

PM₁₀ ve SO₂ Hava Kirleticilerinin Makine Öğrenmesi Modelleri ile Modellenmesi ve Gelecek Tahmini: Türkiye'deki Büyükşehirler Örneği

İbrahim Edibali Atalay^{1,*}, Abdullah Altay², Serdar Neslihanoglu²

¹Artvin Çoruh Üniversitesi, Lisansüstü Eğitim Enstitüsü, Merkez Yerleşke, 08000, Artvin

²Eskişehir Osmangazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Meşelik Yerleşkesi, 26480, Eskişehir

Özet

Son yıllarda Türkiye'deki hava kirliliğinin nüfus yoğunluğu ve sanayileşmedeki artış nedeniyle büyükşehirlerde halk sağlığı ve çevre yönetimi üzerinde ciddi bir sorun haline geldiği görülmektedir. Hava kirliliğine neden olan maddeleri kaynaklarına ve atmosferdeki oluşum süreçlerine göre partikül madde (PM₁₀) ve kükürt dioksit (SO₂) gibi doğrudan bir kaynaktan yayılan maddeler birincil kirleticiler olarak tanımlanmaktadır. Bu çalışmada, Türkiye'nin 7 farklı coğrafi bölgelerindeki nüfus yoğunluğu en yüksek olan 7 büyükşehir (Ankara, Antalya, Erzurum, İstanbul, İzmir, Samsun ve Şanlıurfa) ait Ocak 2016-Şubat 2026 tarihleri aralığındaki günlük PM₁₀ ve SO₂ verilerine odaklanılmıştır. Bu verilerin modelleme ve gelecek 1 yıllık tahmini GBM, HistGBM, LightGBM, Random Forest ve XGBoost makine öğrenmesi modelleri ile yapılmıştır. Sonuç olarak, XGBoost modelinin PM₁₀ ve SO₂ verilerinin modelleme performansı konusunda, GBM modelinin ise verilerin gelecek tahmini konusunda daha etkin olduğu görülmüştür. Buna ek olarak, makine öğrenmesi modellerinin verileri modelleme ve gelecek tahmini konusundaki yüksek açıklama yüzdesiyle hava kalitesi araştırmalarında önerilen makine öğrenmesi modellemelerinin gelecek araştırmalar için kullanılması önerilmektedir.

Anahtar Sözcükler

PM₁₀, SO₂, Büyükşehir, Makine Öğrenmesi, Gelecek Tahmini

Modeling and Forecasting PM₁₀ and SO₂ Air Pollutants Using Machine Learning Models: The Case of Major Cities in Turkey

Abstract

In recent years, air pollution in Turkey has become a serious problem for public health and environmental management in metropolitan areas, owing to high population density and industrialization. Substances that cause air pollution are defined as primary pollutants, such as Particulate Matter (PM₁₀) and Sulfur Dioxide (SO₂), which are emitted directly from sources and form in the atmosphere. This study focuses on daily PM₁₀ and SO₂ data from January 2016 to February 2026 for the seven largest cities with the highest population density in Turkey's geographical regions (Ankara, Antalya, Erzurum, Istanbul, Izmir, Samsun, and Sanliurfa). These data were modeled and predicted for the next year using GBM, HistGBM, LightGBM, Random Forest, and XGBoost machine learning models. As a result, the XGBoost model was more effective at modeling PM₁₀ and SO₂ data, whereas the GBM model was more effective at forecasting future data. In addition, given the high explanatory power of machine learning models for data analysis and prediction, it is recommended that the models proposed in air quality studies be used in future research.

Keywords

PM₁₀, SO₂, Metropolitan City, Machine Learning, Future Prediction

1. Giriş

Son yıllarda Türkiye'deki hava kirliliğinin nüfus yoğunluğu ve sanayileşmedeki artış nedeniyle büyükşehirlerde halk sağlığı ve çevre yönetimi üzerinde ciddi bir sorun haline geldiği görülmektedir. Hava kirliliği kükürt dioksit, partikül madde, azot oksitler, karbonmonoksit, ozon ve hidrokarbonlar gibi havayı kirleticiler maddelerin atmosferde limit değerlerin üzerinde bulunması olarak tanımlanmıştır (Ören, 2024). Hava kirliliğine neden olan maddeleri kaynaklarına ve atmosferdeki oluşum süreçlerine göre partikül madde (PM₁₀) ve kükürt dioksit (SO₂) gibi doğrudan bir kaynaktan yayılan maddeler birincil kirleticiler olarak tanımlanmaktadır.

Birincil kirleticiler olarak tanımlanan PM₁₀ ve SO₂ ile ilgili olarak özetle son yıllarda Türkiye'de çok sayıda çalışma yapılmıştır. Örneğin, Özşahin vd. (2016), Edirne İl ve Çevre ve Orman Müdürlüğünü tarafından yapılan 2013 ile 2015 yıllarının kapsayan hava kirliliği ölçüm verilerini incelemiştir. SO₂ oranının sınır değeri üzerinde olduğunu ve bu durumun ısınma için kullanılan yakıtların kalitesiz olduğu sonucuna ulaşmıştır. Demirarslan ve Akıncı (2016), ArcGIS yazılımını kullanarak 2010 ile 2015 yılları arasındaki saatlik PM₁₀ verilerini kullanmışlardır. Ters Mesafe Ağırlıklı Enterpolasyon Yöntemi (IDW) uygulanarak kirlilik dağılım haritalarını üretmişlerdir. Sonuç olarak kış aylarında ısınma,

diğer mevsimlerde ise trafik (çizgisel kaynaklar) kirliliğin ana sebepleri olarak tespit etmişlerdir. Ayrıca, [Demirarslan ve Akıncı \(2016\)](#), SO₂ konsantrasyon dağılımlarını Artvin, Bayburt, Giresun, Gümüşhane, Ordu, Rize ve Trabzon illerini kapsayan kirlilik dağılım haritalarını üretmek amacıyla 2011 ile 2016 yılları arasındaki verilerini kullanmışlardır. Çalışma sonucunda bölgedeki hava kalitesini en çok etkileyen faktörlerin kışın ısınma, diğer zamanlarda ise trafik ve kısmen sanayi olduğunu tespit etmişlerdir. [Atalay ve Neslihanoglu \(2021\)](#), animasyonlar sonucunda Türkiye geneli PM₁₀ miktarlarının zamanla azalmasına rağmen Türkiye geneli birçok ilde PM₁₀ miktarının T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı (CSB), Çevre Yönetimi Genel Müdürlüğü tarafından Hava Kalitesi Değerlendirme ve Yönetimi Genelgesinde belirtilen limit değerlerinin üzerinde olduğunu belirlemişlerdir. [Ören \(2024\)](#), Kahramanmaraş ve Hatay merkezli meydana gelen depremin neden olduğu yıkım ve yıkılan bina ve diğer yapıların enkazlarının kaldırılması aşamasında havadaki değişiklikleri incelemiştir. Partikül madde (PM₁₀) dış ortamda havada asılı kalabildiklerinden dolayı canlılar için solunum yolunu tehdit etmesinden dolayı aerosol kirleticilerin en tehlikeli olanı olarak tanımlamıştır. Bu bağlamda, T.C. Çevre, Şehircilik ve İklim Değişikliği Bakanlığı, Ulusal Hava Kalitesi İzleme Ağı tarafından elde ettiği 2022 ile 2023 yıllarını kapsayan PM₁₀ ve SO₂ değerleri ile deprem öncesi ve sonrasında dair incelemelerde bulunmuştur. [Olçay vd. \(2025\)](#), çalışmalarında dergipark ve veri tabanında “hava kirliliği” veya “hava kalitesi” kelimelerini içeren makalelerde Türkiye merkezli yayınların hava kirliliği ile doğal ortam arasındaki ilişkiyi analiz etmişlerdir. Çalışma sonucunda Uşak ve Şırnak gibi iller hakkında belirlenen kriterlerde hiç yayın bulunmadığı tespit edilmiştir. [Olçay ve Çukur \(2025\)](#), Manisa ilinde 2006 ile 2023 dönemlerini kapsayan (sıcaklık, basınç, rüzgâr hızı ve yönü gibi) verileri ile aynı istasyonun aylık ve yıllık SO₂ verilerini analiz etmişlerdir. Çalışma sonucunda, meteorolojik unsurların Manisa'daki SO₂ miktarındaki değişimin yaklaşık %20'sini açıkladığı ve kirlilik ile iklim elemanları arasında anlamlı ilişkiler olduğu tespit edilmiştir. [Aydın ve Raja \(2025\)](#), hava kirliliğini Türkiye'deki büyükşehirler de ciddi bir sorun haline geldiğini ve hava kirliliği tahminlerinde doğruluğu arttırmak için Kriging with External Drift (KED) yöntemini kullanarak Ankara'da PM₁₀ ve SO₂ hava kirleticilerinin mekânsal ve zamansal değişimlerini ve meteorolojik faktörlerin etkisi altında 2011 ile 2014 aylık verilerini incelemişlerdir.

