e-ISSN 2757-5195

Çanakkale Onsekiz Mart University
Journal of Advanced Research in Natural and Applied Sciences
Open Access

doi.org/10.28979/jarnas.981202 2022, Cilt 8, Sayı 2, Sayfa: 201-213 dergipark.org.tr/tr/pub/jarnas

Makine Öğrenme Algoritmaları ile PM10 Konsantrasyon Tahmini

Kahraman Oğuz[1*], Muhammet Ali Pekin[1],

[1]Meteoroloji Genel Müdürlüğü, Araştırma Dairesi Başkanlığı, Ankara, Türkiye

Makale Tarihçesi
Gönderim: 10.08.2021
Kabul: 19.10.2021
Yayım: 10.06.2022
Araştırma Makalesi

Öz – Partikül madde (PM) kirliliği önemli çevresel sorunlara sebep olmaktadır. PM kirliliğinin olumsuz etkileri, canlı sağlığına yönelik riskleri nedeniyle yaygın bir sorun haline gelmiştir. PM kirliliğinin tüm bu olumsuz etkileri ve atmosferdeki karmaşık etkileşimi sebebiyle, daha fazla çalışmaya konu olması önemlidir. Özellikle, PM kirliliğinin izlenmesi ve tahmin edilmesi konusunda yapılacak çalışmalar önemlidir. Son yıllarda meteorolojik faktörler göz önüne alınarak PM kirliliğinin tahmin edilmesi çalışmaları artmıştır. Özellikle makine öğrenme yöntemleri ile PM kirliliği tahmini çalışmaları hız kazanmıştır. Bu çalışmada, meteorolojik faktörler göz önüne alınarak çeşitli makine öğrenme algoritmaları ile PM10 kirliliği tahmin edilmiştir. Çalışmada kullanılan meteoroloji verileri Meteoroloji Genel Müdürlüğü Ankara Bölge istasyonundan (enlem:39,9727, boylam:32,8637, rakım:891 m.) elde edilmiştir. PM10 kirlilik verileri ise Çevre, Şehircilik ve İklim Değişikliği Bakanlığı Ankara Keçiören-Sanatoryum hava kalitesi istasyonundan (enlem: 39,999, boylam: 32,856, rakım: 1009 m.) elde edilmiştir. Makine öğrenme çalışması aşamasında, sıcaklık, çiğ noktası sıcaklığı, yağış, bağıl nem, rüzgar hızı, basınç, bulut kapalılığı ve bir önceki güne ait PM10 ölçümleri göz önüne alınarak, farklı makine öğrenme (karar ağacı regresyonu, destek vektör regresyonu, lasso regresyonu ve yapay sinir ağı) algoritmalarıyla ayrı ayrı çalışma yapılmış ve bu algoritmaların tutarlılıkları karşılaştırılmıştır. Tutarlılıklarının incelenmesi aşamasında çeşitli istatistiksel metrikler kullanılmıştır. Sonuçta, test bölümü göz önüne alındığında, yapay sinir ağı algoritmasının belirleme katsayısı 0,6, kök ortalama kare hatası 18 ve ortalama mutlak hata 12 olarak bulunmuş ve yapay sinir ağı algoritmasının diğer algoritmalara göre daha iyi sonuç verdiği görülmüştür.

Anahtar Kelimeler – *Ankara-Keçiören, Makine öğrenme algoritmaları, Meteorolojik faktörler, PM10 kirlilik tahmini*

Estimation of PM10 Concentration with Machine Learning Algorithms

[1]Turkish State Meteorological Service, Research Department, Ankara, Türkiye

Article History
Received: 10.08.2021
Accepted: 19.10.2021
Published: 10.06.2022
Research Article

Abstract – Particulate matter (PM) pollution causes significant environmental problems. The adverse effects of PM pollution have become a common problem due to its risks to living health. Due to all these negative effects of PM pollution and its complex interaction in the atmosphere, it is important that it be the subject of more studies. In particular, studies on monitoring and estimating PM pollution are important. In recent years, studies on estimating PM pollution have increased by considering meteorological factors. Especially with machine learning methods, PM pollution estimating has accelerated. In this study, PM10 pollution is estimated with various machine learning algorithms considering meteorological factors. The meteorological data used in the study were obtained from the Ankara Regional Station of Turkish State Meteorological Service (latitude: 39,9727, longitude: 32,8637, altitude: 891 m.). PM10 pollution data were obtained from the Ministry of Environment, Urbanization and Climate Change Ankara Keçiören-Sanatorium air quality station (latitude: 39,999, longitude: 32,856, altitude: 1009 m.). In the machine learning phase, different machine learning (decision tree regression, support vector regression, lasso regression and neural network) were used, considering temperature, dew point temperature, precipitation, relative humidity, wind speed, pressure, cloud cover and PM10 measurements of the previous day. Algorithms were studied separately and the consistencies of these algorithms were compared. Various statistical metrics were used to examine their consistency. As a result, considering the test section, the determination coefficient was found to be 0,6, root mean square error 18, and mean absolute error 12 for artificial neural network algorithm, and it was seen that the artificial neural network algorithm gave better results than other algorithms.

Keywords – *Ankara-Keçiören, Machine learning algorithms, Meteorological factors, PM10 pollution forecast.*

[1] koguz@mgm.gov.tr*
[2] mapekin@mgm.gov.tr
*Sorumlu Yazar / Corresponding Author

1. Giriş

Hava kirliliği, sanayileşmenin ve nüfus artışının kaçınılmaz etkisiyle beraber önemli çevresel ve atmosferik sorunlardan birisi haline gelmiştir. Hava kirliliği, kentsel ve endüstriyel alanların önemli konularından biridir. Hava kirleticilerinin olumsuz etkileri, canlı sağlığına yönelik riskleri nedeniyle yaygın bir sorun haline gelmiştir. Partikül madde (PM) kirliliği, ABD Çevre Koruma Ajansı (US EPA) tarafından önemli hava kirletici kriterlerinden birisi olarak tanımlanır (Özdemir ve Taner, 2014). PM kirliliği, kalp ve akciğer hastalıklarına neden olur, atmosferik görüş mesafesini azaltır, gölleri ve akarsuları asidik hale getirir, nehir havzalarında besin dengesini değiştirir, ormanlara ve tarım ürünlerine zarar verir, ekosistem çeşitliliğini etkiler, asit yağmurlarına neden olur, heykeller gibi kültürel açıdan önemli nesnelere zarar verir ve neticede maddi kayıplara da neden olur (US EPA, 2021).

PM konsantrasyonunu etkileyen temel faktörler arasında emisyon kaynakları ve meteorolojik faktörler bulunmaktadır. Meteorolojik faktörlerle PM konsantrasyonu arasında önemli bir ilişkinin varlığı, pek çok çalışma tarafından ele alınmıştır (Hrdlickova vd., 2008; Ei-Sharkawy vd., 2015; Oğuz, 2020). PM'nin taşınımı, kimyası ve çökelmesi, meteorolojik faktörler tarafından kontrol edilmektedir. Meteorolojik faktörler birbirini etkilemekle birlikte, PM ile yakından bağlantılı bir sistem oluşturur. Meteorolojik koşulların PM konsantrasyonu üzerindeki etkileri oldukça karmaşıktır (Qin vd., 2019). PM kirliliğinin tüm bu olumsuz etkileri ve atmosferdeki karmaşık etkileşimi sebebiyle, daha fazla çalışmaya konu olması önemlidir. PM konsantrasyonunun tahmini amaçlı çeşitli yöntemler bulunmaktadır. Makine öğrenme yöntemi, son yıllarda sıkça kullanılan tahmin yöntemlerinden birisidir.

