini toplayarak hava kalitesi seviyeleri arasındaki ilişkiyi analiz etmişlerdir. KA, NB ve YSA kullanarak geliştirdikleri hava kalitesi tahmin modelini üç fiziksel test ortamı kullanılarak değerlendirmişlerdir. En yüksek hava kalitesi tahmin başarısını KA yöntemiyle Kyoto'ya ait ikinci seviye test ortamında %96,66 ile elde etmiştir. Tokyo ve Tulum test ortamında da KA yöntemiyle sırasıyla %91,31 ile %89,46 doğruluk oranı takip etmiştir. Bu çalışmanın gelecekte geliştirilmesi ile hava kalitesi tahmin

edildiğinde, otomatik havalandırma ve havanın saflaştırılması gibi görevleri otomatikleştirmek için kullanılabileceği önerisinde bulunulmuştur [51].

Zhao ve arkadaşları, SO_2 ve NO_2 tahmininde, orijinal veri setinden bir faktör alt veri setini seçmek için Genetik Algoritmanın (GA) kullanıldığı ve GA tarafından seçilen faktörlerin modelleme ve test için YSA'ya beslendiği GA-YSA adı verilen geliştirilmiş bir YSA modeli önermiştir. 2003-2006 yılları arasında Çin'in Tianjin kentine ait hava kalitesi izleme verileri ve meteorolojik veriler modelin eğitimi için, 2007 yılı verileri ise performans değerlendirmesi için test verisi olarak kullanılmıştır. GA-YSA'nın korelasyon katsayıları, izleme ve tahmin arasında hesaplanan değerler, SO_2 (kükürt dioksit) ve NO_2 (azot dioksit) tahmininde diğer kıyaslanan modellerden daha yüksektir. Elde edilen sonuçlar, GA-YSA modelinin hava kalitesi tahmininde diğer iki modelden daha iyi performans sergilediğini göstermektedir [52].

Kaplan ve arkadaşları, Kütahya iline ait saatlik alınan PM_{10} , SO_2 , rüzgar ve nem verilerinden oluşan veri setine üç katmandan oluşan YSA uygulamasıdır. Tanjant ve sigmoid aktivasyon fonksiyonu olarak seçilmiştir. Ayrıca ağırlıkların değişimi için Marquart öğrenme yöntemini kullanmışlardır. Kullanılan bu modelde gerçekleştirilen tahmin sonucu PM_{10} , SO_2 için sırasıyla OKH değerleri 0,0161 ve 0,0372 'dir. Geliştirdikleri bu modelin diğer kirletici konsantrasyonlarının tahmininde de faydalı olabileceği çıkarımında bulunmuşlardır [42].

Rybarczyk ve Zalakeviciute, Quito kentine ait iki bölgedeki $\text{PM}_{2,5}$, yağış ve rüzgar değişkenlerini ele alarak makine öğrenmesi algoritmasına dayalı bir model geliştirmiştir. $\text{PM}_{2,5}$ seviyesini tahmin etmeye çalıştıkları bu modelde karar ağacı algoritması kullanarak %65 doğruluk elde etmişlerdir [53].

Yu ve arkadaşları, çalışmalarında kentsel algılama sistemlerinde hava kalitesini tahmin etmek için makine öğrenmesi yöntemlerinden Rastgele Orman yaklaşımı ile geliştirdikleri bir modeli önermişlerdir. Kentsel algılama tarafından üretilen veriler, meteoroloji verilerini, yol bilgilerini, gerçek zamanlı trafik durumunu içermektedir. Shenyang'da bulunan toplam 11 hava kalitesi izleme istasyonundan alınan kirletici verilerine uygulanan modelde Rastgele Orman algoritması, HKİ tahmini için %81'lik başarı ile sonuçlanmıştır. Geliştirilen modelin diğer şehirler için kolaylıkla uygulanabileceği çıkarımında bulunmuşlardır [54].

Kök ve arkadaşları, Nesnelerin İnterneti (IoT) tabanlı akıllı şehirlerde hava kalitesi tahmini için O_3 ve NO_2 parametrelerini kullanarak bir derin öğrenme modeli olan LSTM'i kullanmışlardır. Kullandıkları bu model ile O_3 kirleticisi için KOKH 3,26, OMH 2,81, NO_2 kirleticisi için KOKH 3,76, OMH 3,11 değerlerini elde etmişlerdir. Önerilen modelin değerlendirme sonuçlarının umut verici olduğu ve modelin diğer akıllı şehir tahmin problemlerinde de kullanılabileceği beklentilerini ifade etmişlerdir [55].

Bui ve arkadaşları, derin öğrenme yaklaşımı olan LSTM kullanarak Güney Kore'de hava kirliliğini tahmin etmek için yeni bir model geliştirmişlerdir. Yaptıkları bu çalışmada kodlayıcı-kod çözücü (Encoder-Decoder (En-De)) modeli ile RNN ve LSTM kullanarak $PM_{2,5}$ seviyelerini tahmin etmeye çalışmışlardır. En-De, modelin doğruluğunu artırmak için birden fazla RNN katmanının üst üste istiflenmesine izin verir. Bu model, 2008'den 2018'e kadar birden fazla kaynaktan alınan verilerle oluşturulan iki veri seti üzerinde uygulanmış ve En-De kullanılarak 24 saatlik veride KOKH 13,54 olarak ölçülmüştür [44].

Reddy ve arkadaşları, Pekin'deki kirlilik ve meteorolojik bilgilerinden oluşan verilerine dayanarak gelecekte tahmin yapmak için LSTM modelini kullanmışlar ve elde ettikleri sonuçlar doğrultusunda tek bir zaman adımı için temel destek vektörü regresyonuna kıyasla gelecekteki zaman adımlarını tahmin ederken eşdeğer doğruluk ürettiğini gözlemlemişlerdir. Kullanılan veriler 35 istasyona ait 2010-2017 yılları arasında olup 2010-2012 yılları arasındaki veriler modelin eğitilmesinde, 2013 verileri modelin konfigürasyonunda, 2014 ve 2015-2017 iki ayrı veri seti olmak üzere modelin test edilmesinde kullanılmıştır. Uygulanan model sonucunda KOKH=44,15 (5saatlik) $R^2=0,689$ (5 saatlik), KOKH=74,8 (10 saatlik) $R^2=0,58$ (10 saatlik) değerleri elde edilmiştir. Elde edilen sonuçlar üzerine gelecek çalışmalarda öğrenme katsayısı, veri boyutu, LSTM hücre sayısı gibi birçok parametrenin daha düşük KOKH değerleri döndürmesi amacıyla optimize edilebileceği çıkarımında bulunmuşlardır [56].

Huang ve Kuo, çalışmalarında $PM_{2,5}$ kirletici konsantrasyonunu izlemek ve tahmin etmek için CNN ve LSTM algoritmalarını birleştirerek hibrit bir model kullanmışlardır. Makine öğrenmesi teknikleri (DVM, RO, KA, CNN, MLP, LSTM) ile kıyaslamışlar ve yaptıkları bu çalışmada önerilen CNN-LSTM modelinin tahmin doğruluğunun $R^2=0,95$ KOKH=24,22, OKH=14,63 değerleri ile en yüksek olduğunu tespit etmişlerdir. Bu çalışmanın gelecekte, akıllı şehirlerde hava kirliliğinin tahmin edilmesi yeteneğini geliştirmede büyük katkı

sağlayacağı düşünülmektedir. Ayrıca $PM_{2,5}$ kirleticisinin önlenmesi ve kontrol altına alınmasında uygulanabileceği öngörülmektedir [57].

Liu ve arkadaşları, uydu tabanlı yansımadan elde edilen $PM_{2,5}$ ve meteorolojik değişkenleri kullanarak $PM_{2,5}$ seviyesini tahmin eden Rastgele Orman algoritması modelini oluşturmuşlardır. Oluşturulan model ile saatlik $PM_{2,5}$ kirleticisi için $R^2=0,86$ doğruluk oranı elde edilmiştir. Bu çalışmanın günlük $PM_{2,5}$ konsantrasyonunun tahmininde ve bölgesel kirlilik süreçlerinin izlenmesinde faydalı olabileceği çıkarımında bulunmuşlardır [58].

Gültepe, Kastamonu iline ait hava kirliliği ve meteorolojik parametrelerden oluşan veri setine çeşitli makine algoritmaları uygulayarak hava kirliliği tahmininde kullanılacak modeller geliştirmiştir. 2015-2018 yıllarına ait 2500 veriden oluşan veri setinin yaklaşık %70'ini oluşturan 1750 veri eğitim, geriye kalan 750 (%30) veri ise test veri seti olarak kullanılmıştır. YSA, RO, KEYK, NB, KA, Lineer Regresyon ve Lojistik Regresyon tekniklerini uygulamış ve en yüksek başarıyı %99 doğruluk oranı ile RO ve KA algoritmalarında elde etmiştir. En kötü performansı %30 doğruluk oranı ile Lineer Regresyon (LR) algoritmasında elde etmiştir [13].

Ayturan, kirlilik konsantrasyonları ve meteorolojik faktörlerden oluşan 17 parametrelili veri seti ile ileriye yönelik $PM_{2,5}$ konsantrasyonlarının tahmini için Kapalı Tekrarlayan Ünite (Gate Recurrent Unit- GRU) ve Tekrarlayan Sinir Ağları (TSA-Recurrent Neural Network- RNN) derin öğrenme yöntemlerini kullanarak yeni bir model geliştirmiştir. Geliştirilen hibrit modelin veri setine uygulanması sonucu %83 başarı elde etmiştir. Veri setinin boyutunun arttırılması ve kullanılan modelde yapılan bazı değişiklikler sonucu daha yüksek başarı elde edilebileceğini öngörmektedir. Ayrıca bu çalışmanın, gelecekte yapılacak olan daha uzun süreli tahmin performansına sahip modellerin geliştirilmesine katkı sağlayabileceği düşünülmektedir [59].

Xayasouk ve arkadaşları, uzun kısa süreli bellek (LSTM) ve derin otomatik kodlayıcı (DAE) yöntemlerini kullanarak PM konsantrasyonlarını ($PM_{2,5}$ ve PM_{10}) tahmin etmek için modeller geliştirmişler ve model sonuçlarını ortalama karekök hatası (KOKH-RMSE) ile karşılaştırmışlardır. 1 Ocak 2015 ile 31 Aralık 2018 tarihleri arasındaki Güney Kore Seul 'de bulunan 25 istasyondan alınan saatlik hava kalitesi verilerine modeller uygulamış ve LSTM modeli biraz daha iyi performans göstererek, ince PM konsantrasyonlarını etkili bir şekilde tahmin etmiştir. Deneysel sonuçlara göre, LSTM ve DAE modellerinin tahmin

sonuçları; LSTM modelinin minimum KOKH değerleri PM_{10} için 11,113 ve $PM_{2.5}$ için 12,174 iken DAE modeli için PM_{10} için minimum KOKH değerleri 15,038 ve $PM_{2.5}$ için 15,431 olarak ölçülmüştür [60].

Bozdağ ve arkadaşları, Ankara ilindeki 7 istasyondan elde edilen PM_{10} konsantrasyonlarından oluşan veri seti ile seçmiş oldukları makine öğrenmesi algoritmalarından (LASSO, DVM, RO, KEYK, xGBoost, YSA) oluşturdukları modeli eğitmişlerdir. Ankara ilindeki 6 istasyonun 2009-2017 yıllarına ait PM_{10} konsantrasyonları girdi olarak vermişler ve yedinci istasyonun 2018 yılına ait PM_{10} konsantrasyonlarını makine öğrenmesi algoritmaları kullanılarak tahmin edilmeye çalışmışlardır. En iyi sonuç $R^2 = 0,58$, KOKH = 20,8 değerleri ile YSA' da sağlanmıştır. Tahmini konsantrasyon sonuçlarının mekânsal dağılımı Coğrafi Bilgi Sistemi (CBS) aracılığıyla sağlanmış ve hava kirliliğini arazi kullanımı üzerinden iyileştirmeye yönelik mekânsal stratejiler oluşturulmuştur [61].

Faizi, Kabil şehrine ait $PM_{2.5}$, PM_{10} , CO, SO₂, NO₂, O₃ kirleticileri ve sıcaklık, çiy noktası, basınç, nem ve rüzgar hızı meteorolojik parametreleri kullanarak bir veri seti oluşturmuştur. 2019 Ocak ile 2021 Mart tarihleri arasında toplanan 536 günlük kirletici ve meteoroloji verilerinin %70'i eğitim, %30'u test verisi olarak rastgele seçilmiştir. Bu veri seti kullanılarak LR, YSA, DVM, KA ve Gauss Süreç Regresyonu (GSR) algoritmalarıyla modellenmiş ve modellerin başarısı kıyaslanmıştır. Eğitim verilerinde en yüksek $R^2 = 0,97$ değeri ile GSR yöntemi ile elde edilmiştir. Test verileri için en başarılı $R^2 = 0,91$ değeri ile DVM, GSR ve LR yöntemleri olmuştur. Ayrıca tüm model sonuçları incelendiğinde en uyumlu model 0,95 R^2 değeriyle GSR yöntemi olmuştur. En kötü performansı ise LR yöntemi sergilemiştir. Faizi, gelecek çalışmalarda oluşturulacak modele hava kalitesini etkilediği bilinen, ısınma faaliyetleri, sanayi faaliyetleri, nüfus, trafik vb. parametreler giriş değeri olarak eklendiğinde oluşturulan modelin başarısının artacağı öngörüsünde bulunmuştur [62].

Gökçek ve arkadaşları, Kayseri ilindeki üç istasyondan alınan, 2010-2018 yıllarına ait PM_{10} kirletici verilerini kullanarak makine öğrenmesi algoritmalarıyla (KEYK, RO, DVR, YSA, LR) eğitilen bir model geliştirmişlerdir. 2010-2018 yıllarına ait PM_{10} kirletici verileri girdi olarak modele verilmiş ve 2019 yılına ait PM_{10} değerleri tahmin edilmeye çalışılmıştır. En yüksek sonucu DVR ile elde etmişlerdir. Trafik istasyonu için $R^2 = 0,85$, KOKH = 17,57,

OMH = 10,17; Hürriyet istasyonu için $R^2 = 0,73$, KOKH = 34,91, OMH = 24,61 ve OSB istasyonu için $R^2 = 0,82$, KOKH = 41.71, OMH = 21,62 olarak ölçülmüştür. Bu çalışmanın, nüfusu kalabalık şehirlerin hava kalitesinin makine öğrenmesi algoritmaları ile modellenmesi ve arazi kullanım planlamasıyla alakalı öneriler sunduğu suretiyle literatüre katkıda bulunduğu belirtilmiştir [63].

Ünaldı ve Yalçın, İstanbul- Başakşehir ilçesinde 2016-2021 yıllarına ait havada ölçülen günlük kirletici konsantrasyonları (PM_{10} , CO, SO_2 , NO_2 , O_3) ve meteorolojik parametre (rüzgar yönü ve hızı, basınç) değerlerinden oluşan bir veri seti oluşturmuşlardır. Oluşturdukları bu veri seti üzerinde Çoklu Doğrusal Regresyon (ÇDR), DVM, KEYK, KA, RO ve YSA makine öğrenme yöntemlerini kullanarak PM_{10} , CO, SO_2 , NO_2 ve O_3 konsantrasyonlarını tahmin etmişlerdir. PM_{10} , CO, SO_2 ve O_3 kirletici konsantrasyonlarında RO yöntemi sırasıyla; $R^2 = 0,68$, $R^2 = 0,97$, $R^2 = 0,96$, $R^2 = 0,79$, $R^2 = 0,79$, NO_2 de ise ÇDR yöntemi $R^2 = 0,99$ en iyi sonucu vermiştir. Uygulanan modelin geliştirilmesi aşamasında, veri seti trafik ve daha fazla meteoroloji verileriyle genişletilerek doğruluğu artırılıp kirlilik seviyesinin tahmininde gerçek zamanlı tahminler gerçekleştirebileceği öngörülmektedir [64].

Çizelge 2.2. Hava kirliliği tahmini için makine öğrenmesi kullanan çalışmaların kıyaslanması

Seçilen Model	Kullanılan Parametreler	Değerlendirme Ölçütü ve Performansı	Çalışma
KA, NB, YSA	CO ₂	Doğruluk sırasıyla (KA): %89,46(Tulum) %91,31(Tokyo) %96,66(Kyoto)	[51]
Genetik Algoritma-YSA	SO ₂ , NO ₂	Mape: 18,60	[52]
YSA Levenberg-Marquardt	Rüzgar, nem, PM ₁₀ , SO ₂	OKH= PM ₁₀ =0,01 SO ₂ =0,03	[42]
KA	PM _{2,5} , yağış ve rüzgar	Doğruluk=%65	[53]
RO, NB, Lojistik Regresyon, KA, YSA	CO, NO ₂ , PM ₁₀ , SO ₂ , O ₃ , PM _{2,5} , HKİ ve meteorolojik veriler	Doğruluk sırasıyla: %81,5, %52,1, %66,2, %77,4, %71,8	[54]
LSTM	O ₃ , NO ₂	KOKH=3.26 (O ₃) OMH=2.81 (O ₃) KOKH=3.76 (NO ₂) OMH=3,11 (NO ₂)	[55]
LSTM	PM _{2,5}	KOKH 13,54 (24 saat)	[44]

Çizelge 2.2. (devam) Hava kirliliği tahmini için makine öğrenmesi kullanan çalışmaların kıyaslanması

Seçilen Model	Kullanılan Parametreler	Değerlendirme Ölçütü ve Performansı	Çalışma
LSTM	PM _{2,5}	KOKH=44,15 (5saatlik) R ² =0,689 (5 saatlik) KOKH=74,8 (10 saatlik) R ² =0,58 (10 saatlik)	[56]
CNN, LSTM	PM _{2,5}	KOKH sırasıyla: 26,37, 25,27	[57]
RO	PM _{2,5}	R ² =0,86	[58]
YSA, RO, KEYK, Lojistik regresyon, KA, LR, NB	PM ₁₀ , SO ₂ , sıcaklık, Basınç, Nem, Rüzgar yönü ve hızı	R ² sırasıyla: %87, %99, %97, %87, %99, %30, %94	[13]
RNN-GRU	PM _{2,5} ve meteorolojik parametreler	Doğruluk=%83	[59]
LSTM-DAE	PM ₁₀ , PM _{2,5} ve meteorolojik parametreler	LSTM için KOKH: PM ₁₀ =11,113 PM _{2,5} = 12,174 DAE için KOKH: PM ₁₀ =15,038 PM _{2,5} = 15,431	[60]
LASSO, DVM, RO, KEYK, xGBoost, YSA	PM ₁₀	YSA için: R ² = 0,58 KOKH = 20,8	[61]
YSA, DVM, KA, GSR	PM _{2,5} , PM ₁₀ , CO, SO ₂ , NO ₂ , O ₃ kirleticileri ve sıcaklık, çiy noktası, basınç, nem ve rüzgar hızı	GSR için: R ² =0,95	[62]
KEYK, RO, DVR, YSA, LR	PM ₁₀	Trafik istasyonu için DVR ile; R ² = 0,85, KOKH = 17,57, OMH = 10,17; Hürriyet istasyonu için DVR ile; R ² = 0,73, KOKH = 34,91, OMH = 24,61 ve OSB istasyonu için R ² = 0,82, KOKH = 41.71, OMH = 21,62	[63]
ÇDR, DVM, KEYK, KA, RO ve YSA	PM ₁₀ , CO, SO ₂ , NO ₂ , O ₃ ve meteorolojik parametreler	PM ₁₀ R ² =0,68(RO) CO R ² =0,97(RO) SO ₂ R ² =0,96(RO) NO ₂ R ² =0,99(ÇDR) O ₃ R ² =0,79(RO)	[64]

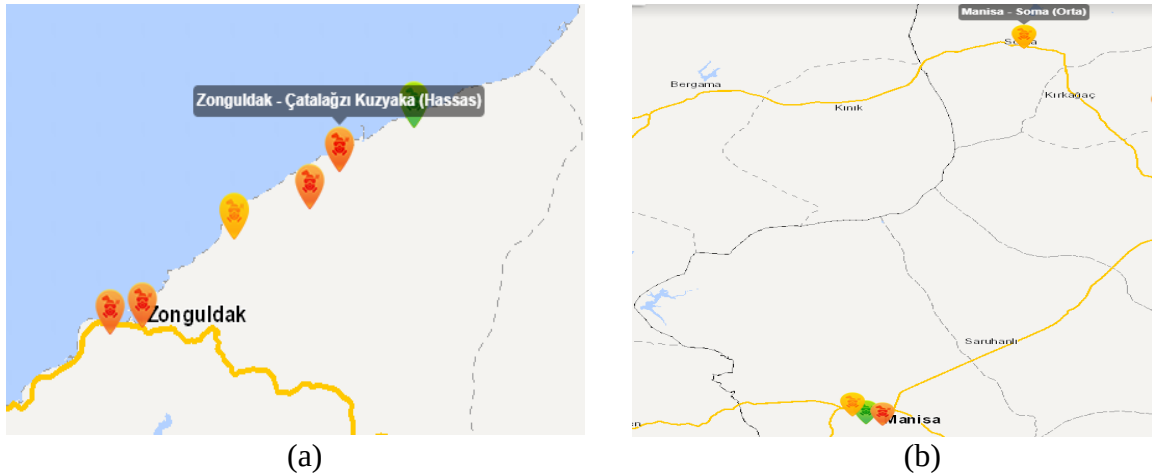
3. MATERYAL VE YÖNTEM

3.1. Veri Seti ve Özellikleri

Bu bölümde kullanılan verinin özellikleri, veri ön işleme süreçlerinden verinin seçilme, temizleme ve normalleştirme süreçleri açıklanmıştır.

