Konya Mühendislik Bilimleri Dergisi, c. 10, s. 4, 807-826, 2022
Konya Journal of Engineering Sciences, v. 10, n. 4, 807-826, 2022
ISSN: 2667-8055 (Elektronik)
DOI: 10.36306/konjes.1082866

ATMOSFERİK PARTİKÜL MADDELERİN MAKİNE ÖĞRENMESİ İLE TAHMİNİ: BEŞİKTAŞ, İSTANBUL ÖRNEĞİ

Ece YAĞMUR

Konya Teknik Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Endüstri Mühendisliği Bölümü, Konya, TÜRKİYE
ecyagmur@ktun.edu.tr

Geliş/Received: 04.03.2022; Kabul/Accepted in Revised Form: 05.08.2022

ÖZ: Hava kirliliği, insan sağlığına ve çevreye olumsuz etkileri nedeniyle uzun yıllardır tartışılmakta olan bir problemdir. Bu problemi çözmek ve gereken önlemleri almak amacıyla hava kalitesinin değerlendirilmesi önem arz etmektedir. Hava kalitesi değerlendirilirken kirletici konsantrasyonları analiz edilerek, toplum açısından herkesin anlayabileceği bir indeks sistemi kullanılmaktadır. Ulusal Hava Kalitesi İndeksi kapsamında kalite indeksi hesaplanan beş temel kirleticiden biri, ciddi solunum yolu hastalıklarına sebep olan atmosferik partikül maddelerdir. Bu çalışmada çapı 2,5 mikrondan küçük olan ve $PM_{2,5}$ olarak adlandırılan atmosferik partikül maddelerin oluşumunda trafik yoğunluğu, meteorolojik koşullar ve NOX, SO_2, PM_{10} hava kirleticilerinin etkisi araştırılmıştır. Bu amaçla İstanbul Büyükşehir Belediyesi tarafından farklı alanlarda verilerin paylaşıldığı açık veri portalından yararlanılarak Beşiktaş bölgesindeki hava kalitesi izleme istasyonu incelenmiştir. Atmosferik partikül maddelerin tahmininde Çoklu Doğrusal Regresyon (ÇDR), Rassal Orman (RO), Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) kullanılmıştır. Regresyon denkleminde farklı bağımsız değişkenlerin incelendiği farklı modeller geliştirilmiştir. Geliştirilen modeller ve kullanılan makine öğrenme algoritmaları determinasyon katsayısı (R^2), düzeltilmiş R^2, ortalama mutlak hata, ortalama hata karesi ve ortalama hata karesi kökü performans ölçütlerine göre karşılaştırılmıştır. Meteorolojik parametreler, trafik yoğunluğu, tarih ve PM_{10} konsantrasyonunun bağımsız değişken olarak kullanıldığı model, incelenen tüm performans ölçütlerine göre diğer modellere üstünlük sağlamıştır. Algoritmalar karşılaştırıldığında ise performans ölçütlerinin modellere göre değişiklik gösterdiği görülmüş ancak en iyi performans ortalamasına sahip teknik RO, en kötü performans ortalamasına sahip teknik ise ÇDR olarak bulunmuştur.

Anahtar Kelimeler: *Hava Kalitesi, Makine Öğrenmesi, Doğrusal Regresyon, Rassal Orman Algoritması, Destek Vektör Makineleri, Yapay Sinir Ağları*

Prediction of Atmospheric Particulate Matter By Machine Learning: A Case Study of Beşiktaş, İstanbul

ABSTRACT: Air pollution is a problem that has been discussed for many years due to its negative effects on human health and the environment. It is important to evaluate air quality to eliminate all these negative effects and take the necessary precautions. When evaluating air quality, pollutant concentrations are analyzed and an index system that can be understood by everyone in the society is used. One of the five main pollutants whose quality index is calculated within the scope of the National Air Quality Index is atmospheric particulate matter, which causes serious respiratory diseases. In the study, the effects of traffic density, meteorological conditions, and NOX, SO_2, PM_{10} pollutants on the formation of atmospheric particulate matter, which is less than 2.5 microns in diameter and called $PM_{2,5}$ were investigated. For this purpose, the air quality monitoring station in Beşiktaş Region was examined by using the open data portal where data in different areas are shared by the Istanbul Metropolitan Municipality. Multiple Linear Regression (MLR), Random Forest (RF), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were used in the prediction of atmospheric particulate matter. Different models have been

developed in which different independent variables were examined in the regression model. The developed models and the machine learning algorithms were compared according to coefficient of determination (R^2), adjusted R^2, mean absolute error, mean squared error and root mean square error performance criteria. The model, in which meteorological parameters, traffic density, date and PM_{10} concentration were used as independent variables, outperformed other models in terms of all performance criteria examined. When the results obtained were examined, it was seen that the algorithm performances varied according to the models. However, according to all performance criteria, the technique with the best average performance was found to be RF, while the technique with the worst performance average was found to be MLR.

Keywords: *Air Quality, Machine Learning, Linear Regression, Random Forest Algorithm, Support Vector Machines, Artificial Neural Networks*

GİRİŞ (INTRODUCTION)

Hava kirliliği tüm dünyayı ilgilendiren ve bu bağlamda ülkelerin, çevre ve insan sağlığını korumaya yönelik eylem planları geliştirmelerine sebep olan küresel bir çevre sorunudur. Sürdürülebilirliğin çevresel boyutu değerlendirildiğinde ise insanları en çok etkileyen, büyük yerleşim merkezleri ve sanayi bölgelerindeki hava kirliliğidir. Özellikle Sanayi Devriminden sonra makineleşmenin başlaması, şehirlerde fabrikaların kurulması ve buna bağlı olarak nüfusun büyük oranda kırsal bölgelerden kente yönelmesiyle bu problem daha da ciddi seviyelere ulaşmıştır.

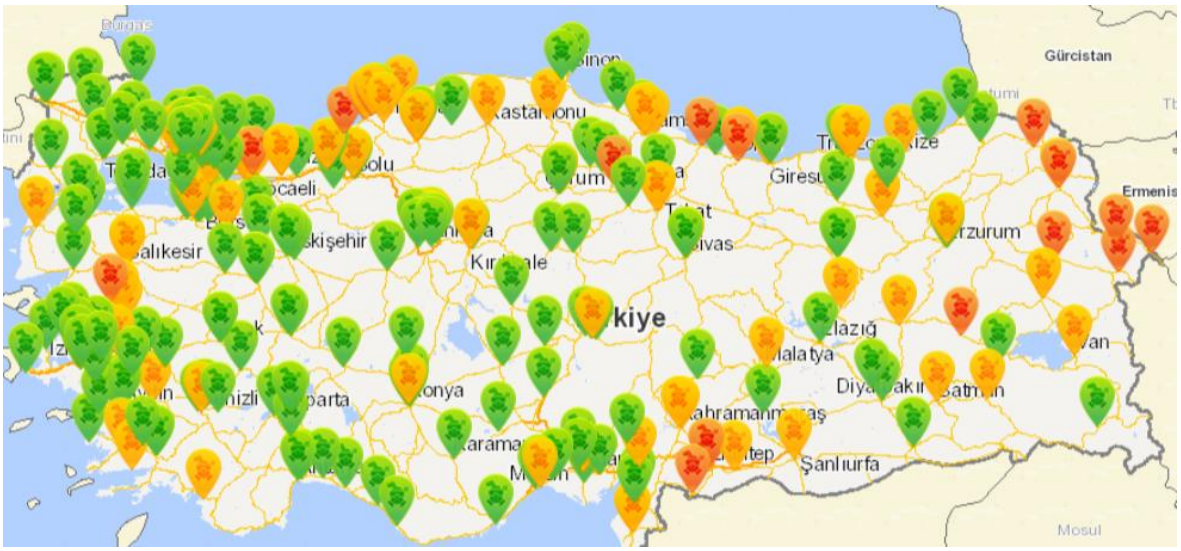
Dünya Sağlık Örgütü (DSÖ), kalp krizi, akciğer kanseri ve kronik solunum yolu hastalıkları nedeniyle yılda yaklaşık 4,2 milyon ölüm gerçekleştiğini ve Dünya nüfusunun yaklaşık %91'inin yaşadığı yerin hava kalitesi seviyesinin DSÖ tarafından belirlenen sınır değerin altında kaldığını raporlamıştır (DSÖ, 2022). Bu durum günümüzde insanların ciddi oranda hava kirleticisine maruz kaldığını göstermektedir. Kentsel hava kirliliği Dünya nüfusunun karşı karşıya olduğu en temel çevresel risklerden birisidir ve asit yağmurları, ozon tabakasının incelmesi, küresel ısınma gibi büyük çapta olumsuz etkileri de beraberinde getirmektedir. Tüm bu olumsuz etkileri önlemek ve daha sağlıklı, sürdürülebilir yaşam alanları oluşturabilmek için hava kirliliğine neden olan faktörler incelenmeli ve gerekli önlemler alınmalıdır.

Uluslararası Telekomünikasyon Birliği (2016) tarafından Akıllı ve Sürdürülebilir Şehirler için Çevresel Sürdürülebilirlik boyutu altında hava kalitesi, CO_2 emisyonu, enerji, iç mekân kirliliği ve su, toprak ve gürültü kirliliği olmak üzere beş performans göstergesi tanımlanmıştır. Bu bağlamda T.C. Çevre ve Şehircilik Bakanlığı Akıllı Şehirler Stratejisi ve Eylem Planı'nda belirtildiği gibi hava kalitesi izleme sistemleri (T.C. Çevre, Şehircilik ve İkim Değişikliği Bakanlığı, Sürekli İzleme Merkezi, www.havaizleme.gov.tr, ziyaret tarihi:23.01.2021) ile hava kalitesine ilişkin verilerin toplanması, değerlendirilmesi, kirlilik analizlerinin yapılması, gerekli önlemlerin alınarak hava kirliliğinin önlenmesi ve hava kalitesinin iyileştirilmesi hedeflenmektedir. T.C. Çevre ve Şehircilik Bakanlığı (2019), tarafından yayınlanan eylem planını hayata geçirmek amacıyla Türkiye genelinde birçok ilde (Ankara, Adana, Diyarbakır, Erzurum, İstanbul, İzmir, Konya ve Samsun) "Temiz Hava Merkezleri" kurulmuştur. Bunun dışında Şekil 1'de görülen mevcut hava kalitesi ölçüm istasyonlarıyla PM_{10}, $PM_{2,5}$, SO_2, CO, NO_2, NOX, O_3 kirleticileri için istenilen bölgeye ait hava kalitesi raporlanabilmektedir.

Tüm dünyada yaygın olarak kullanılan, Hava Kalitesi İndeksi sistemine göre hava kalitesi, havadaki kirleticilerin konsantrasyonlarına göre iyi, orta, hassas, sağlıksız, kötü ve tehlikeli olarak sınıflandırılmakta ve her bir seviye farklı bir renkle temsil edilmektedir. Şekil 1'de görüldüğü gibi ölçüm istasyonları Hava Kalitesi İndeks değerlerine göre yeşil, sarı ve turuncu ile temsil edilmiştir. Buna göre yeşil renk hava kalitesinin iyi olduğunu, sarı renk ortalama düzeyde olup endişe yaratacak bir durumun olmadığını, turuncu renk ise hava kalitesinin hassas düzeyde olduğunu bu nedenle solunum yolları ile ilgili problem yaşayan insanların olumsuz etkilenebileceğini ifade etmektedir.