Makine öğrenimi, hızla gelişen ve bilgisayarların verilere dayalı olarak öğrenmesini sağlayan bir alandır. Veriler, fiziksel deneyler, bilgisayar modelleri veya her ikisinin birleşimi dahil olmak üzere çeşitli kaynaklardan gelebilir. Makine öğreniminin birçok alanda başarılı uygulamaları olmuştur (Panda vd., 2020). Son otuz yılda, makine öğrenimi tekniklerine dayalı istatistiksel modeller geliştirilmesi ve çoklu ve karmaşık veri setlerini keşfetme, analiz etme ve tahminlerde bulunma kabiliyeti nedeniyle hava kalitesi tahmin alanında da giderek daha fazla uygulanmaktadır. Bir makine öğrenimi algoritmasının temel amacı, verilerden bilgi elde etmek ve tahminlerde bulunmak için eğitim aldığı (öğrendiği) veri kümesinin genel özelliklerini ve etkileşimlerini yakalayan bir model sağlamaktır (Alpaydin, 2010; Gagliardi ve Andenna, 2020). Bilinen en yaygın makine öğrenme algoritmalarından biri, değişkenler arasındaki doğrusal olmayan ilişkileri keşfetme yeteneği olan yapay sinir ağlarıdır (YSA). Bu konuda kullanılan diğer bir algoritma, regresyon amacıyla kullanıldığında destek vektör regresyonu (DVR) olarak da adlandırılan destek vektör makinesidir. Bu algoritma, yeni verilere genelleme yapan iyi bir genelleme algoritmasıdır. Üçüncü algoritma, grafiksel bir ters ağaç yapısına sahip, iyi bilinen bir makine öğrenme algoritması olan karar ağacıdır. Regresyon için bir karar ağacı kullanıldığında, buna karar ağacı regresyonu (KAR) denir (Aljanabi, 2020).

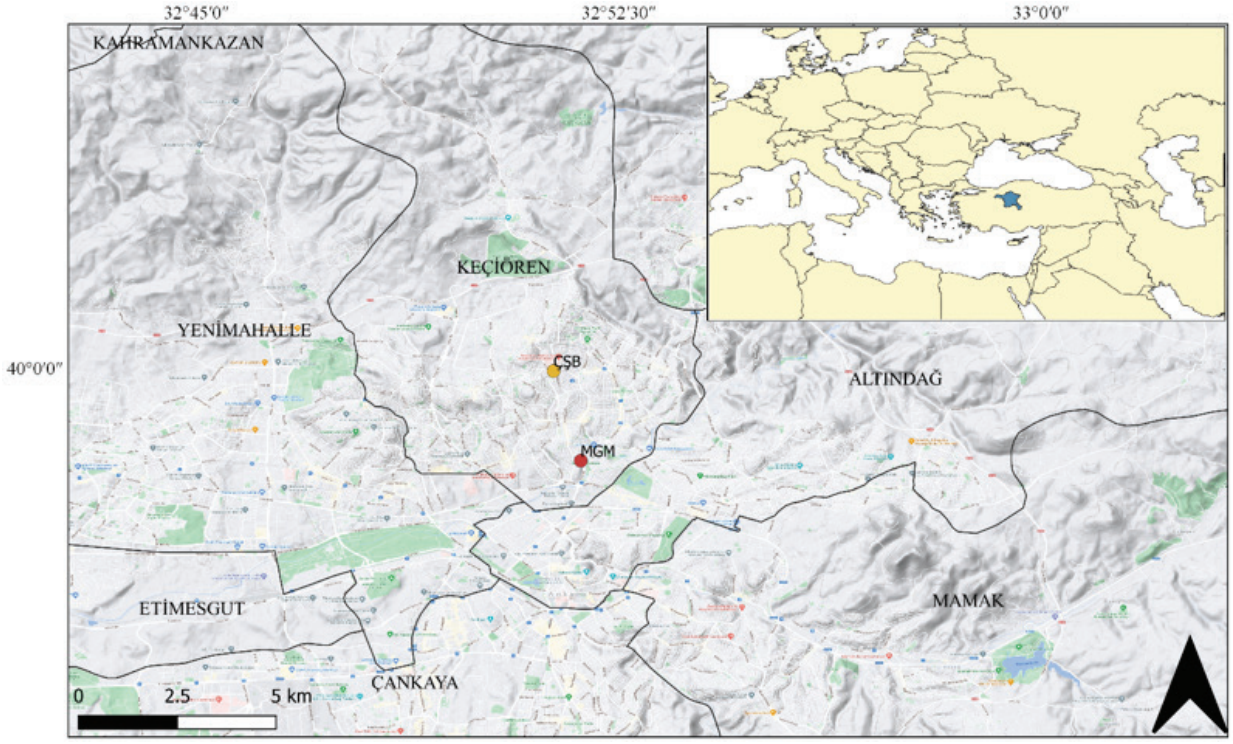
PM konsantrasyonunun tahmini amacıyla makine öğrenme algoritmalarının uygulandığı çeşitli çalışmalar bulunmaktadır. Gültepe (2019) çalışmalarında, Kastamonu ilini ele alarak çeşitli makine öğrenmesi algoritmaları ile hava kirliliğinin tahmininde, bazı meteorolojik değişkenler kullanarak hava kirliliği tahmini yapacak modeller geliştirmişlerdir. Algoritmalardan YSA algoritması için doğru tahmin oranı %87 ve diğer makine öğrenmesi algoritmalarından Rastgele Orman %99 ve KAR %99 değerleri ile en iyi sonuçları verdiği sonucuna ulaşmışlardır. Suleiman ve diğerleri (2019) çalışmalarında, Londra'da çeşitli istasyonlardan elde edilen trafik kaynaklı PM verilerinin makine öğrenme algoritmaları ile tahminini yapmışlardır. YSA algoritmasının trafik kaynaklı PM tahmininde etkin bir şekilde kullanılabileceği sonucuna ulaşmışlardır. Castelli ve diğerleri (2020) çalışmalarında, Kaliforniya'da PM ve hava kalitesini tahmin etmek için makine öğrenme algoritmalarını uygulamışlardır. DVR algoritmasının saatlik kirletici konsantrasyonlarının tutarlı bir şekilde tahmin edilmesini sağladığı sonucuna ulaşmışlardır. Czernecki ve diğerleri (2021) çalışmalarında, Polonya'da çeşitli istasyonlardan elde edilen PM verilerinin makine öğrenme algoritmaları ile tahminini yapmışlardır. KAR tabanlı bir algoritma olan XGBoost algoritmasının PM tahmininde en iyi algoritma olduğu, sonrasında ise YSA algoritmasının geldiği sonucuna ulaşmışlardır.

Bu çalışmada, meteorolojik faktörler göz önüne alınarak çeşitli makine öğrenme algoritmaları ile PM10 konsantrasyonu tahmin edilmiştir. Makine öğrenme algoritmalarının uygulanması aşamasında, sıcaklık, çiğ noktası sıcaklığı, yağış, bağıl nem, rüzgar hızı, basınç, bulut kapalılığı ve bir önceki güne ait PM10 ölçümleri göz önüne alınarak, farklı makine öğrenme (KAR, DVR, Lasso, YSA) algoritmalarıyla ayrı ayrı çalışma yapılmış ve bu algoritmaların tutarlılıkları karşılaştırılmıştır.