Birincil kirleticiler olarak tanımlanan PM₁₀ ve SO₂ ile ilgili olarak yukarıda özetlendiği gibi Türkiye'de son yıllarda çok sayıda çalışma yapılmıştır. Fakat, PM₁₀ ve SO₂ ölçümleri üzerine Türkiye'de makine öğrenmesi ile yapılan çalışmalar incelendiğinde, az sayıda çalışma yapıldığı görülmüştür. Örneğin, [Demir Yetiş vd. \(2025\)](#), Malatya ilindeki partikül madde (PM), azot oksitler (NO_x), ozon (O₃), kükürt oksitler (SO_x), karbonmonoksit (CO), hidrokarbonlar (HC), uçucu organik bileşikler (VOC) ve polisiklik aromatik hidrokarbonlar (PAH) gibi parametrelerin hava kirliliğine sebep olarak hava kirleticileri olarak tanımlamışlardır. Malatya ilindeki hava kalitesini değerlendirmek ve geleceğe yönelik tahminlerde bulunmak amacıyla 1 Ocak 2020 ile 31 Aralık 2024 tarihleri arasındaki beş yıllık süreyi kapsayan verileri analiz etmişlerdir. Çevre, Şehircilik ve İklim Değişikliği Bakanlığının Ulusal Hava Kalitesi İzleme Ağından elde edilen 1827 veri ile Lineer Regresyon (LR), Rastgele Orman (RF), XGBoost, LightGBM, Destek Vektör Regresyonu (SVR), k-En Yakın Komşu (k-NN) ve Yapay Sinir Ağları (ANN) makine öğrenmesi algoritmaları kullanılarak analizler yapılmıştır. 2020-2023 yılları eğitim verisi, 2024 yılı ise test verisi kullanılarak yapılan analiz tahminleri sonucunda XGBoost ve LightGBM algoritmalarının PM₁₀ ve SO₂ tahmin etmede diğer yöntemlere göre daha üstün performans gösterdiğini belirtmişlerdir. [Metin \(2025\)](#), İstanbul'daki hava kirliliği seviyelerini belirleyen temel kirleticilerin PM₁₀, SO₂, NO₂, NO_x ve NO günlük değerlerini makine öğrenmesi, derin öğrenme ve hibrit yöntemler kullanmıştır. Ocak 2013 ile Mayıs 2025 tarihleri arasındaki Çevre, Şehircilik ve İklim Değişikliği Bakanlığındaki Sürekli İzleme Merkezi (SİM) tarafından sağlanan 4500 veri ile Lineer Regresyon (LR), Random Forest (RF), Support Vector Regression (SVR), XGBoost ve çok katmanlı yapay sinir ağları, hibrit modellerde Voting Hybrid Regressor (Oylama Tabanlı), Stacking Hybrid Regressor (Yığınlama Tabanlı) kullanmıştır. Çalışma sonucunda Ağaç tabanlı modeller (Random Forest, XGBoost), doğrusal olmayan ilişkileri yakalamada Lineer Regresyon'a göre çok daha başarılı olduğu tespit etmiştir.

Son yıllarda dünya genelinde, birincil kirleticiler olarak tanımlanan PM₁₀ ve SO₂ ile meteorolojik değişkenlerin (hava sıcaklığı, rüzgâr hızı ve yönü, nem, basınç vb.) ilişkisini inceleyen çok sayıda makine öğrenmesi tabanlı hava kalitesi tahmin çalışması yapılmıştır. Örneğin, [Zhai ve Chen \(2018\)](#), Çin'in Pekin kentindeki günlük Partikül Madde (PM₁₀) konsantrasyonlarını analiz etmek ve birgün sonrasını yüksek doğrulukla tahmin etmek için makine öğrenmesi modeli geliştirmeyi amaçlamışlardır. 2013 ile 2017 yılları arasındaki günlük veriler ile hava kirleticilerinden PM_{2.5}, PM₁₀, NO₂, CO, SO₂ ve O₃ meteorolojik faktörleri (güneşlenme süresi, yağış, sıcaklık, basınç, maksimum ve aşırı rüzgâr hızı/yönü) verileri analiz edilmiştir. Bu çalışmada, LASSO, AdaBoost, XGBoost ve GA-MLP gibi makine öğrenmesi algoritmaları veri analizinde karşılaştırılmıştır. Sonuç olarak, komşu şehirlerden gelen kirleticiler ile yerel meteorolojik faktörlerin Pekin'in hava kirliliği üzerindeki kritik etkisini ortaya koyarak tekil modellerin üstün performans gösterdiğini koymuşlardır. [Suleiman vd. \(2019\)](#), trafik kaynaklı partikül madde (PM₁₀ ve PM_{2.5}) konsantrasyonlarını yönetmek ve azaltmak için makine öğrenimi modellerini incelemişlerdir. Londra'da bulunan 19 farklı hava kalitesi izleme istasyonlarına ait 2007 ile 2012 yılları arasındaki hava durumu ve kirlilik verilerini kullanmıştır. Veri setinin %80'i modelleri eğitmek ve kalan %20'si ise test etmek için ANN, BRT ve SVM makine öğrenmesi modellerini kullanmışlardır. ANN ve BRT modellerinin PM₁₀ ve PM_{2.5} konsantrasyonlarını tahmin etmede çok yüksek bir performansla sahip olduğunu ve iyi eğitilmiş modellerin önceden ölçmede ve çevre koruma politikalarında başarılı olacaklarını belirlemişlerdir. [Li vd. \(2025\)](#), Çin'in Qingdao kentinde hava kirliliğinin takibi için makine öğrenimi ile desteklenen yenilikçi bir lazer (Doppler lidar-CDL) yöntemini sunmuşlardır. Atmosferdeki PM_{2.5} ve PM₁₀ dikey dağılımını doğru bir şekilde profillemek amacıyla rüzgâr hızı ve nem gibi meteorolojik verileri hibrit bir model ile birleştirmiştir.

ET, RF, XGBoost, LightGBM, GBR, KNN ve DT makine öğrenmesi modelleri ile yaptıkları analizler sonrasında en performanslı model Ekstra Ağaçlar Regresyonu (ET) olmuştur. [Shang \(2026\)](#), Çin'deki orta yaşlı ve yaşlı nüfusta hava kirliliği ile kalp hastalığı riski arasındaki bağlantıyı kapsamlı bir şekilde incelemişlerdir. SO₂ maruz kalan insanların sağlık üzerindeki olumsuz etkilerini doğrulamak için makine öğrenimi algoritmalarını ve çok değişkenli lojistik regresyon analizlerini birleştirmişlerdir. SO₂, CO, PM₁₀, PM_{2.5}, NO₂ ve O₃ gibi altı farklı hava kirleticisini AdaBoost, DT, LightGBM, XGBoost, RF ve GBDT makine öğrenmesi modellerini kullanarak yaptıkları analiz sonucunda RF'nin diğerlerine göre en iyi performansı gösterdiğini ve SO₂ konsantrasyonundaki artışın kalp hastalığı olasılığını anlamlı düzeyde yükseldiğini tespit etmişlerdir.

Literatür araştırması sonucunda birincil hava kirleticileri konusunda çok dünya literatüründe çok sayıda araştırma olmasına rağmen makine öğrenmesi konusundaki Türkiye'deki araştırmaların sayısının yetersiz olduğu görülmüştür. Bu doğrultuda araştırmanın temel amacı olarak makine öğrenmesi modelleri ile PM₁₀ ve SO₂ hava kirleticilerinin modellenmesi ve gelecek tahmini araştırılmıştır. Bu amaç ile Türkiye'nin 7 coğrafi bölgesindeki nüfus yoğunluğu en yüksek olan 7 büyükşehir (Ankara, Antalya, Erzurum, İstanbul, İzmir, Samsun ve Şanlıurfa) ait Ocak 2016-Şubat 2026 tarihleri aralığındaki günlük PM₁₀ ve SO₂ verilerine odaklanılmıştır. Bu verilerin modelleme ve gelecek 1 yıllık tahmini konusunda GBM, HistGBM, LightGBM, Random Forest ve XGBoost makine öğrenmesi modellerinin performansları karşılaştırılmıştır.

Bu araştırmanın geri kalan kısımları şu şekildedir: Bölüm 2'de araştırmada kullanılan makine öğrenmesi modellerinin metodolojisi tanıtılacaktır. Bölüm 3'te araştırma verilerinin betimleyici istatistikleri ve zaman serisi grafikleri tanıtılırken, araştırma bulguları bölüm 4'te tanıtılacaktır. Sonuçlar ve tartışmalar Bölüm 5'te yer verilecektir.

2. Metodoloji

2.1 GBM (Gradient Boosting Model)

GBM modeli hataları en aza indirmenin yanında büyük veri setleri için yüksek doğruluk sağlamak amacıyla geliştirilmiştir ([Ridgeway 2007](#)). Bu model, durum ağaçların birbirine bağlandığı adımları içermektedir. Bu süreçte, ilk olarak bir adet başlangıç ağacı oluşturulduktan sonra hedef değişkenin gerçek değeri ile ağacın tahmin ettiği değer arasındaki hata payının azaltılması için ikinci bir ağaç oluşturulur ve ilk ağacın tahmin ettiği negatif gradyanını tahmin etmek için kullanılır. GBM modeli Denklem 1'de gösterilmiştir.

$$\hat{f}(x) = \underset{f(x)}{\operatorname{argmin}} \psi(y, f(x)) \quad (1)$$

Burada $\hat{f}(x)$ başlangıç modelidir. argmin ise ψ fonksiyonunu en küçük yapan değer bulunmasını sağlarken, $\psi(y, f(x))$, kayıp fonksiyonu ifade etmektedir. y gerçek değeri, $f(x)$ modelin tahmin ettiği değeri ve ψ ise tahmin ile gerçek arasındaki hatayı ölçen kural olarak belirtilir.