Hava kirliliğinin, insan faktörü olmadan doğal yollarla (orman yangınları, çöl tozları, yanardağların çevresinde oluşan gaz bulutları vb.) oluşabildiği ancak bu oranın oldukça az olduğu bilinmektedir (Kampa ve Castanas 2008). Günümüzde hava kirliliğine sebep olan faktörlerin büyük çoğunluğu yapay

yollarla yani insanlar tarafından gerçekleştirilen faaliyetlerin sonucu olarak ortaya çıkmaktadır. Yerleşim merkezlerinde hava kirliliğinin ağırlıklı olarak ısınmadan, ulaşımdan ve sanayiden kaynaklandığı söylenebilir (Pénard-Morand ve Annesi-Maesano 2004). Bunun dışında sıcaklık, nem, yağış miktarı, rüzgâr hızı ve yönü gibi meteorolojik koşulların da hava kirliliği üzerinde etkili olduğu bilinmektedir (Pearce ve diğ. 2011). Ayrıca dizel partikül maddesinin kentsel alanlardaki çoğu küçük boyutlu partiküler maddeden sorumlu olduğu belirtilmiştir (Hsu ve diğ. 2019).



Şekil 1. Hava Kalitesi Ölçüm İstasyonları ve Ulusal Hava İndeksi Verileri
(Ziyaret tarihi: 11.01.2022 11:00:00)
Figure 1. *Air Quality Measurement Stations and National Air Index Data*

Ulusal Hava Kalitesi İndeksi kapsamında kalite indeksi hesaplanan beş temel kirleticiden biri, ciddi solunum yolu hastalıklarına sebep olan atmosferik partikül maddelerdir. Çapı 10 mikrondan ve 2,5 mikrondan küçük olan ve sırasıyla PM_{10} ve $PM_{2,5}$ olarak adlandırılan bu partiküller, çok küçük ve hafif olduğundan uzun süre havada kalma eğilimindedir. Burun ve boğazdan rahatlıkla geçebilen bu partiküller ciğerlere nüfuz ederek ciddi solunum ve dolaşım sistemi rahatsızlıklarına neden olmaktadır (Kampa ve Castanas 2008). Literatürde atmosferik partikül maddeleri değerlendirmek ve tahmin etmek için pek çok çalışma yapılmıştır. Bozdağ ve diğ., (2020), 2009-2017 yılları arasında Ankara'da yer alan altı farklı istasyon verisiyle 2018 yılı için PM_{10} konsantrasyonunu farklı makine öğrenme algoritmalarıyla tahmin etmiş en iyi sonuç R^2 değeri 0,58 olarak yapay sinir ağları ile elde edilmiştir. Zickus ve diğ., (2002) PM_{10} konsantrasyonu tahmini için yağış, rüzgar hızı, nem ve bulut oranını değişkenlerini kullanmış ve farklı makine öğrenme algoritmalarını karşılaştırarak karar ağaçlarının performansını diğer yöntemlere göre yetersiz bulmuşlardır. Suleiman ve diğ., (2019) Londra'da 2007-2012 yılları arasında dokuz istasyondan elde ettikleri verilerle yapay sinir ağlarını kullanarak trafikten kaynaklı partikül madde konsantrasyonunu incelemiş ve araçların yakıt tüketimi ve türlerine göre farklı senaryoları karşılaştırmışlardır. Önerilen yöntem sonucunda, model tahminleri ile gözlem değerleri arasındaki korelasyonun 0,8 olduğu güçlü bir tahmin modeli elde edilmiştir. Chen ve diğ., (2018), meteoroloji ve arazi kullanım durumu bilgileriyle Çin'de 2005-2016 yılları arasındaki günlük $PM_{2,5}$ konsantrasyonunu tahmin etmiş ve günlük modele ilişkin R^2 değerini 0,83 olarak bulmuşlardır. Aynı model için zaman periyodunu ay ve yıl olarak incelediklerinde ise R^2 değerinin arttığını vurgulamışlardır. Görüldüğü gibi özellikle son yıllarda çevresel sürdürülebilirlik adına yapılan çalışmalar literatürde de karşılığını bulmuş ve Dünya'da farklı bölgelerde farklı hava kirleticilerine ilişkin tahmin çalışmaları yapılmış ve halen de yapılmaya devam etmektedir.

Türkiye'de hava kirleticilerine ait sınır değerler, AB standartlarına göre belirlenmekte olup $PM_{2,5}$ konsantrasyonu için bu değerin DSÖ tarafından belirlenen standardın iki katının üzerinde olduğu

bilinmektedir (Avrupa Çevre Ajansı, 2022). Ayrıca Sağlık ve Çevre Birliği (HEAL), Halk Sağlığı Uzmanları Derneği (HASUDER) ve Kocaeli Üniversitesi tarafından yürütülen Çevre İklim ve Sağlık için İş birliği Projesi (ÇİSİP) için yayınlanan bilgi notunda Türkiye genelinde hava kalitesi izleme istasyonlarının bazılarında donanım eksikliği nedeniyle $PM_{2,5}$ kirleticisinin izlenemediği belirtilmiştir (ÇİSİP Bilgi Notu, 2022). Tüm bu durumlar göz önünde bulundurularak hem $PM_{2,5}$ kirleticisine neden olan faktörlerin belirlenmesi hem de gerekli önlemlerin alınabilmesi için uzun vadede doğru tahminler sağlayan modellerin geliştirilmesi önem arz etmektedir.

Ayrıca anlık ölçümlerden ziyade gerçeğe yakın tahminlerin yapılmasıyla toplumda özellikle kalp ve solunum yolu hastalıkları açısından risk altında olan bireylerin önceden önlem alması da kolaylaşacaktır. Vücudun filtre sistemi tarafından süzülemediği için akciğer bariyerini geçerek kana karışabilen bu partiküllerin, kardiyovasküler ve solunum yolu hastalıkları ile kanserlere neden olduğu düşünüldüğünde yine başarılı tahminler üreten modellerin geliştirilmesi insan sağlığı açısından hayati bir önem taşımaktadır.

Tüm bunların bir sonucu olarak bu çalışmada $PM_{2,5}$ madde konsantrasyonu için trafik yoğunluğu, meteorolojik koşullar ve NOX, SO_2, PM_{10} hava kirleticileri konsantrasyonlarının kullanıldığı bir tahmin modeli geliştirilmiştir. Tahmin gücü yüksek bir modelin elde edilmesini sağlamak için üç farklı veri seti kullanılmıştır. Hava kalitesi ölçüm istasyonlarından kirletici konsantrasyonları; meteoroloji ölçüm istasyonlarından sıcaklık, nem, rüzgâr hızı ve yağış bilgileri; saatlik trafik yoğunluk istasyonlarından ise ortalama hız ve araç sayısı bilgileri çekilmiştir. Veri setlerinin alınmış olduğu bu istasyonlar farklı lokasyonlarda yer aldığından trafik durumu ve meteorolojik koşulların değerlendirilebilmesi için aynı bölgede yer alan ölçüm istasyonlarının mümkün olduğunca birbirine yakın konumlanmasına dikkat edilmiştir. Ayrıca eksik veriler açısından da analiz sonuçlarının etkilenmemesi adına kayıp veri oranının %10'u geçmemesine dikkat edilmiştir. Her iki koşulu da sağlaması nedeniyle pilot bölge olarak Beşiktaş Bölgesi seçilmiştir. Bunun yanı sıra trafik yoğunluğu açısından da Yıldız/Beşiktaş trafiğin en yoğun olduğu semtlerden birisi olması sebebiyle pilot bölge olarak seçilmeye uygun bulunmuştur.
Bilindiği kadarıyla Türkiye'de $PM_{2,5}$ konsantrasyonu tahmini için literatürde daha önce meteorolojik koşulların, trafik durumunun, zamanın ve hava kirleticilerinin eş zamanlı olarak ele alındığı başka bir çalışma bulunmamaktadır.

Tahmin modelleri geliştirilirken ÇDR, RO, DVM ve YSA yöntemleri kullanılmıştır. Geliştirilen modellerde test ve eğitim verisi ile çalışılıp modeller, determinasyon katsayısı (R^2), düzeltilmiş R^2, ortalama mutlak hata, ortalama hata karesi ve kök ortalama hata karesi performans ölçütlerine göre karşılaştırılmıştır.

Çalışmanın bundan sonraki kısmı şu şekildedir: Materyal ve Yöntem bölümünde makine öğrenme algoritmalarının genel işleyişi sırasıyla verilerin toplanması, veri ön işleme, modelin eğitilmesi, modelin değerlendirilmesi ve tahmin alt başlıklarıyla incelenmiştir. Sonuç ve Tartışma bölümünde ise genel bulgular ve geleceğe yönelik değerlendirmeler yapılmıştır.

MATERYAL ve YÖNTEM (MATERIAL and METHOD)

Makine Öğrenmesi (Machine Learning) bilgisayar sistemlerinin bir işlemi gerçekleştirmek için açıkça programlanmadan, istatistiksel modeller ve algoritmalar ile eğitildiği bir yapay zeka uygulamasıdır (Samuel, 1988). Kullanılan eğitim verileri aracılığıyla makine öğrenmesi algoritmaları verileri algılayarak yorumlayabilir ve bağımlı değişkene ilişkin tahminler yapabilir. Literatürde makine öğrenme algoritmalarının hem tahmin hem de sınıflandırma modellerinde sıklıkla kullanıldığı görülmektedir. Büyük verilerin sınıflandırılmasına biyotıp (Ünlü ve diğ., 2022), sosyal medya (Öztürk ve diğ., 2020), pazarlama (Kaynar ve diğ., 2017) vb. gibi çok çeşitli alanlarda ihtiyaç duyulmaktadır. Tahmin modelleri ise gelecekte yaşanabilecek belirsizliklere karşın etkin bir planlama aracı olarak oldukça geniş bir uygulama alanına sahiptir. Enerji (Demolli ve diğ., 2019, Kuşkapan ve diğ., 2022), çevre (Gültepe 2019, Çelik ve Arıcı 2021), meteoroloji (Başakın ve diğ., 2019), finans (Namlı ve diğ., 2019, Özmaden ve Erdal 2020), kestirimci bakım (Dündar ve diğ., 2021) vb. gibi alanlarda, makine öğrenme algoritmalarıyla tahmin gücü yüksek modeller geliştirilerek geleceğe yönelik kararlar etkin bir şekilde alınabilmektedir. Bu

bölümde makine öğrenme algoritmalarının çalışma adımları sırasıyla açıklanarak ilgili veri setine nasıl uygulandığı detaylı bir şekilde açıklanmıştır. Temel olarak makine öğrenme algoritmalarının uygulama adımları verilerin toplanması, veri önişleme, modelin seçilmesi, modelin eğitilmesi, modelin değerlendirilmesi ve tahmin olmak üzere altı adımdan oluşmaktadır. Yapılan çalışmada bu adımlar uygulanırken içerisinde veri bilimi için faydalı olan kütüphanelerin yer aldığı Anaconda Navigator programı aracılığı ile Jupyter Notebook ve Python programlama dili ile Pandas, Matplotlib, Scikit-Learn ve Seaborn kütüphaneleri kullanılmıştır.

