Avrupa Bilim ve Teknoloji Dergisi
Sayı 16, S. 8-15, Ağustos 2019
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 16, pp. 8-15, August 2019
Copyright © 2019 EJOSAT
Research Article

Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine Karşılaştırmalı Bir Değerlendirme

Yasemin Gültepe[1*]

[1] Kastamonu Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Kastamonu, Türkiye (ORCID: 0000-0002-8684-9907)

(İlk Geliş Tarihi 21 Şubat 2019 ve Kabul Tarihi 17 Nisan 2019)

(**DOI:** 10.31590/ejosat.530347)

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

Öz

Hava kirliliği, günümüzün en büyük sorunlarından birini teşkil etmektedir. Hava kirliliği, nüfusun artması, kentsel gelişme ve büyüme, endüstrinin gelişmesiyle giderek artan bir önem arz etmektedir. Genellikle hava kirleticilerinin insanlara, canlılara ve çevreye zararlı etkileri zaman, mekan, etki süresi, konsantrasyon ve diğer karakteristiklerine bağlı olarak karmaşık dağılım şekilleri göstermektedir. Bu karmaşıklık, kirletici örnekleri ve eğilimleri modelleme veya ölçmede, ayrıca insanların maruz kaldığı seviyeleri tahmin etmenin zor olduğu anlamına gelmektedir. Hava kirliliğini önleme çalışmaları arasında en önemli adımlardan biri hava kirlenmesi olayının bir model içerisinde değerlendirilmesidir. Bu çalışmada Kastamonu ili ele alınarak, meteoroloji ve çevre uygulamalarında oldukça yeni ve başarılı sonuçlar elde edilen çeşitli makine öğrenmesi algoritmaları ile hava kirliliğinin tahmininde, bazı meteorolojik değişkenler kullanılarak hava kirliliği tahmini yapacak modeller geliştirilmiştir. Minimum-Maksimum (Min-Max) normalizasyon tekniği, öğrenme yöntemleri ile birlikte kullanılmıştır. Tahmin modellerinde, Yapay Sinir Ağları (YSA), Rastgele Orman (Random Forest), K-En Yakın Komşu (K-Nearest Neighborhood), Lojistik Regresyon (Logistic Regression), Karar Ağacı (Decision Tree), Lineer Regresyon (Linear Regression) ve Basit Bayes (Naive Bayes) yöntemleri kullanılmıştır. Çalışmada elde edilen performans değerleri, literatürdeki benzer çalışmalarla kıyaslanarak problemin çözümüne ilişkin en uygun tahmin algoritması tespit edilmiştir. Veri setinin %70'i eğitim ve %30'si test verisi olarak ayrılmıştır. Çalışma sonucunda, YSA modeli için doğru tahmin oranı %87 ve diğer makine öğrenmesi modellerinden Rastgele Orman doğruluk oranı %99 ve Karar Ağacı doğruluk oranı %99 değerleri ile tahminlemede en başarılı sonuçları verdiği görülmüştür. Lineer Regresyon yöntemi %30'lık doğruluk oranı ile oldukça kötü performans sergilemektedir. KastamonuDataSet üzerinde kullanılan yöntemlerin performans değerlendirmelerinde Açıklayıcılık Katsayısı (R^2), Ortalama Karesel Hata (Mean Squared Error-MSE), Ortalama Hata Kare Kökü (Root Mean Square Error-RMSE) ve Ortalama Mutlak Hata (Mean Absolute Error-MAE) metrikleri bakımından istatistiksel önemli farklılıkların bulunduğu tespit edilmiştir.

Anahtar Kelimeler: Yapay Sinir Ağları, Hava Kirliliği, Hava Kirliliği Tahmini, Makine Öğrenme Algoritmaları.