3.1.1. Çalışma alanı ve istasyon konumu

Hava Kalitesi İzleme İstasyonu'ndan Zonguldak-Kuzyaka ve Manisa-Soma istasyonlarına ait PM_{10} ve SO_2 kirletici konsantrasyon verileri Çevre ve Şehircilik Bakanlığı'nın Ulusal Hava İzleme Ağı veri bankası kullanılarak elde edilmiştir. Termik santrallerin yer aldığı sanayi alanlarını içeren Zonguldak-Kuzyaka ve Manisa-Soma, PM_{10} ve SO_2 kirliliğinin yoğun yaşandığı bölgelerdir. Şekil 3.1'de hava kalitesi verilerinin temin edildiği istasyonlara ait konum görselleri verilmiştir.



Şekil 3.1. a,b) Zonguldak ve Manisa ili hava kalitesi verilerinin temin edildiği istasyonlar

3.1.2. Çalışma alanına ait iklimsel özellikler

Manisa'nın iklimi, yaz mevsimi kurak ve sıcak, kış mevsimi genellikle yağışlı ve ılıman geçmektedir ve Ege kıyılarında bulunan diğer illere göre daha serittir. Genellikle dağlık alanlardan oluşan kuzey bölgesi, yazları serin olup kışları soğuk geçirmektedir. Manisa-Soma ilçesinde özellikle kış ayları PM_{10} ve SO_2 kirliliği gözlenmektedir [65].

Zonguldak genellikle Karadeniz'in ılıman iklimine sahiptir. Aralık ve Ocak aylarında yoğun yağış almaktadır. Haziran-Ağustos ayları arasında ise en güneşli günler yaşanmaktadır. Deniz kıyılarından iç bölgelere geçildikçe iklim sertleşir. Yıllık ortalama sıcaklık değerlerinde önemli bir fark gözlemlenmemektedir. Kış aylarında düşen sıcaklık nedeniyle hava kirliliğinde ciddi bir artış mevcuttur. Kış ve sonbahar aylarında rüzgarlar daha fazla belirgin yazın rüzgarın etkisi azalmaktadır. Soğuk havalarda artan basıncın etkisi ile yerleşim merkezlerinin kirli hava kütleleri uzun süre etkisini sürdürebilmektedir [66].

3.1.3. Veri seti ve temini

Bu çalışmada kullanılan hava kirleticilerine ait 2016 Aralık ve 2021 Nisan ayları dahil tüm veriler Çevre Şehircilik Bakanlığının Hava Kalitesi İzleme İstasyonu Sürekli İzleme Merkezi'nin paylaştığı istasyon bazlı veri indirme bankasından saatlik olarak alınmıştır [5]. Meteorolojik veriler ise Meteoroloji Genel Müdürlüğü'nden temin edilmiştir[6]. Veriler tarih, PM₁₀ (µg/m³), SO₂(µg/m³), CO(µg/m³), NO₂(µg/m³), NO_x(µg/m³), O₃(µg/m³) konsantrasyonları, hava sıcaklığı(°C), basınç (mbar), rüzgar hızı(m/s), yağış (kg/m²) ve nem (%) şeklindedir. İstasyondan alınan veriler “.xsl” uzantılı excel dosyası olup verilerin bir örneği tabloda verilmiştir.

Çizelge 3. 1. İstasyondan alınan verilerin örneği

Parametreler	Manisa-Soma	Zonguldak-Kuzyaka
Tarih-saat	1.01.2017 9:00	1.01.2017 9:00
PM ₁₀ (µg/m ³)	155,35	42,03
SO ₂ (µg/m ³)	516,31	24,49
CO(µg/m ³)	1531,42	29,18
NO ₂ (µg/m ³)	139,02	23,36
NO _x (µg/m ³)	220,34	447,25
O ₃ (µg/m ³)	2,44	10,21
Hava sıcaklığı(°C)	12,1	13,51
Basınç (mbar)	1006,2	26,88
Rüzgar hızı(m/s)	2	1,5
Nem (%)	78,7	92
Yağış(kg/m ²)	0,1	14,8

Python programlama dili kullanılarak veri birleştirme, temizleme, dönüştürme vb. veri ön işleme operasyonları gerçekleştirilmiştir. Öncelikle txt uzantısıyla MGM’ den alınan meteorolojik veriler aynı tarih ve saat baz alınarak tek bir dosyada birleştirilmiştir. Daha sonra xlsx uzantısı ile ÇŞB veri bankasından elde edilen kirleticilerle xlsx uzantılı bir dosyada birleştirilmiştir. Bu işlemler sırasında Pandas kütüphanesinden faydalanılmıştır. Çekilen verilerde bulunan şehir ve ilçe bilgileri ihtiyaç olmaması sebebiyle veri setinden çıkarılmıştır.

```
import os,sys
import json
import pandas as pd
from datetime import timedelta

file_name = "manisaSonVersiyon.xlsx"
dfs = pd.read_excel(file_name, sheet_name=None)
dfs_df = dfs['Sheet1']

dfs_df['Tarih'] = dfs_df['Tarih'].apply(
    lambda x: x.replace(minute=0, second=0))

#import pdb; pdb.set_trace()
src_path = "C:\\Users\\busra\\Desktop\\django\\data\\"
data_file_list = os.listdir(src_path)
#print(data_file_list)

for i in data_file_list:
    df = pd.read_csv(f"{src_path}{i}", delimiter=",")
    df = df[df['Istasyon_No'] == 18040].copy()
    #df = pd.read_csv(f"{src_path}202106290616-Saatlik Toplam Yağış (mm=kg÷m²) OMĞİ.txt", delimiter=",")

    df['date'] = df['YIL'].astype(
        str) + "/" + df['AY'].astype(str) + "/" + df['GUN'].astype(str) + "/" + df['SAAT'].astype(str)

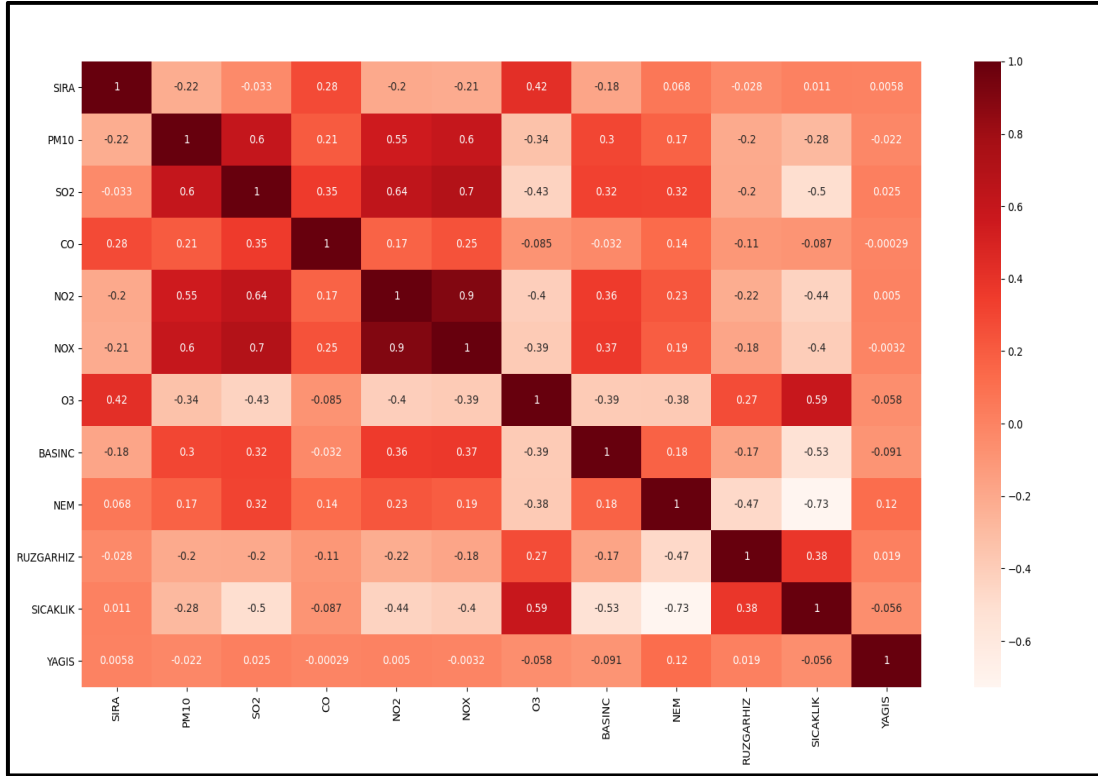
    df["Tarih"] = pd.to_datetime(df['date'], format=""%Y/%m/%d/%H"")

    #import pdb; pdb.set_trace()
    selected_cols = ["Tarih", df.columns[-3]]
    dfs_df = dfs_df.merge(df[selected_cols], on="Tarih",
        how='left').drop_duplicates()

    print(dfs_df)
dfs_df.to_excel('manisa.xlsx')
```

Şekil 3.2. Manisa iline ait verilerin birleştirilmesi

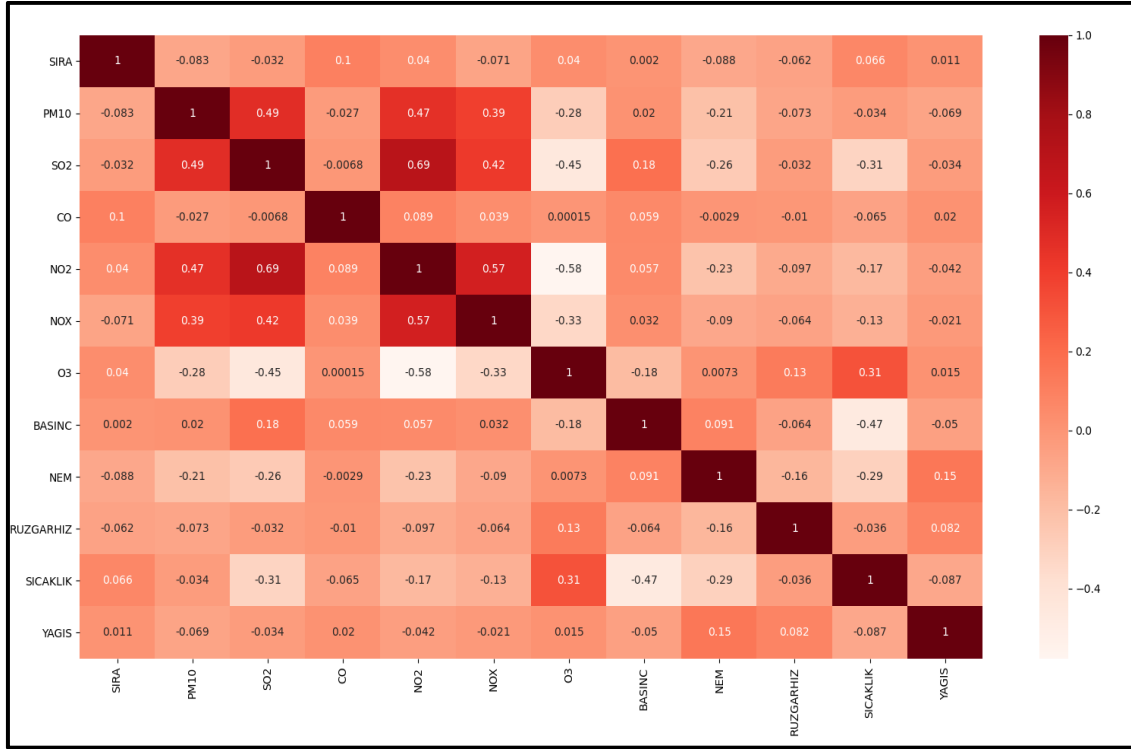
Daha sonra Manisa ve Zonguldak illerine ait iki veri seti kullanılarak korelasyon matrisleri yardımıyla parametreler arasındaki ilişkiler incelenmiştir. (Şekil3.3 ve Şekil 3.4)



Şekil 3.3. Manisa iline ait korelasyon matrisi

Şekil 3.3'te bulunan parametreler arasındaki ilişki güçleri $[-1,1]$ aralığında gösterilmiştir. Parametreler arasındaki ilişkiler incelendiğinde Manisa iline ait PM_{10} parametresine ait en güçlü ilişki 0,60 değeriyle SO_2 ile olan ilişkisine ait olduğu görülmektedir. Daha sonra sırasıyla NO_X ve NO_2 takip etmektedir. En düşük korelasyon ilişkisi ise -0,022 değeri ile yağış parametresine aittir.

SO_2 parametresinin diğer parametrelerle ilişkisine bakıldığında en güçlü ilişki 0,70 değeri ile NO_X 'e aittir. Daha sonra sırasıyla NO_2 ve PM_{10} takip etmektedir. Negatif yönde en yüksek korelasyon ilişkisi ise -0,50 değeri ile sıcaklık parametresine aittir. Yani hava sıcaklığı düştükçe atmosfere salınan SO_2 miktarı artmaktadır.



Şekil 3.4. Zonguldak iline ait korelasyon matrisi

Şekil 3.4’te bulunan parametreler arasındaki ilişki güçleri $[-1,1]$ aralığında gösterilmiştir. Parametreler arasındaki ilişkiler incelendiğinde Zonguldak iline ait PM_{10} parametresinin en güçlü ilişki 0,49 değeri ile SO_2 ile olan ilişkisine ait olduğu görülmektedir. Daha sonra sırasıyla NO_2 ve NO_x takip etmektedir. Negatif yönde en yüksek korelasyon ilişkisi ise -0,28 değeriyle O_3 parametresine aittir. SO_2 parametresinin diğer parametrelerle ilişkisine bakıldığında en güçlü ilişki 0,69 değeriyle NO_2 ’e aittir. Daha sonra sırasıyla PM_{10} ve NO_x takip etmektedir. Negatif yönlü en yüksek korelasyon ilişkisi ise -0,45 değeriyle O_3 parametresine aittir.

Veri setinin istatistiksel nitelikleri veri ön işleme ve analizinde büyük önem taşımaktadır. Pandas kütüphanesi kullanılarak parametrelere ait ortalama, standart sapma, minimum ve maksimum değerler hesaplanmıştır. Manisa ve Zonguldak illeri 2016-2021 yılları arasındaki hava kirleticileri ve meteorolojik parametrelere ait tanımlayıcı istatistikler Çizelge 3.2 ve Çizelge 3.3’te verilmiştir.

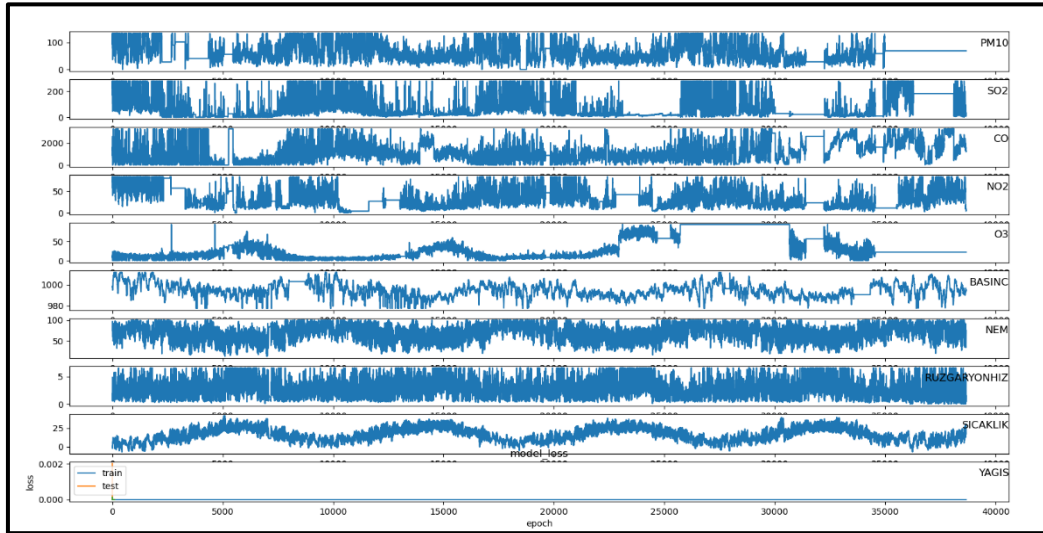
Çizelge 3.2. Manisa kirletici parametrelere ait tanımlayıcı istatistikler

Parametreler	Ortalama	Standart sapma	Maksimum	Minimum
PM ₁₀ (µg/m ³)	66,99	45,08	1524,29	0
SO ₂ (µg/m ³)	88,05	113,52	1335,90	0
CO(µg/m ³)	1165,18	940,13	26478,87	0
NO ₂ (µg/m ³)	35,06	25,29	223,4	0
NO _x (µg/m ³)	49,60	43,60	759,04	0,14
O ₃ (µg/m ³)	21,06	19,45	208,16	0
Basınç (mbar)	994,68	6,43	1017,20	957,20
Nem (%)	69,71	19,40	100	13
Rüzgar hızı(m/s)	2,78	1,67	13,60	0
Hava sıcaklığı(°C)	16,03	8,98	41,80	-6
Yağış(kg/m ²)	0,06	0,49	24,90	0

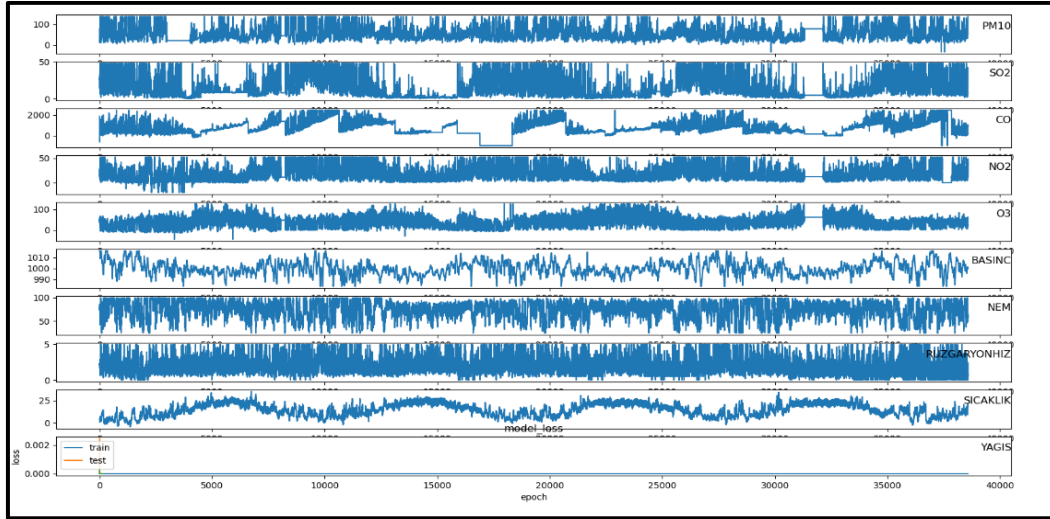
Çizelge 3.3. Zonguldak kirlетici parametrelerе ait tanımlayıcı istatistikler

Parametreler	Ortalama	Standart sapma	Maksimum	Minimum
PM ₁₀ (µg/m ³)	58,50	37,60	321,64	-161,31
SO ₂ (µg/m ³)	17,19	85,64	574,82	0,03
CO(µg/m ³)	3793,88	609,94	9999,99	-1040,44
NO ₂ (µg/m ³)	19,37	16,49	138,52	-51,88
NO _x (µg/m ³)	26,50	36,76	829,13	-100,36
O ₃ (µg/m ³)	42,43	28,47	254,79	-80,03
Basınç (mbar)	999,98	6,23	1023,770	975,50
Nem (%)	74,58	18,39	100	8
Rüzgar hızı(m/s)	2,17	1,32	16,90	0
Hava sıcaklığı(°C)	14,48	7,10	35,5	-4,50
Yağış(kg/m ²)	0,13	0,83	64	0

Verilerin ham hali işlendikten sonra Şekil 3.5 ve Şekil 3.6’ da Manisa ve Zonguldak illerine ait meteorolojik ve kirlетici parametrelerin grafiksel gösterimi verilmiştir.