2.2 HistGBM (Histogram-Based Gradient Boosting Model)

HistGBM (Histogram tabanlı GBM) özellikle büyük veri setlerinde GBM'den daha hızlı ve etkin performanslı olarak tasarlanmıştır. HistGBM modeli, veriyi işlemeyen önce aralıklara ayırarak algoritmanın hızını artıran ve diğer makine öğrenmesi modelleri gibi değişkenleri sıralamak yerine, onları belirli bir sayıda aralığa sıkıştırarak eğitim süresinin ciddi oranda düşmesine olanak sağlamıştır. HistGBM modeli Denklem 2'de gösterilmiştir. ([Nhat-Duc & Van-Duc 2023](#); [Mert 2025](#)).

$$\operatorname{bin}(x) = \lfloor \frac{x - x_{\min}}{x_{\max} - x_{\min}} \cdot n_{\text{bins}} \rfloor \quad (2)$$

Burada $x_{\max}$, x ve $x_{\min}$ sırasıyla özelliğin maksimum, minimum ve gerçek değerini temsil eder. n_{bins} , ayırıştırma için kullanılan bölme sayısını belirtirken, $\lfloor \cdot \rfloor$, taban fonksiyonu en yakın tam sayıya yuvarlar. Bu uyarlama, yöntemin histogramları ve bölme noktalarını etkili bir şekilde oluşturmasını sağlar ve böylece bellek kullanımı ve eğitim süresi azalır.

2.3 LightGBM (Light Gradient Boosting Machines)

LightGBM modeli, Microsoft tarafından büyük veri ve çok boyutlu özelliklerini hızlı bir şekilde işleyip tahmin yapılabilmesi için tasarlanmıştır. LightGBM modeli, karar ağaçları öğreniminde seviye odaklı (level-wise or depth-wise) veya yaprak odaklı (leaf-wise) olarak iki strateji kullanılabilir (Eşidir, 2025). LightGBM modeli Denklem 3'te gösterilmiştir.

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (3)$$

Burada, $L^{(t)}$ t adımıdaki toplam kaybı, l kayıp fonksiyonunu, y_i i . gerçek değeri ve $\hat{y}_i^{(t-1)}$, $t - 1$ adımıdaki tahmini, $f_t(x_i)$ ise t adımıdaki model tarafından eklenen yeni fonksiyonu ve $\Omega(f_t)$ yeni fonksiyonun cezasını temsil etmektedir.

2.4 Random Forest

Random Forest, tekil karar ağaçlarının yapısal zaafılarını gidermek amacıyla geliştirilmiş topluluk öğrenme (ensemble learning) algoritmasıdır. Modelin temelinde orijinal veri setinden iadeli örnekleme yoluyla çok sayıda alt veri kümesinin oluşturulmasını sağlayan Bagging tekniği kullanılmaktadır. Tek bir karar ağacı modeline karşın; Random Forest bu ağaçlardan oluşan bir orman kurarak hatanın minimize edilmesine olanak sağlamaktadır. Buna ek olarak, her düğümde en iyi ayrımı bulmak için tüm özelliklere bakmak yerine, rastgele seçilen bir alt kümeye bakarak ağaçlar arasındaki korelasyonu düşürüp modelin genelleme kapasitesini artırır. Random Forest modeli Denklem 4'te gösterilmiştir.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B f_b(x) \quad (4)$$

Burada, $\hat{y}$ modelin ortalama sonucu ifade etmektedir. B ise toplam model karar ağacı sayısını ifade ederken, $\frac{1}{B}$ toplamın model sayısına bölünerek aritmetik ortalamasının alınması sağlar. $f_b(x)$ ise bağımsız temel modelin (karar ağacı) tek başına ürettiği tahmin değeridir.

2.5 XGBoost (Extreme Gradient Boosting)

XGBoost, Gradient Boosting Machine (GBM) modelinin optimize edilmiş (Yeşilyurt & Dalkılıç, 2021) ve yüksek performanslı bir ağaç tabanlı gradyan artırma kütüphanesidir (Chen & Guestrin, 2016). XGBoost, sık karar ağaçlarını sıralı bir şekilde eğiterek güçlü bir tahminleyici oluşturma prensibine dayanır. Burada, her yeni ağaç, bir önceki ağacın hatalarını telafi edecek şekilde eğitilir. XGBoost yüksek tahmin gücü, azaltılmış karmaşıklık, eksik verilerin yönetilmesi, aşırı öğrenimin önlenmesi ve çalışma hızı gibi önemli avantajları sağlamaktadır. XGBoost modeli Denklem 5-7'de gösterilmiştir.

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (5)$$

$$L(\theta) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (6)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (7)$$

Burada, $\hat{y}_i$ i . dönemdeki tahmin edilen değer, K toplamdaki veri sayısı, $f_k(x_i)$ x_i için k . ağaç tahminini temsil etmektedir. L modelin hedef değişkeni ne kadar iyi tahmin ettiğini ölçen kayıp fonksiyonu, $\Omega(f_k)$ modelin karmaşıklığını cezalandıran düzeltme terimidir. Ayrıca, T ağaçtaki yaprak sayısını, γ ve $\Omega(f)$ modelin karmaşıklığını kontrol eden düzenleme parametresidir.

2.6 Model Kıyaslama Kriterleri

Modelleme ve gelecek tahmini performanslarını karşılaştırmak için kullanılan Ortalama Mutlak Hata (Mean Absolute Error, MAE), Kareler Ortalama Kök Hatası (Root Mean Square Error, RMSE) ve Belirleme Katsayısı (R^2) kriterleri Denklem 8-10'da gösterilmiştir.

$$MAE_i = \sum_{t=1}^N \frac{|\hat{y}_{it} - y_{it}|}{N} \quad (8)$$

$$RMSE_i = \sqrt{\sum_{t=1}^N \frac{(\hat{y}_{it} - y_{it})^2}{N}} \quad (9)$$

$$R_i^2 = 1 - \frac{\sum_{t=1}^N (y_{it} - \hat{y}_{it})^2}{\sum_{t=1}^N (y_{it} - \bar{y}_{it})^2} \quad (10)$$

Burada, y_{it} ve $\hat{y}_{it}$ sırasıyla i . varlığın t . zamandaki gözlem değerini ve modelleme (gelecek tahmin) tahmini değerini göstermektedir. Ayrıca modeller kıyaslanırken, en düşük MAE (RMSE) ve en yüksek R^2 değerlerine sahip modellerin, modelleme ve tahmini performansının daha iyi olduğu söylenebilir. Bu çalışmada Python programlama dilinde yer alan Scikit-learn, XGBoost, LightGBM, LazyPredict, Matplotlib, Seaborn, Plotly ve tqdm paketleri kullanılmıştır.

3. Veri

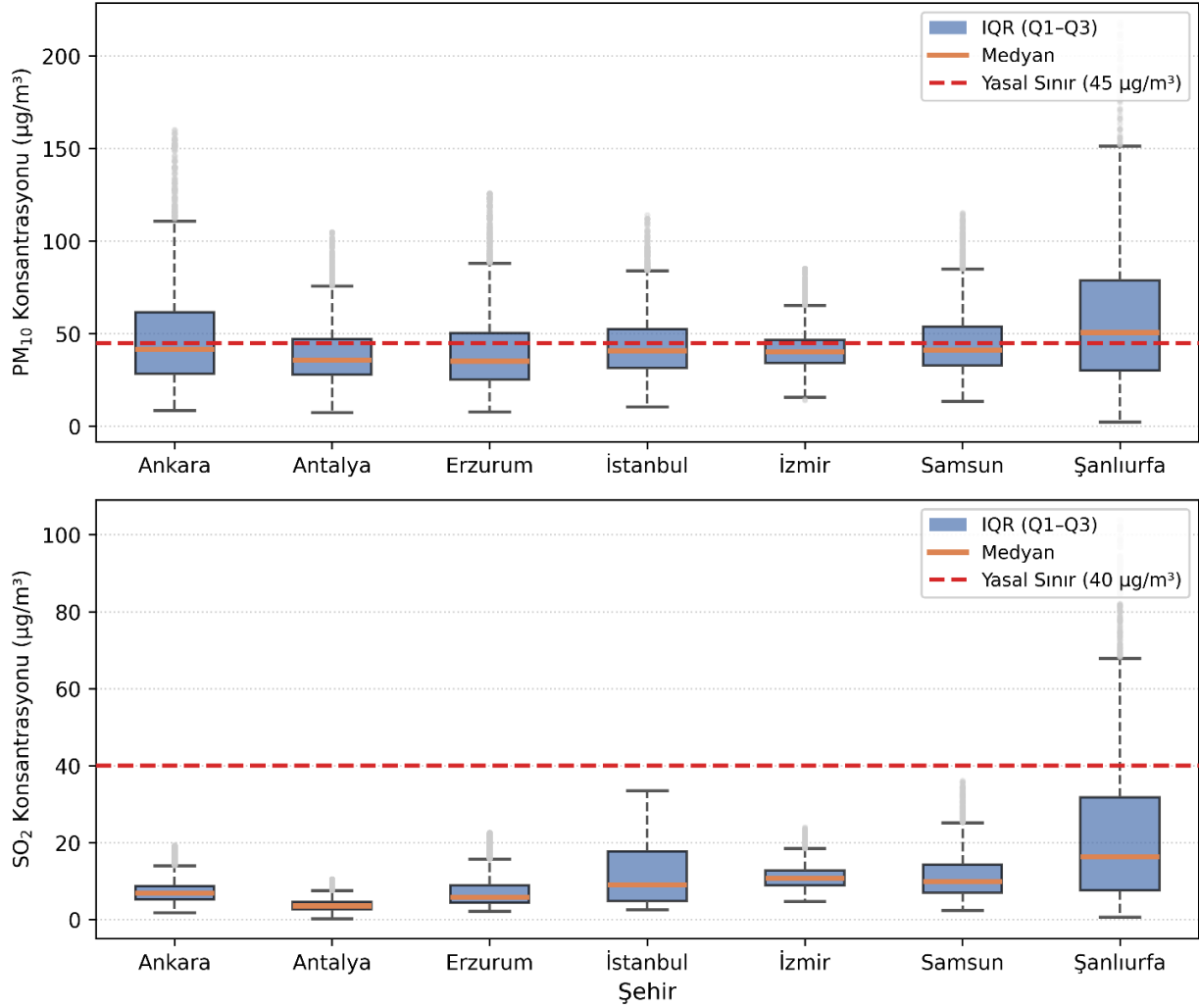
Bu çalışmada, Türkiye'deki 7 farklı coğrafi bölgedeki nüfus yoğunluğu öncelikli tercih sebebi olarak ve eksik gözlemler 7 Büyükşehir'in (Ankara, Antalya, Erzurum, İstanbul, İzmir, Samsun ve Şanlıurfa) PM₁₀ (Partikül Madde) ve SO₂ (Kükürtdioksit) verileri Çevre, Şehircilik ve İklim Değişikliği Bakanlığı resmi internet sitesi elde edilmiştir. Büyükşehirlerde hava kirliliği ölçüm sensörlerinden elde edilen bütün veriler çalışmaya dahil edilmiştir. Ayrıca, Ocak 2016-Şubat 2026 tarihleri aralığındaki günlük veriler bu çalışmada kullanılmıştır.