A Comparative Assessment on Air Pollution Estimation by Machine Learning Algorithms

Abstract

Air pollution is one of the biggest problems of today. Air pollution, population growth, urban development and growth are increasingly important with the development of industry. Generally, the harmful effects of air pollutants on humans, animals and the environment show complex distribution patterns depending on time, space, duration of action, concentration and other characteristics. This complexity means that modeling and measurement of pollutant samples and trends is also difficult to predict the levels of

[*] Sorumlu Yazar: Kastamonu Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Kastamonu, Türkiye, ORCID: 0000-0002-8684-9907, yasemingultepe@kastamonu.edu.tr

pollution to which people are exposed. One of the most important steps in prevention of air pollution is the evaluation of contamination in a model. In this study, it is aimed to model air pollution by using some meteorological parameters in the estimation of air pollution by various machine learning algorithms which give new and successful results in meteorology and environment applications. Minimum-Max (Min-Max) normalization technique was used with learning methods. The performance values obtained in the study are compared with the similar studies in the literature and the most appropriate classification algorithm for the solution of the problem has been determined. Separate models were designed and analyzed by using methods such as Artificial Neural Networks (ANN), Random Forest, K-Nearest Neighborhood (K-NN), Logistic Regression, Decision Tree, Linear Regression and Naive Bayes. The performance values obtained in the study were compared with similar studies in the literature and the most appropriate estimation algorithm for the solution of the problem was determined. In this case, 70% of the data set is used for training and 30% for testing. As a result of the study, it was seen that the correct estimation rate for the ANN model was 87% and the other machine learning models gave the best results in the estimation with 99% of the Random Forest accuracy rate and 99% of the Decision Tree accuracy rate. The Linear Regression method performs poorly with a 30% accuracy rate. Performance evaluation of methods used on KastamonuDataSet in terms of the Explanatory Coefficient (R^2), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics.

Keywords: Artificial Neural Networks, Air Pollution, Air Pollution Estimation, Machine Learning Algorithms.

1. Giriş

Günümüzde insanlığın en büyük sorunlarından birisi hava kirliliği haline gelmiştir. Hava kirliliği, nüfusun artması, kentsel gelişme ve büyüme, endüstrinin gelişmesiyle giderek artan bir önem arz etmektedir. Genellikle hava kirleticilerinin insanlara, canlılara ve çevreye zararlı etkileri zaman, mekan, etki süresi, konsantrasyon ve diğer ayrıcı niteliklere bağlı olarak karmaşık dağılım şekilleri göstermektedir. Bu karmaşıklık, kirletici örnekleri ve eğilimleri modelleme veya ölçmede, ayrıca insanların maruz kaldığı seviyeleri tahmin etmenin zor olduğu anlamına gelmektedir. Hava kirliliğinin önlenmesinde en önemli adımlardan biri, kontaminasyon olayının bir model içindeki değerlendirilmesidir (Kunt, 2007). Hava kalitesi düzenli olarak ölçülürse, bölgedeki kirlilik seviyesini belirlemek mümkün olur. Böylelikle temiz hava planları çıkarılır, hava kirliliği haritaları oluşturulur ve dağıtım modelleri yapılabilir. Bu bağlamda üretilen hava kalitesi ölçümlerine dayanarak, elde edilen sonuçlara dayalı çözümler, hava kalitesini ve standartlarını iyileştirmek için çözümler daha sağlıklı, daha gerçekçi ve kolay bir şekilde oluşturulabilir (Aydınlar ve ark., 2009).

Günümüz yapay zeka teknolojilerinden makine öğrenmesi ile hava kirliliğine yönelik objektif ve daha hassas sonuçlar elde edilmiştir. Makine öğrenmesi, karmaşık örüntü algılama ve veriye dayalı karar verebilme özellikleriyle ele alınan problemin çözümünü kendi kendine öğrenebilen bilgisayar algoritmalarının genel adıdır. Varolan veri setleri ve kullanılan makine öğrenmesi yöntemleri ile oluşturulan model, en yüksek performansı elde etmek üzerine kurulmaktadır. Bu nedenle birçok makine öğrenmesi yöntemi geliştirilmiş olup bunlardan bazıları; Destek Vektör Makineleri, Lojistik Regresyon, Lineer Regresyon, Basit Bayes, K-En Yakın Komşu, Rastgele Orman ve Karar Ağacı'dır.

Literatüre bakıldığında hava kirliliği üzerine farklı ülkelerde farklı şehirlere ait verisetleri ve çeşitli makine öğrenmesi yöntemleri kullanılarak çeşitli çalışmalar yapılmıştır. Bunlardan bazıları Tablo 1'de karşılaştırılmıştır. Hava kirleticiler, gaz (SO_2, NOx, HC, CO, CO_2, O_3) ve toz şeklindeki kirletici maddeler olmak üzere genel olarak iki alt grupta toplanmaktadır. Makine öğrenmesi yöntemlerine ait performans sonuçlarında RMSE değerleri bakımından istatistiksel olarak anlamlı farklılıklar bulunmaktadır.