Şekil 3.5. Manisa iline ait meteorolojik ve kirlетici parametrelerin grafiksel gösterimi



Şekil 3.6. Zonguldak iline ait meteorolojik ve kirletici parametrelerin grafiksel gösterimi

3.1.4. Veri setinin hazırlanması ve normalleştirilmesi

Manisa ve Zonguldak illeri için iki ayrı veri seti oluşturulmuştur. Zonguldak iline ait veri setinde 2016 yılının sadece Aralık ayında ölçüm yapılmasından dolayı Manisa iline ait veri seti de 2016 Aralık ayından başlatılmıştır. İlk olarak veri setinde bulunan tüm özellikler fit fonksiyonuna sokulmuştur. Verilerdeki uyuşmazlığı gidermek amacıyla tüm veriler float64 tipine çevrilmiştir. (Şekil 3.7)

```
RangeIndex: 38576 entries, 0 to 38575
Data columns (total 12 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   SIRA             38576 non-null  int64
1   PM10            35792 non-null  float64
2   SO2             36823 non-null  float64
3   CO              35078 non-null  float64
4   NO2             35521 non-null  float64
5   NOX             35521 non-null  float64
6   O3              36807 non-null  float64
7   BASINC          38487 non-null  float64
8   NEM             38418 non-null  float64
9   RÜZGARHIZ       38560 non-null  float64
10  SICAKLIK        38550 non-null  float64
11  YAGIS           38502 non-null  float64
```

Şekil 3.7. Zonguldak iline ait özniteliklerin formatı

Oluşturulan veri setlerinde eksik değerler olduğu görülmüş ve eksik değere sahip tüm satırlar veri setinden çıkarılmıştır. Örnek olarak Zonguldak iline ait veri setinden eksik veriler

çıkarılmadan önce ve çıkarıldıktan sonra özniteliklerin özellikleri paylaşılmıştır. (Şekil 3.8.a-c)

	SIRA	PM10	SO2	CO	NO2	NOX	O3	BASINC	NEM	RUZGARHIZ	SICAKLIK	YAGIS
count	38576.000000	35792.000000	36823.000000	3.507800e+04	35521.000000	35521.000000	36807.000000	38487.000000	38418.000000	38560.000000	38550.000000	38502.000000
mean	19287.500000	58.509794	17.192086	3.793388e+04	19.373044	26.508892	42.431074	999.982573	74.587771	2.171131	14.486327	0.132881
std	11136.076329	40.174995	20.018364	6.099411e+05	16.491388	36.763172	28.478596	6.231349	18.399610	1.329902	7.100366	0.831272
min	0.000000	-161.310000	0.030000	-1.040449e+04	-51.880000	-100.360000	-80.030000	975.500000	8.000000	0.000000	-4.500000	0.000000
25%	9643.750000	31.750000	4.170000	4.007700e+02	7.570000	9.380000	19.705000	995.800000	63.000000	1.200000	8.600000	0.000000
50%	19287.500000	47.340000	9.750000	7.408400e+02	14.210000	17.010000	41.130000	999.600000	78.000000	1.900000	14.500000	0.000000
75%	28931.250000	72.992500	22.850000	1.221425e+03	26.650000	32.190000	61.380000	1003.900000	89.000000	2.700000	20.400000	0.000000
max	38575.000000	853.190000	568.280000	9.999999e+06	138.520000	829.130000	254.790000	1023.700000	100.000000	16.900000	35.500000	64.000000

a)

	Eksik Değerler	% Değeri
SIRA	0	0.000000
PM10	2784	7.216922
SO2	1753	4.544276
CO	3498	9.067814
NO2	3055	7.919432
NOX	3055	7.919432
O3	1769	4.585753
BASINC	89	0.230713
NEM	158	0.409581
RUZGARHIZ	16	0.041477
SICAKLIK	26	0.067399
YAGIS	74	0.191829

	Eksik Değerler	% Değeri
SIRA	0	0.0
PM10	0	0.0
SO2	0	0.0
CO	0	0.0
NO2	0	0.0
NOX	0	0.0
O3	0	0.0
BASINC	0	0.0
NEM	0	0.0
RUZGARHIZ	0	0.0
SICAKLIK	0	0.0
YAGIS	0	0.0

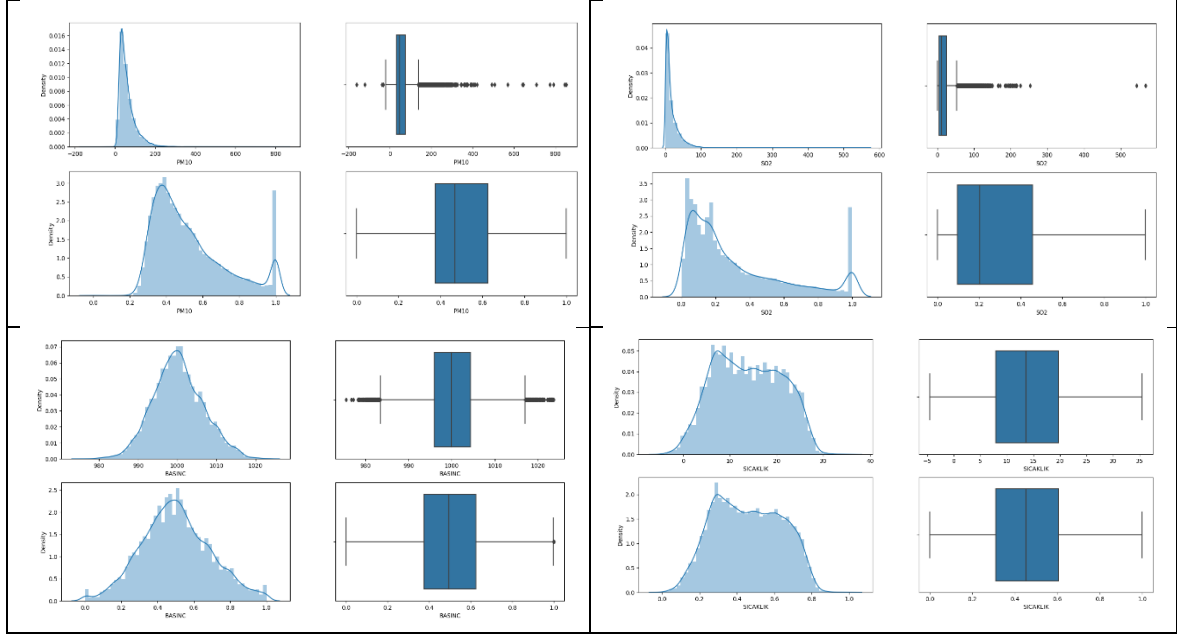
b)

	SIRA	PM10	SO2	CO	NO2	NOX	O3	BASINC	NEM	RUZGARHIZ	SICAKLIK	YAGIS
count	29937.000000	29937.000000	29937.000000	29937.000000	29937.000000	29937.000000	29937.000000	29937.000000	29937.000000	29937.000000	29937.000000	29937.0
mean	19159.596486	56.964156	16.545486	833.221899	18.823443	22.934325	41.725382	1000.255142	74.062364	2.162904	13.845913	0.0
std	11334.772113	33.249909	15.083602	657.807468	14.588850	17.893944	28.140889	6.376090	18.827361	1.217392	7.136484	0.0
min	6.000000	-31.750000	0.000000	-812.850000	-20.250000	-23.550000	-43.750000	983.400000	21.500000	0.000000	-4.500000	0.0
25%	9622.000000	31.700000	5.000000	412.800000	7.800000	9.600000	18.800000	996.000000	62.000000	1.300000	7.900000	0.0
50%	18991.000000	47.600000	10.700000	766.800000	14.300000	17.200000	40.600000	1000.000000	77.000000	2.000000	13.600000	0.0
75%	28705.000000	74.000000	24.100000	1229.900000	26.500000	31.700000	60.500000	1004.400000	89.000000	2.800000	19.700000	0.0
max	38575.000000	137.450000	52.750000	2455.550000	54.550000	64.850000	123.050000	1017.000000	100.000000	5.050000	35.500000	0.0

c)

Şekil 3.8. a, b, c.) Zonguldak iline ait veri setinden eksik veriler çıkarılmadan öncesi ve sonrası

Aykırı değerler kutu grafiği yöntemi kullanılarak tespit edilip alt ve üst sınır değerlerini aşan tüm değerler alt ve üst sınır değerine eşitlenerek veri seti güncellenmiştir [67]. Örnek olarak Zonguldak ili veri setine ait PM₁₀, SO₂, basınç ve sıcaklık parametrelerine ait kutu grafikleri Şekil 3.9’da gösterilmiştir.



Şekil 3.9. Zonguldak ili veri setine ait PM₁₀, SO₂, basınç ve sıcaklık parametrelerine ait kutu grafikleri

Literatürde minimum maksimum, z-skör, sigmoid ve ondalık ölçeklendirme gibi birden fazla normalizasyon çeşidi bulunmaktadır. Bu çalışmada veri setinde bulunan veriler minimum-maksimum (min-max) metodu kullanılarak 0-1 değer aralığında normalize edilmiştir.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.1)$$

x' : x değişkenin yeni değeri,

x : x değişkeninin şimdiki değeri,

$\min(x)$: veri setinde bulunan en küçük x değeri,

$\max(x)$: veri setinde bulunan en büyük x değeri.

Normalize edilen verilerden örnek bir kısım Şekil 3.10'da gösterilmiştir.

	SIRA	PM10	SO2	CO	NO2	NOX	O3	BASINC	NEM	RUZGARYONHIZ	SICAKLIK	YAGIS
6	0.000000	0.173189	0.047334	0.614293	0.329832	0.130744	0.410693	0.865145	0.706522	0.142012	0.1700	0.0
7	0.000026	0.175653	0.056308	0.610654	0.332983	0.132462	0.417264	0.858921	0.750000	0.124260	0.1550	0.0
8	0.000052	0.183440	0.043815	0.597566	0.317752	0.128223	0.427718	0.842324	0.771739	0.142012	0.1425	0.0
9	0.000078	0.191621	0.044871	0.620490	0.358193	0.135442	0.396655	0.831950	0.815217	0.112426	0.1300	0.0
10	0.000104	0.196156	0.048566	0.630254	0.369748	0.137390	0.375149	0.821577	0.706522	0.159763	0.1450	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
38571	0.999896	0.213504	0.034137	0.658665	0.412815	0.151484	0.409797	0.508299	0.532609	0.029586	0.6175	0.0
38572	0.999922	0.241005	0.017244	0.658570	0.381828	0.143119	0.422342	0.518672	0.597826	0.029586	0.5850	0.0
38573	0.999948	0.267620	0.030618	0.687786	0.472689	0.162026	0.358423	0.518672	0.521739	0.142012	0.6050	0.0
38574	0.999974	0.304682	0.058420	0.728005	0.529937	0.174401	0.292712	0.520747	0.586957	0.029586	0.5775	0.0
38575	1.000000	0.329128	0.073729	0.717396	0.494748	0.168901	0.314516	0.518672	0.608696	0.041420	0.5700	0.0

Şekil 3.10. Normalize edilen verilerin örnek satır görünümü

3.1.5. Performans değerlendirme ölçütleri

Bu çalışmada oluşturulan modelinin tahmin performansını değerlendirmek, gözlemlenen gerçek değerler ile model tarafından tahmin edilen değerler arasındaki farkların kıyaslanması için makine ve derin öğrenme yöntemlerinin performans değerlendirmelerinde sıklıkla kullanılan R^2 , KOKH, OMH göstergeleri kullanılmıştır [68]. Tahmin hatasını hesaplamak için kullanılan OMH ve KOKH değerlerinin 0'a yakın değerde olması kullanılan modelin başarısı hakkında fikir verebilir. Gözlemlenen değerler ile tahmin değerleri arasındaki uyumu temsil eden regresyon denklemi ile hesaplanan R^2 değerinin 1'e yakınlığı gerçek ve tahmin değerleri arasında yüksek bir benzerlik olduğunu gösterir.

Tahmin performansını değerlendirmek ve model özelliklerini daha iyi öğrenmek için Çizelge 3.4'te sunulan performans metrikleri kullanılmıştır: n , değerlendirilecek verinin uzunluğudur; belirli zamanı temsil eden t olmak üzere F_t , tahmin edilen değerdir; A_t , gözlemlenen değeri temsil eder. Denklem 3.2 de hata formülü verilmiştir.

$$e_t = A_t - F_t \quad (3.2)$$

Çizelge 3.4. Performans Metrikleri

Performans Metrikleri	Denklemleri
R^2	$R^2 = 1 - \frac{\sum_1^n (\hat{y}_i - y_i)}{\sum_1^n (y_i - \bar{y}_i)} \quad (3.4)$
OMH	$OMH = \frac{1}{n} \sum_{i=1}^n e_t \quad (3.5)$
KOKH	$KOKH = \sqrt{\frac{1}{n} \sum_{i=1}^n e_t^2} \quad (3.6)$

4. MAKİNE ÖĞRENMESİ YÖNTEMLERİ VE BULGULARI

Veri setinde gerçekleştirilen ön işlem aşamasından sonra PM_{10} ve SO_2 tahmini için Manisa-Soma ve Zonguldak-Kuzyaka istasyonlarından alınan sırasıyla 38664 ve 38576 veri örneği elde edilmiştir. İki istasyondan alınan veri setleri, oluşturulan modelleri eğitmek ve değerlendirmek için rastgele eğitim ve test olmak üzere ikiye ayrılmıştır. Veri setinin %80'i modelleri eğitmek için kullanılırken %20'si modelleri test etmek için ayrılmıştır.

4.1. Manisa-Soma İstasyon Sonuçları

Tez çalışması kapsamında kullanılan makine öğrenmesi yöntemleri için test setleri kullanılarak elde edilen performans metrik değerleri Manisa-Soma istasyonu için Çizelge 4.1 ve 4.2'de özet olarak verilmiştir.

Çizelge 4.1. Manisa-Soma istasyonuna ait PM_{10} tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları

Kullanılan Yöntemler	OMH	KOKH	R^2
KA	0,093	0,171	0,48
RO	0,080	0,146	0,63
DVM	0,083	0,153	0,59
KEYK	0,076	0,151	0,60
NB	0,105	0,174	0,47
LSTM	8,934	12,340	0,84

Çizelge 4.2. Manisa-Soma istasyonuna ait SO_2 tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları

Kullanılan Yöntemler	OMH	KOKH	R^2
KA	0,047	0,152	0,75
RO	0,047	0,136	0,80
DVM	0,058	0,128	0,82
KEYK	0,033	0,125	0,83

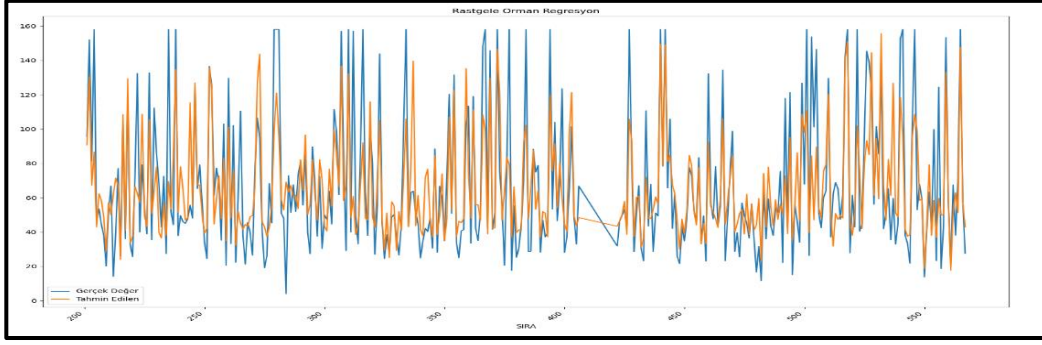
Çizelge 4.2. (devam) Manisa-Soma istasyonuna ait SO₂ tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları

Kullanılan Yöntemler	OMH	KOKH	R ²
NB	0,093	0,167	0,70
LSTM	8,677	12,070	0,85

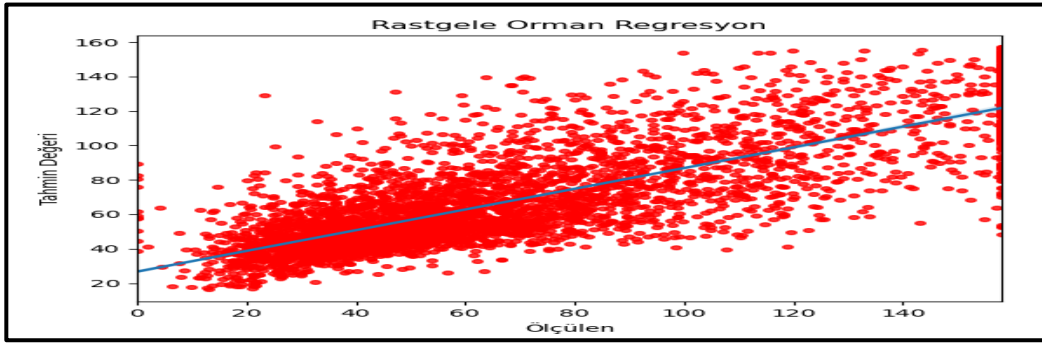
4.1.1. RO makine öğrenmesi sonuçları

Manisa -Soma istasyonuna ait veri setine RO regresyon modeli uygulanmış ve sonuçlar Çizelge 4.1 ve Çizelge 4.2 ‘de özetlenmiştir. RO regresyon modeli uygulanırken optimum değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Yöntemde kullanılan parametrelere ait en uygun değerler sırasıyla maksimum_derinlik = 8, minimum_onek_yaprak =10 ve minimum_ornek_bolunme =10 olarak belirlenmiştir.

PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,080, KOKH=0,146, R² =0,63 olarak ölçülmüştür. RO yöntemi kullanarak gözlenen değerlere karşı tahmin edilen PM₁₀ kirletici konsantrasyon grafikleri Şekil 4.1.a ve Şekil 4.1.b’de verilmiştir.



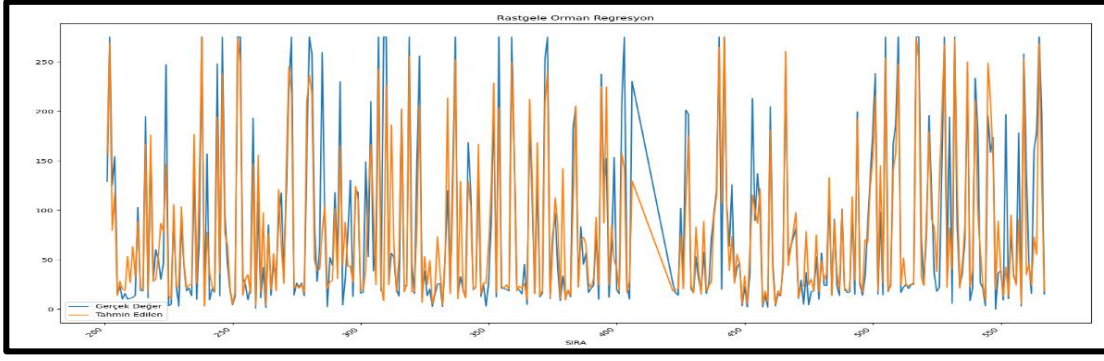
a)



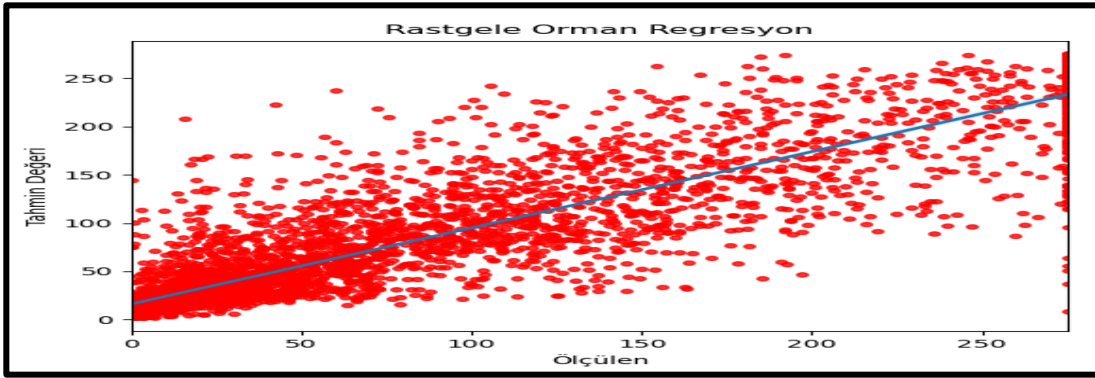
b)

Şekil 4. 1. a) Gerçek ve tahmin PM_{10} değerlerin karşılaştırma grafiği b) RO kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,047$, $KOKH=0,136$, $R^2=0,80$ olarak ölçülmüştür. RO yöntemi kullanarak gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.2.a ve Şekil 4.2.b’de gösterilmektedir.



a)

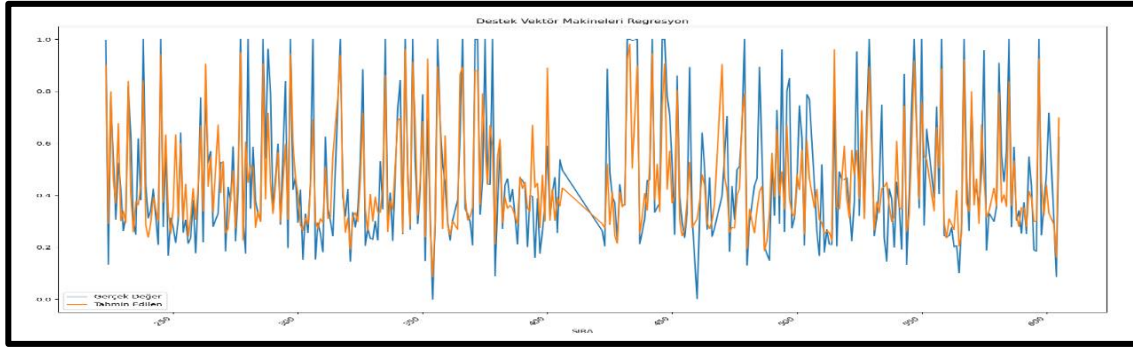


b)

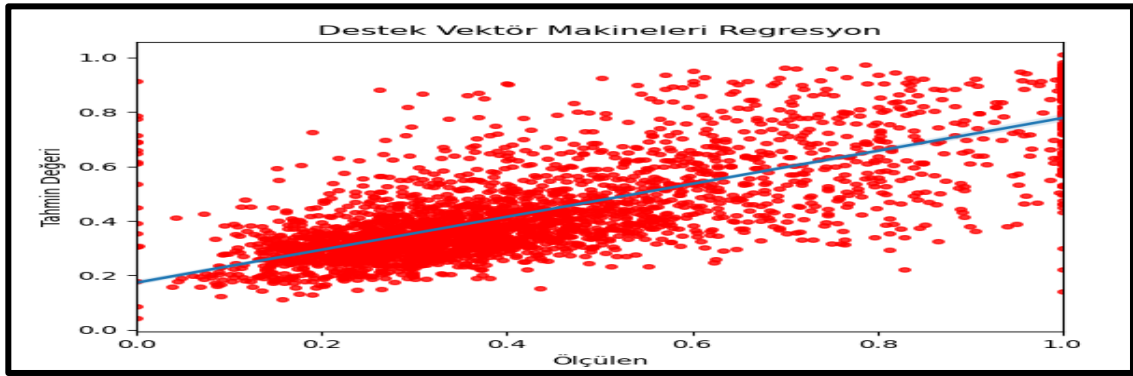
Şekil 4.2. a) Gerçek ve Tahmin SO₂ Değerlerin karşılaştırma grafiği b) RO kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

4.1.2. DVM makine öğrenmesi sonuçları

Manisa -Soma istasyonuna ait veri setine DVM regresyon modeli uygulanmış ve sonuçlar Çizelge 4.1. ve Çizelge 4.2 ‘de özetlenmiştir. DVM regresyon modeli uygulanırken en uygun değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi kullanılarak elde edilmiştir. Elde edilen deneysel sonuçlar incelendiğinde C parametresi için en iyi parametre değerleri [0,1, 1, 10, 100, 1000] olduğu görülmüştür. Bu değerlerden en uygun ve en iyi kombinasyonu sağlayan 1000 değeri seçilmiştir. Model uygulandığında, PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,083, KOKH=0,153, $R^2=0,59$ olarak ölçülmüştür. DVM yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM₁₀ konsantrasyon değerinin grafikleri Şekil 4.3.a ve Şekil 4.3.b’de gösterilmektedir.



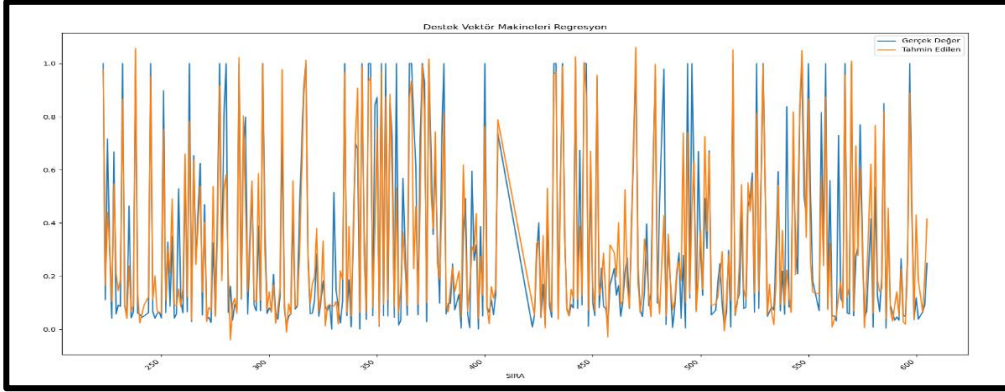
a)



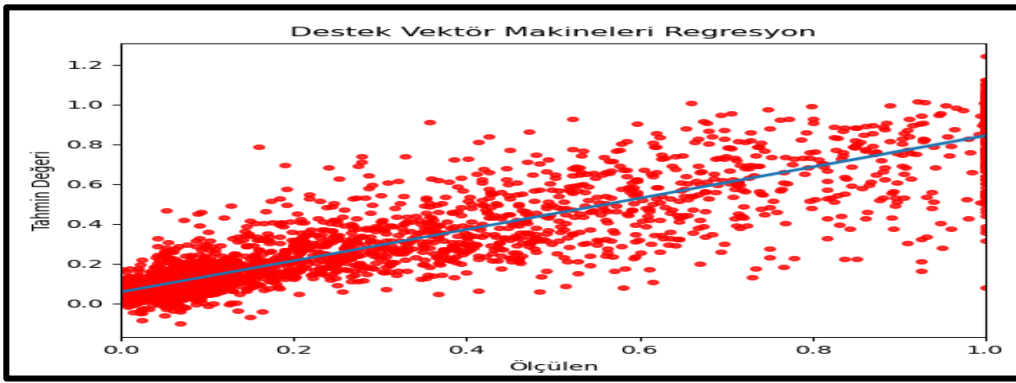
b)

Şekil 4.3. a) Gerçek ve tahmin PM_{10} değerlerin karşılaştırma grafiği b) DVM kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,058$, $KOKH=0,128$, $R^2=0,82$ olarak ölçülmüştür. DVM regresyon modellemesi kullanarak gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.4.a ve Şekil 4.4.b’de verilmiştir.



a)

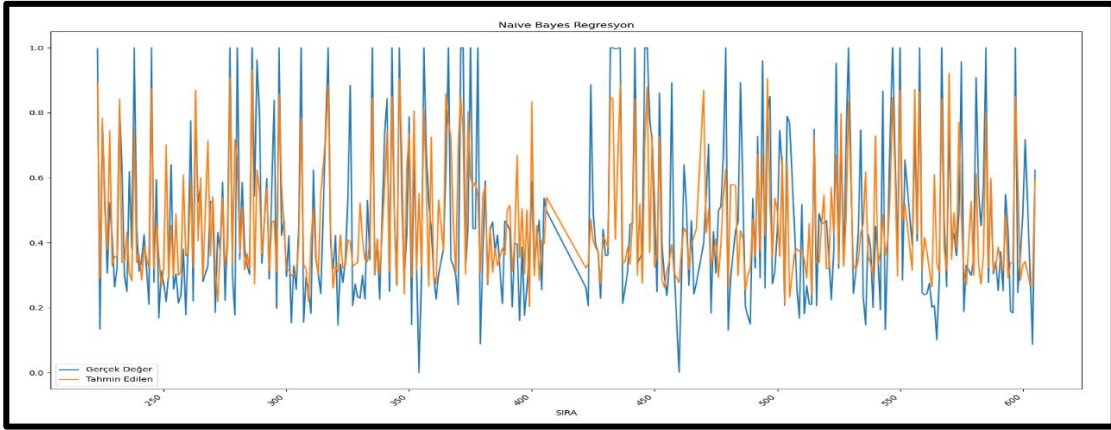


b)

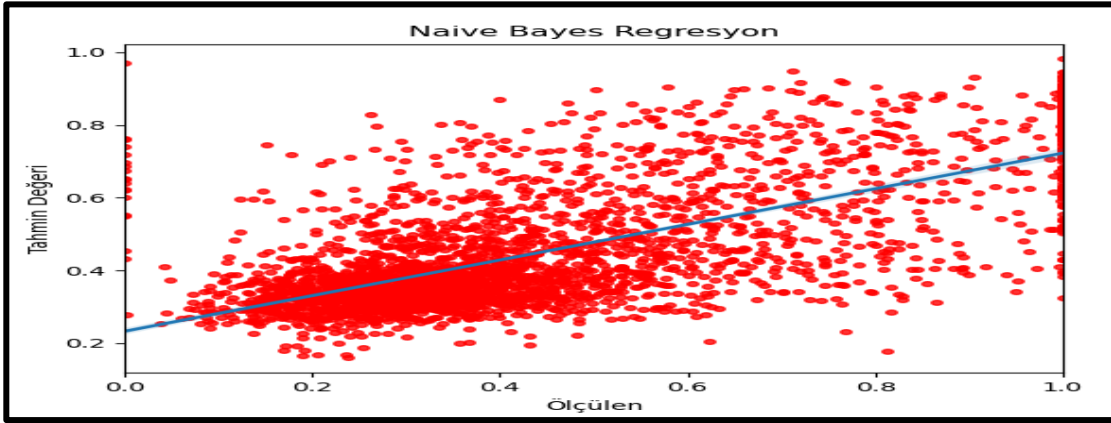
Şekil 4.4. a) Gerçek ve tahmin SO_2 değerlerin karşılaştırma grafiği b) DVM kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

4.1.3. NB makine öğrenmesi sonuçları

Manisa -Soma istasyonuna ait veri setine NB regresyon modeli uygulanmış ve sonuçlar Çizelge 4.1. ve Çizelge 4.2 ‘de özetlenmiştir. NB regresyon modeli uygulanırken optimum değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Elde edilen deneysel sonuçlar incelendiğinde en iyi parametre değeri $n_iter=10$ olduğu görülmüştür. Model uygulandığında, PM_{10} parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,105$, $KOKH=0,174$, $R^2=0,47$ olarak ölçülmüştür. NB yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM_{10} konsantrasyon değerlerinin grafikleri Şekil 4.5.a ve Şekil 4.5.b’de gösterilmektedir.



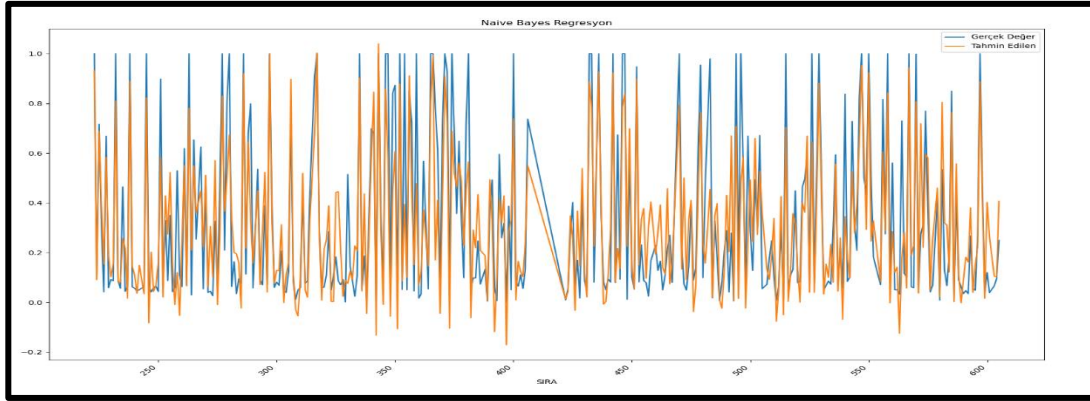
a)



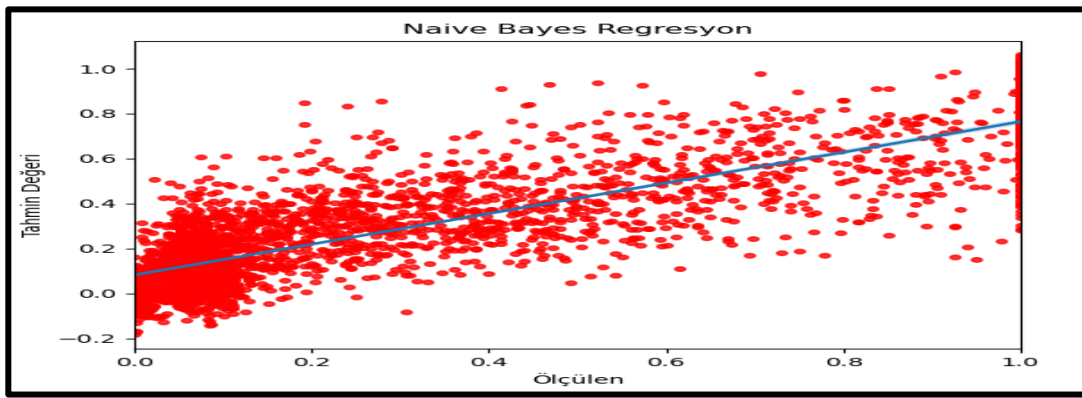
b)

Şekil 4.5. a) Gerçek ve tahmin PM_{10} değerlerin karşılaştırma grafiği b) NB kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,093$, $KOKH=0,167$, $R^2=0,70$ olarak ölçülmüştür. NB yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.6.a ve Şekil 4.6.b’de gösterilmektedir.



a)

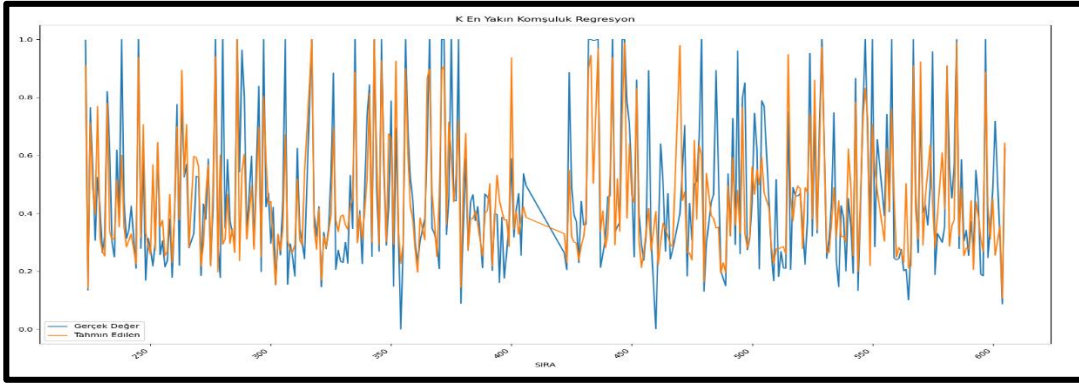


b)

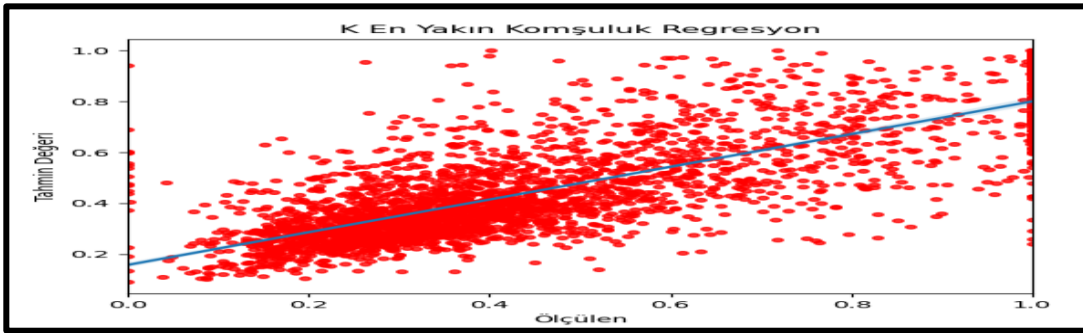
Şekil 4.6. a) Gerçek ve tahmin SO₂ değerlerin karşılaştırma grafiği b) NB kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

4.1.4. KEYK makine öğrenmesi sonuçları

Manisa -Soma istasyonuna ait veri setine KEYK regresyon modeli uygulanmış ve sonuçlar Çizelge 4.1 ve Çizelge 4.2’de özetlenmiştir. KEYK regresyon modeli uygulanırken optimum değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Elde edilen deneysel sonuçlar incelendiğinde $n_neighbors$ parametresi için en iyi parametre değeri 10 olduğu görülmüştür. Model uygulandığında, PM_{10} parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,076$, $KOKH=0,151$ $R^2=0,60$ olarak ölçülmüştür. KEYK yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM_{10} konsantrasyon değerlerinin grafikleri Şekil 4.7.a ve Şekil 4.7.b’de gösterilmektedir.