2. Materyal ve Yöntem

2.1. Çalışma Alanı ve Veri

Bu çalışmada, günlük meteoroloji verileri (sıcaklık, çiğ noktası sıcaklığı, yağış, bağıl nem, rüzgar hızı, basınç, bulut kapalılığı) Meteoroloji Genel Müdürlüğü Ankara Bölge istasyonundan (enlem:39,9727, boylam:32,8637, rakım:891 m.) elde edilmiştir (MGM, 2021). PM10 kirlilik verileri (günlük) ise Çevre, Şehircilik ve İklim Değişikliği Bakanlığı Ankara Keçiören-Sanatoryum hava kalitesi istasyonundan (enlem: 39,999, boylam: 32,856, rakım: 1009 m.) elde edilmiştir (ÇŞBHKİ, 2021). Çalışmada kullanılan verilerden yağış verisi günlük toplam iken, diğer tüm veriler günlük ortalama olarak elde edilmiştir. Ankara Keçiören-Sanatoryum hava kalitesi istasyonu, meteoroloji istasyonuna yakın olması sebebiyle ve düzenli ölçümlerinin bulunması sebebiyle tercih edilmiştir. Bu istasyonların konumları Şekil 1'de gösterilmektedir. Ele alınan veriler Ocak 2018-Temmuz 2021 dönemini kapsamaktadır. Makine öğrenme yöntemiyle uygulama yapılırken, veri setinin rastgele (randomize) seçilen %80'lik kısmı (888 satır) eğitim verisi, %20'lik kısmı (223 satır) test verisi olarak kullanılmıştır.



Şekil 1. Çalışmada verileri kullanılan ölçüm istasyonlarının konumu (Sarı nokta: Çevre, Şehircilik ve İklim Değişikliği Bakanlığına ait hava kalite istasyonunu, kırmızı nokta: Meteoroloji Genel Müdürlüğüne ait meteoroloji istasyonunu gösterir).

Tablo 1'de çalışma dönemi için meteorolojik parametrelerin ve PM10'un istatistiki bilgileri görülmektedir. PM10 ortalaması 45,9 µg/m³, minimum ve maksimum değerleri ise sırasıyla 5,2 µg/m³ ve 295,5 µg/m³ olarak bulunmuştur. Çalışmada kullanılan verilerin birbiri ile olan korelasyonu Tablo 2'de görülmektedir. Şekil 2'de ise meteorolojik parametrelerin PM10 ile saçılım grafiği görülmektedir. PM10 ile en yüksek korelasyon katsayısının (r=-0,319) rüzgar hızı ile olduğu görülmüştür. Bunu ise basınç (r=0,233) takip etmiştir. PM10 ile rüzgar hızı arasında negatif yönlü zayıf bir ilişkinin varlığından söz edilebilir. PM10 ile basınç arasında ise pozitif yönlü zayıf bir ilişkinin varlığından söz edilebilir. PM10 diğer parametreler arasında ise korelasyon katsayısı r<0,2 olup, ilişkinin çok zayıf olduğundan söz edilebilir.

Tablo 1

Çalışmada kullanılan verilerin tanımlayıcı istatistikleri

	PM10	T	TD	PRC	RH	W	Q	C
Ortalama	45,9	13,755	3,27	1,096	55,511	1,935	913,303	3,135
Ortalamanın Standart Hatası	0,923	0,251	0,162	0,094	0,515	0,029	0,144	0,067
Medyan	38,5	13,6	2,9	0,0	55,8	1,8	913,2	2,9
Mod	21,1	17,7	2,2	0,0	54,5	1,3	910,7	0,0
Standart Sapma	30,75	8,376	5,394	3,119	17,181	0,967	4,808	2,25
Çarpıklık	2,256	-0,051	-0,127	4,652	0,076	1,219	0,111	0,216
Çarpıklığın Standart Hatası	0,073	0,073	0,073	0,073	0,073	0,073	0,073	0,073
Minimum	5,2	-8,4	-13,2	0,0	13,5	0,3	896,2	0,0
Maksimum	295,5	31,9	15,1	32,6	97,8	7,0	928,5	8,0

T: Temperature-Sıcaklık (°C), TD: Dew point temperature-Çiğ noktası sıcaklığı (°C),
PRC: Precipitation-Yağış (mm), RH: Relative humidity-Bağıl nem (%), W: Wind-Rüzgar hızı (m sñ¹),
Q: Pressure-Basınç hpa, C: Ceiling-Bulut kapalılığı (1/8)

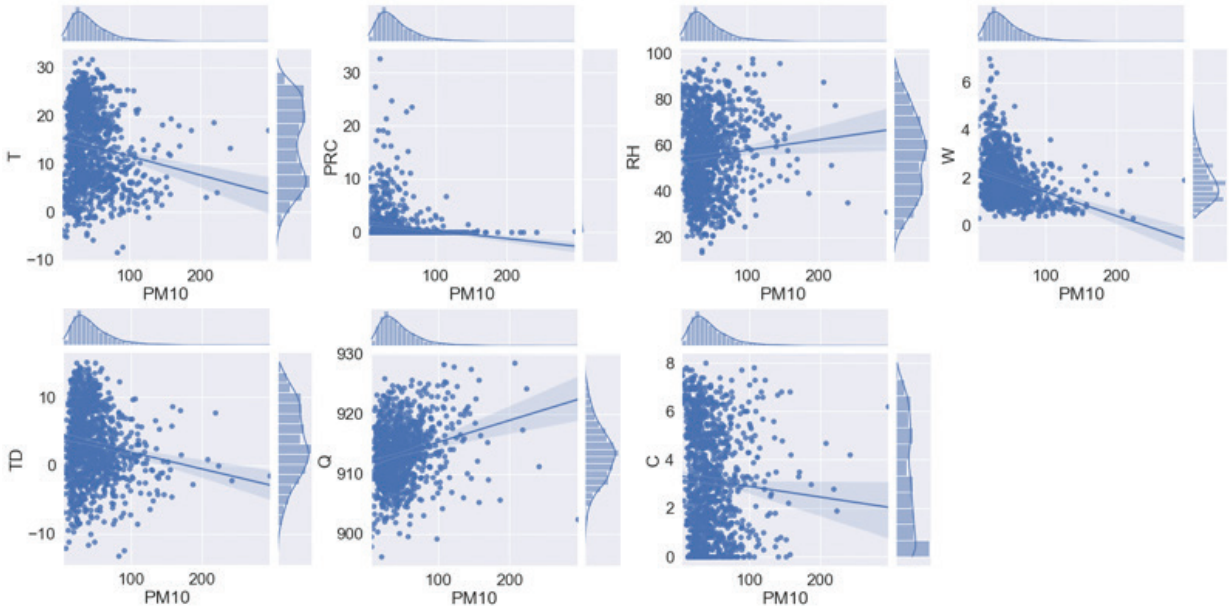
Tablo 2

Çalışmada kullanılan verilerin birbiri ile olan korelasyonu

Parametre		PM10	T	TD	PRC	RH	W	Q	C
PM10	Pearson's r	—							
	p-value	—							
T	Pearson's r	-0,146	—						
	p-value	< ,001	—						
TD	Pearson's r	-0,140	0,719	—					
	p-value	< ,001	< ,001	—					
PRC	Pearson's r	-0,147	-0,145	0,132	—				
	p-value	< ,001	< ,001	< ,001	—				
RH	Pearson's r	0,080	-0,718	-0,061	0,394	—			
	p-value	0,008	< ,001	0,042	< ,001	—			
W	Pearson's r	-0,319	0,293	0,173	-0,072	-0,265	—		
	p-value	< ,001	< ,001	< ,001	0,017	< ,001	—		
Q	Pearson's r	0,233	-0,312	-0,350	-0,260	0,063	-0,125	—	
	p-value	< ,001	< ,001	< ,001	< ,001	0,037	< ,001	—	
C	Pearson's r	-0,060	-0,422	0,045	0,371	0,668	-0,107	-0,283	—
	p-value	0,044	< ,001	0,134	< ,001	< ,001	< ,001	< ,001	—

2.2. Makine Öğrenme Algoritmaları

Bu çalışmada, çeşitli makine öğrenme algoritmaları kullanılarak uygulama yapılmıştır. Ele alınan algoritmalar yapay sinir ağı, destek vektör regresyonu, karar ağacı regresyonu ve lasso algoritmalarından