Tablo 1. Hava Kirliliği Tahminleri İçin Makine Öğrenme ile Yapılan Önceki Çalışmaların Karşılaştırılması

Yayınlar	Bölge	Hava Kirletici	RMSE	Metot
Alimissis ve ark., 2018	Athens/Greece	NO_2, NO, O_3,CO, SO_2,	27.28, 55.15, 16.6, 1.06, 11.97, 25,60, 44.58, 16.41, 0.81, 5.74	Çok Değişkenli Doğrusal Regresyon, İleri Beslemeli Ağlar
Hu ve Rahman, 2017.	Sydney	CO	0.64, 0.63, 0.61	Destek Vektör Regresyon, Karar Ağacı Regresyon, Rastgele Orman
Huang & Kuo, 2018	Beijing/China	$PM_{2.5}$	26.37, 25.27	Konvolüsyonel Sinir Ağları, Uzun-Kısa Süreli Bellek
Martínez-España ve ark., 2018	Murcia/Spain	O_3	10.99 10.83 10.20 11.19 12.33	Örnekleme, Rastgele Topluluk, Rastgele Orman, M5P Karar Ağacı, En Yakın K-Komşu
Tamas ve ark., 2016	Corsica/France	PM10, O_3, NO_2,	8.72, 24.22, 17.57 8.28, 22.56, 15.42	Hiyerarşik Kümeleme, K-Ortalamalar Kümeleme
Zaree ve Honarvar, 2018.	Brasov/Romania	O_3	0.632	K-Ortalamalar Kümeleme

Dünya Sağlık Örgütü'nün hazırladığı hava kirliliği ile ilgili açıklanan listede Avrupa'da öne çıkan büyük şehirler arasında İstanbul birinci sıradadır. Saygın ve ark., (2018) İstanbul'daki hava kirliliğini tahmin etmek için Eşik Seviyesini aşan Değer Modelini (Peaks over Threshold Method, ESAD) kullanmışlardır. ESAD yönteminde, eşik seviyesi olarak belirlenen bir değerden büyük tüm değerler eşiği aşan değer olarak kabul edilerek, bu değerler esas alınarak tahminler gerçekleştirilir. Yazarlar, POT yöntemlerinin İstanbul'daki hava kirliliğinin (SO_2 ve PM10 hava kirliteciler) limit aşımlarının oluşumu hakkında yararlı bilgiler sağlayabildiğini ve bu modellerin limit aşımları hakkında kısa vadeli tahminler yapmak için kolaylıkla kullanılabileceğini belirtmişlerdir. Zhao ve Hasan çalışmalarında, Hong Kong şehrindeki $PM_{2.5}$ konsantrasyon seviyesini tahmin etmek için YSA ve Destek Vektör Makineleri (Support Vector Machines, DVM) algoritmalarını geliştirmeyi hedeflemişlerdir. Tahmin modelinin performansı YSA için %75-79 değerleri arasında ve DVM için ise %80-82 değerleri arasında başarı sağlamıştır (Zhao & Hasan, 2013).

Bu çalışmanın amacı, Türkiye'nin Kastamonu ilini dikkate alarak meteoroloji ve çevre uygulamalarında yeni ve başarılı sonuçlar veren makine öğrenme modelleriyle bazı meteolojik parametreler kullanılarak hava kirliliğinin tahmin elde edilmesidir. Kastamonu ili hava kirliliği ile ilgili mevcut veriseti üzerinde Rastgele Orman, K-En Yakın Komşu, Lojistik Regresyon, Karar Ağacı, Lineer Regresyon ve Basit Naives makine öğrenmesi algoritmaları kullanılmıştır. Çalışma kapsamında verilerin %70'i eğitim ve %30'si test verisi olarak ayrılmıştır ve her bir yöntemin tahminleme performansı elde edilmiştir. Elde edilen sonuç değerleri literatürdeki benzer çalışmalarla kıyaslanarak hava kirliliği tahmininde diğer makine öğrenme yöntemleri, YSA'na göre daha başarılı olmuştur. Makine öğrenmesi yöntemleriyle ortaya çıkacak modelin erken tahmin başta olmak üzere birçok avantajı da beraberinde getireceği söylenebilir.