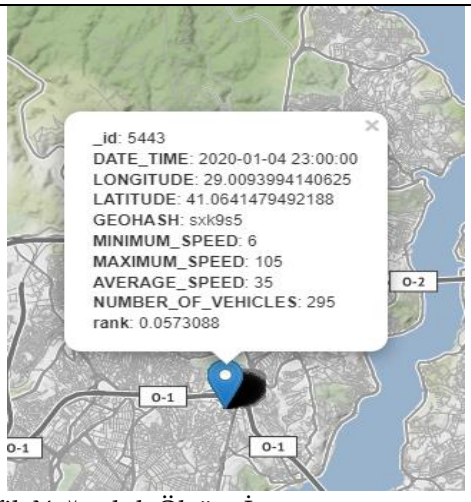
i.Verilerin Toplanması (Data Collection)

Makine öğrenmesinin ilk aşaması olan verilerin toplanması aşaması geliştirilen modelin doğru sonuçlar üretmesi açısından oldukça önemlidir. Verilerin miktarı, kalitesi ve probleme özgü veri seçimi, makine öğrenmesi algoritmalarının başarısı için kritik önem taşımaktadır. Verilerin toplanması aşamasında İstanbul Büyükşehir Belediyesi (2022) tarafından farklı alanlarda verilerin paylaşıldığı açık veri portalından yararlanılmış ve 01.01.2020-01.01.2021 dönemine ait Şekil 2'de konum bilgileri verilen istasyonlardan elde edilen veri setleri kullanılmıştır. Bu amaçla öncelikle hava kalitesi ölçüm istasyonları incelenmiştir (Sürekli İzleme Merkezi, 2022). Ölçümler İstanbul genelinde farklı bölgelerde yer alan 36 istasyon tarafından gerçekleştirilmektedir. Her bir istasyona ilişkin belirli bir dönem için saatlik, günlük, haftalık, aylık ve yıllık olmak üzere farklı periyotlarda veriler raporlanabilmektedir. Meteorolojik veriler için portalda yer alan Meteoroloji Gözlem İstasyonu Veri Seti (İBB Meteoroloji Gözlem İstasyonu Veri Seti, https://data.ibb.gov.tr/dataset/meteorology-observation-station-data-set, ziyaret tarihi: 11.01.2022), trafik yoğunluğuna ait veriler için ise Saatlik Trafik Yoğunluk Veri Seti (İBB Saatlik Trafik Yoğunluk Veri Seti, https://data.ibb.gov.tr/dataset/hourly-traffic-density-data-set, ziyaret tarihi: 11.01.2022) kullanılmıştır.

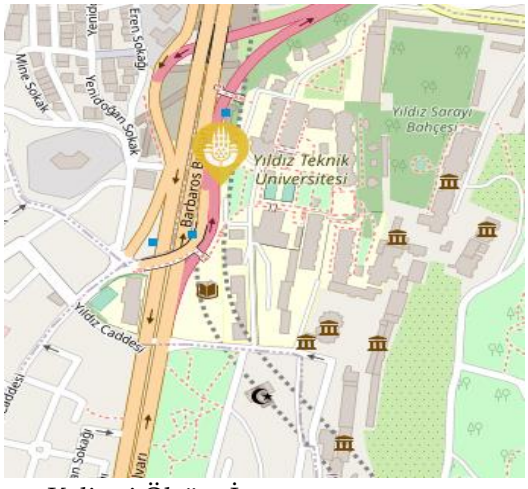
Veri setlerinin alınmış olduğu bu istasyonlar farklı lokasyonlarda yer aldığından trafik durumu ve meteorolojik koşulların değerlendirilebilmesi için aynı bölgede yer alan ölçüm istasyonlarının mümkün olduğunca birbirine yakın konumlanmasına dikkat edilerek pilot bölge olarak Beşiktaş bölgesi, veri periyotu olarak saatlik periyot seçilmiştir. Çizelge 1'de 01.01.2020- 31.12.2020 tarihleri arasında Beşiktaş Hava Kalitesi Ölçüm İstasyonu'ndan elde edilen saat bazındaki verilere (8784 adet) ait tanımlayıcı istatistikler verilmiştir. Her bir kirleticiye ait maksimum, minimum, ortalama ve standart sapma değerleri ile elde edilen veri yüzdeleri Çizelge 1'de görülmektedir. Buna göre olması gereken veri 1 yıl boyunca veri kaybı olmadığı durumda elde edilmesi gereken saatlik veri sayısını ifade ederken; gelen veri, ölçüm istasyonlarındaki problemlerden dolayı kayıp verilerin olduğu durumda elde edilen nihai veri sayısını ifade etmektedir. Bu durumda veri yüzdesi gelen verinin olması gereken veriye oranıyla hesaplanmaktadır.

ii.**Veri Ön İşleme (Data Preprocessing)**

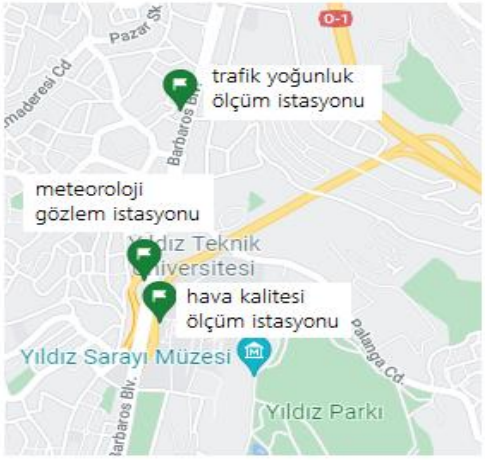
Bu adımda birinci adımda toplanan ham veri, modelde kullanılmak üzere işlenir. Ön işleme aşaması veri hazırlama ve veri indirgeme olmak üzere iki alt aşamadan oluşmaktadır. Ön işleme aşamasında süreci yönetmek için Çizelge 2'de verilen soru listesinden yararlanılır (García ve diğ., 2015).



Trafik Yoğunluk Ölçüm İstasyonu



Hava Kalitesi Ölçüm İstasyonu



_id	Sira No	Istasyon Adi	Istasyon Lokasyonu	Istasyon Koordinatlari
8	beus8	Beşiktaş-Yıldız	Yıldız	41º 3′ 12.67″ N, 29º 0′ 31.81″ E

Meteoroloji Gözlem İstasyonu

Şekil 2. Beşiktaş Bölgesi Ölçüm İstasyonları Konum Bilgileri

Figure 2. *Location Information of Measurement Stations in Beşiktaş*

Çizelge 1. Hava kirleticilerine ilişkin tanımlayıcı istatistikler

Table 1. *Descriptive Statistics for Air Pollutants* (μg/m³)

	PM_{10}	**$PM_{2,5}$**	**SO_2**	**NOX**
Maksimum	144,5	91,9	314,8	918,5
Minimum	0,6	0,4	0,2	5,4
Ortalama	26,4	18,8	3,1	101,9
Standart Sapma	16,8	10,8	4,3	82,6
Gelen Veri	8203	8229	8769	7969
Olması Gereken Veri	8784	8784	8784	8784
Veri Yüzdesi	%93,4	%93,7	%99,8	%90,7

Çizelge 2. Veri Ön İşleme Aşaması
Table 2. *Data Preparation Phase*

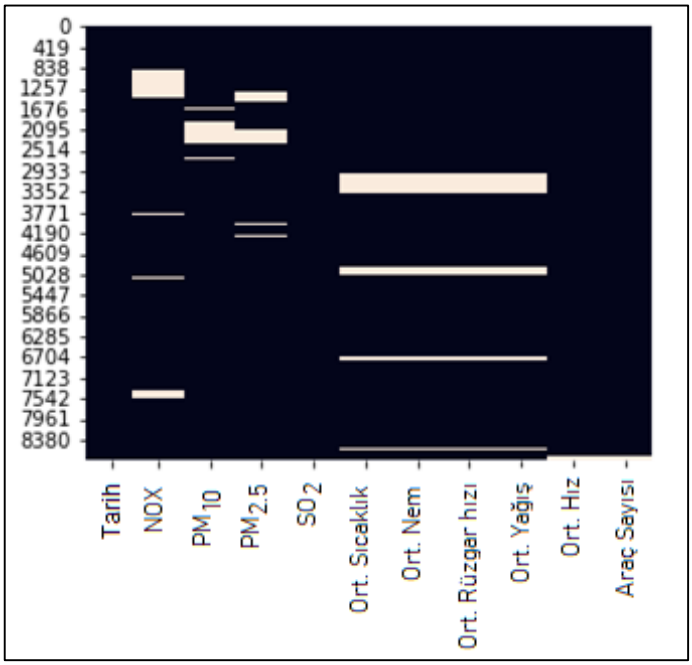
Veri Hazırlama	**Soru**
Veri Temizleme	Verileri nasıl temizlerim?
Veri Dönüşümü	Verileri belirli bir formata nasıl getirebilirim?
Veri Entegrasyonu	Verileri nasıl birleştiririm?
Veri Normalleştirme	Verileri nasıl ölçeklendiririm?
Eksik Veri Tahmini	Eksik verileri nasıl işleyebilirim?
Gürültü Tanımlama	Gürültüyü nasıl tespit eder ve yönetirim?
Veri İndirgeme	**Soru**
Öznitelik Seçimi	Veri boyutunu nasıl azaltabilirim?
Örnek Seçimi	Gereksiz ve/veya çelişkili örnekleri nasıl kaldırabilirim?
Ayrıklaştırma	Bir özniteliğin etki alanını nasıl basitleştirebilirim?
Öznitelik çıkarma	Ham veriyi özelliklerine göre indirgeyerek nasıl daha yönetilebilir gruplar oluşturabilirim?

- Verilerin temizlenmesi: Hatalı verilerin düzeltilmesi, yanlış verilerin filtrelenmesi ve gereksiz veri ayrıntılarının azaltılması işlemlerinden oluşur. Meteorolojik veri seti incelendiğinde ölçümün normal olarak yapılamadığı durumlarda bazı değerlerin "-99" olarak girildiği görülmüş ve bu değerler veri setinden temizlenerek yanlış veriler filtrelenmiştir.
- Verilerin dönüştürülmesi: Veriler, modelin ihtiyaç duyacağı uygun formatlara dönüştürülür. Her üç veri dosyası virgülle ayrılmış değer (.csv) uzantılı dosya formatına dönüştürülmüştür.
- Verilerin entegrasyonu: Farklı veri setlerinden elde edilen veriler birleştirilir. Bu aşamada hava kalitesi, meteorolojik durum ve trafik yoğunluğu için kullanılan üç veri dosyası, verinin elde edildiği tarih ortak sütun olacak şekilde birleştirilmiş ve veriler tarih sütununa göre sıralanmıştır. Pandas kütüphanesinde tarih formatı "yyyy-MM-dd hh:mm:ss" şeklinde kullanılmıştır. Formattaki ifadeler sırasıyla yıl, ay, gün, saat, dakika ve saniyeyi ifade etmektedir.
- Verilerin normalleştirilmesi: Bu aşamada tüm verilerin aynı ölçüm biriminde incelenmesi için ortak bir ölçek veya aralık kullanılır. Yapılan çalışmada incelenen parametre birimleri birbirinden farklı olduğundan verilerin standart forma dönüştürülmesinde Sklearn kütüphanesindeki MinMaxScaler modülü kullanılmıştır.