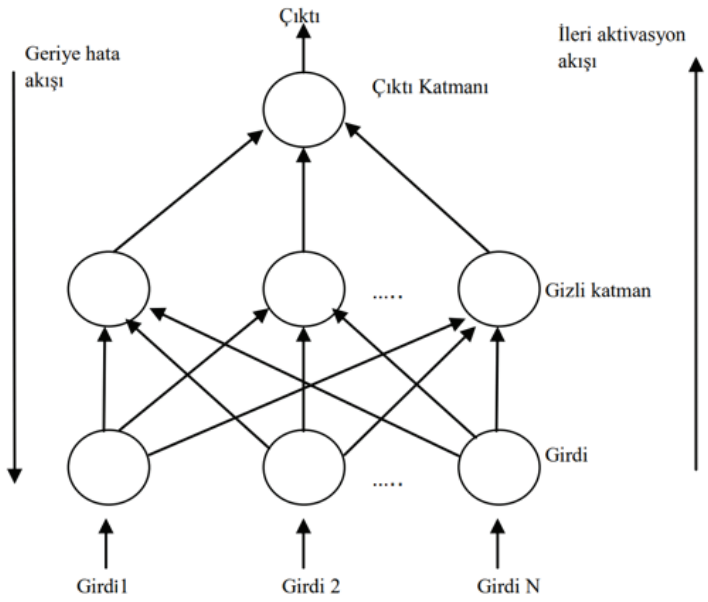
Şekil 2. Çalışmada kullanılan verilerin saçılım ve yoğunluk grafikleri

oluşmaktadır. Makine öğrenme algoritmalarının hiperparametreleri deneme yanılma yöntemi ile belirlenmiştir. Makine öğrenmede verilerin normalize edilmesi, tahmin başarısını artırmaktadır (Singh ve Singh, 2019). Bu çalışmada da maksimum-minimum normalizasyon yöntemi uygulanarak veriler ön işlemden geçirilmiştir.

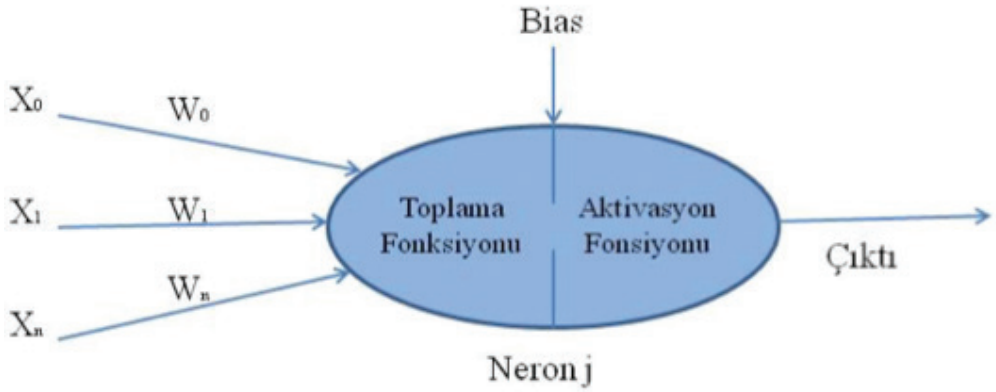
2.2.1. Yapay Sinir Ağı

Yapay sinir ağı (YSA), insan beyninin sinirsel yapısını taklit eden bir veri işlemedir. Girdiler ve çıktılar arasında ilişkiler kurar. İnsan sinir sistemi gibi paralel veri işleme mimarisine sahiptir (Haykin, 1999). İnsan sinir sisteminin temel öğesi, dört temel bileşeni olan bir nörondur. Nöronlar ağırlıklı girdiler alır, bunları birleştirir, doğrusal olmayan işlem uygular ve çıktıyı verir. Dolayısıyla bir YSA'nın temel işlem elemanı olan yapay nöronun, doğal nöronlar gibi dört işlevi vardır. Bu yapay nöronların kümelenmesi yapay sinir ağını oluşturur. Bu kümeleme, daha sonra birbiriyle ilişkilendirilen katmanlar oluşturarak gerçekleşir. Çoğu YSA uygulamasında uygulama, giriş, gizli ve çıkış katmanları olmak üzere birbirine bağlı üç katmana sahiptir. Şekil 3'de bir YSA mimarisi görülmektedir. Bu YSA'lar çok katmanlı algılayıcı (MLP) olarak bilinir (Yaseen vd., 2015). Katmanlar arasındaki bağlantı, bir YSA'nın en önemli özelliklerinden biridir. İleri beslemeli veya geri bildirimli olabilir. Bir YSA'nın en bilinen avantajlarından biri, öğrenebilmesidir. Öğrenme, tahmin edilen ve gözlenen değerler arasındaki hatayı en aza indirmek için ağırlıkları ayarlayarak gerçekleşir (Alizamir vd., 2020).

YSA algoritma katmanları, bağlantılarla birbirine bağlı yapay sinir hücreleri aracılığıyla oluşur. Yapay sinir hücresi, Şekil 4'de gösterilen girdi, toplama fonksiyonu, aktivasyon fonksiyonu ve çıktı kısımlarından oluşur. Girdi kısmında (X), dışardan elde edilen veriler yapay sinir hücresine giriş sağlar. Ağırlık (W) ise, yapay sinir hücresi için önemini içeren değerleri ifade etmektedir. Tüm girdiler hücre açısından önem ve etkilerine göre bir ağırlığa sahip olur. Yönüne göre pozitif veya negatif değerler alabilir. Sonraki adım olan toplama fonksiyonu, hücreye gelen net girdinin analiz edildiği kısımdır. Girdilerle bunların ağırlıklarının çarpımları sonucu elde edilen değerlerin toplamları, sonrasında eşik değerlerle toplanır. Elde edilen değerler aktivasyon fonksiyonuna girdi oluşturur. Aktivasyon fonksiyonu, toplama fonksiyonundan aldığı net girdiyi işleyerek çıktıyı oluşturacak değerleri üretir. Aktivasyon fonksiyonu doğrusal olmamakla birlikte çeşitlilik gösterebilmektedir. Toplama ve Aktivasyon fonksiyonları nöron içerisinde bulunur. Tüm bu verilerin işlenmesi sonucunda elde edilen değerlerin ulaştığı yer ise çıktı kısmıdır. Bir sinir hücresinin girdi sayısı çoklu olsa da sadece bir tane çıktısı olabilir (Aydoğan ve Zırhlıoğlu, 2017). Bu çalışmada, 2 farklı nöron sayısı (YSA1= 150x100 ve YSA2= 100x100) göz önüne alınarak YSA ile tahmin yapılmıştır.



Şekil 3. Yapay sinir ağı algoritması (Karaatlı, 2012).



Şekil 4. Yapay sinir hücresinin bileşenleri

2.2.2. Destek Vektör Regresyonu

Destek Vektör Regresyonu (DVR), birçok farklı regresyon probleminde iyi sonuçlar veren, regresyon ve fonksiyon yaklaşımı için en gelişmiş algoritmalardan biridir (Smola ve Scholkopf, 2004). DVR algoritmaları, yalnızca verilerin hata tahminlerini değil, aynı zamanda regresyon modelinin genelleştirilmesini de hesaba kattıkları için çok çeşitli regresyon problemleri için çok kullanışlıdır (Carro-Calvo vd., 2017). DVR, 2-boyutlu uzayda lineer olarak ayrılamaz olsalar bile kullanılır. DVR, sınıfları lineer olarak ayrılabilir hale getirmek amacıyla daha yüksek boyutlu uzaya dönüştüren bir hile tanıtarak elde eder ve bu hileye çekirdek hilesi denir. DVR'de sınıflandırma, sınıflar arasında yapılması yerine, temelde belirli bir eşiğin üzerinde veya altında olan regresyon hatalarının sınıflandırılması mantığına dayanır (Abuella ve Chowdhury, 2016). DVR işlemi, sınıflandırma adımlarında oluşturulan her sınıflamada uygulanır. DVR, lineer olmayan girdiyi, gerçek dünyadaki problemler nedeniyle boyutun daha yüksek olduğu özellik alanına dönüştürmek için çekirdek fonksiyonlarını kullanır ve genellikle nadiren lineer olarak ayrılabilir.