Bu çalışmanın ilk bölümünde detaylı olarak hava kirliliği tahmin problemi için literatürdeki makine öğrenmesi yöntemleri tanıtılmış ve başarımları tablo halinde sunulmuştur. İkinci bölümde kullanılan veri seti, tahmin modellerinde kullanılan makine öğrenmesi yöntemleri ve makine öğrenmesi adımları hakkında özet bilgiler verilmiştir. Üçüncü bölümde bu çalışmada elde edilen sonuçlar hakkında değerlendirme ve tartışma yer almaktadır. Son bölümde ise, çalışmadan elde edilen sonuçlar yorumlanmıştır.

2. Materyal ve Metot

2.1. Veri Seti

Türkiye'nin birçok ilinde olduğu gibi Kastamonu merkezinde zaman zaman hava kirliliği problemi rahatsız edici duruma gelmektedir. Hava kirliliğinin sebeplerinin araştırılması yanında çözüm yollarının ortaya konulması son derece önemlidir. Kastamonu ilinde hava kirliliği ölçümleri düzenli olarak 2002 yılında başlamıştır. Bir kaynaktan çıkan kirleticilerin atmosferdeki dağılımları rüzgar hızı ve yönü, sıcaklık, güneş ışığı oranı, bulutluluk ve yağışlılık gibi meteorolojik koşullara bağlı olarak değişkenlik göstermektedir (Demirarslan ve ark., 2008). Günümüzde de hava kalitesinin düzeyini ölçmek amacıyla kurulan Hava Kalitesi Gözlem İstasyonları vasıtası ile saatlik ortalama olarak SO_2, PM10 ve hava sıcaklığı, hava basıncı, nem, rüzgar yönü ve rüzgar hızı ölçümleri yapılmaktadır.

YSA ve diğer makine öğrenmesi yöntemleri kadar kullanılan veri setindeki öznitelikler de bu yöntemlerin başarısını etkileyen önemli unsurlardan biridir. Bu çalışmada kullanılan veriler, Çevre ve Şehircilik Bakanlığı Hava Kalitesi İzleme İstasyonları web sitesinden Kastamonu il merkezi için alınmıştır. Veri setinde çok büyük veri ile birlikte çok sayıda özellik bulunuyorsa ve bunları yorumlayacak denklem veya fonksiyonlar yok ise bir anlam çıkarmak için makine öğrenmesi kullanılabilir. Çözülmesi gereken problemi çözmek için doğru veri ile algoritmayı beslemek çok önemlidir. İyi veri olsa bile, yararlı bir ölçekte, formatta ve anlamlı özelliklerin dahil olması gerekir. Fakat çeşitli sebeplerden dolayı istasyonların ölçümleri belirli zaman aralıklarında NULL değerler bulunduğu ve kayıp değerler olduğu görülmektedir. Veri setindeki kayıp değerler için o niteliğe ait ortalama değer ile doldurulması yöntemi uygulanmıştır. KastamonuDataSet eğitim seti olarak; 4 Ekim 2015 ile 30 Ekim 2018 tarihleri arasında ortalama 2500 gün olarak belirlenen örnek veriler kullanılmıştır. Bu veri setinde 7 farklı öznitelik (PM10, SO_2, hava sıcaklığı, hava basıncı, nem, rüzgar yönü ve rüzgar hızı) bulunmaktadır.