Veri ön işleme süreci şu şekildedir: çok istasyonlu şehirler için tarih bazlı ortalama alınmış ve veriler MinMax ölçeklendirme yöntemiyle (MinMaxScaler) [0,1] aralığına normalize edilmiştir. Aykırı değer temizleme işlemleri ön işleme safhasında üç katlı çeyrek arası aralık (3xIQR) yöntemiyle ayıklanmış ve oluşan boşluklar zamansal sürekliliği korumak adına ardışık doldurma (bfill+ffill) teknikleriyle tamamlanmıştır. Hava kalitesi verilerinin doğasında bulunan yüksek otokorelasyon ve zamansal bağımlılıkları modelleyebilmek adına, PM₁₀ ve SO₂ hava kirliliğinin 1-3 günlük gecikmeli değerleri ile takvim bazlı özniteliklerden oluşan 21 değişkenli bir özellik seti türetilmiştir. Modeller, 2016–2025 dönemini kapsayan eğitim seti üzerinde zaman serisine duyarlı çapraz doğrulama (TimeSeriesSplit, k=5) ve hiperparametreler GridSearchCV ile optimize edilmiş; ardından 1.095 gün (3 yıl) sabit pencere kayan pencere yürüyüşlü doğrulama (rolling window walk-forward validation) yaklaşımı ile 2025–2026 yıllarını kapsayan bir yıllık test periyodu boyunca her gün için tek adım ileriye (one-step-ahead) yönelik yeniden eğitilerek test edilmiştir. Veri kapsamı, her şehir için istasyon sayısı, şehir temsil yöntemi, eksik veri oranları ve aykırı değer (outlier) oranları Ek 1'de verilmiştir. PM₁₀ ve SO₂ verilerinin betimleyici istatistikleri Tablo 1'de verilmiştir.

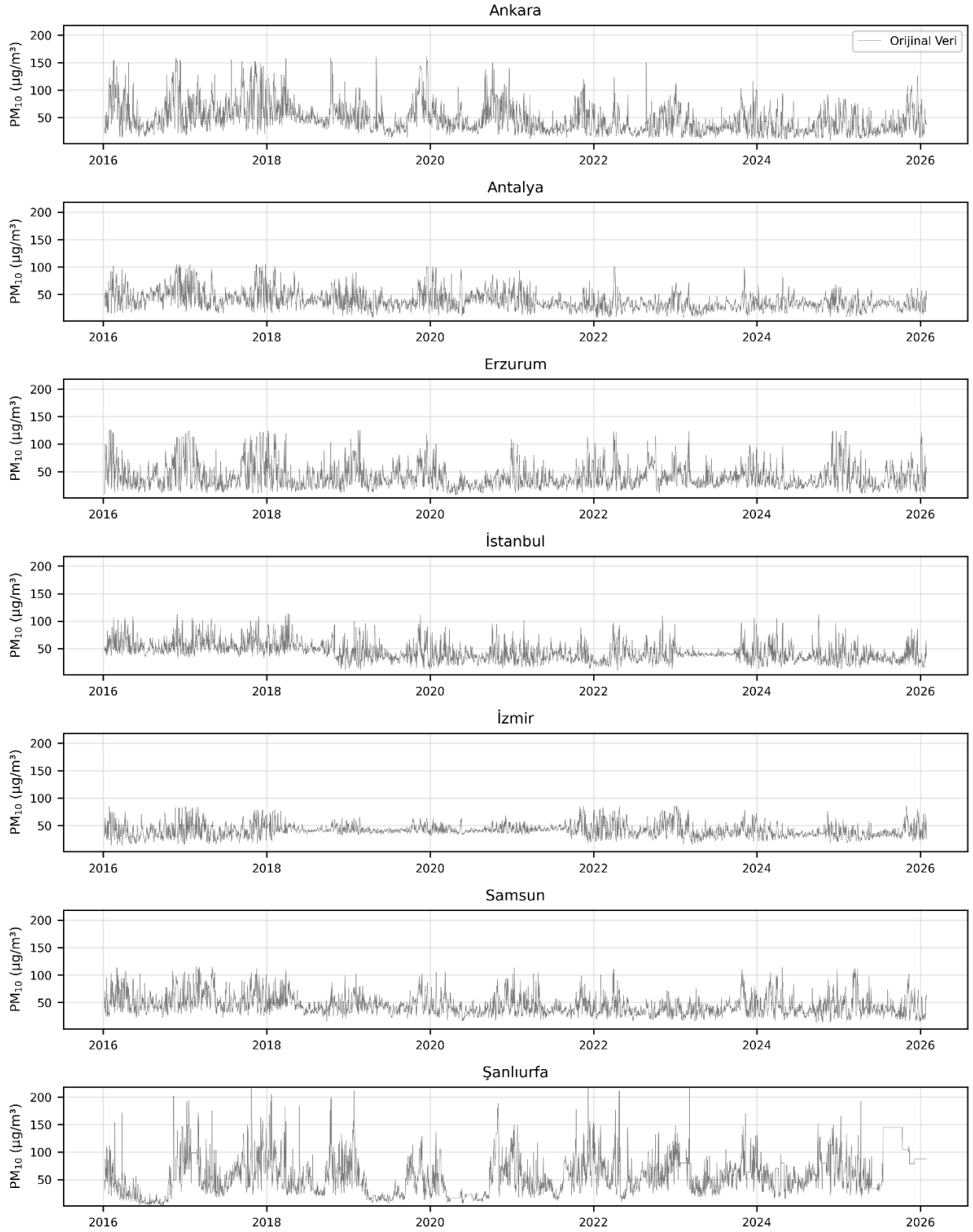
Tablo 1: PM₁₀ ve SO₂ verileri betimleyici istatistikleri

PM ₁₀ (µg/m ³)							
	Ankara	Antalya	Erzurum	İstanbul	İzmir	Samsun	Şanlıurfa
Ortalama	46,8	39	41,2	43,7	41,4	45,5	57,7
Standart Sapma	26,5	16,7	22,3	17,2	11,8	18,2	36,3
Minimum	7,3	7,4	7,7	10,5	13,9	13,5	2,3
Q1	27,4	27,9	25,3	31,5	34,1	32,9	30,1
Medyan	39,7	35,7	35,1	40,8	40,1	41,2	50,7
Q3	59,3	47,1	50,4	52,6	46,6	53,7	78,6
Maksimum	155,9	104,8	125,8	114,1	84,9	115,1	217,8
SO ₂ (µg/m ³)							
	Ankara	Antalya	Erzurum	İstanbul	İzmir	Samsun	Şanlıurfa
Ortalama	7,4	3,9	7,4	11,1	11,2	11,3	23,3
Standart Sapma	3,1	2,1	4,4	6,8	3,3	5,7	21,1
Minimum	1,9	0,2	2,2	2,5	4,7	2,4	0,6
Q1	5,2	2,6	4,4	4,9	8,9	7,0	7,6
Medyan	6,9	3,6	5,8	9,0	10,8	9,9	16,3
Q3	8,8	4,6	8,9	17,8	12,7	14,2	31,8
Maksimum	19,4	10,5	22,6	33,5	23,9	36,1	103,8