Temel çekirdek fonksiyonları lineer, polinom, gauss fonksiyonlarıdır. Lineer fonksiyon (Denklem 2.1):

$$k(x, y) = x^T y + C \quad (2.1)$$

Polinom fonksiyon (Denklem 2.2):

$$k(x, y) = (\alpha x^T y + C)^d \quad (2.2)$$

Gauss fonksiyonu (Denklem 2.3):

$$k(x, y) = \exp(-¥ \|x - y\|^2) \quad (2.3)$$

denklemi ile çözümlenir. Burada, x ve y, eğitim seti vektörlerini ifade ederken, ¥ ise kernel parametresini ifade eder. C ve ¥ parametreleri, DVR'nin hiper parametreleridir. Eğitim vektörleri, α fonksiyonu tarafından daha yüksek boyutlu bir uzaya eşlenir. Her çekirdek işlevi, öncelikle atanması gereken parametre değerine sahiptir. Parametre C değeri, lineer çekirdek işlevine atıfta bulunur. Parametre C, y, T ve d değeri polinom çekirdek fonksiyonuna atıfta bulunurken, ¥ parametresi gauss çekirdek fonksiyonuna atıfta bulunur. Parametre değerleri, elde edilen DVR algoritması üzerinde büyük etki sağlar. Daha optimal parametre, daha iyi sonuçlanmış tahmin anlamına gelir (Adhani vd., 2013).

2.2.3. Karar Ağacı Regresyonu

Karar Ağacı, akış şeması benzeri bir ağaç yapısı kullanan veya kararlar ve çıktılar, girdi maliyetleri ve fayda dahil tüm olası sonuçlarından oluşan bir model gibi olabilen bir makine öğrenme algoritmasıdır. Denetimli öğrenme algoritmalarının bir türüdür. Karar ağaçları, kategorik veya sürekli olmak üzere her iki çıktı değişkeni türü için de kullanılabilir. Karar ağacı regresyon algoritması, bir nesnenin özelliklerini tespit eder. Karar ağacı regresörü, ağaç benzeri bir oluşumda bir sistemi eğitir ve anlamlı sürekli çıktıya sahip olmak için geleceğe yönelik verileri tahmin eder. Sürekli çıktının anlamı, bilinen değerler veya sayılar kümesiyle gösterilmediği anlamına gelir (Badarpura vd., 2020). Bir karar ağacı algoritmasının ana bileşenleri düğümler ve dallardır ve bir sistem oluşturmanın en önemli adımları ise bölme, durdurma ve budamadır. Karar ağacı regresyonu (KAR) uygulamasındaki en temel sorun, her seviyede kök düğüm için özniteliği seçmektir. Bu problemin üstesinden gelmek için bilgi kazancı ve gini indeksi olmak üzere iki nitelik seçim ölçütü vardır. Bilgi kazancı seçim (Denklem 2.4) ve gini indeksi ölçütü (Denklem 2.5) formülasyonları aşağıdaki şekildedir (Hariskumar vd., 2020).

$$\text{Bilgi Kazancı (G, A)} = \text{Entropi (S)} - \sum_{v \in Değerler(A)} \frac{|Sv|}{|S|} \text{Entropi(Sv)} \quad (2.4)$$

burada S: örnekler kümesi, A: bir nitelik, Sv: S'nin alt kümesi ve değerler(A): A'nın tüm olası değerlerinin kümesidir.

$$\text{Gini} = 1 - \sum_{i=1}^{n} (pi)^2 \quad (2.5)$$

burada pi: bir nesnenin belirli bir sınıfa sınıflandırılma olasılığıdır.

2.2.4. Lasso Regresyonu

Lasso regresyonu, elde edilen istatistiksel algoritmanın tahmin doğruluğunu ve yorumlanabilirliğini geliştirmek için hem değişken seçimini hem de düzenlileştirmeyi gerçekleştiren bir regresyon analizi yöntemidir. Lasso regresyonu, küçülmeyi kullanan bir tür lineer regresyondur. Küçülme, veri değerlerinin (ortalama gibi) merkezi bir noktaya doğru küçüldüğü yerdir. Lasso basit, seyrek modelleri (yani daha az parametreli modelleri) teşvik eder. Bu özel regresyon türü, çoklu bağlantı gösteren modeller için veya değişken seçimi/parametre elenmesi gibi model seçiminin belirli kısımlarını otomatikleştirmek istendiğinde

oldukça faydalıdır. Lasso regresyonu özetle, örneklem büyüklüğüne kıyasla tahmin edicilerin sayısı büyük olduğunda bir regresyon modeline uyma sorununu çözmek için geliştirilmiştir. Lasso tahminleri, katsayı tahminlerinin mutlak değerlerinin toplamına bir sınır ile karesel hataların toplamını en aza indirerek tanımlanır (Sun vd., 2013). Her biri p ortak değişkenden ve n olaydan oluşan bir örnek varsayarsak, Lasso aşağıdaki formülasyonla (Denklem 2.6) hesaplanır:

$$\sum_{i=1}^{n}(yi - \sum_j xij.\beta j)^2 + \lambda \sum_{j=1}^{p} |\beta j| \quad 2.6$$

Burada βj, farklı bağımsız değişkenler için katsayı tahminlerini temsil eder ve sırasıyla her bir özelliğe eklenen ağırlıkları veya büyüklüğü tanımlar. Bunun yanında, yi hedef vektörü, xij i.durum için ortak değişken vektörünü ifade eder. λ ise, daha yüksek λ değerleri için artan küçülme yoluyla küçülme miktarını kontrol eden ve negatif olmayan bir ayar parametresidir. Kullanıcı tanımlı sabiti ifade eder (Musoro, 2014).

2.3. Tahmin Değerlendirme Metrikleri

Farklı makine öğrenme yöntemleri ile gerçekleştirilen tahmin sonuçlarının değerlendirilmesinde çeşitli istatistiksel metrikler kullanılmıştır. Bu metrikler korelasyon katsayısı (CC), belirleme katsayısı (R2), kök ortalama kare hatası (RMSE) ve ortalama mutlak hata (MAE)'dır.

CC, veriler arasındaki ilişkinin ne kadar güçlü olduğunu belirlemek amacıyla kullanılır. -1 ve 1 arasında değişir. 1 değeri güçlü pozitif ilişkiyi gösterirken, -1 değeri güçlü negatif ilişkiyi gösterir. Aşağıdaki formül (Denklem 2.7) ile hesaplanır:

$$CC = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} \quad 2.7$$

burada xi ve yi, i'inci veri için x ve y değişkenlerinin değerleridir.

R2, verilerin uygun regresyon çizgisine ne kadar yakın olduğunu gösteren istatistiksel bir ölçüdür. 0-1 arasında değişir. Aşağıdaki formül (Denklem 2.8) ile hesaplanır:

$$R2 = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{(n\sum x^2-(\sum x)^2)(n\sum y^2-(\sum y)^2)}} \quad 2.8$$

burada x ve y değişkenleri ifade etmektedir.

RMSE, tahminlerin performansını değerlendirmek için yaygın olarak kullanılan bir istatistiksel hatadır. Tahmin hatalarının standart sapmasını ifade eder. Aşağıdaki formül (Denklem 2.9) ile hesaplanır:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y-x)^2} \quad 2.9$$

burada n veri sayısını, y gerçek veriyi, x tahmin edilen veriyi ifade eder.