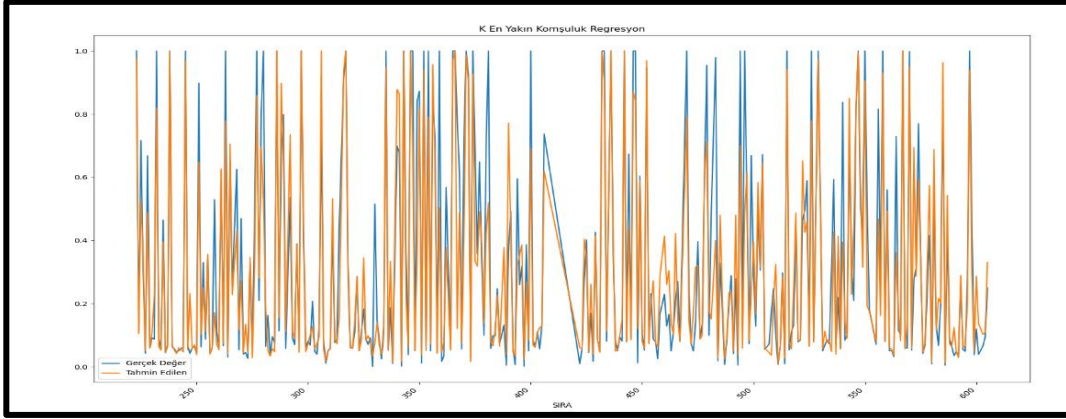
a)



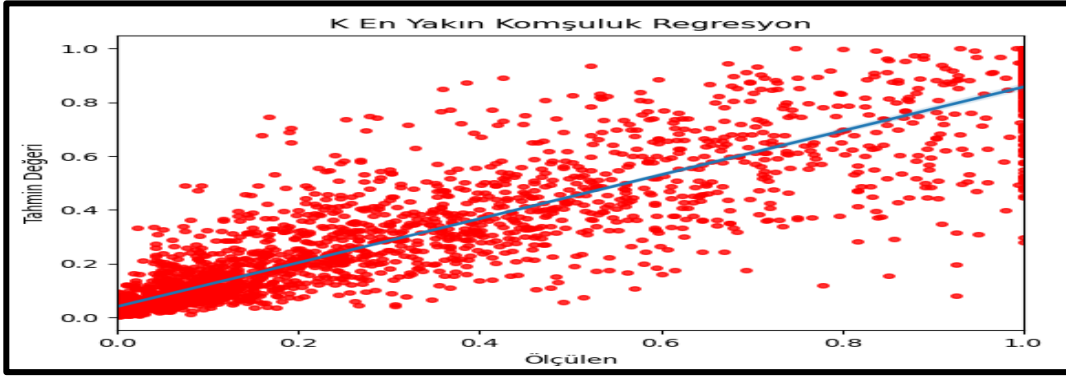
b)

Şekil 4.7. a) Gerçek ve Tahmin PM_{10} değerlerin karşılaştırma grafiği b) KEYK kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,033$, $KOKH=0,15$, $R^2=0,83$ olarak ölçülmüştür. KEYK yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.8.a ve Şekil 4.8.b’de verilmiştir.



a)

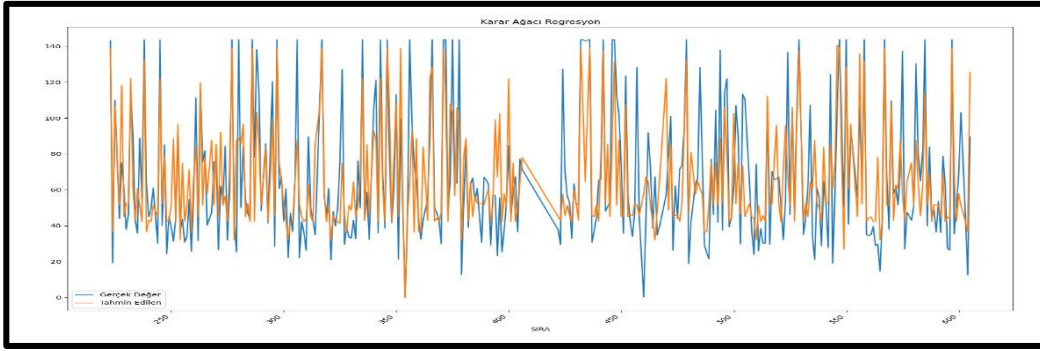


b)

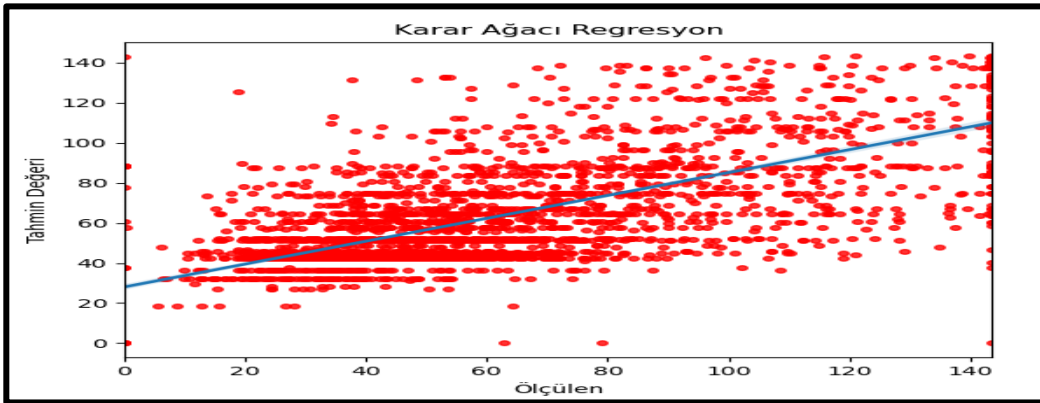
Şekil 4.8. a) Gerçek ve tahmin SO₂ değerlerin karşılaştırma grafiği b) KEYK kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

4.1.5. KA makine öğrenmesi sonuçları

Manisa -Soma istasyonuna ait veri setine KA regresyon modeli uygulanmış ve sonuçlar Çizelge 4.1., Çizelge 4.2’de özetlenmiştir. KA regresyon modeli uygulanırken optimum değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Kullanılan parametrelere ait optimum değerler sırasıyla maksimum_derinlik = 10, minimum_ornek_yaprak =10 ve minimum_ornek_bolunme =2 olarak belirlenmiştir. Model uygulandığında, PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,093, KOKH=0,171, $R^2=0,48$ olarak ölçülmüştür. KA yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM₁₀ konsantrasyon değerlerinin grafikleri Şekil 4.9.a ve Şekil 4.9.b’de verilmiştir.



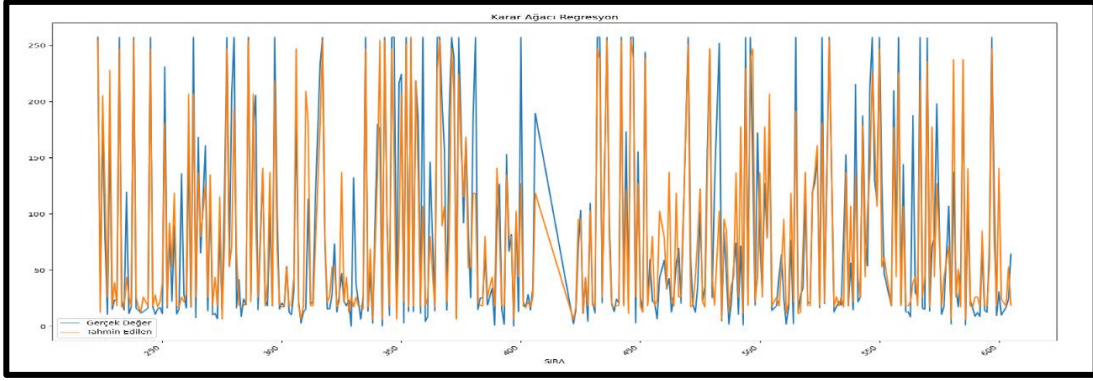
a)



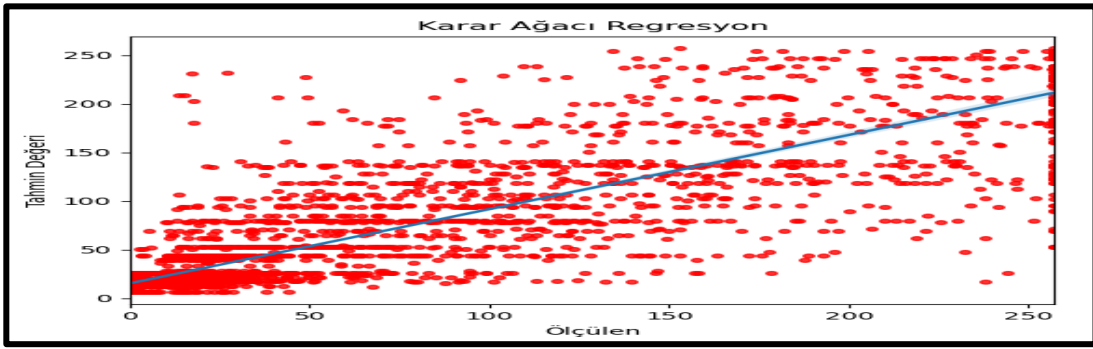
b)

Şekil 4.9. a) Gerçek ve Tahmin PM_{10} değerlerin karşılaştırma grafiği b) KA kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,047$, $KOKH=0,152$, $R^2=0,75$ olarak ölçülmüştür. KA yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.10.a ve Şekil 4.10.b’de verilmiştir.



a)



b)

Şekil 4.10. a) Gerçek ve tahmin SO₂ değerlerin karşılaştırma grafiği b) KA kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

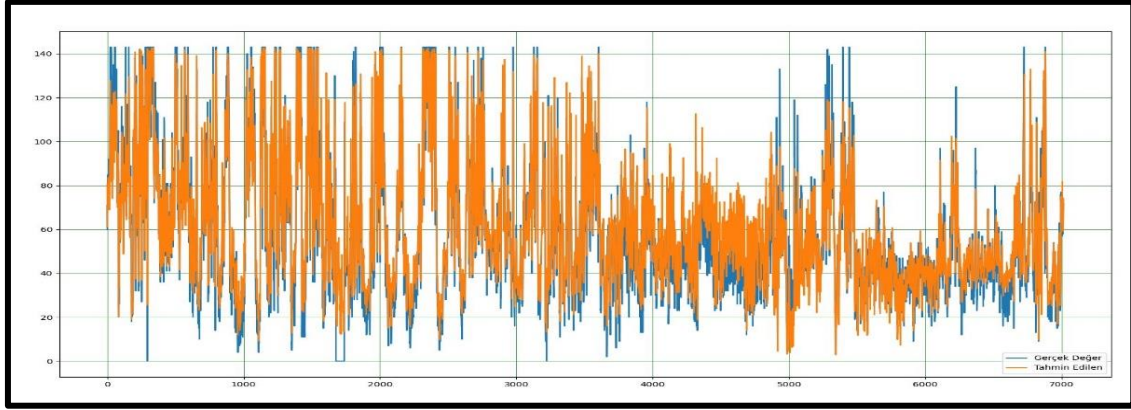
4.1.6. LSTM makine öğrenmesi sonuçları

Manisa-Soma istasyonuna ait veri setine LSTM modeli uygulanmış ve sonuçlar Çizelge 4.1. ve Çizelge 4.2’de özetlenmiştir. LSTM modeli uygulanırken veri seti %80 eğitim ve %20 test verisi olarak ayrılmıştır. Kullanılan parametrelere ait optimum değerlerin belirlenmesi büyük önem taşımaktadır. Oluşturulan modelin eğitim sürecinde, farklı parametre değerleri kullanılarak en optimum değerler bulunmaya çalışılmıştır. Bu amaçla farklı yığın boyutu (50,64,72), seyreltme (0,15,0,25), LSTM ünite sayısı (16,32,64,128), katman (1,2) ve tekrarlama sayıları (50,60,70,100) ile eğitilerek modelin başarı oranı yükseltilmeye çalışılmıştır. Uygulanan LSTM tahmin modeline ait en iyi sonuç veren parametre değerleri Çizelge 4.3’te verilmiştir.

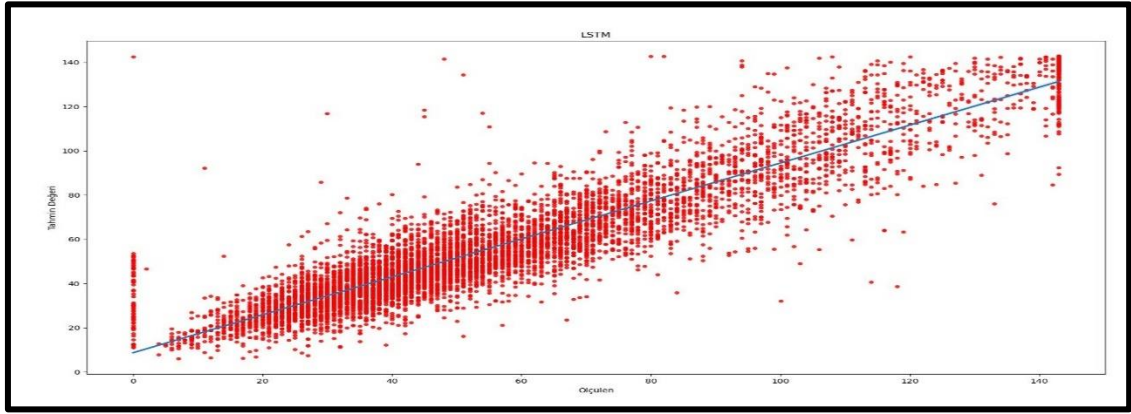
Çizelge 4.3. Uygulanan LSTM tahmin modeline ait parametrelerin değerleri

Modele Ait Parametreler	Değerler
Girdi parametre sayısı(n)	11
LSTM ünite sayısı(lstm_unit_num)	128,16
Yığın boyutu(batch_size)	72
Tekrarlama sayısı(epoch_num)	50
Öğrenme oranı(learning_rate)	0,01
Gizli katman sayısı(hidden_layer_num)	2
Seyreltme (drop_out)	0,25

PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,060, KOKH=0,119, R²=0,84 olarak ölçülmüştür. LSTM yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM₁₀ konsantrasyon değerlerinin grafikleri Şekil 4.11.a ve Şekil 4.11.b’de verilmiştir.



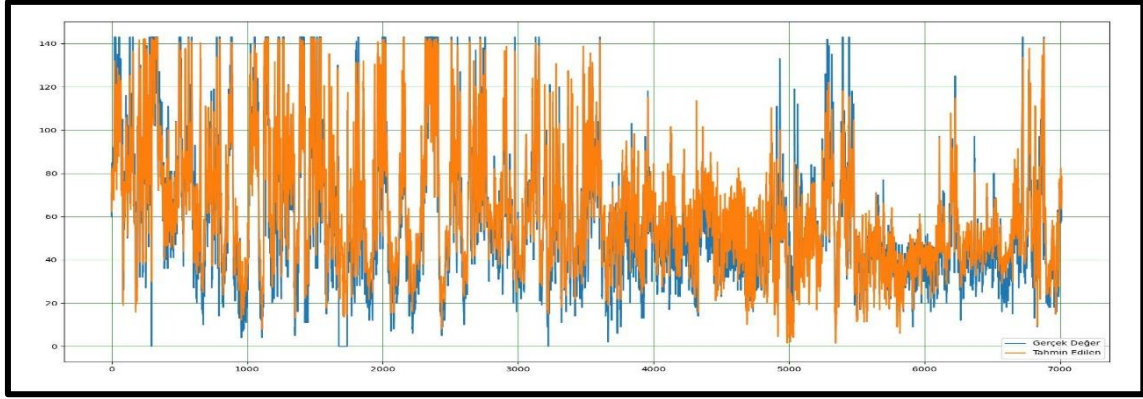
a)



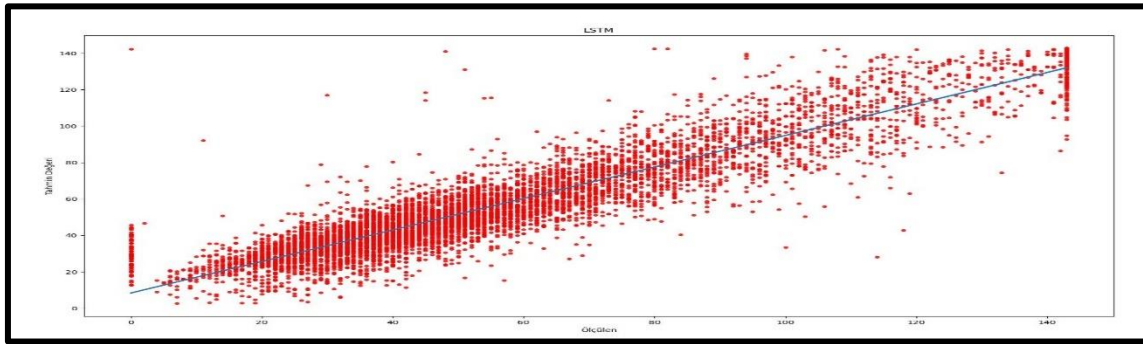
b)

Şekil 4.11. a) Gerçek ve tahmin PM_{10} değerlerin karşılaştırma grafiği b) LSTM kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,050$, $KOKH=0,117$, $R^2=0,85$ olarak ölçülmüştür. LSTM yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.12.a ve Şekil 4.12.b’de verilmiştir.



a)



b)

Şekil 4.12. a) Gerçek ve tahmin SO_2 değerlerin karşılaştırma grafiği b) LSTM kullanılarak tahmin edilen SO_2 değerlerinin dağılım grafiği

4.2. Zonguldak-Kuzyaka İstasyon Sonuçları

Tez çalışması kapsamında kullanılan makine öğrenmesi yöntemlerinde test kümeleri kullanılarak elde edilen performans değerleri Zonguldak-Kuzyaka istasyonu için Çizelge 4.4 ve Çizelge 4.5'te özet olarak verilmiştir.