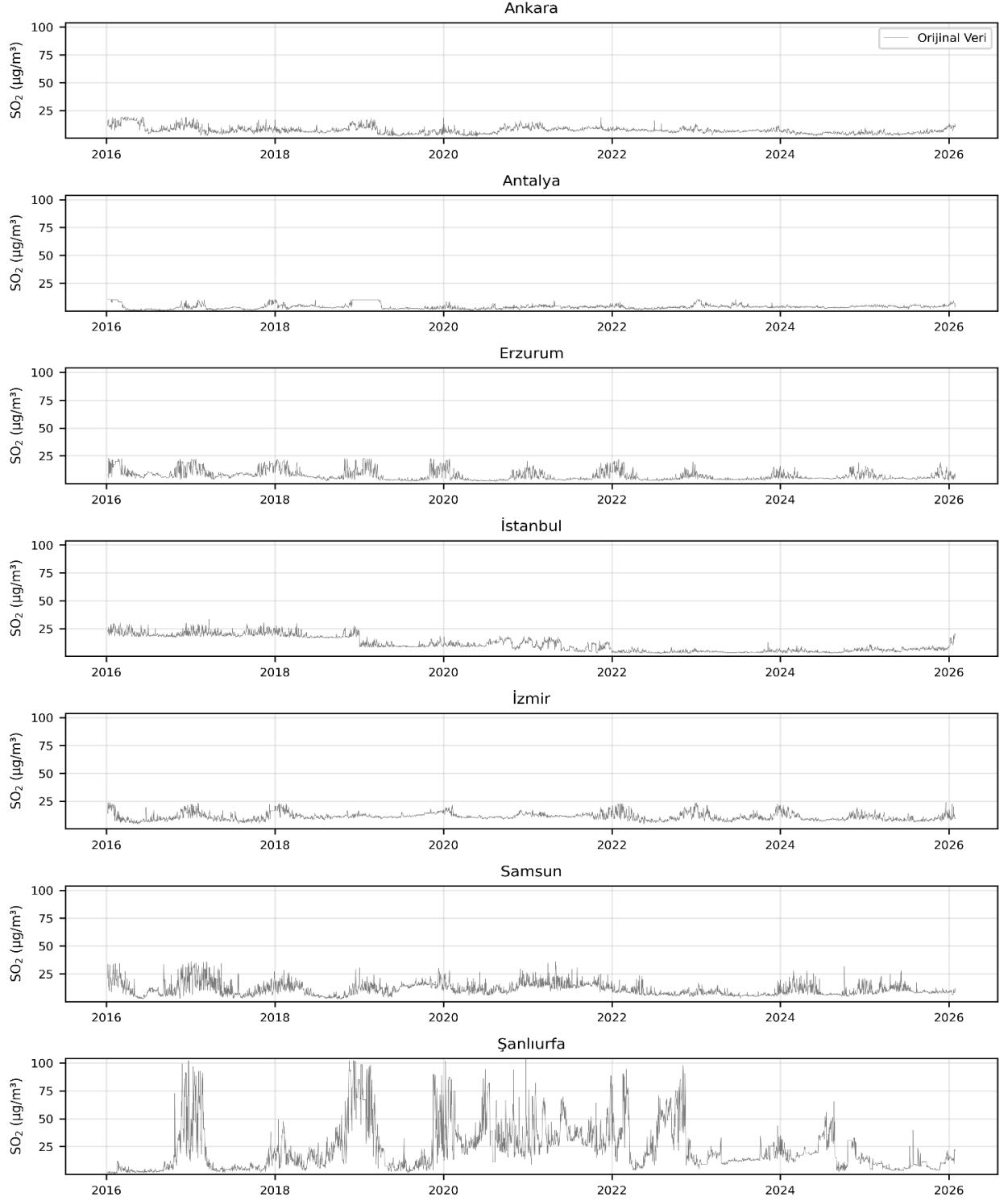
PM₁₀ ve SO₂ ortalama değerlerine göre iller bazında en yüksek değere Şanlıurfa sahip iken, Antalya da en düşük değer gözlemlenmiştir. Araştırma periyodu boyunca en yüksek PM₁₀ ve SO₂ standart sapmalarının Şanlıurfa ilinde gözlemlendiği, kısaca PM₁₀ ve SO₂ hava kirleticisi miktarlarının en fazla dalgalandığı şehir Şanlıurfa olarak Şekil 2 ve Şekil 3’de görülmektedir.

Şekil 1: İllere göre PM₁₀ ve SO₂ Kutu Grafikleri

PM₁₀ konsantrasyonu yasal sınırı 45 µg/m³ ve SO₂ yasal sınırı 40 µg/m³ olarak belirlenmiştir (DSÖ 2021). Şekil 1'e göre PM₁₀ konsantrasyonu kutu grafiği incelendiğinde şehirlerin çoğunda medyan değerleri yasal sınıra çok yakın olduğu görülmektedir. En yüksek medyan değerine sahip Şanlıurfa ili 150 µg/m³ seviyesi ile PM₁₀ konsantrasyonu bakımından en kirli şehir olarak tespit edilmiştir. Antalya ve Erzurum şehirleri diğer illere göre medyan değeri yasal sınır olan 45 µg/m³ altında bulunduğu için diğerlerine göre daha temiz havaya sahip gibi görünselerde aykırı değerlere sahiplerdir. SO₂ konsantrasyonu grafiği incelendiğinde bütün şehir medyan değerlerinin yasal sınırın altında olduğu görülmektedir. En yüksek medyan değerine sahip Şanlıurfa ili SO₂ konsantrasyonu bakımından medyan değerine göre yasal sınırın altında yer almaktadır. Antalya şehri ise diğerlerine göre medyan değeri yasal sınır olan 40 µg/m³ altında bulunduğu için daha temiz hava kalitesine sahip olduğu söylenebilir.



Şekil 2: İllere göre PM₁₀ verileri zaman serisi grafikleri

Şekil 3: İllere göre SO₂ verileri zaman serisi grafikleri

4. Bulgular ve Tartışma

4.1 Modelleme

Bu bölümde, iller bazındaki PM₁₀ ve SO₂ verilerinin modellenmesi için GBM, HistGBM, LightGBM, Random Forest ve XGBoost makine öğrenmesi modellerinin performansları MAE, RMSE ve R² model kıyaslama kriterleri ile karşılaştırılmıştır. Söz konusu modellerin hiperparametreleri grid search yöntemiyle optimize edilmiştir. Bu modellerin tercih edilen en iyi hiperparametre tahminleri Ek 2 ve Ek 3’de yer almaktadır.

Tablo 2: Makine Öğrenmesi Modelleri ile PM₁₀ verileri modelleme performansı karşılaştırılması

MAE								
Model	Ankara	Antalya	Erzurum	İstanbul	İzmir	Samsun	Şanlıurfa	Ortalama
GBM	0,071	0,078	0,076	0,071	0,073	0,076	0,055	0,071
HistGBM	0,055	0,060	0,058	0,054	0,056	0,057	0,042	0,055
LightGBM	0,055	0,061	0,058	0,055	0,056	0,057	0,042	0,055
Random Forest	0,031	0,033	0,033	0,031	0,033	0,033	0,024	0,031
XGBoost	0,022	0,025	0,025	0,024	0,023	0,024	0,017	0,023
RMSE								
GBM	0,097	0,109	0,104	0,100	0,103	0,105	0,083	0,100
HistGBM	0,075	0,083	0,079	0,075	0,078	0,079	0,063	0,076
LightGBM	0,075	0,083	0,079	0,076	0,078	0,078	0,063	0,076
Random Forest	0,045	0,047	0,048	0,044	0,048	0,047	0,038	0,045
XGBoost	0,032	0,034	0,035	0,034	0,031	0,033	0,026	0,032
R ²								
GBM	0,726	0,639	0,697	0,641	0,621	0,654	0,744	0,674
HistGBM	0,839	0,790	0,828	0,797	0,782	0,805	0,851	0,813
LightGBM	0,837	0,789	0,827	0,790	0,782	0,807	0,850	0,812
RF	0,940	0,933	0,937	0,930	0,919	0,929	0,945	0,933
XGBoost	0,970	0,964	0,966	0,958	0,965	0,965	0,975	0,966

Tablo 2’de iller bazında PM₁₀ verilerinin modelleme performansı MAE ve RMSE model kıyaslama kriterleri sonuçlarına göre XGBoost modeli diğer kullanılan modellere göre en düşük ortalama değerlere sahip olduğu için PM₁₀ verilerini modellemede daha etkin olduğu söylenebilir. Ayrıca, XGBoost modeli PM₁₀ verilerinin %96,60 açıklama yüzdesi ile en yüksek değere sahiptir.

Tablo 3: Makine Öğrenmesi Modelleri ile SO₂ verileri modelleme performansı karşılaştırılması

MAE								
Model	Ankara	Antalya	Erzurum	İstanbul	İzmir	Samsun	Şanlıurfa	Ortalama
GBM	0,048	0,042	0,056	0,027	0,048	0,062	0,045	0,047
HistGBM	0,038	0,033	0,042	0,021	0,037	0,048	0,034	0,036
LightGBM	0,038	0,033	0,042	0,021	0,037	0,048	0,034	0,036
Random Forest	0,021	0,018	0,025	0,012	0,022	0,028	0,020	0,021
XGBoost	0,016	0,014	0,016	0,009	0,015	0,018	0,013	0,014
RMSE								
GBM	0,071	0,063	0,089	0,041	0,072	0,092	0,077	0,072
HistGBM	0,055	0,047	0,063	0,031	0,054	0,069	0,056	0,054
LightGBM	0,055	0,046	0,064	0,031	0,054	0,070	0,057	0,054
Random Forest	0,032	0,029	0,042	0,019	0,034	0,042	0,037	0,033
XGBoost	0,022	0,020	0,024	0,013	0,021	0,027	0,022	0,021
R ²								
GBM	0,846	0,918	0,855	0,967	0,828	0,729	0,876	0,860
HistGBM	0,908	0,954	0,926	0,981	0,905	0,848	0,933	0,922
LightGBM	0,908	0,956	0,925	0,981	0,905	0,846	0,932	0,922
Random Forest	0,969	0,983	0,968	0,993	0,962	0,944	0,971	0,970
XGBoost	0,985	0,992	0,990	0,997	0,985	0,978	0,990	0,988

Tablo 3’te iller bazında SO₂ verilerinin modelleme performansı MAE ve RMSE model kıyaslama kriterleri sonuçlarına göre XGBoost modeli diğer kullanılan modellere göre en düşük ortalama değerlere sahip olduğu için SO₂

verilerini modellemede daha etkin olduğu söylenebilir. Ayrıca, XGBoost modeli PM₁₀ verilerinin %98,80 açıklama yüzdesi ile en yüksek değere sahiptir.

Sonuç olarak, PM₁₀ ve SO₂ verilerinin modelleme konusunda en etkin makine öğrenme modelinin XGBoost olduğunu söylenebilir.