X_{min} ve X_{max} sırasıyla veri setindeki en küçük ve en büyük değeri; X' ise ölçeklendirilmiş veriyi ifade etmek üzere; Min-Max ölçekleme formülü Eşitlik (1)'de verilmiştir. Buna göre veri setindeki en küçük değer 0, en büyük değer 1 olacak şekilde veriler [0,1] aralığında ölçeklendirilmiştir.

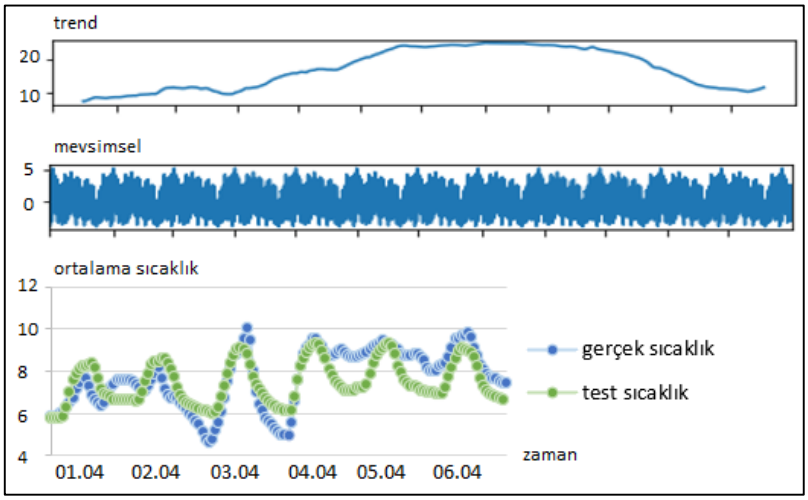
$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

- Eksik veri tahmini: Veri setlerinde sıklıkla karşılaşılan problemlerden birisi olan eksik verilerin görselleştirilmesi için farklı teknikler kullanılmaktadır. Bunlardan birisi ısı haritalarıdır. Şekil 3'te görüldüğü gibi açık renkli alanlar ilgili sütuna ait eksik verileri tasvir etmektedir. Schafer (1999), %5 ve altındaki eksik veri oranının ihmal edilebilir olduğunu; Bennett (2001) ise %10'dan fazla eksik verinin istatistiksel analizin yanlı olmasına sebep olacağını belirtmiştir. Bu değerlendirmeler göz önünde bulundurulduğunda eksik veri oranı çalışılan veri seti için ihmal edilebilir düzeydedir. Eksik veriler göz ardı edilerek ilgili satır silinebilir ancak bu şekilde veri yanlı hale gelerek kalitesiz sonuçlar üretilebilir (Little ve Rubin 2019). Eksik veriler incelendiğinde Şekil 3'te de görüldüğü gibi meteorolojik koşullara ait veriler için iki veri arasındaki boşluğun oldukça büyük olduğu görülmektedir. Bu boşluğun büyüklüğü kayıp veri sayısı ile ölçülmektedir. Ortalama hız ve araç sayısına ilişkin ise böyle bir durum söz konusu değildir. Bu nedenle eksik verilerin doldurulması aşamasında farklı stratejiler uygulanmıştır.



Şekil 3. Isı haritası
Figure 3. *Heat Map*

Meteorolojik veriler için zaman serisinin özelliklerini incelemek amacıyla ilk aşamada mevsimsel ayrıştırma yapılmıştır. Şekil 4'te görüldüğü gibi ortalama sıcaklık değerlerine ilişkin trend grafiği incelendiğinde verilerin durağan olmadığı ve mevsimsel etkilerin görüldüğü söylenebilir. (Box ve Jenkins 1970) tarafından durağan olmayan zaman serileri için ARIMA modeli geliştirilmiştir. ARIMA modeline mevsimselliğin eklenmesiyle SARIMA (Mevsimsel Otoregresif Entegre Hareketli Ortalama) süreci geliştirilmiştir (Hyndman ve Athanasopoulos 2018). Böylece meteorolojik özelliklere ilişkin verilerin doldurulmasında SARIMA yöntemi kullanılmıştır. SARIMA yöntemi özellikle mevsimsel davranışlar gibi daha gerçekçi dinamiklerin tahmine dahil edildiği bir yöntem olup yalnızca komşu değerlerden bilgi almakla kalmaz aynı zamanda zaman periyodu boyunca değişim düzenliliğini de kontrol eder (Li ve diğ., 2018). Literatürde eksik verilerin doldurulmasında SARIMA yönteminin kullanıldığı pek çok çalışma bulunmaktadır (Spławińska, 2015; Layanun ve diğ., 2017). $SARIMA(p,d,q)(P,D,Q)s$ şeklinde ifade edilen modelde p otoregresyon seviyesini, d fark alma seviyesini, q hareketli ortalamalar seviyesini ifade ederken; P, D ve Q sırasıyla p,d,q parametrelerinin sezonsal seviyesini göstermektedir. Mevsimsel periyotun 24 saat olarak alındığı $SARIMA(2,1,0)(1,0,1)24$ modelinde parametreler pilot koşumlara göre belirlenmiştir. Şekil 4'te 4-31 Mart tarihleri arasındaki dört haftalık sıcaklık değerleri eğitim verisi olarak kullanılarak 1-7 Nisan tarihleri arasındaki sıcaklık değerleri tahmin edilmiştir.



Şekil 4. SARIMA yöntemi ile eksik verilerin doldurulması
Figure 4. *Imputing missing data by SARIMA*

Ortalama hız ve araç sayısına ilişkin eksik verilerin doldurulmasında ise K-En Yakın Komşu Algoritması kullanılmıştır. Algoritmaya göre Öklid uzaklıklar aracılığıyla en yakın k adet eksik değer içermeyen komşu veri ile eksik değer arasındaki benzerlikler hesaplanarak eksik veri doldurulur (Pujianto ve diğ., 2019). Bu çalışmada K-En Yakın Komşu algoritması için Scikit-Learn (Sklearn) kütüphanesinden yararlanılmıştır. K-En Yakın Komşuluk algoritmasında komşu sayısı algoritma performansını etkileyen önemli bir parametredir. K değerinin çok düşük seçilmesi gürültü etkisini artırarak sonuçları daha az genellenebilir yaparken; çok yüksek seçilmesi yerel etkilerin önemi azalarak daha uzak tahminler yapılmasına sebep olmaktadır. Bu durum göz önünde bulundurularak çalışmada K değeri pilot koşumlara göre 5 olarak seçilmiştir.

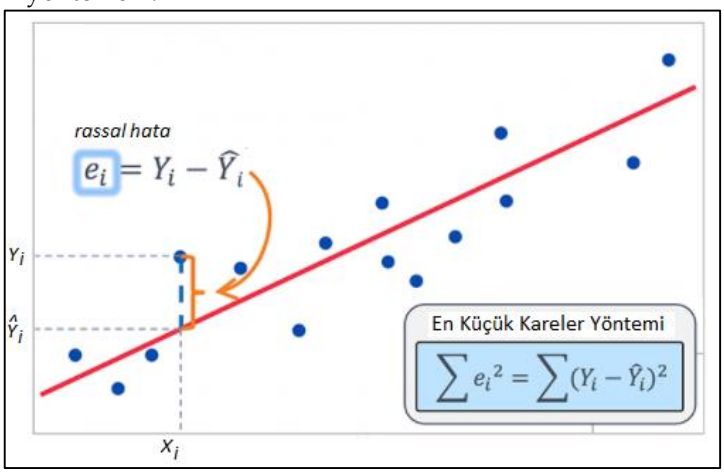
- Öznitelik seçimi: Bu aşama modelin eğitim süresini etkileyeceğinden veri seti içerisindeki özniteliklerin yeterli sayıda ve doğru olarak belirlenmesi gerekir. Yapılan çalışmada veri seti içerisindeki gereksiz öznitelikler uzaklaştırılmıştır. Literatürde filtreleme ve sarmal yöntem olarak adlandırılan iki temel öznitelik seçim yöntemi vardır. Filtreleme yöntemlerinde istatistiksel analizlere başvurularak anlamlılık düzeylerine göre öznitelikler belirlenirken, sarmal yöntemlerde, iteratif bir arama süreci gerçekleştirilerek farklı alt kümeler ile denemeler yapılır ve en iyi öznitelik kümesi seçilir. Bu çalışmada filtreleme yöntemlerinden "korelasyon tabanlı öznitelik seçim yöntemi" kullanılmıştır. Yöntemin temel mantığı iyi bir öznitelik alt kümesindeki özelliklerin hedef değişkenle yüksek korelasyona sahip olması ancak mümkün olduğunca birbiriyle ilişkili olmamasıdır (Hall 1999). Buna göre her özellik ayrı ayrı değerlendirilerek önem düzeyine göre hangi özelliklerin nihai özellik alt kümesine dahil edilmesi gerektiğine karar verilmiştir(Koprinska ve diğ., 2015). Meteorolojik veri seti içerisindeki rüzgar yönü, hissedilen sıcaklık ve yol sıcaklığı parametrelerine ilişkin değerler; trafik yoğunluğu veri setinde ise minimum ve maksimum hız parametrelerine ilişkin değerler modelden uzaklaştırılmıştır.
- Örnek seçimi: Bu aşamada orijinal veri seti içerisinden gereksiz ve/veya çelişkili veriler kaldırılarak veri seti yönetilebilir bir boyuta indirgenir.
- Ayrıklaştırma: Bu süreçte sürekli değişkenler kesikli özniteliklere dönüştürülür. Yapılan çalışmada verinin alındığı tarihin etkisini gözlemlemek amacıyla tarih değişkeni ay gün ve saat olacak şekilde kategorik niteliklere dönüştürülmüştür.

iii.Modelin Seçilmesi (Model Selection)

Makine öğrenmesinde kullanılan algoritmaların performansı her problem için aynı olmayıp veri boyutu, veri özellikleri ve problemin yapısına göre farklılıklar göstermektedir. Bu bölümde çalışma kapsamında kullanılan yöntemler ayrıntılı olarak açıklanmıştır.

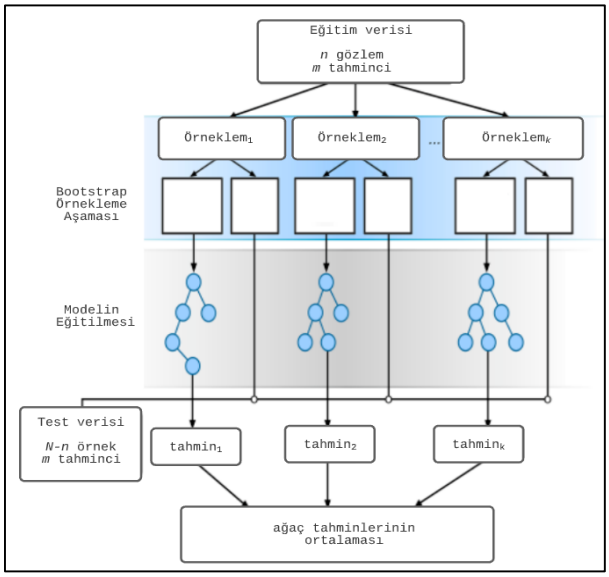
- Çoklu Doğrusal Regresyon: Bağımlı ve bağımsız değişkenler arasındaki matematiksel ilişkiyi

incelemek için kullanılan en basit makine öğrenme tekniğidir. Yapılan çalışmada ÇDR için Sıradan En Küçük Kareler (SEKK) yöntemi kullanılmıştır. Şekil 5'te görüldüğü gibi SEKK yöntemi hata kareleri toplamını yani tahmin değerleri ile hedef değerler arasındaki uzaklıkların kareleri toplamını en küçük yapmayı amaçlayan bir yöntemdir.