Çizelge 4.4. Zonguldak-Kuzyaka istasyonuna ait PM₁₀ tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları

Kullanılan Yöntemler	OMH	KOKH	R ²
KA	0,087	0,149	0,42
RO	0,074	0,131	0,55
DVM	0,079	0,137	0,51
KEYK	0,070	0,135	0,53
NB	0,096	0,156	0,36
LSTM	7,627	11,065	0,87

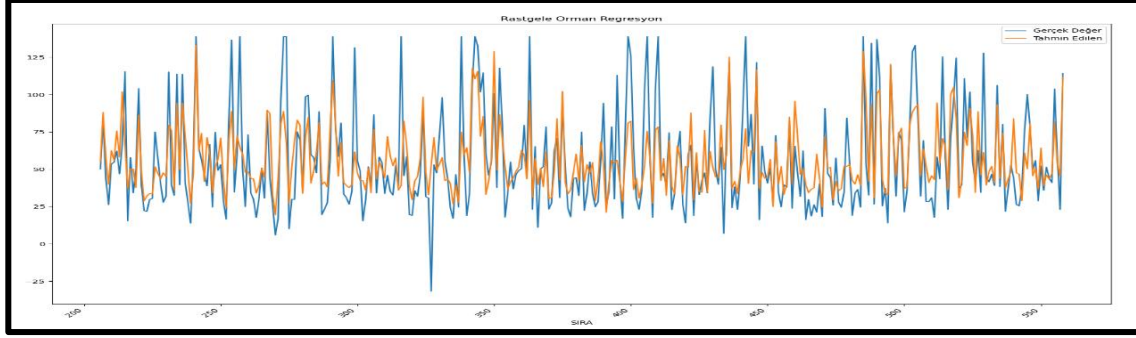
Çizelge 4.5. Zonguldak-Kuzyaka istasyonuna ait SO₂ tahmininde kullanılan makine öğrenmesi algoritmalarının sonuçları

Kullanılan Yöntemler	OMH	KOKH	R ²
KA	0,073	0,157	0,69
RO	0,056	0,130	0,80
DVM	0,069	0,132	0,78
KEYK	0,054	0,132	0,78
NB	0,097	0,165	0,66
LSTM	7,658	11,219	0,87

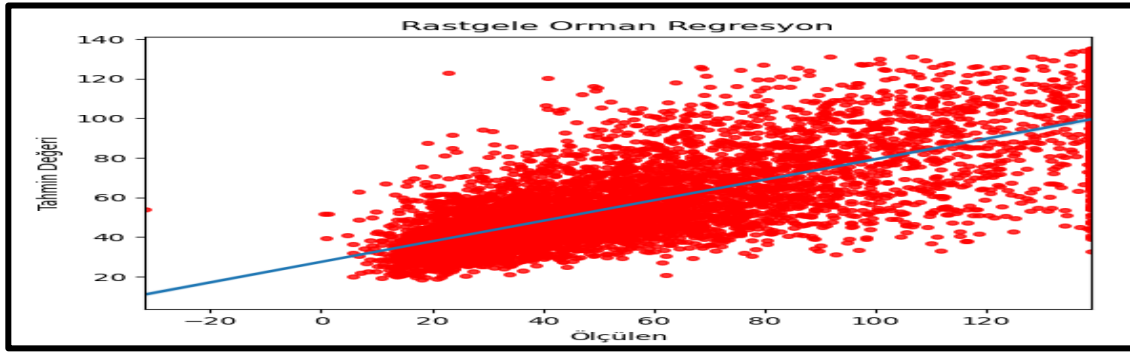
4.2.1. RO makine öğrenmesi sonuçları

Zonguldak-Kuzyaka istasyonuna ait veri setine RO regresyon modeli uygulanmış ve sonuçlar Çizelge 4.4 ve Çizelge 4.5’te özetlenmiştir. RO regresyon modeli uygulanırken optimum değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Kullanılan parametelere ait optimum değerler sırasıyla maksimum_derinlik = 8, minimum_ornek_yaprak =10 ve minimum_ornek_bölme= 10 olarak belirlenmiştir.

PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,074, KOKH=0,131, $R^2=0,55$ olarak ölçülmüştür. RO yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM₁₀ konsantrasyon değerlerinin grafikleri Şekil 4.13.a ve Şekil 4.13.b’de verilmiştir.



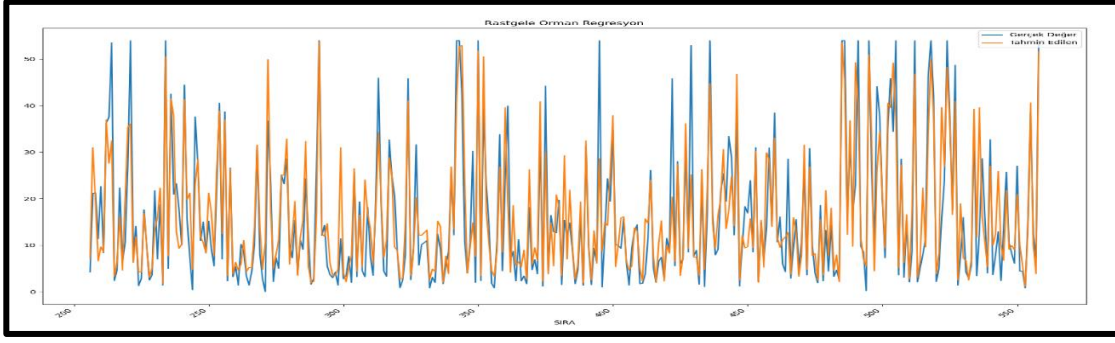
a)



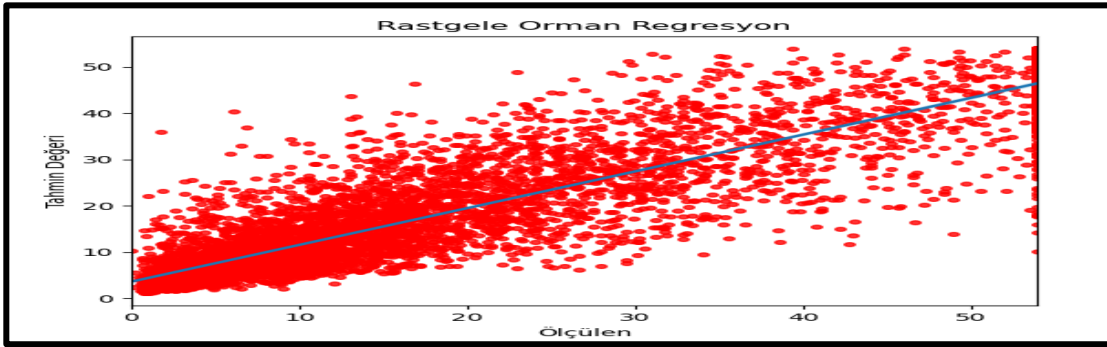
b)

Şekil 4.13. a) Gerçek ve tahmin PM₁₀ değerlerin karşılaştırma grafiği b) RO kullanılarak tahmin edilen PM₁₀ değerlerinin dağılım grafiği

SO₂ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,056, KOKH=0,130 $R^2=0,80$ olarak ölçülmüştür. RO yöntemiyle gözlenen değerlere karşı elde edilen tahmini SO₂ konsantrasyon değerinin grafikleri Şekil 4.14.a ve Şekil 4.14.b’de gösterilmektedir.



a)

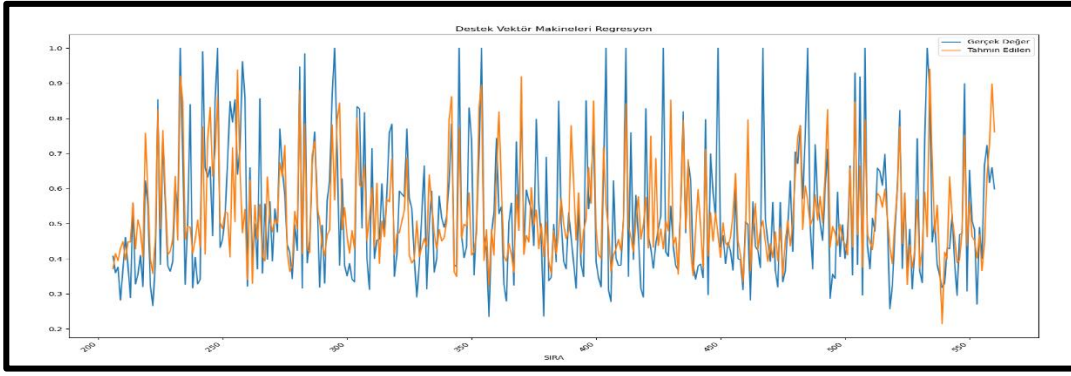


b)

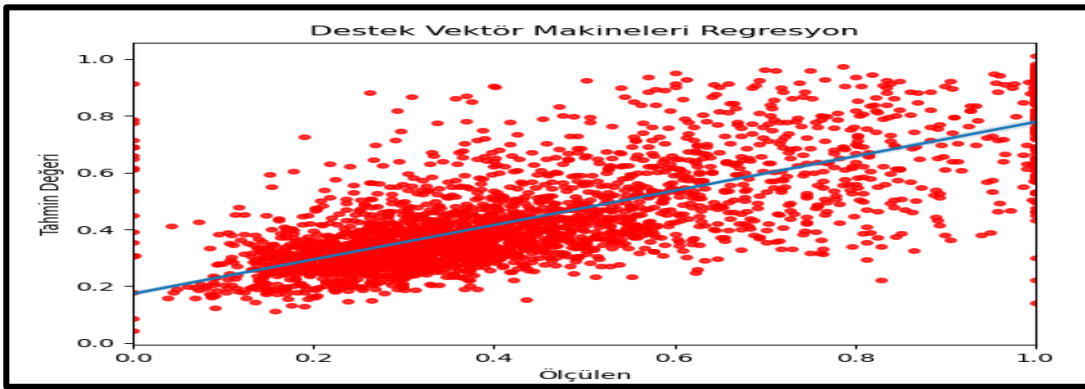
Şekil 4.14. a) Gerçek ve Tahmin SO₂ Değerlerin karşılaştırma grafiği b) RO kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

4.2.2. DVM makine öğrenmesi sonuçları

Zonguldak-Kuzyaka istasyonuna ait veri setine DVM regresyon modeli uygulanmış ve sonuçlar Çizelge 4.4 ve Çizelge 4.5'te özetlenmiştir. DVM regresyon modeli uygulanırken optimum değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Elde edilen deneysel sonuçlar incelendiğinde C parametresi için en iyi parametre değerleri [0,1, 1, 10, 100, 1000] olduğu görülmüştür. Bu değerlerden en uygun ve en iyi kombinasyonu sağlayan 1000 değeri seçilmiştir. Model uygulandığında, PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,079 KOKH=0,137 $R^2 = 0,51$ olarak ölçülmüştür. DVM yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM₁₀ konsantrasyon değerlerinin grafikleri Şekil 4.15.a ve Şekil 4.15.b'de gösterilmektedir.



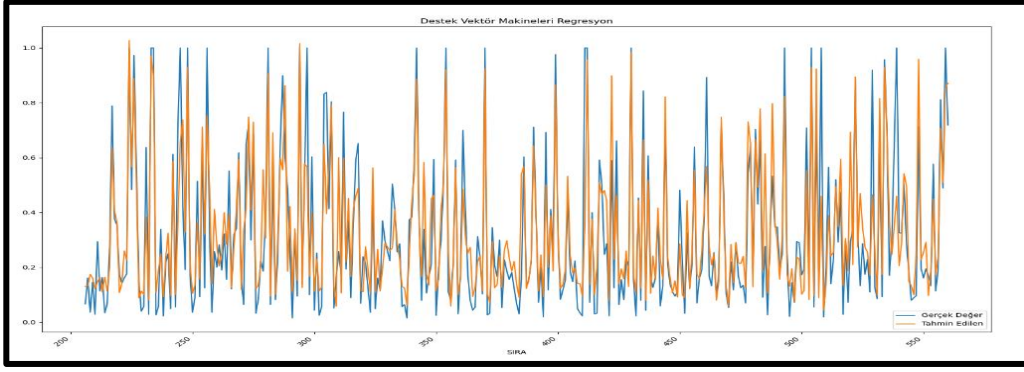
a)



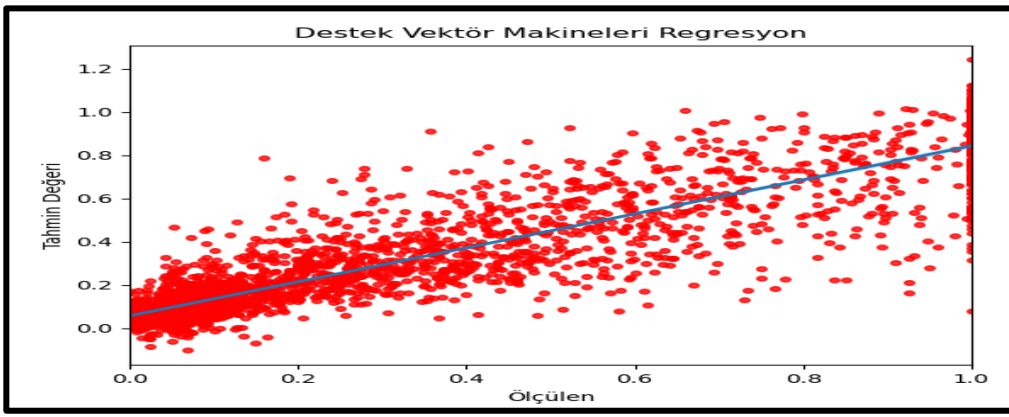
b)

Şekil 4.15. a) Gerçek ve tahmin PM_{10} değerlerin karşılaştırma grafiği b) DVM kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,069$, $KOKH=0,132$, $R^2=0,78$ olarak ölçülmüştür. DVM yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.16.a ve Şekil 4.16.b’de verilmiştir.



a)

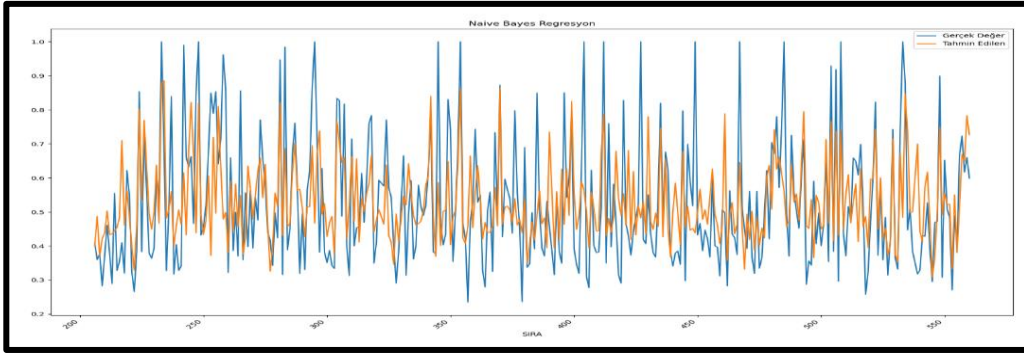


b)

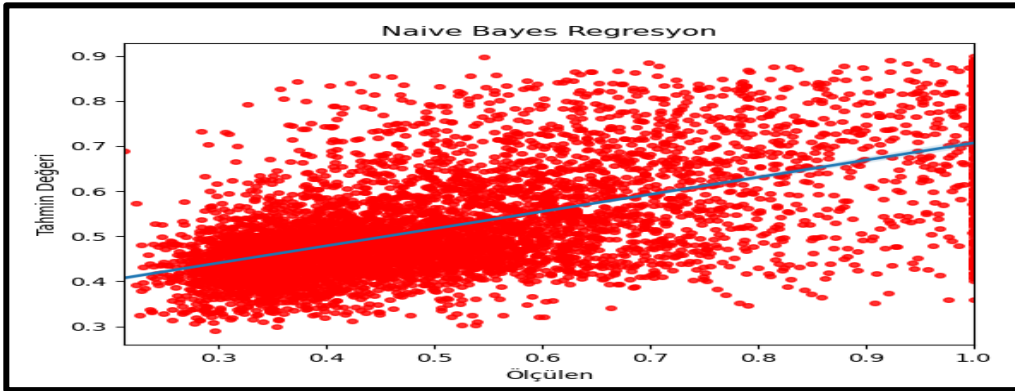
Şekil 4.16. a) Gerçek ve tahmin SO₂ değerlerin karşılaştırma grafiği b) DVM kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

4.2.3. NB makine öğrenmesi sonuçları

Zonguldak-Kuzyaka istasyonuna ait veri setine NB yöntemi uygulanmıştır ve elde edilen sonuçlar Çizelge 4.4 ve Çizelge 4.5’te özetlenmiştir. NB yöntemi kullanılırken en uygun değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Elde edilen deneysel sonuçlar incelendiğinde en iyi parametre değeri $n_iter=10$ olduğu görülmüştür. Model uygulandığında, PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,096, KOKH=0,156, $R^2=0,36$ olarak ölçülmüştür. NB modeliyle gözlemlenen değerlere karşı elde edilen tahmini PM₁₀ konsantrasyon değerlerinin grafikleri Şekil 4.17.a ve Şekil 4.17.b’de verilmiştir.



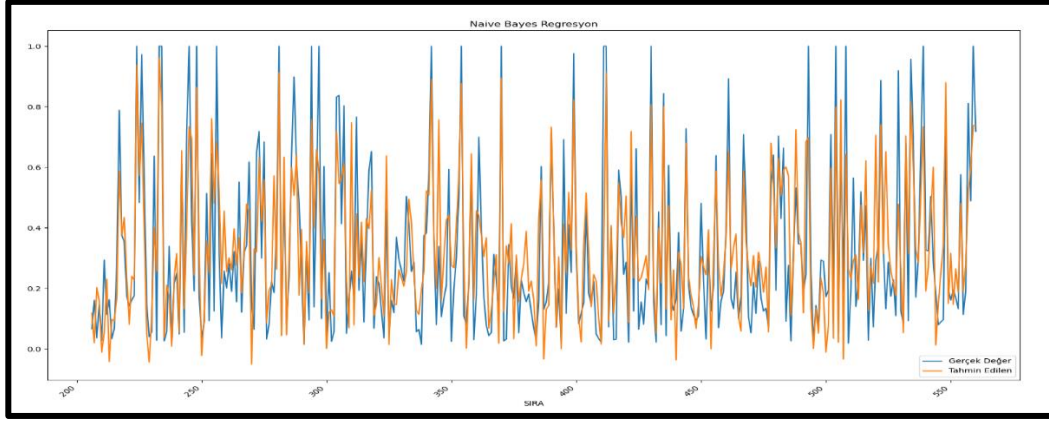
a)



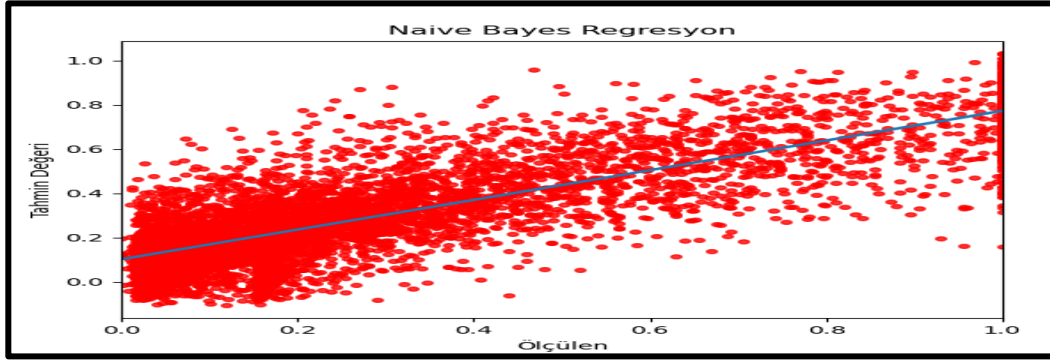
b)

Şekil 4.17. a) Gerçek ve Tahmin PM10 Değerlerin karşılaştırma grafiği b) NB kullanılarak tahmin edilen PM10 değerlerinin dağılım grafiği

SO₂ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,097, KOKH=0,165, $R^2=0,66$ olarak ölçülmüştür. NB yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO₂ konsantrasyon değerlerinin grafikleri Şekil 4.18.a ve Şekil 4.18.b’de verilmiştir.



a)

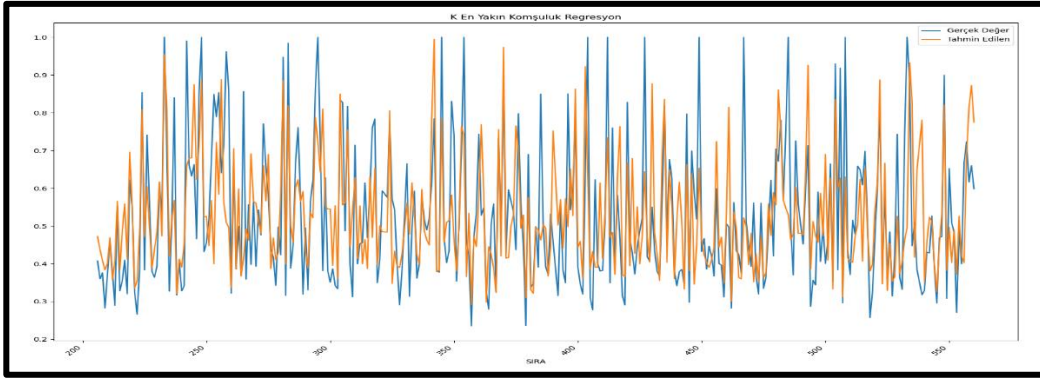


b)

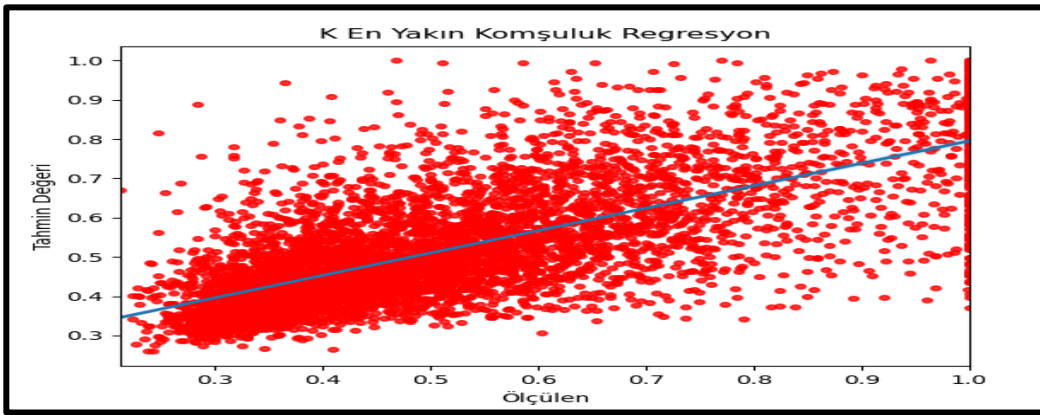
Şekil 4.18. a) Gerçek ve tahmin SO₂ değerlerin karşılaştırma grafiği b) NB kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

4.2.4. KEYK makine öğrenmesi sonuçları

Zonguldak-Kuzyaka istasyonuna ait veri setine KEYK yöntemi uygulanmış ve sonuçlar Çizelge 4.4 ve Çizelge 4.5'te özetlenmiştir. KEYK yöntemi kullanılırken en uygun değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Elde edilen deneysel sonuçlar incelendiğinde $n_neighbors$ parametresi için en iyi parametre değeri 10 olduğu görülmüştür. Model uygulandığında, PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,070, KOKH=0,135, $R^2=0,53$ olarak ölçülmüştür. KEYK yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM₁₀ konsantrasyon değerlerinin grafikleri Şekil 4.19.a ve Şekil 4.19.b'de verilmiştir.