Tablo 2. Hava Kirletici Parametreler İçin İstatistiksel Oranlar

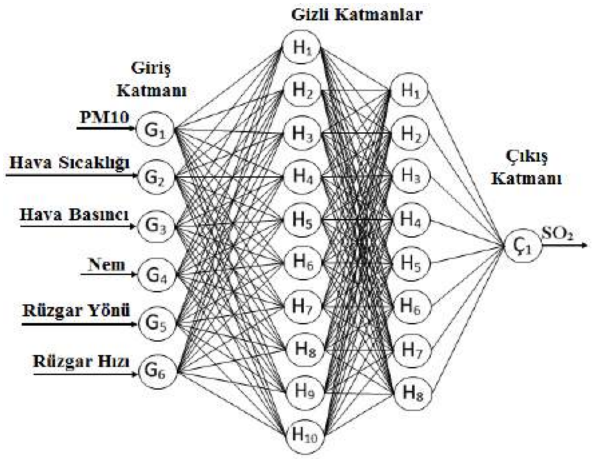
Hava Kirleticiler	**Özellikler**			
	Ortalama	**Standard Sapma**	**Maksimum**	**Minimum**
SO_2 (µg/m³)	5.52	7.89	173.53	-2.87
PM10 (µg/m³)	48.06	47.33	4925,31	0
Hava Sıcaklığı (°C)	9.12	10.25	41,57	-17,4
Hava Basıncı (mbar)	967.31	200.68	1031,1	15,69
Bağıl Nem (%)	68.98	23.09	97,13	-1,17
Rüzgar Yönü (degree)	195.02	121.12	359,71	0,08
Rüzgar Hızı (m/s)	143	1.88	81,05	0,37

Veri setinin istatistiksel özellikleri, veri analizi ve ön işleme için çok önemlidir. İstatistiksel özellikler ortalama, standart sapma, varyans, hipotez testleri vb. ile temsil edilir. Tablo 2'de Kastamonu ili 2015-2018 yılları arasında hava kirleticilere ait olan istatistiksel

özellikler sırasıyla ortalama, standart sapma, maksimum ve minimum değerleri özetlenmiştir. Her meteorolojik parametrenin istatistiksel değişim aralığı incelenmiştir. İstatistikler incelendiğinde özellikle çevreyi ve insan sağlığını olumsuz etkileyen kesim ortalama PM10 ve SO_2 seviyelerinin sırasıyla ≥48.06 µg/m³ ≥5.52 µg/m³ değerleridir.

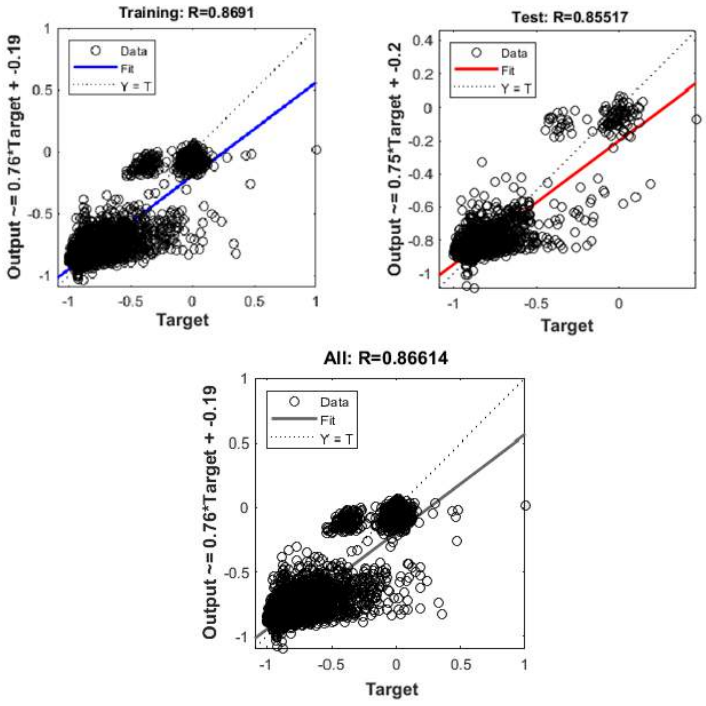
2.2. Yapay Sinir Ağları Kullanılarak Hava Kirliliği Tahmini

YSA ile oluşturulan modeller genellikle istatistiksel veri analizi veya sistem optimizasyonu için kullanılır. Ayrıca regresyon ve diğer klasik yöntemler ile tahminlemede de yaygın olarak kullanılmaktadır. Bu çalışmada hava kalitesi modelinin YSA yaklaşımı ile tahmin edilmesi ve modellenmesi için ağa sunulacak olan veri kümesi, KastamonuDataSet'ten elde edilen verilerdir. Bu veriler Min-Max normalizasyon tekniği ile normalize edildikten sonra günlük sıralı veriler olarak ağa sunulmuştur. Tablo 2'de bulunan bir kirlilik parametresi (PM10) ve günlük olarak ölçülen 5 farklı meteorolojik faktör (hava sıcaklığı, hava basıncı, nem, rüzgar yönü ve rüzgar hızı), modelin giriş parametreleri olarak kullanılıp, çıktı olarak bir gün sonraki SO_2 konsantrasyonu tahmin edilmiştir. Böylelikle Kastamonu il merkezi hava kalitesi modeli, yapay sinir ağları ile oluşturulmuştur.