Şekil 5. Sıradan En Küçük Kareler Yöntemi
Figure 5. *Ordinary Least Squares (OLS) Method*

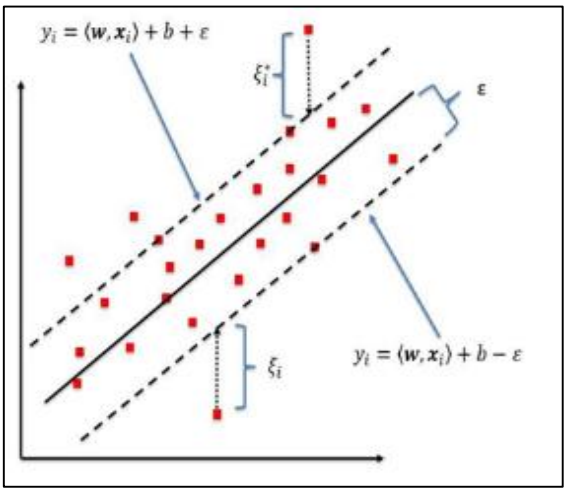
- Rassal Orman: Breiman (2001)'in "bagging" fikrinin bir uzantısı olarak geliştirilmiştir. Birden fazla karar ağacının birlikte çalıştığı bir kolektif öğrenme algoritmasıdır. Hem veri seti hem de öznitelik setinden rassal olarak seçilen alt kümelerle modeller eğitilir (Cutler ve diğ., 2012). Şekil 6'da görüldüğü gibi birbiri arasında korelasyon olmayan modeller birlikte çalışarak sonuçlar birleştirilir. Düğümde bölünme yapılacak olan öznitelik kümesinin rassal olarak seçilmesi, ağaçlar arasındaki korelasyonun düşük olmasını sağlamaktadır. Hem regresyon hem de sınıflandırma problemlerine uygulanabilir olması, diğer yöntemlere göre daha hızlı eğitilmesi ve tahmin hızının daha yüksek olması nedeniyle literatürde sıklıkla kullanılmaktadır.



Şekil 6. Rassal Orman Algoritması (Rodriguez-Galiano ve diğ.,, 2016)
Figure 6. *The Random Forest Algorithm*

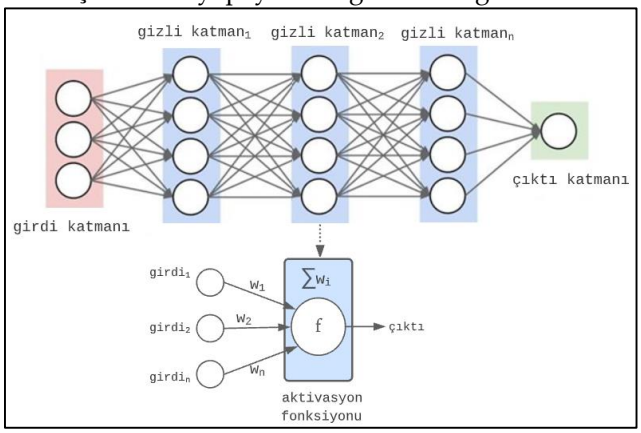
- Destek Vektör Makineleri: Genellikle sınıflandırma problemlerinde kullanılmaktadır. Regresyon problemlerinde kullanılan versiyonu Destek Vektör Regresyonu olarak adlandırılmaktadır (Smola ve Schölkopf, 2004). Yöntemde veriler bir düzlemle iki bölüme ayrılmakta olup amaç oluşturulan bu düzlemin iki sınıfa da mümkün olduğunca eşit uzaklıkta olmasını sağlamaktır. Düzlemin genişliği Şekil

7'de görüldüğü gibi ε parametresiyle belirlenmektedir.



Şekil 7. Destek Vektör Regresyonu (Parbat ve Chakraborty, 2020)
Figure 7. *Support vector regression*

- Yapay Sinir Ağları: İnsan beyninde öğrenme, nöronlar arasındaki sinaptik ilişki ile gerçekleşmektedir. YSA'nı oluşturan yapay sinir hücreleri çeşitli katmanlar içerisinde paralel bağlantılar kurarak beynin bir işlevi yerine getirme yöntemini taklit etmektedir. Bir YSA, bir girdi katmanı, bir çıktı katmanı ve en az bir gizli katmandan oluşmakta olup bu katmanlar birbirinden bağımsızdır yani belirli bir katman herhangi bir sayıda düğüme sahip olabilir (Abiodun ve diğ., 2018). Girdi katmanında bilgiler alınıp gizli katmana iletilir, gizli katmanda ise gelen bilgiler işlenerek çıktı katmanına aktarılır. Sinir ağlarında kaç tane gizli katman kullanılacağı ve her bir gizli katmanda kaç nöron olacağı problem yapısına göre değişiklik göstermektedir. Şekil 8'de yapay sinir ağlarına ait genel mimari verilmiştir.



Şekil 8. Yapay Sinir Ağı Mimarisi
Figure 8. *The Architecture of Artificial Neural Network*

Yapılan çalışmada yapay sinir ağı modelini geliştirmek için bir sinir ağı kütüphanesi olan Keras Tensor Flow kullanılmıştır.

Çalışmada kullanılan yöntemlere ilişkin hiperparametre değerleri ise Çizelge 3'te verilmiştir. Hiperparametreler kullanılan yöntemlerin performansını önemli düzeyde etkilediğinden literatürde hiperparametre optimizasyonu alanında pek çok çalışmanın yapıldığı görülmüştür (Edali ve Yucel 2018, Feurer ve Hutter 2019). Bu çalışma kapsamına hiperparametreler için bir optimizasyon çalışması dahil edilmemiş ancak pilot koşumlarla parametrelerin farklı seviyeleri değerlendirilerek Çizelge 3'teki parametreler belirlenmiştir.

Çizelge 3. Algoritmaların Hiperparametre Değerleri
Table 3. *Hyperparameters of Algorithms*

Algoritma	Hiperparametre	Değer
Rassal Orman Algoritması	Ağaç sayısı	100
	Maksimum özellik	Auto
	Maksimum derinlik	None
Destek Vektör Regresyonu	C (Ceza parametresi)	1
	ε (Sapma için eşik değer)	0,1
	γ (Yayılma parametresi)	0,1
	Kernel fonksiyonu	"rbf" (Radyal tabanlı çekirdek fonksiyonu)
Yapay Sinir Ağları	Küme boyutu	10
	Gizli katman sayısı	2
	Gizli katmandaki düğüm sayısı	11
	Girdi katmandaki düğüm sayısı	15
	Çıktı katmandaki düğüm sayısı	1
	Aktivasyon fonksiyonu	ReLU (Doğrultulmuş doğrusal birimler)
	Kayıp fonksiyonu	MSE (Ortalama hata karesi)
	Optimizer	"Adam" (Adaptif moment tahmini)
	Epochs	100

Çizelge 3'te görüldüğü gibi, RO' da ormandaki ağaç sayısı 100 olarak seçilmiştir. Bir düğümü ayırırken dikkate alınacak maksimum özellik sayısı Sklearn kütüphanesinde "Auto" olarak belirtilen modeldeki özellik sayısı kadar olacak şekilde belirlenmiştir. Düğümlerin bölünmesini durdurmak için kullanılan maksimum derinlik için Sklearn kütüphanesinin varsayılan değeri olan "None" seçeneği kullanılarak bir sınır değeri atanmamıştır. Bu durumda düğümler tüm yapraklar saf olana kadar genişletilmektedir. ÇDR' de eğitim setindeki tüm veriler dikkate alınırken, DVM' de dikkate alınacak verilerin oluşturulacak regresyon doğrusundan en az ε uzaklıkta olması sağlanmaktadır. Düzenlileştirme parametresi C ise hataların uyarlanması ile regresyon fonksiyonunun düzlüğü arasındaki dengeyi sağlayan bir parametredir. Bu çalışmada ε değeri 0,1, C değeri ise 1 olarak alınmıştır. Kernel fonksiyonu olarak ise radyal tabanlı fonksiyon (rbf) kullanılmıştır. Yapay sinir ağı iki gizli katmandan oluşmakta olup, küme boyutu ve Epochs sırasıyla 10 ve 100 olarak belirlenmiştir. Optimize edici algoritma olarak Kingma ve Ba (2014) tarafından geliştirilen ve performansından dolayı literatürde de oldukça sık kullanılan "Adam" algoritması kullanılmıştır. YSA'da doğrusal olmayan durumların yorumlanabilmesi için yapılan işlemleri doğrusal olandan doğrusal olmayan yapıya dönüştürülmesini sağlayan aktivasyon fonksiyonu olarak "relu" kullanılmıştır. Kayıp fonksiyonu olarak ise tahmin modellerinde sıklıkla kullanılan fonksiyon olan ortalama hata karesi kullanılmıştır. Gizli katmandaki nöron sayısı belirlenirken ise Karsoliya (2012) tarafından önerilen durumlar dikkate alınmıştır. Buna göre gizli katmandaki nöronlarının sayısı, girdi katmanı boyutu ile çıktı katmanı boyutu arasında ve girdi katmanının yaklaşık 2/3'ü olmalıdır. Bunun sonucunda girdi katmanı 15, gizli katman 11 ve çıktı katmanı ise 1 nörondan oluşacak şekilde katmanlar düzenlenmiştir.

iv.Modelin Eğitilmesi (Model Training)

Modelin eğitilmesi aşaması makine öğrenmesindeki en önemli adımdır. Bu aşamada modelin kalıplar bularak tahminler yapması için veriler modele iletilir. Zaman ilerledikçe ve model eğitildikçe daha iyi hale gelir. Yapılan çalışmada her üç yöntem için verilerin %80'i modelin eğitilmesi için kalan %20'si ise test için kullanılmıştır. Bu oran literatürde de sıklıkla kullanılmaktadır (Joseph 2022).

v.Modelin Değerlendirilmesi (Model Evaluation)

Geliştirilen modeller determinasyon katsayısı (R^2), düzeltilmiş R^2, ortalama mutlak hata, ortalama hata karesi ve ortalama hata karesi kökü performans ölçütlerine göre karşılaştırılmıştır.