MAE, tahmin edilen değerler ile gözlemlenen değerler arasındaki farkları belirleyerek tahmin doğruluğunu ölçen istatistiksel bir ölçüdür. Aşağıdaki formül (Denklem 2.10) ile hesaplanır:

$$\text{MAE} = \frac{1}{n}\sum|y-x| \quad 2.10$$

burada n veri sayısını, y gerçek veriyi, x tahmin edilen veriyi ifade eder.

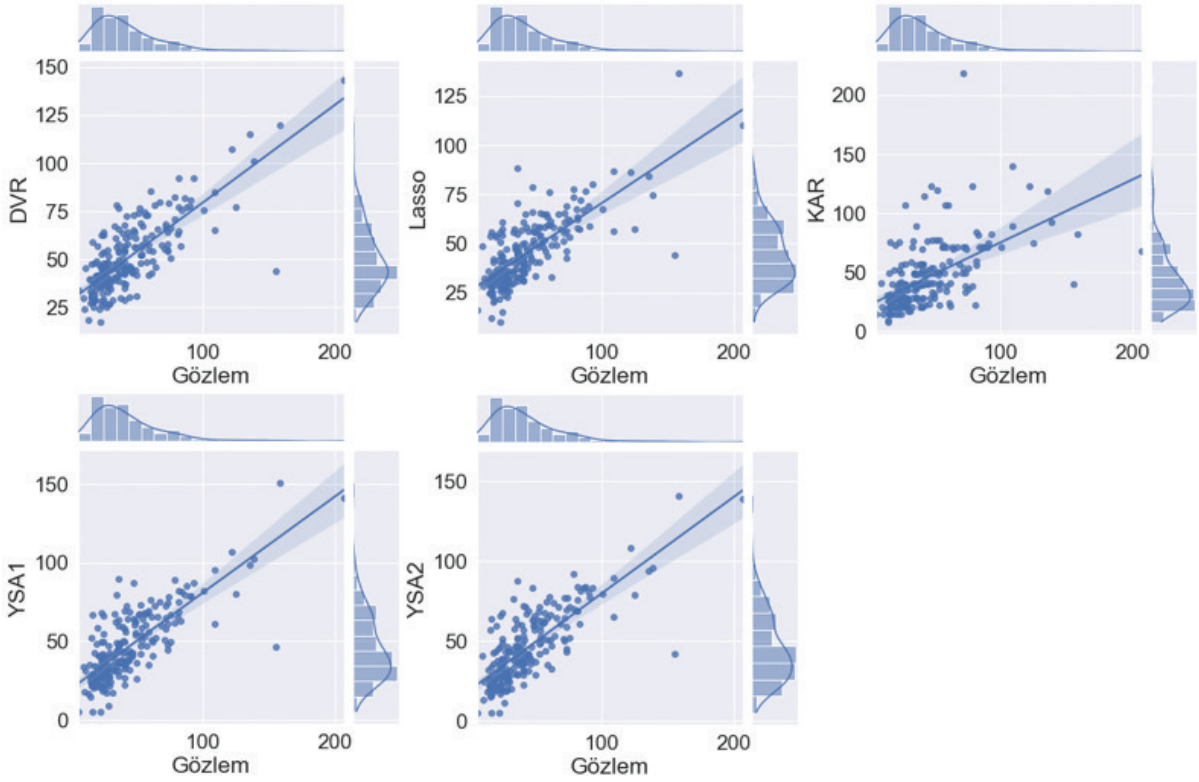
3. Bulgular ve Tartışma

Bu çalışmada, meteorolojik faktörler göz önüne alınarak çeşitli makine öğrenme algoritmaları ile PM10 konsantrasyonu tahmin edilmiştir. Makine öğrenme aşamasında, 2018 ocak – 2021 temmuz döneminin sıcaklık, çiğ noktası sıcaklığı, yağış, bağıl nem, rüzgar hızı, basınç, bulut kapalılığı ve bir önceki güne ait PM10 ölçümleri göz önüne alınarak, farklı makine öğrenme (KAR, DVR, Lasso, YSA) algoritmalarıyla ayrı ayrı çalışma yapılmış ve bu algoritmaların tutarlılıkları karşılaştırılmıştır. Tablo 3'de makine öğrenme algoritmalarının eğitim ve teste tabi tutulan kısımları için performans sonuçları gösterilmiştir. Buna göre, eğitim kısmı için en yüksek R2'nin 0,919 değeri ile KAR algoritmasında olduğu, bunu ise 0,624 ve 0,609 değerler ile YSA1 ve YSA2 algoritmalarının takip ettiği görülmektedir. Eğitim kısmında en düşük hata değerleri (RMSE: 8,87, MAE: 5,86) ise yine KAR algoritmasında görülmektedir. Bunu yine YSA1 ve YSA2 algoritmaları takip etmiştir. Test kısmı için ise en yüksek R2'nin 0,606 ve 0,603 değerleri ile YSA1 ve YSA2 algoritmalarında olduğu görülmüştür. En düşük hata değerleri de yine bu algoritma sonuçlarında görülmektedir. KAR algoritması ise test kısmında oldukça düşük performans göstermiştir. Makine öğrenme algoritmalarının test bölümü için saçılım grafikleri Şekil 5'de gösterilmektedir.

Tablo 3

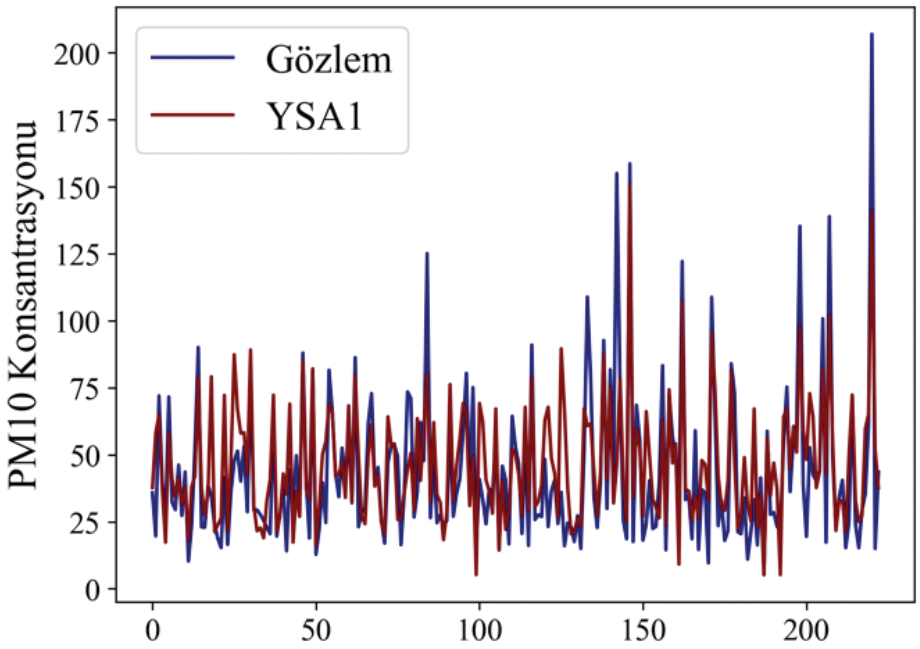
Makine öğrenme algoritmaları performans sonuçları

	Eğitim			**Test**		
	R2	RMSE	MAE	R2	RMSE	MAE
YSA1	0,624	19,13	12,78	**0,606**	**17,94**	**12,51**
YSA2	0,609	19,53	12,86	**0,603**	**18,01**	**12,38**
Lasso	0,476	22,59	14,77	0,534	19,51	13,32
DVR	0,576	20,34	15,40	0,527	19,66	15,41
KAR	**0,919**	**8,87**	**5,86**	0,143	26,47	16,31