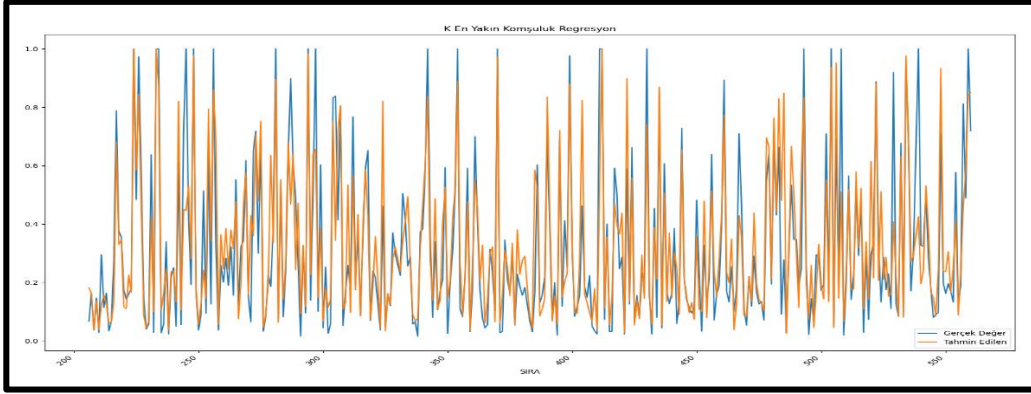
a)



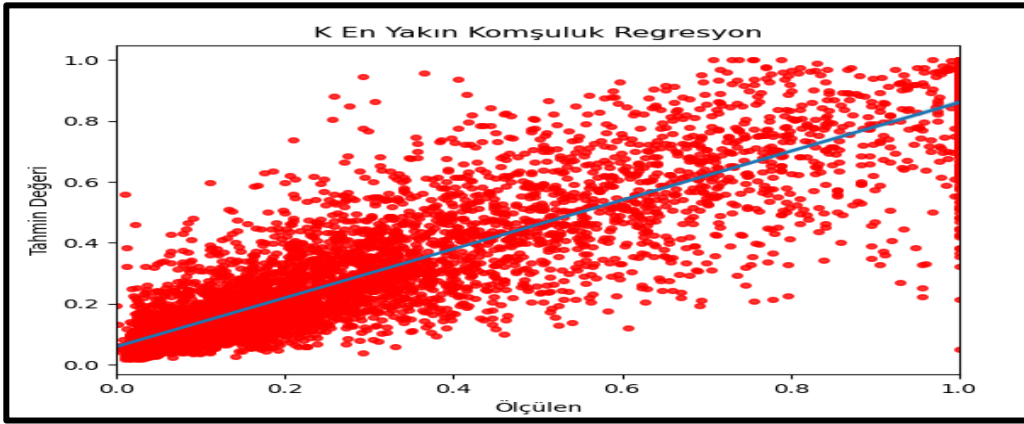
b)

Şekil 4.19. a) Gerçek ve tahmin PM_{10} değerlerin karşılaştırma grafiği b) KEYK kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,054$, $KOKH=0,132$, $R^2=0,78$ olarak ölçülmüştür. KEYK yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.20.a ve Şekil 4.20.b’de verilmiştir.



a)

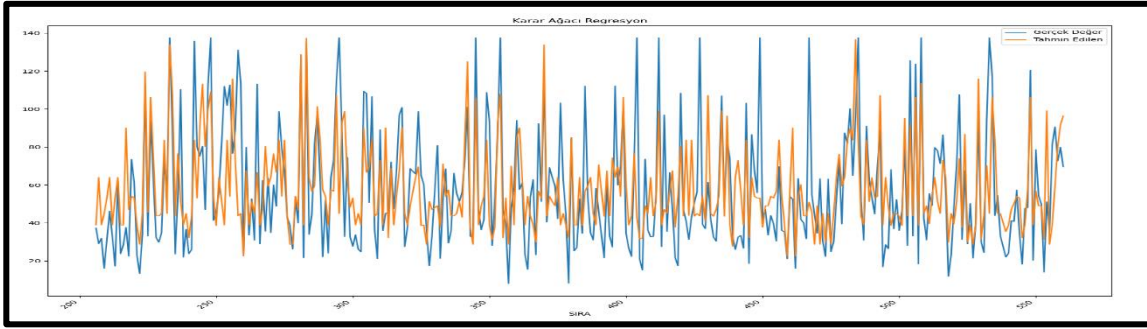


b)

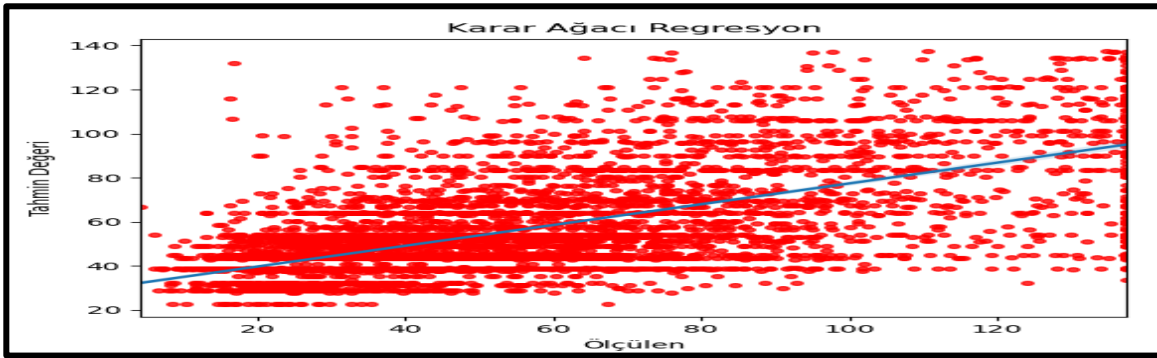
Şekil 4.20. a) Gerçek ve tahmin SO₂ değerlerin karşılaştırma grafiği b) KEYK kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

4.2.5. KA makine öğrenmesi sonuçları

Zonguldak-Kuzyaka istasyonuna ait veri setine KA yöntemi uygulanmış ve sonuçlar Çizelge 4.4 ve Çizelge 4.5'te özetlenmiştir. KA yöntemi uygulanırken en uygun değerler, Python dilindeki “Grid Search” (ızgara arama) fonksiyonu kullanılarak beş kat çapraz doğrulama yöntemi ile elde edilmiştir. Kullanılan parametrelere ait en uygun değerler sırasıyla maksimum_derinlik = 10, minimum_ornek_yaprak = 10 ve minimum_ornek_bolme = 2 olarak belirlenmiştir. Model uygulandığında, PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,087, KOKH=0,149, R² =0,42 olarak ölçülmüştür. KA yöntemiyle gözlemlenen değerlere karşı tahmini PM₁₀ konsantrasyon değerlerinin grafikleri Şekil 4.21.a ve Şekil 4.21b'de verilmiştir.



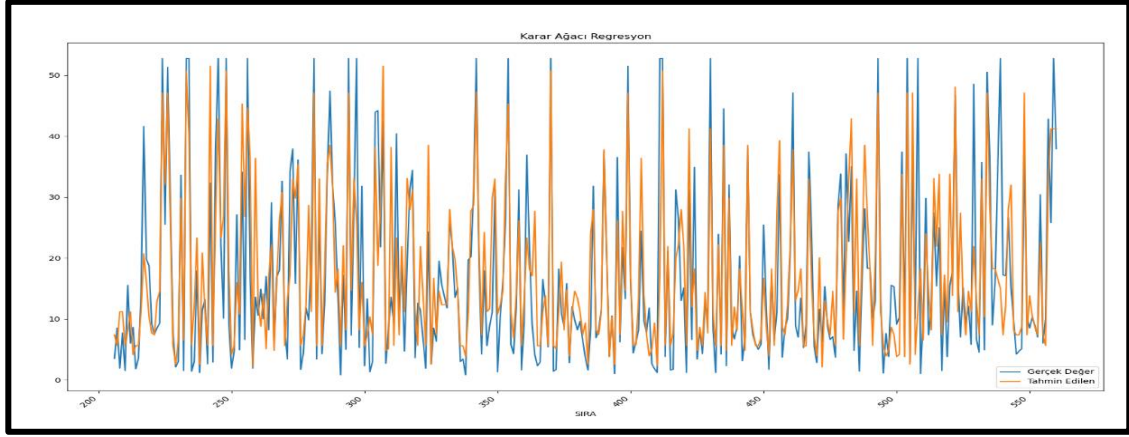
a)



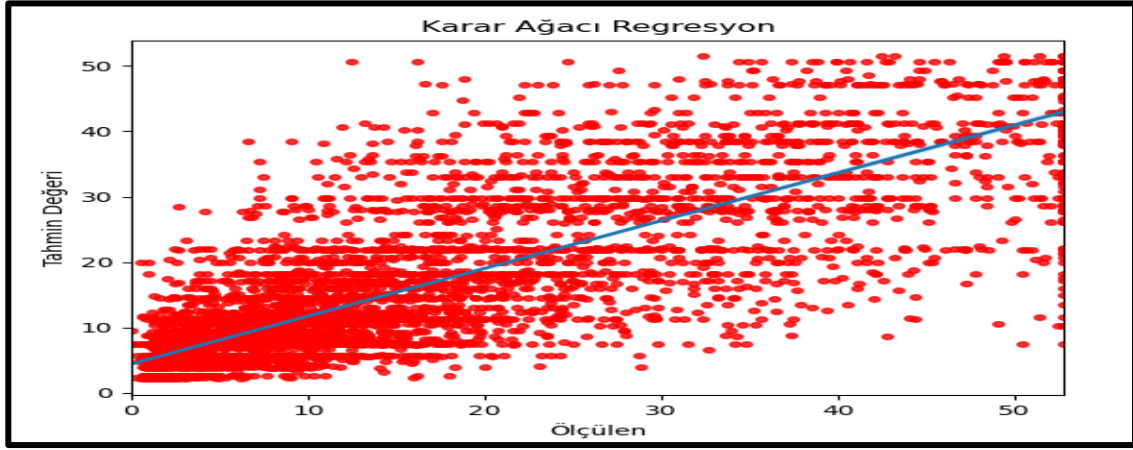
b)

Şekil 4.21. a) Gerçek ve tahmin PM_{10} değerlerin karşılaştırma grafiği b) KA kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,073$, $KOKH=0,157$, $R^2=0,69$ olarak ölçülmüştür. KA yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.22.a ve Şekil 4.22.b’de verilmiştir.



a)



b)

Şekil 4.22. a) Gerçek ve tahmin SO₂ değerlerin karşılaştırma grafiği b) KA kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

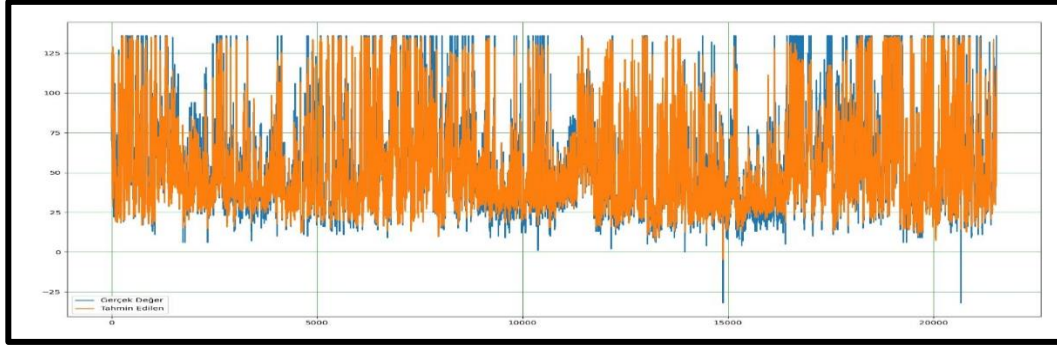
4.2.6. LSTM makine öğrenmesi sonuçları

Zonguldak-Kuzyaka istasyonuna ait veri setine LSTM modeli uygulanmış ve sonuçlar Çizelge 4.4 ve Çizelge 4.5'te özetlenmiştir. LSTM yöntemi kullanılırken veri seti %80 eğitim ve %20 test verisi olarak ayrılmıştır. Kullanılan parametrelere ait en uygun değerlerin belirlenmesi büyük önem taşımaktadır. Oluşturulan modelin eğitim sürecinde, farklı parametre değerleri kullanılarak en iyi değerler bulunmaya çalışılmıştır. Bu amaçla farklı yığın boyutu (50,64,72), seyreltme (0,15,0,25), LSTM ünite sayısı (16,32,64,128), katman (1,2) ve tekrarlamaya sayıları (50,60,70,100) ile eğitilerek modelin başarı oranı yükseltilmeye çalışılmıştır. Uygulanan LSTM tahmin modeline ait en iyi sonuç veren parametre değerleri Çizelge 4.6'da verilmiştir.

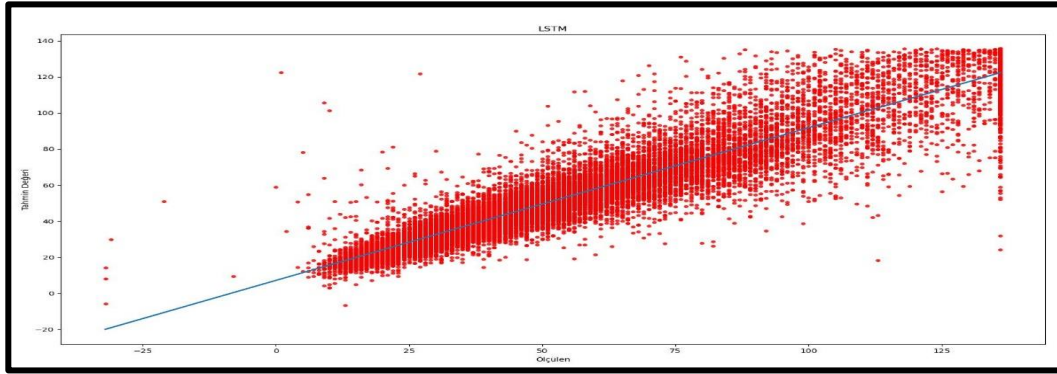
Çizelge 4.6. Uygulanan LSTM tahmin modeline ait parametrelerin değerleri

Modele Ait Parametreler	Değerler
Girdi parametre sayısı (n)	11
LSTM ünite sayısı (lstm_unit_num)	128,16
Yığın boyutu (batch_size)	72
Tekrarlama sayısı (epoch_num)	50
Öğrenme oranı (learning_rate)	0,01
Gizli katman sayısı (hidden_layer_num)	2
Seyreltme (drop_out)	0,25

PM₁₀ parametresinin tahmininde elde edilen performans değerleri sırasıyla OMH=0,048, KOKH=0,115, R²=0,87 olarak ölçülmüştür. LSTM yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini PM₁₀ konsantrasyon değerlerinin grafikleri Şekil 4.23.a ve Şekil 4.23.b’de gösterilmektedir.



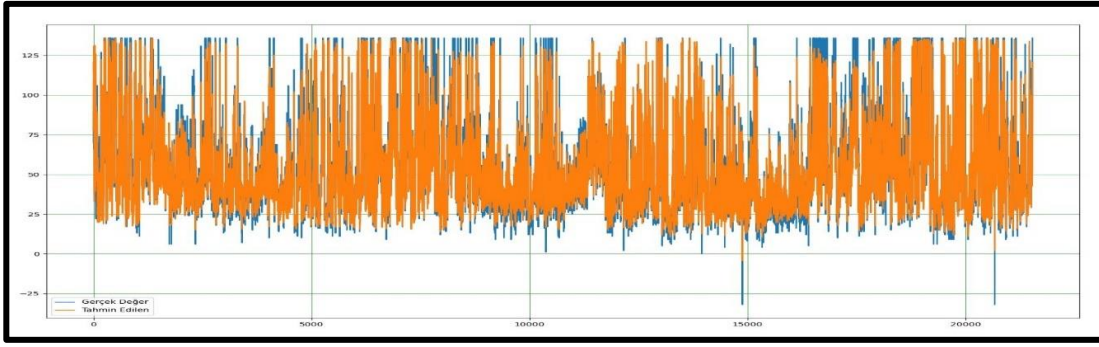
a)



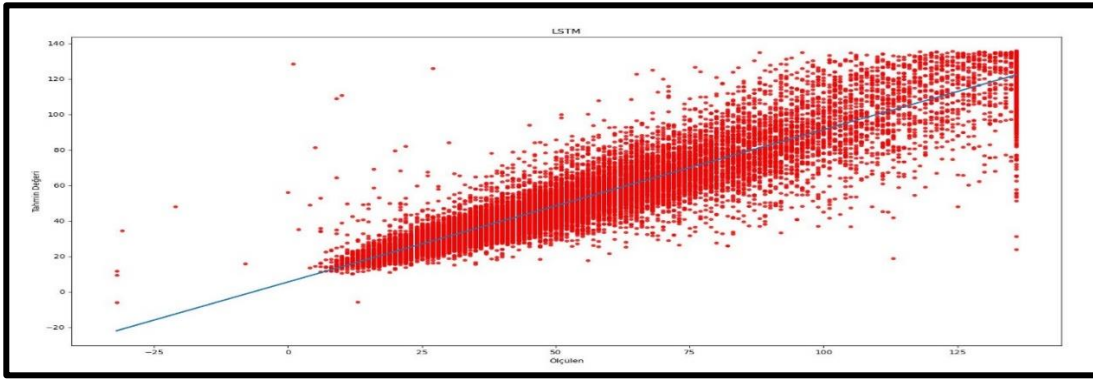
b)

Şekil 4.23. a) Gerçek ve tahmin PM_{10} değerlerin karşılaştırma grafiği b) LSTM kullanılarak tahmin edilen PM_{10} değerlerinin dağılım grafiği

SO_2 parametresinin tahmininde elde edilen performans değerleri sırasıyla $OMH=0,047$, $KOKH=0,114$, $R^2=0,87$ olarak ölçülmüştür. LSTM yöntemiyle gözlemlenen değerlere karşı elde edilen tahmini SO_2 konsantrasyon değerlerinin grafikleri Şekil 4.24.a ve Şekil 4.24.b’de verilmiştir.



a)



b)

Şekil 4.24. a) Gerçek ve tahmin SO₂ Değerlerin karşılaştırma grafiği b) LSTM kullanılarak tahmin edilen SO₂ değerlerinin dağılım grafiği

5. SONUÇ VE ÖNERİLER

Hava kirliliğinin yoğun yaşandığı termik santral bölgelerinde önceden önlem alınabilmesi, sağlık üzerindeki kritik etkilerin azaltılabilmesi gibi sebeplerle hava kirliliğinin önceden tahmini büyük önem taşımaktadır. Bu amaçla gelecek kirlletici değerlerinin tahmininde kullanılması için makine öğrenmesine dayalı altı model önerilmiştir. Önerilen altı modelle, hava kirliliğine neden olan PM_{10} ve SO_2 kirlletici gazlarının gelecek değerleri tahmin edilmektedir.