- Determinasyon Katsayısı (R^2): Korelasyon katsayısının karesi olan determinasyon katsayısı, bağımlı değişkendeki değişkenliğin ne kadarının bağımsız değişkenlerdeki değişkenlikle açıklanabildiğini ifade eder. İki değişken arasındaki doğrusal ilişkinin gücünü belirleyen determinasyon katsayısı 0-1 aralığında değer almakta olup ilişkinin gücü 1'e yaklaştıkça artmaktadır. R^2'nin 1'e eşit olması regresyon tahminlerinin gerçekleşen değerlere tam olarak uyduğu anlamına gelmektedir.
- Düzeltilmiş R^2: Modele yeni bir bağımsız değişken eklendiğinde R^2 değeri genellikle artma eğiliminde olduğundan, modelde birden fazla bağımsız değişkenin olduğu durumda tek başına R^2 değerini yorumlamak yeterli değildir. Böylece gereksiz eklenen bağımsız değişkenler cezalandırılarak düzeltilmiş R^2 formülü oluşturulmuştur.
- Ortalama Mutlak Hata: Model tahmini ile hedef değer arasındaki mutlak farkın ortalamasını verir.
- Ortalama Hata Karesi: Model tahmini ile hedef değer arasındaki farkların karelerinin ortalamasını verir.
- Kök Ortalama Hata Karesi: Tahmin hatalarının standart sapmasıdır.

Çizelge 4 ve Çizelge 5'te sırasıyla performans ölçütlerinde kullanılan notasyonlar ve formüller verilmiştir.

Modeller oluşturulurken, bağımsız değişken sayısı çok fazla olduğundan tüm kombinasyonların denenmesi oldukça zaman alıcı bir süreçtir. Mitchell ve Beauchamp (1988) mümkün olan en az tahmin edici ile yorumlanabilir bir modele sahip olmak için değişken seçiminin yapılması gerektiğini vurgulamıştır. Yanıt ve tahmin ediciler arasındaki ilişkiyi olabildiğince basit bir şekilde ifade etmek ve tahmin maliyetini azaltmak amacıyla değişken seçimi yapılarak Çizelge 6'da görüldüğü gibi yedi farklı model oluşturulmuştur. Öncelikle meteorolojik değişkenler, trafik yoğunluğu ve tarih değişkenleri için literatürde de sıklıkla kullanılan Aşamalı (Step-Wise) değişken seçim prosedürü uygulanarak Model1, Model2 ve Model3 oluşturulmuştur. Ardından meteorolojik değişkenler, trafik yoğunluğu ve tarih değişkenlerinin birlikte kullanıldığı Model3'e sırasıyla NOX, SO2 ve PM10 değişkenleri dahil edilerek Model4, Model5 ve Model6 oluşturulmuştur. Kirleticilerin bireysel olarak modele dahil edilmesinin nedeni bağımsız değişkenlerin mümkün olduğunca birbiriyle ilişkili olmaması ilkesini sağlayarak model yorumlanabilirliğini artırmaktır. Son olarak PM10 kirleticisi, NOX ve SO2'ye göre daha iyi bir tahminci olduğu için PM10'un bireysel olarak PM2,5 oluşumuna etkisini değerlendirmek amacıyla Model7 oluşturulmuştur.

Çizelge 4: Performans Ölçütlerinde Kullanılan Notasyonlar

Table 4. *Notations Used in Performance Criterion*

Notasyon	Açıklama
y_i:	i. gözlemin hedef değeri
$\hat{y}_i$:	i. gözlemin tahmin edilen değeri
$\bar{y}$:	Hedef değerlerin ortalaması
p:	Bağımsız değişken sayısı
n:	Örneklem büyüklüğü
$e_i = y_i - \hat{y}_i$	i. gözleme ilişkin hata

Çizelge 5. Performans Ölçütleri
Table 5. *Performance Criterion*

Performans Ölçütü	Formül
R^2	$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
Düzeltilmiş R^2	$= 1 - \frac{(1 - R^2)\,(n-1)}{n-p-1}$
MAE	$= \frac{1}{n}\sum_{i=1}^{n}\lvert e_i \rvert$
MSE	$= \frac{1}{n}\sum_{i=1}^{n} e_i^2$
RMSE	$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$

Çizelge 6. $PM_{2,5}$ Konsantrasyonu için bağımsız değişkenler
Table 6. *Independent Variables for* $PM_{2,5}$

Model	Bağımsız Değişkenler
$Model_1$	Ortalama sıcaklık, ortalama nem, ortalama rüzgar hızı, ortalama yağış miktarı
$Model_2$	$Model_1$ + ortalama hız, araç sayısı
$Model_3$	$Model_2$ + saat, gün, ay
$Model_4$	$Model_3$ + NOX konsantrasyonu
$Model_5$	$Model_3$ +SO_2 konsantrasyonu
$Model_6$	$Model_3$ + PM_{10} konsantrasyonu
$Model_7$	PM_{10}

Çizelge 7 ve Çizelge 8'de sırasıyla bir yıllık ve üç aylık modellere ilişkin performans ölçütleri raporlanmıştır. $Model_1$ için meteorolojik parametreler incelendiğinde ortalama yağış miktarı istatistiksel olarak anlamlı bulunmazken en fazla katsayı ortalama rüzgar hızına ait olup rüzgar hızı arttıkça $PM_{2,5}$ konsantrasyonu azalmıştır. Trafik yoğunluğunun meteorolojik parametrelerin incelendiği modele dahil edilmesiyle modelin R^2 değerinde bir artış gözlenmemiş olup $PM_{2,5}$ konsantrasyondaki değişimin sadece meteorolojik koşullar ve trafik yoğunluğu parametreleriyle açıklanamadığı görülmüştür. Kategorik veriler, sonlu bir seçim kümesinden bir veya daha fazla ögeyi temsil eden girdi özelliklerini ifade etmektedir. Verilerin elde edildiği saat, gün ve ay parametrelerinin kategorik olarak modele dahil edilmesiyle R^2 değerinin arttığı gözlenmiştir. Son olarak $PM_{2,5}$ ile ilişkili olduğu düşünülen NOX, SO_2 ve PM_{10} konsantrasyonları incelenmiş ve en yüksek R^2 değeri meteorolojik parametreler, trafik yoğunluğu, tarih ve PM_{10} konsantrasyonunun bağımsız değişken olarak kullanıldığı $Model_6$ için elde edilmiştir. Böylece performans ölçütleri karşılaştırıldığında gerçeğe en uygun model $Model_6$ olarak bulunmuştur. Doğrusal regresyon modelinin sonuçları incelendiğinde, $Model_6$ için $PM_{2,5}$ değerinin; sıcaklık, nem, yağış miktarı, ortalama hız, araç sayısı ve PM_{10} konsantrasyonu arttıkça arttığı görülmüştür. Ortalama rüzgar hızı, saat ve ay değişkenleri ise $PM_{2,5}$ ile negatif yönlü ilişkiye sahiptir. Çizelge 7'de verilen algoritma performansları değerlendirildiğinde özellikle $Model_3$ için RO diğer yöntemlere göre büyük üstünlük sağlamıştır. En yüksek R^2 değerine sahip $Model_6$ için ise algoritma performanslarının sıralaması, RO, DVM, YSA ve ÇDR olarak bulunmuştur.

Çizelge 7. Geliştirilen modellerin karşılaştırılması (01.01.2020- 31.12.2020)

Table 7. *Comparison of developed models (01.01.2020- 31.12.2020)*

	Performans Ölç.	Model$_1$	Model$_2$	Model$_3$	Model$_4$	Model$_5$	Model$_6$	Model$_7$
Doğrusal Regresyon	R^2	0,047	0,044	0,094	0,214	0,101	**0,504**	0,480
	Düzeltilmiş R^2	0,047	0,043	0,093	0,213	0,100	**0,503**	0,480
	MAE	0,088	0,085	0,087	0,074	0,080	**0,064**	0,068
	MSE	0,014	0,013	0,014	0,009	0,023	**0,007**	0,008
	RMSE	0,118	0,114	0,119	0,099	0,152	**0,087**	0,093

	Performans Ölç.	Model$_1$	Model$_2$	Model$_3$	Model$_4$	Model$_5$	Model$_6$	Model$_7$
Rassal Orman Algoritması	R^2	0,137	0,173	0,644	0,591	0,657	**0,675**	0,395
	Düzeltilmiş R^2	0,137	0,173	0,643	0,591	0,657	**0,674**	0,395
	MAE	0,075	0,074	0,048	0,052	0,048	**0,047**	0,064
	MSE	0,011	0,010	0,004	0,005	0,004	**0,004**	0,007
	RMSE	0,105	0,103	0,067	0,072	0,066	**0,064**	0,088

	Performans Ölç.	Model$_1$	Model$_2$	Model$_3$	Model$_4$	Model$_5$	Model$_6$	Model$_7$
Destek Vektör Regresyonu	R^2	0,125	0,133	0,325	0,454	0,331	**0,638**	0,488
	Düzeltilmiş R^2	0,125	0,132	0,324	0,453	0,330	**0,637**	0,488
	MAE	0,077	0,078	0,069	0,064	0,069	**0,054**	0,062
	MSE	0,011	0,011	0,008	0,006	0,008	**0,005**	0,006
	RMSE	0,106	0,103	0,092	0,084	0,092	**0,068**	0,080

	Performans Ölç.	Model$_1$	Model$_2$	Model$_3$	Model$_4$	Model$_5$	Model$_6$	Model$_7$
Yapay Sinir Ağları	R^2	0,125	0,115	0,291	0,431	0,341	**0,623**	0,500
	Düzeltilmiş R^2	0,125	0,115	0,290	0,430	0,341	**0,622**	0,500
	MAE	0,078	0,078	0,072	0,064	0,066	**0,052**	0,059
	MSE	0,011	0,011	0,009	0,007	0,008	**0,005**	0,006
	RMSE	0,106	0,106	0,095	0,085	0,091	**0,069**	0,079

Bunun dışında veri boyutu azaltılarak 01.01.2020- 31.03.2020 tarihleri arasındaki üç aylık periyot incelenmiş ve sonuçlar Çizelge 8'de verilmiştir. Elde edilen sonuçlara göre 01.01.2020- 31.03.2020 tarihleri arasında Model$_6$ için R^2 değeri DVM ile 0,769; RO ile de 0,765 olarak elde edilmiş olup bu değer $PM_{2,5}$ konsantrasyonundaki değişimin yaklaşık olarak %76'sının modeldeki değişkenlerle açıklanabildiği anlamına gelmektedir.