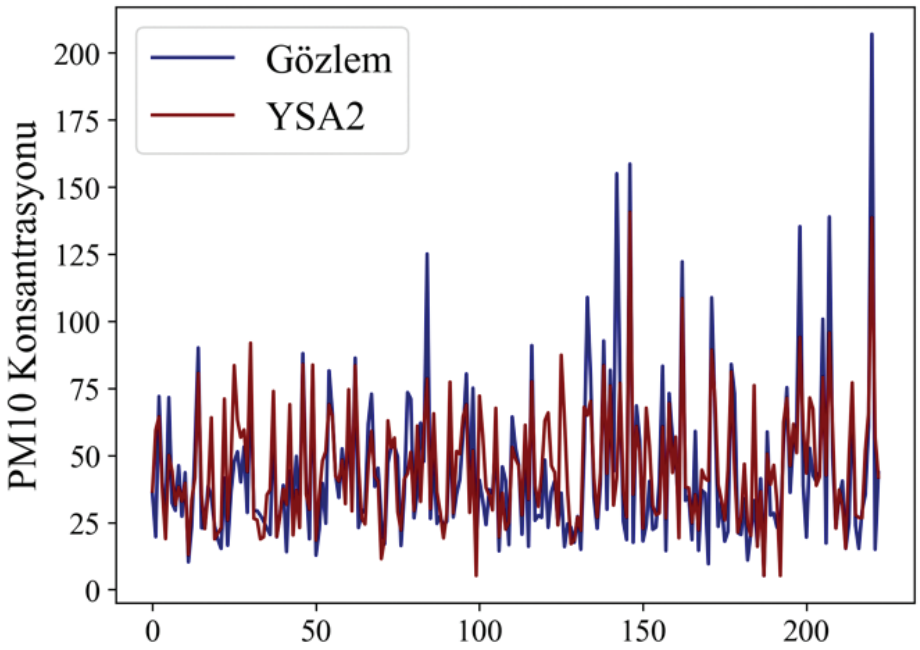


Şekil 5. Makine öğrenme algoritmaları saçılım ve dağılım grafikleri (test bölümü)

Test bölümünde en iyi performansı gösteren YSA1 ve YSA2 ile elde edilen PM10 tahminlerinin, gözlem değerleri ile karşılaştırması sırasıyla Şekil 6 ve Şekil 7'de gösterilmektedir. Her iki grafikte de algoritma sonucunun genel eğilimi yakaladığı ve gözlem değerleri ile çoğu yerde örtüştüğü görülmektedir. Bu ise YSA1 ve YSA2 tahminlerinin başarılı sonuçlar ürettiğini göstermektedir.



Şekil 6. YSA1 ile elde edilen PM10 tahmin değerlerinin gözlem değerleri ile karşılaştırması



Şekil 7. YSA2 ile elde edilen PM10 tahmin değerlerinin gözlem değerleri ile karşılaştırması

Makine öğrenme algoritmaları kullanılırken bazı hiperparametreler (katsayılar) tanımlanır. Bu parametrelerin belirlenmesinde sabit bir kural olamamakla beraber sınırsız farklı değer atanabilir ve genellikle deneme yanılma yöntemi ile belirlenir ve bu parametreler algoritmaların başarısını doğrudan etkiler. Algoritma ba-

şarısını etkileyen bir diğer önemli etmen ise çözülen problemin doğasıdır. Bu çalışmada kullanılan parametrelerin sınırsız seviyede yükselmesi mümkün değildir, dolayısı ile parametreler parabol davranışı sergiler. Bu yüzden kullanılan algoritmaların farklı tutarlılıklar göstermesi beklenilen bir durumdur.

Ayrıca, PM10 konsantrasyonu, emisyon kaynaklarına bağlı olarak değişiklik göstermektedir. Dolayısıyla, emisyon kaynaklarının PM10 seviyelerinin değişiminde ve dolayısıyla makine öğrenme algoritmalarının tutarlılıklarında önemli bir katkısı bulunmaktadır. Sisteme öğretilmemiş ve beklenmedik şekilde gerçekleşen farklı kaynakların kirliliklerinin (çöl tozu taşınımı, kış ayı ve yaz ayı evsel ısınma farklılıkları) etkili olduğu dönemlerde makine öğrenme yöntemleri ile elde edilen sonuçlar istikrarlı olmayacaktır. Bu nedenle makine öğrenme yöntemleri ile PM10 konsantrasyonlarının tahmini, genellikle emisyonların çok fazla değişmediği kaynaklar ve dönemler için daha başarılı olacaktır.

Elde edilen sonuçlar, literatürdeki diğer sonuçlar ile kıyaslandığında bazı farklılıkların ortaya çıktığı görülmektedir. Bu çalışmada, test bölümü göz önüne alındığında en iyi sonucu YSA algoritması vermiştir. Bunu ise Lasso ve DVR algoritmaları takip etmiştir. Gültepe (2019) tarafından yapılan çalışmada ise, KAR algoritmasının en iyi sonucu verdiği tespit edilmiştir. Castelli ve diğerleri (2020) tarafından yapılan çalışmada, DVR algoritmasının tutarlı bir şekilde tahmin sonucu sağladığı görülmüştür. Son olarak Czernecki ve diğerleri (2021) tarafından yapılan çalışmada ise, KAR tabanlı bir algoritma olan XGBoost algoritmasının en tutarlı olduğu, sonrasında ise YSA algoritmasının geldiği sonucuna ulaşmışlardır. Bu sonuçlar makine öğrenme algoritmalarının, bölgesel olarak farklı tutarlılıklar gösterdiğini işaret etmektedir. Bu farklılıklardaki en önemli etken, farklı bölgelerin emisyon kaynaklarının ve meteorolojik koşullarının farklılık göstermesidir. Ayrıca verilerin makine öğrenme sistemine öğretilmesi aşamasında olası bazı farklılıklar içermesi de diğer önemli etkenlerden birisidir.

4. Sonuçlar

Bu çalışmada, farklı makine öğrenme algoritmalarının (KAR, DVR, Lasso, YSA), PM10 kirliliği tahminindeki performansları karşılaştırılmıştır. Algoritmalar göz önüne alınarak yapılan çalışma aşamasında, sıcaklık, çiğ noktası sıcaklığı, yağış, bağıl nem, rüzgar hızı, basınç, bulut kapalılığı ve bir önceki güne ait PM10 ölçümleri (Ocak 2018 – Temmuz 2021 dönemi) göz önüne alınmıştır. Meteoroloji verileri Meteoroloji Genel Müdürlüğü Ankara Bölge istasyonundan, PM10 kirlilik verileri ise Çevre, Şehircilik ve İklim Değişikliği Bakanlığı Ankara Keçiören-Sanatoryum hava kalitesi istasyonundan elde edilmiştir. Eğitim bölümünde, en yüksek R2 ve en düşük hata oranları ile KAR algoritması en iyi performansı göstermiştir. Ancak KAR algoritmasının performansı test bölümünde oldukça düşük olarak bulunmuştur. Test bölümünde KAR algoritması hariç diğer tüm algoritmaların 0,5'in üzerinde R2 değerleri vermiştir. YSA1 ve YSA2 algoritmaları en yüksek R2 ve en düşük hata oranları ile en iyi performansı göstermiştir. Çalışma sonuçları, test bölümü göz önüne alındığında, YSA algoritmasının diğer algoritmalara göre daha iyi sonuç verdiğini göstermiştir. Elde edilen sonuçlar ışığında daha detaylı çalışmalar yapılması, makine öğrenme yöntemleriyle PM10 tahminine yönelik olarak entegre hava kalitesi erken uyarı sisteminin geliştirilmesine yardımcı olabilir. Gelecek çalışmalarda makine öğrenme algoritmalarının performansları, daha uzun dönemi kapsayan verilerle ve daha fazla bölge için incelenebilir. Bu yolla algoritmaların bölgesel performanslarının da karşılaştırılması imkanı olacaktır.