4.2 Gelecek Tahmini

Bu bölümde, iller bazındaki PM₁₀ ve SO₂ verilerinin gelecek 1 yıllık tahminleri için makine öğrenmesi modellerinin performansları MAE, RMSE ve R² model kıyaslama kriterleri ile karşılaştırılmıştır.

Tablo 4: Makine öğrenmesi modelleri ile PM₁₀ verileri gelecek tahmin performansı karşılaştırılması

MAE								
Model	Ankara	Antalya	Erzurum	İstanbul	İzmir	Samsun	Şanlıurfa	Ortalama
GBM	0,064	0,057	0,079	0,082	0,073	0,089	0,069	0,073
HistGBM	0,066	0,059	0,082	0,084	0,076	0,090	0,098	0,079
LightGBM	0,066	0,059	0,082	0,084	0,076	0,090	0,094	0,079
Random Forest	0,068	0,059	0,081	0,084	0,076	0,090	0,080	0,077
XGBoost	0,071	0,062	0,088	0,089	0,083	0,095	0,091	0,083
RMSE								
GBM	0,087	0,078	0,116	0,113	0,102	0,121	0,099	0,102
HistGBM	0,089	0,080	0,122	0,118	0,108	0,124	0,154	0,114
LightGBM	0,088	0,080	0,121	0,117	0,109	0,123	0,140	0,111
Random Forest	0,091	0,079	0,122	0,116	0,107	0,122	0,119	0,108
XGBoost	0,095	0,083	0,132	0,124	0,115	0,130	0,133	0,116
R ²								
GBM	0,566	0,482	0,603	0,400	0,487	0,556	0,663	0,537
HistGBM	0,536	0,448	0,562	0,351	0,422	0,539	0,186	0,435
LightGBM	0,551	0,455	0,573	0,357	0,410	0,544	0,329	0,460
Random Forest	0,524	0,462	0,565	0,373	0,428	0,548	0,518	0,488
XGBoost	0,471	0,419	0,493	0,280	0,341	0,492	0,398	0,413

Tablo 4'teki PM₁₀ verilerinin iller bazında 1 yıllık gelecek tahmini MAE ve RMSE model kıyaslama kriterleri performans sonuçlarına göre GBM modeli diğer kullanılan modellere göre düşük ortalama değerlere sahip olduğu için GBM modelinin PM₁₀ verileri için gelecek tahmininde daha etkin olduğu görülmektedir. Buna ek olarak, GBM modeli %53,70 açıklama yüzdesi ile PM₁₀ verilerinin gelecek tahminini açıklama oranı açısından en yüksek değer olarak karşımıza çıkmaktadır.

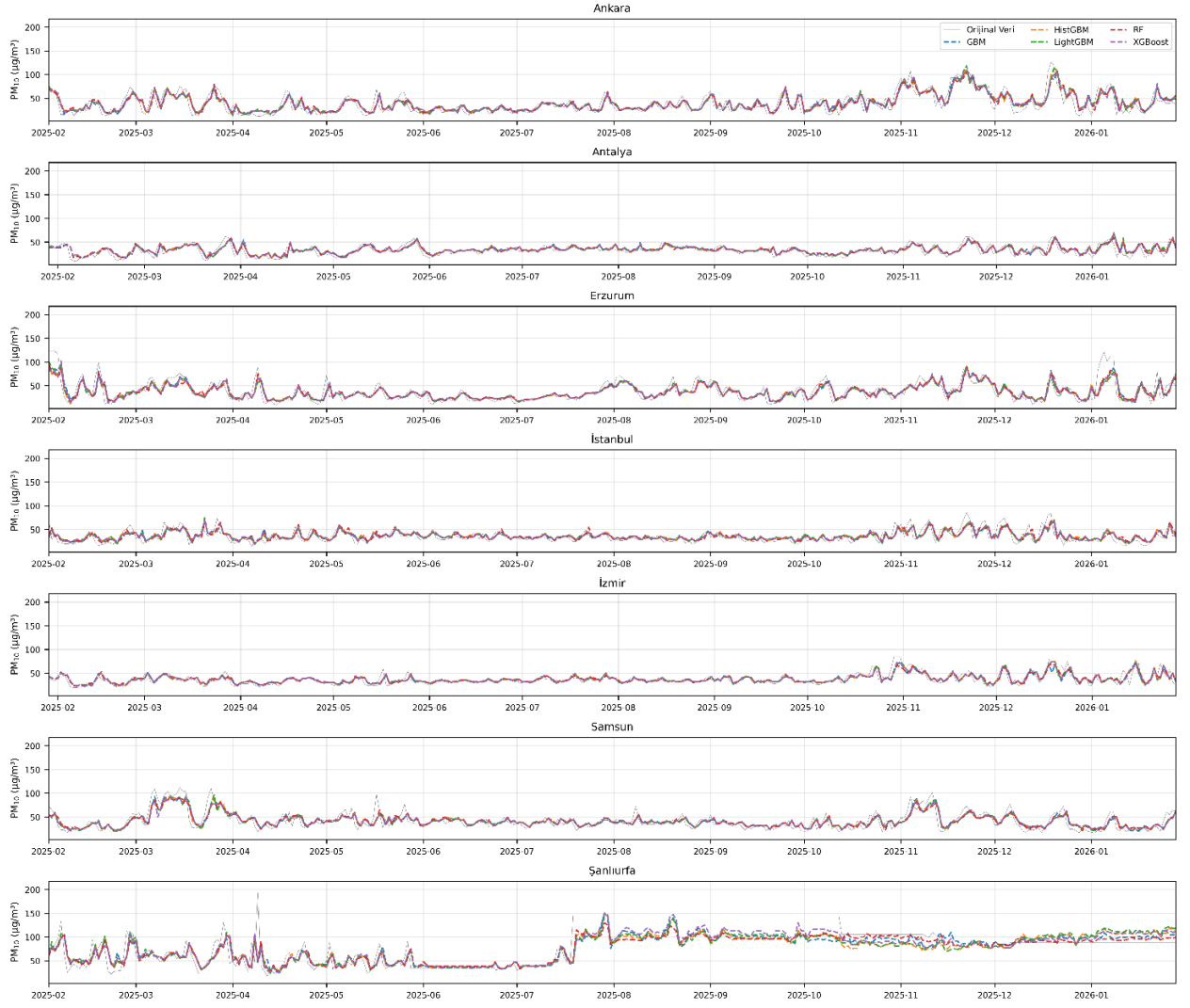
Tablo 5: Makine öğrenmesi modelleri ile SO₂ verileri gelecek tahmin performansı karşılaştırılması

Model	MAE							
	Ankara	Antalya	Erzurum	İstanbul	İzmir	Samsun	Şanlıurfa	Ortalama
GBM	0,037	0,037	0,048	0,021	0,054	0,062	0,022	0,040
HistGBM	0,039	0,041	0,050	0,022	0,057	0,066	0,023	0,043
LightGBM	0,039	0,040	0,050	0,022	0,058	0,064	0,024	0,042
Random Forest	0,039	0,038	0,050	0,023	0,057	0,066	0,023	0,042
XGBoost	0,043	0,044	0,053	0,023	0,062	0,071	0,025	0,046
Model	RMSE							
	Ankara	Antalya	Erzurum	İstanbul	İzmir	Samsun	Şanlıurfa	Ortalama
GBM	0,050	0,050	0,075	0,030	0,077	0,097	0,039	0,060
HistGBM	0,052	0,056	0,078	0,031	0,084	0,103	0,042	0,064
LightGBM	0,052	0,055	0,079	0,031	0,084	0,101	0,043	0,063
Random Forest	0,053	0,053	0,079	0,031	0,080	0,101	0,043	0,063
XGBoost	0,058	0,059	0,085	0,032	0,088	0,107	0,047	0,068
Model	R ²							
	Ankara	Antalya	Erzurum	İstanbul	İzmir	Samsun	Şanlıurfa	Ortalama
GBM	0,764	0,695	0,653	0,834	0,701	0,339	0,852	0,691
HistGBM	0,742	0,629	0,629	0,827	0,644	0,269	0,822	0,652
LightGBM	0,744	0,641	0,619	0,828	0,639	0,295	0,817	0,655
Random Forest	0,732	0,663	0,615	0,821	0,676	0,289	0,819	0,659
XGBoost	0,680	0,583	0,558	0,808	0,611	0,210	0,782	0,605