Çizelge 8. Geliştirilen modellerin karşılaştırılması (01.01.2020- 31.03.2020)
Table 8. *Comparison of developed models (01.01.2020- 31.03.2020)*

	Performans Ölç.	$Model_1$	$Model_2$	$Model_3$	$Model_4$	$Model_5$	$Model_6$	$Model_7$
Doğrusal Regresyon	R^2	0,095	0,096	0,236	0,329	0,336	**0,649**	0,556
	Düzeltilmiş R^2	0,093	0,093	0,232	0,325	0,332	**0,647**	0,556
	MAE	0,109	0,103	0,094	0,090	0,089	**0,069**	0,075
	MSE	0,019	0,019	0,017	0,016	0,016	**0,008**	0,011
	RMSE	0,138	0,139	0,131	0,127	0,126	**0,092**	0,108

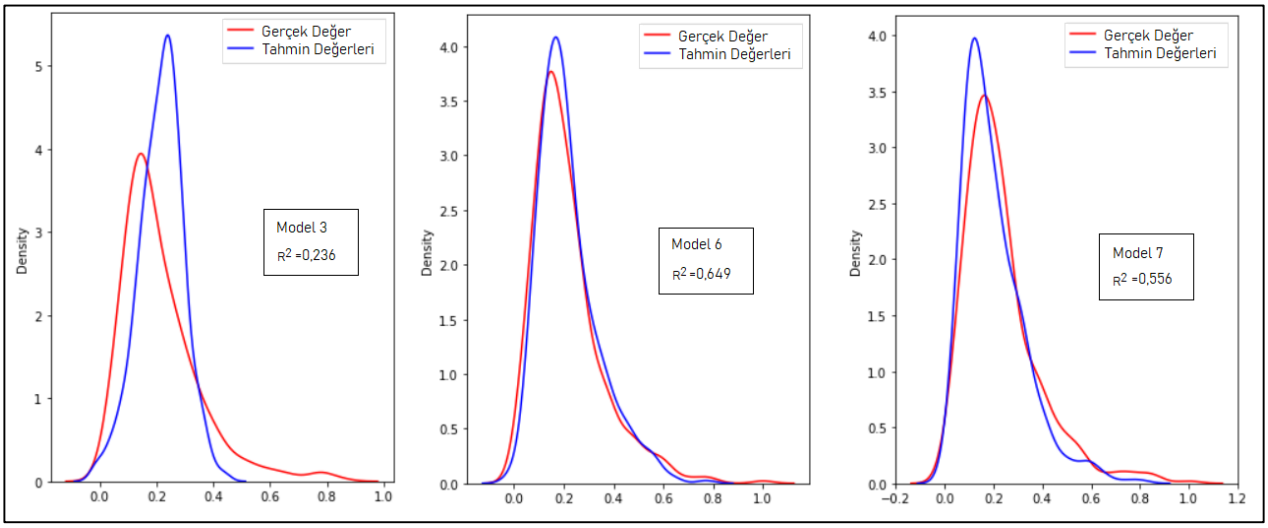
	Performans Ölç.	$Model_1$	$Model_2$	$Model_3$	$Model_4$	$Model_5$	$Model_6$	$Model_7$
Rassal Orman Algoritması	R^2	0,217	0,163	0,575	0,672	0,740	**0,765**	0,428
	Düzeltilmiş R^2	0,216	0,160	0,573	0,670	0,739	**0,764**	0,428
	MAE	0,095	0,099	0,065	0,060	0,052	**0,049**	0,077
	MSE	0,018	0,019	0,009	0,007	0,006	**0,005**	0,013
	RMSE	0,134	0,139	0,099	0,087	0,077	**0,074**	0,115

	Performans Ölç.	$Model_1$	$Model_2$	$Model_3$	$Model_4$	$Model_5$	$Model_6$	$Model_7$
Destek Vektör Regresyonu	R^2	0,177	0,186	0,434	0,598	0,596	**0,769**	0,575
	Düzeltilmiş R^2	0,175	0,184	0,432	0,596	0,594	**0,768**	0,575
	MAE	0,100	0,098	0,082	0,072	0,070	**0,057**	0,072
	MSE	0,019	0,019	0,013	0,009	0,009	**0,005**	0,009
	RMSE	0,138	0,137	0,114	0,096	0,097	**0,073**	0,099

	Performans Ölç.	$Model_1$	$Model_2$	$Model_3$	$Model_4$	$Model_5$	$Model_6$	$Model_7$
Yapay Sinir Ağları	R^2	0,192	0,180	0,457	0,580	0,542	**0,756**	0,585
	Düzeltilmiş R^2	0,190	0,178	0,454	0,578	0,540	**0,755**	0,585
	MAE	0,104	0,103	0,078	0,070	0,076	**0,053**	0,071
	MSE	0,018	0,019	0,013	0,009	0,010	**0,005**	0,009
	RMSE	0,137	0,138	0,112	0,098	0,103	**0,076**	0,098

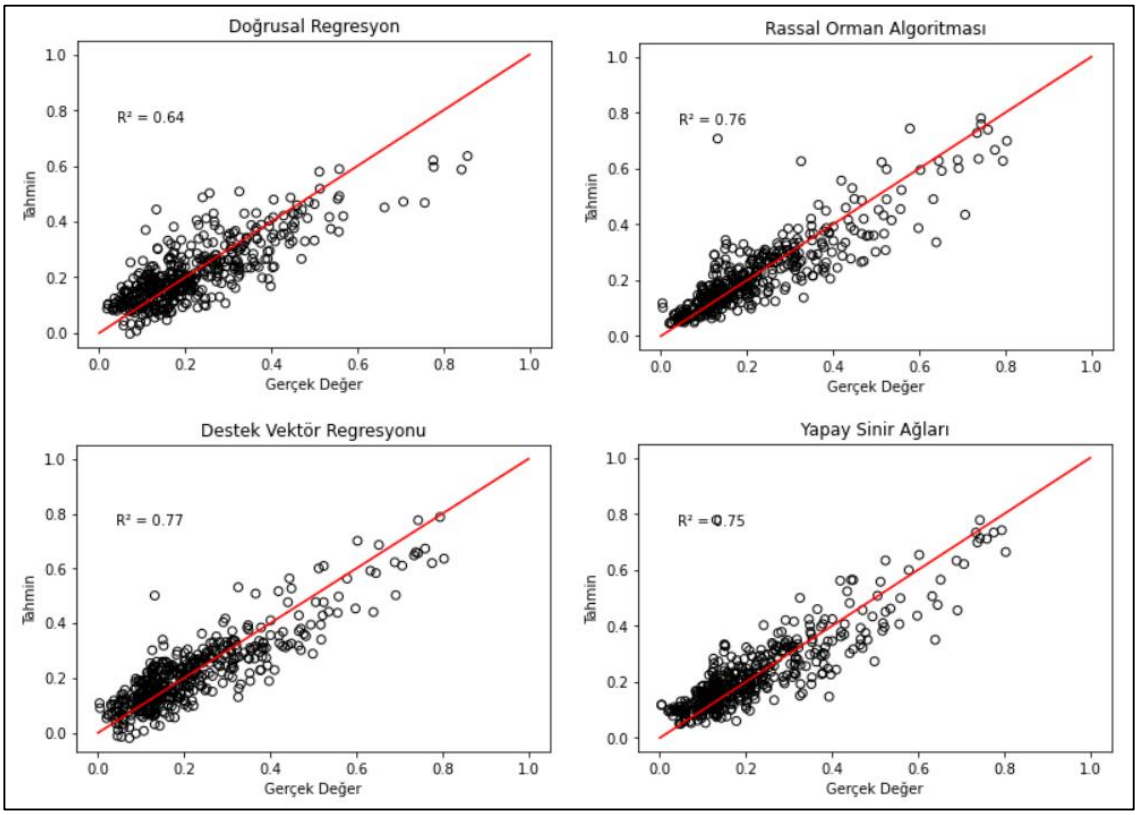
vi.Tahmin **(Prediction)**

Çok değişkenli modeller için modelin uygunluğuna bakmanın bir yolu dağılım grafiğini değerlendirmektir. Bu şekilde modelin ürettiği tahmin değerlerinin dağılımı ile gerçek değerlerin dağılımı karşılaştırılır (Sadriev ve Kamanev, 2020). Şekil 9'da Çizelge 7'de R^2 değerleri verilen doğrusal regresyona ait $Model_3$, $Model_6$ ve $Model_7$ için dağılım grafikleri verilmiştir. Kırmızı eğri gerçek dağılımı ifade ederken mavi eğri modelin üretmiş olduğu tahmin değerlerinin dağılımını göstermektedir. Şekil 9 incelendiğinde meteoroloji, trafik yoğunluğu ve tarih parametrelerinin incelendiği $Model_3$'te tahmin değerlerinin gerçek değerlerle çok örtüşmediği düşük R^2 değerini de açıklamaktadır. PM_{10} konsantrasyonunun modele eklenmesiyle oluşan $Model_6$'da ise iki eğri birbiriyle oldukça örtüşmektedir.



Şekil 9. Dağılım Grafikleri
Figure 9. *Distribution Plots*

Şekil 10'da ise kullanılan yöntemlerin performanslarını değerlendirmek amacıyla her bir algoritma için x ekseninde gerçek, y ekseninde ise tahmin değerlerinin olduğu grafik verilmiştir. Buna göre DVM, RO ve YSA yöntemlerinin performansları birbirine oldukça yakın bulunurken, ÇDR yöntemi diğer yöntemlere göre daha başarısız bulunmuştur.



Şekil 10. Tahmin Değerleri ve Gerçek Değerlerin Karşılaştırılması
Figure 10. *Comparison of Real and Predicted Values*

SONUÇ ve TARTIŞMA (RESULT and DISCUSSIONS)

Yapılan çalışmada İstanbul Beşiktaş Bölgesindeki atmosferik partikül madde $PM_{2.5}$ konsantrasyonu farklı makine öğrenme algoritmaları ile tahmin edilmiştir. Makine öğrenmesinin altı aşaması bireysel olarak ele alınmış ve her aşamada kullanılan yöntemler detaylı bir şekilde açıklanmıştır. İBB tarafından 01.01.2020- 31.12.2020 tarihleri arasında saat bazında paylaşılan verilerin (8784 adet) %80'i modelin eğitilmesinde %20'si ise modeli test etmek için kullanılmıştır.

Çalışmanın ilk aşamasında bağımsız değişkenler açısından öznitelik seçim sürecine bağlı olarak yedi farklı tahmin modeli oluşturulmuştur. Bu çalışmada performans ölçütleri incelendiğinde, $PM_{2,5}$ konsantrasyondaki değişimin sadece meteorolojik koşullar ve trafik yoğunluğu parametreleriyle açıklanamadığı görülmüştür. Bu nedenle çalışma kapsamına zaman ve hava kirleticilerinin etkisi de dahil edilmiştir. Bilindiği kadarıyla Türkiye'de $PM_{2,5}$ konsantrasyonu tahmini için literatürde daha önce meteorolojik koşulların, trafik durumunun, zamanın ve hava kirleticilerinin eş zamanlı olarak ele alındığı başka bir çalışma bulunmamaktadır. Verilerin elde edildiği saat, gün ve ay parametrelerinin kategorik olarak modele dahil edilmesiyle R^2 değerinin arttığı gözlenmiştir. Son olarak $PM_{2,5}$ ile ilişkili olduğu düşünülen NOX, SO_2 ve PM_{10} konsantrasyonları incelenmiş ve en yüksek R^2 değeri meteorolojik parametreler, trafik yoğunluğu, tarih ve PM_{10} konsantrasyonunun bağımsız değişken olarak kullanıldığı model için elde edilmiştir. Kullanılan makine öğrenme algoritmaları değerlendirildiğinde ise performans ölçütlerinin modellere göre değişiklik gösterdiği görülmüş ancak en iyi performans ortalamasına sahip teknik RO, en kötü performans ortalamasına sahip teknik ise ÇDR olarak bulunmuştur. Bunun dışında veri boyutu azaltılarak yöntemlerin performansları aynı modeller üzerinde yeniden denenmiştir. Veri boyutunun azalmasıyla algoritma performansları genel olarak artma eğilimi göstermiştir.