Yazar Katkıları

Kahraman OĞUZ: Makalenin tasarımı, yorumu ve yazımı.

Muhammet Ali PEKİN: Verilerin toplanması ve hesaplamaların yapılması.

Çıkar Çatışması

Yazarlar çıkar çatışması bildirmemişlerdir.

Kaynaklar

Abuella, M. ve Chowdhury, B. (2016). Solar Power Forecasting Using Support Vector Regression. *American Society for Engineering Management International Annual Conference*, USA. Erişim adresi: https://arxiv.org/abs/1703.09851

Adhani, G., Buono, A. ve Faqih, A. (2013). Support Vector Regression modelling for rainfall prediction in dry season based on Southern Oscillation Index and NINO3.4. *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Sanur Bali, Indonesia. https://doi:10.1109/ICACSIS.2013.6761595

Alizamir, M., Kisi, O., Ahmed, A.N., Mert, C., Fai, C.M., Kim, S., Kim, N.W. ve El-Shafie, A. (2020). Advanced machine learning model for better prediction accuracy of soil temperature at different depths. *PLoS ONE*, *15*(4), 1-25. https://doi.org/10.1371/journal. pone.0231055

Aljanabi, M., Shkoukani, M. ve Hijjawi, M. (2020). Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan. *International Journal of Automation and Computing*, *17*(5), 667-677. https://doi:10.1007/s11633-020-1233-4

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, Cambridge, MA, USA. Erişim adresi: https://mitpress.mit.edu/books/introduction-machine-learning

Aydoğan, İ. ve Zırhlıoğlu, G. (2018). Öğrenci Başarılarının Yapay Sinir Ağları ile Kestirilmesi. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *15*(1), 577-610. Erişim adresi: https://dergipark.org.tr/tr/pub/yyuefd/issue/40566/495631

Badarpura, S., Jain, A., Gupta, A. ve Patil, D. (2020). Rainfall Prediction using Linear approach & Neural Networks and Crop Recommendation based on Decision Tree, *International Journal of Engineering Research & Technology*, *09*(04), 394-399, http://dx.doi.org/10.17577/IJERTV9IS040314

Carro-Calvo, L., Casanova-Mateo, C., Sanz-Justo, J., Casanova-Roqueb, J.L. ve Salcedo-Sanz, S. (2017). Efficient prediction of total column ozone based on support vector regression algorithms, numerical models and Suomi-satellite data. *Atmosfera*, *30*(1), 1-10, https://doi:10.20937/ATM.2017.30.01.01

Castelli, M., Clemente, F.C., Popovič, A., Silva, S. ve Vanneschi, L. (2020). A Machine Learning Approach to Predict Air Quality in California. *Complexity 2020*(2020), 1-23. https://doi.org/10.1155/2020/8049504

Czernecki, B., Marosz, M. ve Jędruszkiewicz, J. (2021). Assessment of Machine Learning Algorithms in Short-term Forecasting of PM10 and PM2.5 Concentrations in Selected Polish Agglomerations. *Aerosol Air Qual. Res.*, *21*(7), 1-18. https://doi.org/10.4209/aaqr.200586

ÇŞBHKİ, Çevre, Şehircilik ve İklim Değişikliği Bakanlığı Ulusal Hava Kalite İzleme Ağı, (2021). Erişim tarihi: 16.09.2021, https://www.havaizleme.gov.tr/

Ei-Sharkawy MF. ve Zaki G.R. (2015). Effect of meteorological factors on the daily average levels of particulate matter in the Eastern Province of Saudi Arabia: a cross-sectional study. *J Sci Technol*, *5*(1), 18–29. Erişim adresi: https://dergipark.org.tr/tr/pub/tojsat/issue/22636/241852

Gagliardi, R.V. ve Andenna, C. (2020). A Machine Learning Approach to Investigate the Surface Ozone Behavior. *Atmosphere*, *11*(11), 1-16. https://doi:10.3390/atmos11111173

Gültepe, Y. (2019). Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine Karşılaştırmalı Bir Değerlendirme. *Avrupa Bilim ve Teknoloji Dergisi*, *16*, 8-15. https://10.31590/ejosat.530347

Harishkumar, K. S., Yogesh, K. M. ve Gad, I. (2020). Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. *Procedia Computer Science*, *171*, 2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

Haykin S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, USA. Erişim adresi: https://dl.acm.org/doi/book/10.5555/521706

Hrdlickova, Z., Michalek, J., Kolar, M. ve Vesely, M. (2008). Identification of factors affecting air pollution by dust aerosol PM10 in Brno City, Czech Republic. *Atmos Environ*, *42*(37), 8661–8673. https://doi:10.1016/j.atmosenv.2008.08.017

Karaatlı, M., Helvacıoğlu, Ö., Ömürbek, N. ve Tokgöz, G. (2012). Yapay Sinir Ağları Yöntemi İle Otomobil Satış Tahmini. *Uluslararası Yönetim İktisat ve İşletme Dergisi*, *8*(17), 87-100. https://10.11122/ijmeb.2012.8.17.290

MGM, Meteoroloji Genel Müdürlüğü, (2021). Erişim tarihi: 16.09.2021, https://mevbis.mgm.gov.tr/mevbis/ui/index.html#/Workspace

Musoro, J.Z., Zwinderman, A.H., Puhan, M.A., Riet, G. ve Geskus, R.B. (2014). Validation of prediction models based on lasso regression with multiply imputed data. *BMC Med Res Methodol*, *14*(116), 1-13. https://doi.org/10.1186/1471-2288-14-116

Oğuz, K. (2020). Nevşehir İlinde Hava Kalitesinin ve Meteorolojik Faktörlerin Hava Kirliliği Üzerine Etkilerinin İncelenmesi. *Doğal Afetler ve Çevre Dergisi*, *6*(2), 391-404. https://doi:10.21324/dacd.686052

Özdemir, U. ve Taner, S. (2014). Impacts of Meteorological Factors on PM10: Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) Approaches. *Environmental Forensics*, *15*(4), 329–336. https://doi:10.1080/15275922.2014.950774

Panda, N., Osthus, D., Srinivasan, G., O'Malley, D., Chau, V., Oyen, D. ve Godinez, H. (2020). Mesoscale informed parameter estimation through machine learning: A case-study in fracture modeling. *Journal of Computational Physics*, *420*, 1-15. https://doi.org/10.1016/j.jcp.2020.109719

Qin, Y.-G., Yi, C., Dong, G.-L. ve Min, J.-Z. (2019). Investigating the influence of meteorological factors on particulate matters: A case study based on path analysis. *Energy & Environment*, *31*(3), 1-13. https://doi:10.1177/0958305x19876696

Singh, D. ve Singh, B. (2019). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi:10.1016/j.asoc.2019.105524

Smola, A. J. ve Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. https://doi:10.1023/B:STC0.0000035301.49549.88

Suleiman, A. ve Tight, M.R., Quinn, A.D. (2019). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmospheric Pollution Research*, *10*(1), 134–144. https://doi.org/https://doi.org/10.1016/j.apr.2018.07.001

Sun, Z., Tao, Y., Li, S., Ferguson, K. K., Meeker, J. D., Park, S. K., Batterman, S. A. ve Mukherjee, B. (2013). Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environ Health*, *12*(1), 1-19. https://doi:10.1186/1476-069X-12-85

US EPA, U.S. Environmental Protection Agency, (2021). Erişim tarihi: 02.08.2021, https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm

Yaseen, Z.M., El-Shafie, A., Jaafar, O., Afan, H.A. ve Sayl, K.N. (2015). Artificial intelligence based models for stream-flow forecasting: 2000–2015. *J. Hydrol.*, *530*, 829–844. https://doi.org/10.1016/j.jhydrol.2015.10.038