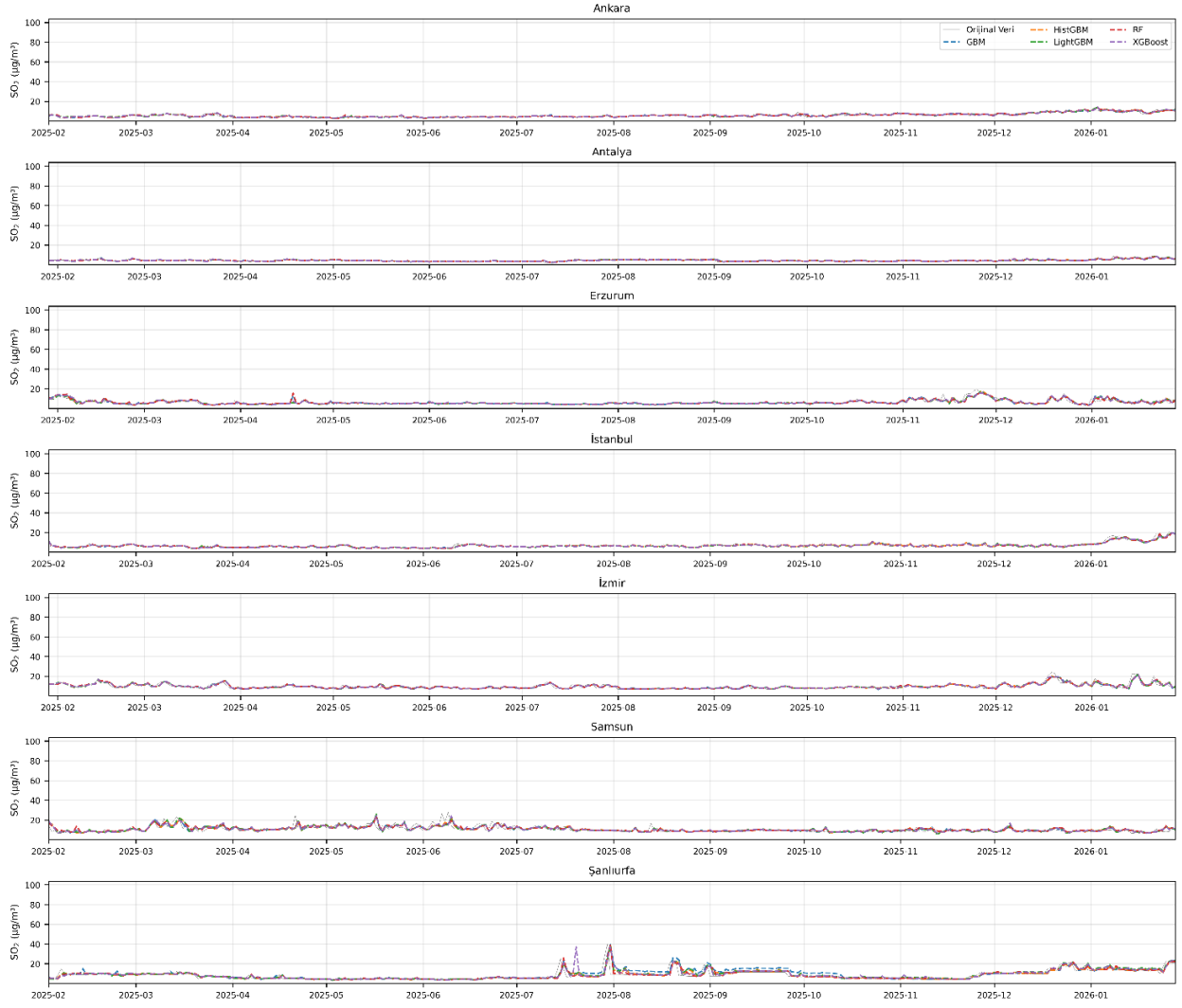
Tablo 5'te iller bazında SO₂ verilerinin 1 yıllık gelecek tahmini MAE ve RMSE model kıyaslama kriterleri performans sonuçlarına göre GBM modeli diğer kullanılan modellere göre her bir kıyaslama kriterine göre en düşük ortalama değerlere sahip olduğu için GBM modelinin SO₂ verileri için gelecek tahmininde daha etkin olduğu söylenebilir. Ayrıca, GBM modeli SO₂ verilerinin %69,10 ile gelecek tahminini açıklama oranı açısından en yüksek değer olarak karşımıza çıkmaktadır.

Sonuç olarak, PM₁₀ ve SO₂ verilerinin gelecek tahmini konusunda en etkin makine modelinin GBM olduğunu söylenebilir.

Şekil 4 ve 5'te söz konusu makine öğrenmesi modelleri genel olarak PM₁₀ ve SO₂ verilerinin gelecek tahmini konusunda birbirlerinin sonuçlarına yakın derece olduğu ve sonuçların Tablo 5'te uygun görülmüştür. Buna ek olarak, modeller hava kirleticisi verilerinin gelecek tahmini konusundaki genel trendleri yakalama da başarılı, ancak özellikle Erzurum ve Şanlıurfa illeri için ani dalgalamaların olduğu gözlemlerde yetersiz kaldığı açık olarak görülmektedir. Bu duruma sebep olarak emisyon envanteri ve üst atmosfer verilerinin eksikliği, şehir bazlı istasyon heterojenliği ile yerel mikroklimatik dinamiklerin modele tam dahil edilememesi gibi kısıtlılıklar söylenebilir.



Şekil 4: İllere göre PM₁₀ verileri gelecek tahminleri

Şekil 5: İllere göre SO₂ verileri gelecek tahminleri

5 Sonuçlar

Bu araştırma, insan sağlığını tehdit eden ve birincil kirleticiler olarak tanımlanan PM₁₀ ve SO₂ hava kirleticilerinin ölçümleri ve değerlendirmeleri yapılarak hava kalitesi konusunda ülkemizdeki gelecek nesilleri ve araştırmacıları bilinçlendirmek amacıyla Türkiye'deki Ocak 2016-Şubat 2026 tarihleri arasındaki 7 Büyükşehir'in (Ankara, Antalya, Erzurum, İstanbul, İzmir, Samsun ve Şanlıurfa) PM₁₀ ve SO₂ değerlerinin modellenmesi ve gelecek tahmininde makine öğrenmesi yöntemlerinden GBM, HistGBM, LightGBM, Random Forest ve XGBoost modelleri ile gerçekleştirilmiştir. Genel olarak, makine öğrenmesi yöntemlerinin hava kirleticileri modellemesi ve gelecek tahmini konusunda etkin bir performansa sahip olduğu görülmüştür. Buna ek olarak, birincil hava kirleticilerini modellemede diğer modellere göre XGBoost ve gelecek tahmini konusunda ise diğer modellere göre GBM modelinin daha etkin performansa sahip olduğu görülmüştür.

Bu araştırma makine öğrenmesi algoritmalarının hava kalitesi tahminindeki potansiyelini ortaya koymakla birlikte; emisyon envanteri ve üst atmosfer verilerinin eksikliği, şehir bazlı istasyon heterojenliği ile yerel mikroklimatik dinamiklerin modele tam dahil edilememesi gibi kısıtlılıklar içermektedir. Özellikle, Erzurum ve Şanlıurfa gibi illerdeki ani rejim değişimleri ve uç değerlerin tahminindeki yetersizlikler, tek adım ileriye yönelik mevcut stratejinin durağan olmayan (non-stationary) ekstrem hava olaylarını yakalamadaki sınırlarını göstermektedir. Veri normalize yöntemleri, aykırı değer modifikasyonları değiştirilerek ve ayrıca değişken setine meteorolojik değişkenler (örn. hava sıcaklığı, rüzgar hızı ve yönü, nem, basınç vb.) eklenerek gelecekteki araştırmalara yön gösterici niteliktedir.

Demir Yetiş vd. (2025), çalışmasında PM₁₀ ve SO₂ parametrelerine ait kalitesi incelemesi sonucunda XGBoost ve LightGBM algoritmalarında yüksek doğruluk gözlemlenmiş ve Random Forest ile ANN modelleri ile yüksek başarı elde etmiştir. Metin (2025) ise İstanbul hava kirliliği seviyelerinin (PM₁₀, SO₂, NO₂, NO_x ve NO) makine öğrenmesi ve hibrit yöntemlerle tahmin edilmesini aşamasında hava kalitesi yönetimi için hibrit modellerin etkinliğini ortaya koyarak,

gelecek araştırmalara meteorolojik veri entegrasyonu, hiperparametre optimizasyonu ve zaman serisi modellerinin kullanımını önermiştir. Xu vd. (2023), Çin'in doğu bölgesindeki SO₂ hava kirlleticilerinin makine öğrenimi modeli ile uzamsal ve zamansal değişimlerini tahmin etmede LightGBM modelinin yüksek performans gösterdiğini vurgulamışlardır. Singh vd. (2025) çalışmasın da Hindistan'daki Hava kirliliği PM_{2.5} için modellenmesi için Histogram Gradient Boosting ve XGBoost konsantrasyonunun tahmininde hem genelleme hem de doğruluk açısından en doğru model olmaktadır. Burada görüldüğü gibi Türkiye ve Dünya'nın farklı bölgelerindeki hava kirleticilerinin makine öğrenmesi ile analizlerinin genel olarak performanslarının etkin olduğu ve bu araştırmadaki bulguların makine öğrenmesi ve hava kirleticileri literatürünü destekleyici nitelikte olduğu görülmüştür.

Bu araştırmanın bilimsel sonuçları ışığında alınacak bütüncül önlemler, hem mevcut hava kalitesinin korunmasını sağlayacak hem de gelecek nesiller için yaşanabilir bir çevrenin teminatı olacaktır. Bu araştırmadaki gibi bilimsel anlamda makine öğrenmesi modellerinin hava kalitesi konularında kullanılması, temiz hava eylem planı oluşturulması, uygun hava kirliliği yönetim planı, evsel ve sanayi ısınmasında yakıt türlerinin değiştirilmesi ve hava kalitesi eylem stratejilerinin geliştirilmesi gerektiğini belirterek toplumun bilinçlendirilmesi için çalışmalara hız verilmesi gerektiği sonucuna ulaşılmıştır. Bu alanda yeni çalışmalar yapılması, hava kirliliği konusunda araştırmalarda disiplinler arası çalışmaların sayısının artırılması gerekliliği önem teşkil etmektedir.