Önerilen tahmin modellerinin makine öğrenmesi yöntemleri kullanılarak geliştirilmesinde; hava kirliliği probleminin birden çok değişkenden etkilenmesi nedeniyle lineer modelleme ile çözülememesi, makine öğrenmesi algoritmalarıyla elde edilen sonuçların yüksek doğruluk değerlerine sahip olması ve geleneksel yaklaşımlara göre daha uygun olması gibi sebepler etkili olmuştur.

RO, KA, DVM, KEYK, NB, LSTM makine öğrenmesi yöntemleri kullanılmıştır. Altı yöntemin, PM_{10} ve SO_2 konsantrasyonun değerlerini tahmin başarısı KOKH, OMH ve R^2 olmak üzere üç performans metriklerine göre kıyaslanmıştır. Manisa ve Zonguldak şehirlerine ait PM_{10} , CO, SO_2 , NO_2 , NO_x , O_3 kirlletici verileri ve Sıcaklık, Nem, Rüzgar Hızı, Basınç ve Yağış meteorolojik parametreleri girdi değerleri olarak alınıp PM_{10} ve SO_2 kirlletici değerleri oluşturulan modellerde çıkış/tahmin değeri olarak kullanılmıştır. Aralık 2016 ile Nisan 2021 tarihleri arasında Manisa ve Zonguldak illerine ait sırasıyla toplam 38664 ve 38576 536 saatlik meteoroloji ve hava kirliliği verilerinin rastgele seçim yapılarak %80' i eğitim verisi olarak, geri kalan %20 test verisi olarak seçilmiştir.

Bu çalışmada önerilen modeller kıyaslanmıştır. İki farklı kirlletici tahmini yapan Makine Öğrenmesi modellerinden en başarılı model $R^2=0,87$ değeriyle LSTM'e dayalı model olmuştur. Oluşturulan LSTM modelinin düşük hata değerleri ve yüksek R^2 ile kullanılan diğer beş yöntem arasından en iyi yöntem olduğu anlaşılmıştır. LSTM modelini yüksek R^2 (0,80) değerine sahip RO modeli takip etmiştir. Manisa-Soma istasyonu SO_2 tahmininde LSTM' den sonra en yüksek açıklama katsayısı değerini ise K-EYK modeli gerçekleştirmiştir. NB, iki istasyon için kullanılan veri setlerinde daha yüksek bir hata oranına ve daha düşük açıklama katsayısı değerine sahip olduğu için altı algoritmadan en

kötüsü olarak seçilmiştir. Elde edilen sonuçlara bakıldığında Makine Öğrenmesine dayalı modellerin hava kirliliği tahmininde başarılı olduğu ortaya konmuştur.

Gelecekte yapılacak çalışmalarda, meteorolojik ve kirletici verilerinin parametre sayısı arttırılarak model eğitilip elde edilen sonuçlar diğer yöntemlerle kıyaslanabilir. kirletici tahminleri için RNN+LSTM, RNN+GRU, LSTM+GRU gibi hibrit yöntemler dikkate alınabilir. Ayrıca makine öğrenmesi yöntemleri ile hibrit oluşturulan yöntemler kıyaslanabilir.

Akıllı şehir sistemlerinde kullanılan sensörlerden elde edilen kirletici verileri alınıp bu sistemlere kirlilik tahmin modelimiz entegre edilebilir. Ayrıca Coğrafi Bilgi Sistemleri ile entegre edilip kirlilik dağılım haritaları oluşturulabilir. Oluşturulan bu haritalar hava kirliliğinin kontrol edilmesinde büyük katkı sağlayabilir.

KAYNAKLAR

1. Atacak, I., Arici, N. and Guner, D. (2017). “Modelling and Evaluating Air Quality with Fuzzy Logic Algorithm-Ankara-Cebeci Sample”, *International Journal of Intelligent Systems and Applications in Engineering*, 5(4), 263–268.
2. Güngör, A. ve Sevindir, C. H. (2014). “İsparta İlindeki Atmosferde Bulunan Kükürt dioksit (SO₂) ve Partikül Madde (PM) Konsantrasyonunun Çoklu Doğrusal Regresyon Yöntemi İle Modellenmesi”, *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dersi.*, 17(1), 95-108.
3. Şimşek, T., and Yılmaz, V. (2020). Hava kalitesi ile covid-19 arasındaki ilişki: G-7 ülkeleri üzerine bir uygulama. *Journal of Turkish Studies*, 15(4), 1353–1366.
4. Yafouz, A., AlDahoul, N., Birima, A. H., Ahmed, A. N., Sherif, M., Sefelnasr, A., Allawi, M. F., and Elshafie, A. (2021). “Comprehensive comparison of various machine learning algorithms for short-term ozone concentration prediction.” *Alexandria Engineering Journal*. 61(6),4607-4622.
5. İnternet: ÇŞB. *Hava kalitesi istasyon veri indirme T.C. çevre ve şehircilik bakanlığı*. Web:https://sim.csb.gov.tr/STN/STN_Report/StationDataDownloadNew Son Erişim Tarihi: 15.03.2022
6. İnternet: MGM. *Meteoroloji Genel Müdürlüğü*. Web: <https://www.mgm.gov.tr/tahmin/il-ve-ilceler.aspx#/> Son Erişim Tarihi: 15.03.2022
7. Altunkaynak, A., Başakın, E. E., and Kartal, E. (2020). “Dalgacık K-En Yakın Komşuluk Yöntemi İle Hava Kirliliği Tahmini.” *Uludağ University Journal of The Faculty of Engineering*, 25(3), 1547–1556.
8. İnternet: DSÖ. “*Dünya Sağlık Örgütü: Hava kirliliği nedeniyle dünyada her yıl 7 milyon kişi ölüyor!*” Web: <https://www.medikalakademi.com.tr/dso-hava-kirliligi-nedeniyle-her-yil-7-milyon-kisi-oluyor/> Son Erişim Tarihi: 20.09.2021
9. İnternet: Zonguldak Çevre ve Şehircilik Müdürlüğü. *Zonguldak İli Temiz Hava Eylem Planı THEP (2014-2019)*. Web: <https://webdosya.csb.gov.tr/db/zonguldak/webmenu/webmenu36983.pdf> Son Erişim Tarihi: 20.09.2021
10. İnternet: Öztürk, M. (2005). *Şehir İçi Bölgelerde Hava Kirliliğinin Sağlık Üzerine Etkileri, Çevre ve Orman Bakanlığı, Ankara*. Web: http://www.cevresehirkutuphanesi.com/assets/files/slider_pdf/k5ZJrB8L2eyV.pdf

11. Duygu Çelik, B. ve Arici, N. (2021). "Covid-19 Salgın Sürecinde Hava Kalitesi Tahmini: Zonguldak Örneği". *Gazi Mühendislik Bilimleri Dergisi*, 7(3), 222–232.
12. İnternet: WHO. "Ambient (outdoor) air pollution." Web: [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health). Son Erişim Tarihi: 06.05.2021
13. Gültepe, Y. (2019). Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine Karşılaştırmalı Bir Değerlendirme. *European Journal of Science and Technology*, 16, 8–15.
14. Kutlay, A., Özgiray, S., Yücecengiz, I. (2016). "Kurumsal Mobil Sağlık Bilgi Sistemi: Yaklaşım ve Deneyimler", *Ulusal Yazılım Mühendisliği Sempozyumu*, 394–399.
15. Wang, L.K., Pereira, N.C. and Hung, T.S. 2004. Air pollution control engineering, volume 1. New Jersey, Humana Press, 100-105.
16. Suárez Sánchez, A., García Nieto, P. J., Riesgo Fernández, P., del Coz Díaz, J. J., and Rodríguez, F. J. (2011). Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *Mathematical and Computer Modelling*, 54(5–6), 1453-1466.
17. Çıldır, İ. (2021). *Balıkesir’de hava kirliliği seviyelerinin meteorolojik parametrelere bağlı olarak solunum yolu hastalıkları üzerindeki etkilerinin incelenmesi*, Yüksek Lisans Tezi, Balıkesir Üniversitesi Fen Bilimleri Enstitüsü, Balıkesir, 20-25.
18. Selici, A. T. (2006). *Enerji kullanımının çevresel etkileri ve sürdürülebilir gelişme: Balıkesir örneği*. Yüksek Lisans Tezi, Balıkesir Üniversitesi Fen Bilimleri Enstitüsü, Balıkesir, 10-17.
19. Radaideh, J. A. (2017). Effect of Meteorological Variables on Air Pollutants Variation in Arid Climates. *Journal of Environmental & Analytical Toxicology*, 07(04), 1-12.
20. Kara, G. (2012). Kentsel Hava Kirleticilerine Meteorolojinin Etkisi. *Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi*, 27(3), 73–86.
21. Sari, M. F. ve Esen, F. (2019). PM10 ve SO2 Konsantrasyonları ve Meteorolojik Parametrelerin Konsantrasyonlar Üzerinde Etkileri. *Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 8(2), 689-697.
22. Tepe, A. M. ve Doğan, G. (2019). Türkiye’nin Güney Sahilinde Yer Alan Dört Şehrin Hava Kalitelerinin İncelenmesi. *Mühendislik Bilimleri ve Tasarım Dergisi*, 7(3), 585–595.
23. Öztürk, D. Ve Bayram, T. (2019). Van İli Kent Merkezinde Hava Kirliliği. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 8(3), 1142–1153.
24. Sindhu, V., Nivedha, S. and Prakash, M. (2020). An Empirical Science Research On Bioinformatics In Machine Learning. *Journal Of Mechanics Of Continua And Mathematical Sciences*, 7(1), 86-94.

25. İnternet: Wikipedi. (2022). *Makine öğrenimi* - Wikipedi. Web: https://tr.wikipedia.org/wiki/Makine_öğrenimi#cite_note-Alpaydin2020-1. Son Erişim Tarihi: 15.09.2021
26. Alpaydin, E. (2020). *Introduction to Machine Learning* (Fourth eddition). England: MIT Press, 4-10.
27. Masih, A. (2019). Machine learning algorithms in air quality modeling. *Global Journal of Environmental Science and Management*, 5(4), 515–534.
28. Breiman, L. (2001). Random Forests. *Machine Learning* , 45(1), 5–32.
29. Huo, J., Shi, T. and Chang, J. (2016). Comparison of Random Forest and SVM for electrical short-Term load forecast with different data sources. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 0*, 1077–1080.
30. Uyanık, T., Karatuğ, Ç., ve Arslanoğlu, Y. (2020). Machine learning approach to ship fuel consumption: A case of container vessel. *Transportation Research Part D: Transport and Environment*, 84, 102389.
31. Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197–227.
32. Kudoh, T. and Matsumoto, Y. (2000). Use of Support Vector Learning for Chunk Identification. *Proceedings of CoNLL-2000 and LLL-2000*, 1,142–144.
33. Mountrakis, G., Im, J. and Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247–259.
34. Deters, J. K., Zalakeviciute, R., Gonzalez, M. and Rybarczyk, Y. (2017). Modeling PM_{2.5} urban pollution using machine learning and selected meteorological parameters. *Journal of Electrical and Computer Engineering*, 1–14.
35. Wu, J., Cai, Z. and Zhu, X. (2013). Self-adaptive probability estimation for Naive Bayes classification. *Proceedings of the International Joint Conference on Neural Networks*.1-8.
36. Yao, Z. and Ruzzo, W. L. (2006). A regression-based k nearest neighbor algorithm for gene function prediction from heterogenous data. *BMC Bioinformatics*, 7(1), 1–11.
37. Singh, A., Thakur, N. and Sharma, A. (2016). A Review of Supervised Machine Learning Algorithms. *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315.
38. Dey, A. (2016). Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, 7(3), 1174–1179.
39. Metz, M., Lesnoff, M., Abdelghafour, F., Akbarinia, R., Massegia, F. and Roger, J. M. (2020). A “big-data” algorithm for KNN-PLS. *Chemometrics and Intelligent Laboratory Systems*, 203.

40. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). Classification and regression trees. *Classification and Regression Trees*. Monterey: CRC Press. 199-215.
41. Liu, H., Wu, H., Lv, X., Ren, Z., Liu, M., Li, Y., and Shi, H. (2019). An intelligent hybrid model for air pollutant concentrations forecasting: Case of Beijing in China. *Sustainable Cities and Society*, 47, 205-231.
42. Kaplan, Y., Saray, U., and Azkeskin, E. (2014). Which cause air pollution PM10 and SO2 using by artificial neural network of estimation and error rate calculation. *Afyon Kocatepe University Journal of Sciences and Engineering*, 14(2), 1–6.
43. Sindhu, V., Nivedha, S. ve Prakash, M. (2020). An Empirical Science Research On Bioinformatics In Machine Learning. *Journal Of Mechanics Of Continua And Mathematical Sciences*, 7(1), 86-94.
44. Bui, T.-C., Le, V.-D., and Cha, S. K. (2018). A Deep Learning Approach for Forecasting Air Pollution in South Korea Using LSTM, *Department of Electrical and Computer Engineering Seoul National University*, 1-6.
45. Lv, J., Guo, C., Shen, Z. P., Zhao, M., and Zhang, Y. (2007). *Summary of Artificial Neuron Model Research*. Industrial Electronics Conference, 677–682.
46. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R. and Schmidhuber, J. (2016). Lstm: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 1-12.
47. İnternet: Olah, C. (2015). *Understanding LSTM Networks* . Web: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> Son Erişim Tarihi: 10.09.2021
48. Lin, T., Horne, B., Tino, P., and Giles. (1996). Learning long-term dependencies in NARX recurrent neural networks, *IEEE Transactions on Neural Networks*, 7(6), 1329–1338.
49. Tsai, Y. T., Zeng, Y. R. and Y. S. Chang (2018). "Air Pollution Forecasting Using RNN with LSTM." 2018 The 16th IEEE International Conference on Pervasive Intelligence and Computing, 1074-1079.
50. Ma, J., Ding, Y., Gan, V. J. L., Lin, C. and Wan, Z. (2019). Spatiotemporal Prediction of PM2,5 Concentrations at Different Time Granularities Using IDW-BLSTM. *IEEE Access*, 7(1), 107897–107907.
51. Deleawe, S., Kuszniir, J., Lamb, B., and Cook, D. J. (2010). Predicting air quality in smart environments. *Journal of Ambient Intelligence and Smart Environments*, 2(2), 145–154.
52. Zhao, H., Zhang, J., Wang, K., Bai, Z. P., and Liu, A. (2010). A GA-ANN model for air quality predicting. *ICS 2010 - International Computer Symposium*, 693–699.

53. Rybarczyk, Y., and Zalakeviciute, R. (2016). Machine learning approach to forecasting urban pollution. *2016 IEEE Ecuador Technical Chapters Meeting, ETCM 2016*.
54. Yu, R., Yang, Y., Yang, L., Han, G., and Move, O. A. (2016). RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. *Sensors*, 16(1),86.
55. K k, I., ŐimŐek, M. U., and  zdemir, S. (2017). A deep learning model for air quality prediction in smart cities. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018-January*, 1983–1990.
56. Reddy, V., Yedavalli, P., and Mohanty, S. (2018). Deep Air: Forecasting Air Pollution in Beijing, China. *Environmental Science*,1,1-8.
57. Huang, C. J., and Kuo, P. H. (2018a). A Deep CNN-LSTM Model for Particulate Matter (PM_{2.5}) Forecasting in Smart Cities. *Sensors 2018*, 18(7), 2220.
58. Liu, J., Weng, F., and Li, Z. (2019). Satellite-based PM_{2.5} estimation directly from reflectance at the top of the atmosphere using a machine learning algorithm. *Atmospheric Environment*, 208, 113–122.
59. Ayturan, Y.(2019). *Derin  ğrenme ile havadaki partik l madde konsantrasyonu tahmini*.Y ksek Lisans Tezi. KTO Karatay  niversitesi Fen Bilimleri Enstit s , Konya.
60. Xayasouk, T., Lee, H. M., and Lee, G. (2020). Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models. *Sustainability 2020, Vol. 12, Page 2570*, 12(6), 2570.
61. Bozdağ, A., Dokuz, Y., and G k ek,  . B. (2020). Spatial prediction of PM₁₀ concentration using machine learning algorithms in Ankara, Turkey. *Environmental Pollution*, 263, 114635.
62. Faızı, S. (2021). *Kabil'in hava kalitesi tahmininde makine  ğrenmesi algoritmalarının kullanılması*.Y ksek Lisans Tezi,Niğde  mer Halisdemir  niversitesi, Fen Bilimleri Enstit s , Niğde, 70-71.
63. G k ek,  .B., ŐaŐa, N., Dokuz, Y., ve Bozdağ, A. (2022). PM₁₀ Parametresinin Makine  ğrenmesi Algoritmaları ile Mek nsal Analizi, Kayseri İli  rneđi.*Dokuz Eyl l  niversitesi Fen ve M hendislik Dergisi*, 24(70), 65-80.
64.  naldı, S., ve Yal ın, N. (2022). Hava Kirliliđinin Makine  ğrenmesi Tabanlı Tahmini: BaŐakŐehir  rneđi. *M hendislik Bilimleri ve AraŐtırmaları Dergisi*, 4(1), 35–44.
65. İnternet: Manisa  evre ve Őehircilik M d rl đ . (2020). *Manisa İli Temiz Hava Eylem Planı THEP(2020-2024)*. Web: https://webdosya.csb.gov.tr/db/manisa/menu/temiz-hava-eylem-planı-2020-2024-57-sayfa_20201231100147.pdf Son EriŐim Tarihi: 20.08.2021

66. İnternet: Zonguldak Çevre ve Şehircilik Müdürlüğü. (2014). *Zonguldak İli Temiz Hava Eylem Planı THEP (2014-2019)*. Web: <https://webdosya.csb.gov.tr/db/zonguldak/webmenu/webmenu36983.pdf> Son Erişim Tarihi: 20.08.2021
67. McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1), 12-16.
68. Karasu, S., Altan, A., Sarac, Z., and Hacıoglu, R. (2018). Prediction of bitcoin prices with machine learning methods using time series data. *26th IEEE Signal Processing and Communications Applications Conference, SIU*, 1–4.



GAZİ GELECEKTİR..