Sonuç olarak bu çalışma ile hem çevresel hem de insan sağlığı açısından büyük bir risk unsuru oluşturan partikül maddelerin farklı model ve yöntemlerle tahmin çalışması yapılarak R^2 değeri 0,76 olan iyi bir tahmin modeli geliştirilmiştir. Basit istatistiksel analizler, büyük veriler üzerinde kısıtlı performans göstereceğinden çalışmada makine öğrenmesi teknolojisinden yararlanılarak hem güncel bir konu ele alınmış hem de gerçeğe yakın tahminlerin yapılması sağlanmıştır. Gelecekte yapılması planlanan çalışmalarda modelin seçimi aşamasında hiperparametrelerin seçimi üzerinde durularak algoritma parametreleri ayrıntılı bir şekilde incelenebilir. Ya da hibrit makine öğrenme teknikleri aynı probleme uygulanabilir.

KAYNAKLAR (REFERENCES)

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., Arshad, H., 2018, "State-of-the-art in artificial neural network applications: A survey". *Heliyon*, Cilt 4, Sayı 11, e00938.

Avrupa Çevre Ajansı, https://www.eea.europa.eu/data-and-maps/figures/air-quality-standards-under-the-1, ziyaret tarihi: 01.06.2022.

Başakın, E. E., Ekmekcioğlu, Ö., Özger, M., 2019, "Makine öğrenmesi yöntemleri ile kuraklık analizi". *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, Cilt 25, Sayı 8, ss. 985-991.

Box, G., Jenkins, G., 1970, "Time series analysis: forecasting and control,(revised edition 1976) Holden-Day". *San Francisco*.

Bozdağ, A., Dokuz, Y., Gökçek, Ö. B., 2020, "Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey". *Environmental Pollution*, Cilt 263, 114635.

Breiman, L., 2001, "Random forests". *Machine learning*, Cilt 45, Sayı 1, ss. 5-32.

Chen, G., Li S., Knibbs, L. D., Hamm, N. A., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M. J., Guo, Y., 2018, "A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information". *Science of the Total Environment*, Cilt 636, ss. 52-60.

Cutler, A., Cutler, D. R., Stevens, J. R., 2012, Random forests. In: Ensemble machine learning. Eds: Springer, ss. 157-75.

Çelik, B., Arici, N., 2021, "Covid-19 Salgın Sürecinde Hava Kalitesi Tahmini: Zonguldak Örneği". *Gazi*

Mühendislik Bilimleri Dergisi, Cilt 7, Sayı 3, ss. 222-232.

ÇİSİP (Çevre İklim ve Sağlık için İş birliği Projesi) Bilgi Notu, https://www.env-health.org/wp-content/uploads/2022/03/Hava_Kirliligi_Bilgi_Notu.pdf, ziyaret tarihi: 01.06.2022.

Demolli, H., Dokuz, A., Gokcek, M., Ecemiş, A., 2019, "Makine Öğrenmesi Algoritmalarıyla Güneş Enerjisi Tahmini: Niğde İli Örneği", *International Turkic World Congress on Science and Engineering, ss*. 783.

Dündar, D., Sariçiçek, İ., Çinar, E., Yazici, A., 2021, "Kestirimci Bakımda Makine Öğrenmesi: Literatür Araştırması". *Eskişehir Osmangazi Üniversitesi Mühendislik ve Mimarlık Fakültesi Dergisi*, Cilt 29, Sayı 2, ss. 256-76.

Dünya Sağlık Örgütü (World Health Organization, WHO), www.who.int/health-topics/air-pollution, ziyaret tarihi: 13.01.2022.

Edali, M., Yucel, G., 2018, "Automated analysis of regularities between model parameters and output using support vector regression in conjunction with decision trees", *Jasss-The Journal Of Artıfıcıal Socıetıes And Socıal Sımulatıon*, Cilt 21, Sayı 4.

Feurer, M., Hutter, F., 2019, Hyperparameter optimization. In: Automated machine learning. Eds: Springer, Cham, ss. 3-33.

García, S., Luengo, J., Herrera F., 2015, Data preprocessing in data mining, Springer, p.

Gültepe, Y., 2019, "Makine öğrenmesi algoritmaları ile hava kirliliği tahmini üzerine karşılaştırmalı bir değerlendirme". *Avrupa Bilim ve Teknoloji Dergisi*, Cilt 16, ss. 8-15.

Hall, M. A., 1999, "Correlation-based feature selection for machine learning".

Hsu, Y-H, Chuang, H-C, Lee, Y-H, Lin, Y-F, Chen, Y-J, Hsiao, T-C, Wu, M-Y, Chiu, H-W, 2019, "Traffic-related particulate matter exposure induces nephrotoxicity in vitro and in vivo", *Free Radical Biology and Medicine*, Cilt 135, ss. 235-44.

Hyndman, R. J., Athanasopoulos, G., 2018, Forecasting: principles and practice, OTexts, p.

İBB Meteoroloji Gözlem İstasyonu Veri Seti, https://data.ibb.gov.tr/dataset/meteorology-observation-station-data-set, ziyaret tarihi: 11.01.2022.

İBB Saatlik Trafik Yoğunluk Veri Seti, https://data.ibb.gov.tr/dataset/hourly-traffic-density-data-set, ziyaret tarihi: 11.01.2022.

İstanbul Büyükşehir Belediyesi, Veri setleri, https://data.istanbul/dataset, ziyaret tarihi: 11.01.2022.

Joseph, V. R., 2022, "Optimal ratio for data splitting". *Statistical Analysis and Data Mining: The ASA Data Science Journal*.

Kampa, M., Castanas, E., 2008, "Human health effects of air pollution". *Environmental pollution*, Cilt 151, Sayı 2, ss. 362-7.

Karsoliya, S., 2012, "Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture". *International Journal of Engineering Trends and Technology*, Cilt 3, Sayı 6, ss. 714-7.

Kaynar, O., Tuna, M. F., Görmez, Y., Deveci, M. A., 2017, "Makine öğrenmesi yöntemleriyle müşteri kaybı analizi", *Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi*, Cilt 18, Sayı 1, ss. 1-14.

Kingma, D. P., Ba, J., 2014, "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*.

Koprinska, I., Rana, M., Agelidis, V. G., 2015, "Correlation and instance based feature selection for electricity load forecasting". *Knowledge-Based Systems*, Cilt 82, ss. 29-40.

Kuşkapan, E., Çodur, M. K., Çodur, M. Y., 2022, "Türkiye'deki Demiryolu Enerji Tüketiminin Yapay Sinir Ağları İle Tahmin Edilmesi". *Konya Mühendislik Bilimleri Dergisi*, Cilt 10, Sayı 1, ss. 72-84.

Layanun, V., Suksamosorn, S., Songsiri, J., 2017, "Missing-data imputation for solar irradiance forecasting in Thailand", *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, ss. 1234-9

Li, L., Zhang, J., Wang, Y., Ran, B., 2018, "Missing value imputation for traffic-related time series data based on a multi-view learning method". *IEEE Transactions on Intelligent Transportation Systems*, Cilt 20, Sayı 8, ss. 2933-43.

Little, R. J., Rubin, D. B., 2019, Statistical analysis with missing data, John Wiley & Sons, p.

Mitchell, T. J., Beauchamp, J. J., 1988, "Bayesian variable selection in linear regression". *Journal of the american statistical association*, Cilt 83, Sayı 404, ss. 1023-32.

Namlı, E., Ramazan, Ü., Ecem, G., 2019, "Fiyat Tahminlemesinde Makine Öğrenmesi Teknikleri Ve Doğrusal Regresyon Yöntemlerinin Kıyaslanması; Türkiye'de Satılan İkinci El Araç Fiyatlarinin Tahminlenmesine Yönelik Bir Vaka Çalışması". *Konya Mühendislik Bilimleri Dergisi*, Cilt 7, Sayı 4, ss. 806-21.

Özmaden, M. Ş., Erdal, M., 2020, "Performance analysis of methods used in the cost estimation of residential buildings". *Konya Mühendislik Bilimleri Dergisi*.

Öztürk, A., Durak, Ü., Badilli, F., 2020, "Twitter verilerinden doğal dil işleme ve makine öğrenmesi ile hastalık tespiti". *Konya Mühendislik Bilimleri Dergisi*, Cilt 8, Sayı 4, ss. 839-52.

Parbat, D., Chakraborty M., 2020, "A python based support vector regression model for prediction of COVID19 cases in India". *Chaos, Solitons & Fractals*, Cilt 138, 109942.

Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J., Tapper, N. J., 2011, "Quantifying the influence of local meteorology on air quality using generalized additive models". *Atmospheric Environment*, Cilt 45, Sayı 6, ss. 1328-36.

Pénard-Morand, C., Annesi-Maesano, I., 2004, "Air pollution: from sources of emissions to health effects". *Breathe*, Cilt 1, Sayı 2, ss. 108-19.

Pujianto, U., Wibawa, A. P., Akbar, M. I., 2019, "K-nearest neighbor (k-NN) based missing data imputation", *2019 5th International Conference on Science in Information Technology (ICSITech)*, ss. 83-8

Sadriev, A. R., Kamaev, B. N., 2020, "Multivariate Prediction Model of Trade Diversity: Brics Countries". *SCMS Journal of Indian Management*, Cilt 17, Sayı 3.

Samuel, A. L., 1988, "Some studies in machine learning using the game of checkers. II—recent progress". *Computer Games I*, ss. 366-400.

Smola, A. J., Schölkopf B., 2004, "A tutorial on support vector regression". *Statistics and computing*, Cilt 14, Sayı 3, ss. 199-222.

Spławińska, M., 2015, "The problem of imputation of the missing data from the continuous counts of road traffic". *Archives of civil engineering*, Cilt 61, Sayı 1.

Suleiman, A., Tight M., Quinn A., 2019, "Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2. 5)". *Atmospheric Pollution Research*, Cilt 10, Sayı 1, ss. 134-44.

Sürekli İzleme Merkezi (2022), https://sim.csb.gov.tr/Services/AirQuality, ziyaret tarihi: 11.01.2022.

T.C. Çevre ve Şehircilik Bakanlığı (2019), 2020-2023 Ulusal Akıllı Şehirler Stratejisi ve Eylem Planı, https://www.akillisehirler.gov.tr/wp-content/uploads/EylemPlani.pdf, ziyaret tarihi: 23.01.2021.

T.C. Çevre, Şehircilik ve İkim Değişikliği Bakanlığı, Sürekli İzleme Merkezi, www.havaizleme.gov.tr, ziyaret tarihi:23.01.2021.

Uluslararası Telekomünikasyon Birliği (International Telecommunication Union, ITU), ITU Shaping smarter more sustainable cities, https://smartnet.niua.org/sites/default/files/resources/t-tut-smartcity-2016-1-pdf-e.pdf, ziyaret tarihi:23.01.2021.

Ünlü, O., Ünlü, H., Atay, Y., 2022, "Kalp Hastalığı Teşhisinde Yapay Zekâ Yöntemlerinin Kullanımı ve Karşılaştırılması". *Konya Mühendislik Bilimleri Dergisi*, Cilt 10, Sayı 2, ss. 396-411.

Zickus, M., Greig, A., Niranjan, M., 2002, "Comparison of four machine learning methods for predicting PM10 concentrations in Helsinki, Finland". *Water, Air and Soil Pollution: Focus*, Cilt 2, Sayı 5, ss. 717-29.