Kaynaklar

- Atalay, İ. E., & Neslihanoglu, S. (2021). Türkiye'deki illerin partikül madde (PM₁₀) miktarının değerlendirilmesi ve R programlama dili ile görselleştirilmesi. *Doğal Afetler ve Çevre Dergisi*, 7(2), 354-361.
- Aydın, O., & Raja, N. B. (2025). Ankara'da Partikül Madde (PM₁₀) ve Kükürt Dioksit (SO₂) Konsantrasyonlarının İklim Parametreleri İle Mekânsal-Zamansal Analizi. *Doğal Afetler Ve Çevre Dergisi*, 11(1), 249-267.
- Chen, T., & Guestrin, C. (2016) "XGBoost: A scalable tree boosting system" in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Vol:13-17, August 2016. Association for Computing Machinery, 8, pp. 785-794.
- Demir Yetiş, A., İlhan, N., & Karakuş, S. (2025). Makine öğrenmesi algoritmaları ile Malatya ilinde PM₁₀ ve SO₂ konsantrasyonlarının tahmini. *Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 30(2), 621-635.
- Demirarslan, K., & Akıncı, H. (2016). Doğu Karadeniz Bölgesindeki Partikül Madde Dağılımlarının Coğrafi Bilgi Sistemleri Yardımıyla Belirlenmesi. *Doğal Afetler ve Çevre Dergisi*, 2(1), 30-45.
- Demirarslan, K. O., & Akıncı, H. (2016). Doğu Karadeniz Bölgesinde Kükürtdioksit (SO₂) Dağılımlarının Coğrafi Bilgi Sistemleri Yardımıyla Belirlenmesi. *Doğal Afetler ve Çevre Dergisi*, 2(2), 81-99.
- DSÖ, Ambient (outdoor) air pollution, 2021. [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), (Erişim Tarihi: 03 Mart 2026)
- Eşidir, K. A. (2025) Makine öğrenimi modelleri ile yetişkin eğitimi analizi: modellerin karşılaştırmalı performansı. *Elektronik Sosyal Bilimler Dergisi*, 24(2), 946-964.
- Li, R., Sun, K., Wang, X., Dai, G., Fan, M., Ma, L., ... & Wu, S. (2025). PM_{2.5} and PM₁₀ vertical distribution retrieval methods based on coherent Doppler lidar via machine learning: For haze and dust in Qingdao. *Atmospheric Environment*, 359, 121351.
- Metin, S. (2025). Hava Kirliliği Değerlerinin Makine Öğrenmesi Ve Hibrit Yöntemler İle Tahmini: İstanbul Örneği. *Malatya Turgut Özal Üniversitesi İşletme ve Yönetim Bilimleri Dergisi*, 6(2), 190-203.
- Mert, İ. (2025). Prediction of Wind Speed Using Tree-Based Ensemble Algorithms: CatBoost, HistGBM, and XGBoost. *International Journal of Multidisciplinary Studies and Innovative Technologies*, 9(1), 145-150.
- Nhat-Duc, H., & Van-Duc, T. (2023). Comparison of histogram-based gradient boosting classification machine, random forest, and deep convolutional neural network for pavement raveling severity classification. *Automation in construction*, 148, 104767.
- Olçay, H., & Çukur, H. (2025). Sıcaklık Terselmesi ve Hava Kirliliği (SO₂) İlişkisi: Manisa Örneği. *Turkish Studies*, 20(2), 1109-1127.
- Olçay, H., Yakın, F., & Gökçe Gündüzoğlu, H.A. (2025). Hava kirliliği ile doğal ortam arasındaki ilişki: Sistematik literatür taraması. *Coğrafi Bilimler Dergisi/ Turkish Journal of Geographical Sciences*, 23 (2), 701-733.
- Ören, S. (2024). Hatay ve Kahramanmaraş merkezli meydana gelen depremler sonrasında dış ortam hava kalitesinde PM₁₀ ve SO₂ Düzeyleri ve Değişimleri. *Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi*, 11(1), 104-115.
- Özşahin, E., Eroğlu, İ., & Pektezeli, H. (2016). Keşan'da Edirne Hava Kirliliği. *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, (36), 83-100.
- Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Update*, 1(1), 2007.
- Yeşilyurt, S. E., & Dalkılıç H. Y. (2021) Xgboost ve gradient boost machine ile günlük nehir akımı tahmini. In 3rd International Symposium of III Engineering Applications on Civil Engineering and Earth Sciences.
- Shang, H., Zhang, X., Cheng, M., Zhao, B., & Zhao, X. (2026). Machine learning prediction of cardiovascular disease risk progression from sulfur dioxide exposure in longitudinal population studies in China. *BMC Cardiovascular Disorders*, 26, 229 (2026). <https://doi.org/10.1186/s12872-026-05591-5>
- Singh, S. K., Jain, R., Palaniappan, D., Parmar, K., T. P., & Gothania, J. (2025). Spatiotemporal analysis and machine learning-based prediction of air quality in Indian urban cities. *Environmental Research and Technology*, 8(4), 809-822.
- Suleiman, A., Tight, M. R., & Quinn, A. D. (2019). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM₁₀ and PM_{2.5}). *Atmospheric Pollution Research*, 10(1), 134-144.
- Zhai, B., & Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average PM_{2.5} concentrations in Beijing, China. *Science of the Total Environment*, 635, 644-658.
- Xu T, Chen B, Ren Y, Zhao L, Hu J, Wang Y, Song Z & Li X (2023). Estimation of the ground-level SO₂ concentration in eastern China based on the LightGBM model and Himawari-8 TOAR. *Frontiers in Earth Science*, 10, 1037719.

Ek 1. Araştırma verileri tablosu

Şehir	İstasyon Sayısı	Şehir Temsil Yöntemi	PM ₁₀ Eksik Veri Oranı (%)	SO ₂ Eksik Veri Oranı (%)	PM ₁₀ Aykırı Değer Oranı (Outlier % - IQR x3)	SO ₂ Aykırı Değer Oranı (Outlier % - IQR x3)
Izmir	23	İstasyonların Ortalaması	4,17	4,17	1,01	1,31
Istanbul	39	İstasyonların Ortalaması (Mean of Stations)	4,17	4,17	0,9	0,87
Ankara	18	İstasyonların Ortalaması (Mean of Stations)	4,17	4,17	0,73	1,17
Gaziantep	6	İstasyonların Ortalaması (Mean of Stations)	3,99	3,89	0,44	2,26
Erzurum	4	İstasyonların Ortalaması (Mean of Stations)	4,16	4,16	1,69	4,41
Samsun	6	İstasyonların Ortalaması (Mean of Stations)	4,17	4,16	0,54	1,47
Antalya	8	İstasyonların Ortalaması (Mean of Stations)	4,12	4,09	1,44	4,82
Sanliurfa	3	İstasyonların Ortalaması (Mean of Stations)	3,52	3,82	0,49	2,37

Ek 2. PM₁₀ için tercih edilen en iyi hiperparametre tahminleri

Şehir	Model	Hiperparametreier
Ankara	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=500, max_depth=5, learning_rate=0.01, l2_regularization=1.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
Antalya	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=300, max_depth=10, learning_rate=0.01, l2_regularization=0.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
Erzurum	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=500, max_depth=5, learning_rate=0.01, l2_regularization=1.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
İstanbul	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=300, max_depth=10, learning_rate=0.01, l2_regularization=0.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=5, learning_rate=0.05, gamma=0.1, colsample_bytree=0.8, random_state=42
İzmir	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=300, max_depth=5, learning_rate=0.01, l2_regularization=0.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=100, min_samples_split=5, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
Samsun	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=500, max_depth=5, learning_rate=0.01, l2_regularization=1.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=100, min_samples_split=5, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
Sanliurfa	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=300, max_depth=5, learning_rate=0.01, l2_regularization=0.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42

Ek 3. SO₂ için tercih edilen en iyi hiperparametreler tahminleri

Şehir	Model	Hiperparametreler
Ankara	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=500, max_depth=5, learning_rate=0.01, l2_regularization=1.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=100, min_samples_split=2, min_samples_leaf=1, max_features=sqrt, max_depth=None, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
Antalya	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=500, max_depth=5, learning_rate=0.01, l2_regularization=1.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=100, min_samples_split=5, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
Erzurum	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=300, max_depth=5, learning_rate=0.01, l2_regularization=0.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=100, min_samples_split=5, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
İstanbul	GradientBoostingRegressor	subsample=1.0, n_estimators=500, min_samples_split=2, max_depth=5, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=500, max_depth=5, learning_rate=0.01, l2_regularization=1.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=127, n_estimators=500, max_depth=7, learning_rate=0.05, colsample_bytree=0.8, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=5, learning_rate=0.1, gamma=0, colsample_bytree=1.0, random_state=42
İzmir	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=500, max_depth=5, learning_rate=0.01, l2_regularization=1.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=500, max_depth=7, learning_rate=0.05, colsample_bytree=0.8, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=1.0, random_state=42
Samsun	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=500, max_depth=5, learning_rate=0.01, l2_regularization=1.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=31, n_estimators=100, max_depth=-1, learning_rate=0.1, colsample_bytree=1.0, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=5, learning_rate=0.05, gamma=0.1, colsample_bytree=0.8, random_state=42
Sanliurfa	GradientBoostingRegressor	subsample=1.0, n_estimators=100, min_samples_split=2, max_depth=3, learning_rate=0.1, random_state=42
	HistGradientBoostingRegressor	max_iter=300, max_depth=5, learning_rate=0.01, l2_regularization=0.0, random_state=42
	LGBMRegressor	subsample=0.8, num_leaves=127, n_estimators=500, max_depth=7, learning_rate=0.05, colsample_bytree=0.8, random_state=42
	RandomForestRegressor	n_estimators=300, min_samples_split=10, min_samples_leaf=4, max_features=1.0, max_depth=20, random_state=42
	XGBRegressor	subsample=0.8, n_estimators=100, max_depth=5, learning_rate=0.05, gamma=0.1, colsample_bytree=0.8, random_state=42