Şekil 1. Sistem için tasarlanan yapay sinir ağı modeli

Şekil 1'de YSA modeli için ileri beslemeli yapay sinir ağı gösterilmiştir. Yapay sinir ağı için iki gizli katman kullanılmıştır. Birinci gizli katmandaki nöron sayısı 10, ikinci gizli katmkanda nöron sayısı 8 olarak belirlenmiştir. Yapay Sinir Ağı parametrelerini belirlemek için çok sayıda deneme yanılma yöntemi kullanıştır ve çok sayıda test yapılmıştır. Öğrenme oranı, yapay sinir ağlarındaki ağın öğrenme performansı ile yakından ilişkilidir. Öğretilen ağın doğruluk oranı karar verme katsayısı R^2 değeri, 0.76 olarak ölçülmüştür.



Şekil 3. Eğitim ve test regresyon grafiği

Şekil 3'de öğrenme sonrası elde edilen eğitim ve test için regresyon grafiği gösterilmiştir. YSA modelinin performansı karşılaştırıldığında sırasıyla korelasyon katsayıları, R_1=0.8691, R_2=0.08552 ve R_3=0.8661 olarak elde edilmiştir. Bu sonuçlar, modelin ve gerçek verinin birbiriyle uyumlu olduğunu göstermektedir. Bu model için genel hava kalitesi endeksini doğru bir şekilde tahmin etme olasılığı %87'dir.

2.3. Makine Öğrenme Yöntemleri Kullanılarak Hava Kirliliği Tahmini

2.3.1. Makine Öğrenmesi Algoritmaları

Makine öğrenimi, yapay zekanın önemli bir alt alandır. Temel amacı, veriden bilgi elde etmek için hesaplama yöntemlerini kullanmaktır. Makine öğrenimi, el yazısı ve konuşma tanıma, robotik ve bilgisayar oyunları, doğal dil işleme, beyin-makine arayüzü vb. dahil geniş bir uygulama yelpazesine sahiptir. Çevre bilimlerinde, veri işleme, model emülasyonu, hava ve iklim tahmini, hava kalitesi tahmini ve hidrolojik tahminlerde makine öğrenme yöntemleri yoğun olarak kullanılmaktadır (Peng, 2013).

Makine öğrenmesi tabanlı hava kirliliği tahmini sistemi için kullanılacak algoritmaların seçimi önemlidir. Bu çalışmada aşağıda açıklanan makine öğrenmesi yöntemleri kullanılmıştır (Dey, 2016).

- **Rastgele Orman:** Rastgele Orman denetimli bir öğrenme algoritmasıdır. Adından anlaşılacağı üzere rastgele bir orman oluşturur. Oluşturulan orman, genellikle "torbalama" yöntemiyle eğitilmiş karar ağaçları topluluğudur. Torbalama yönteminin amacı, öğrenme modellerinin bir kombinasyonunun genel sonucu arttırmasıdır.
- **K-En Yakın Komşu:** Parameterik olmayan bir tekniktir ve sınıflandırmasında en yakın komşularının sayısı olan k'ı grup üyeliğine göre verileri sınıflandırmak için kullanır.
- **Lojistik Regresyon:** Bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi, temel lojistik fonksiyonunu kullanarak olasılıkları tahmin ederek ölçmektedir.
- **Karar Ağacı:** Bir karar ağacı, her düğümün bir özelliği (niteliği) temsil ettiği, her bağlantının (dal) bir kararı temsil ettiği ve her bir yapının bir sonucu olduğu bir ağaçtır. Her ağaç düğüm ve dallardan oluşur. Her düğüm, sınıflandırılacak olan bir gruptaki özellikleri temsil eder ve her dal, düğümün alabilecğei bir değeri temsil eder.
- **Lineer Regresyon:** Bağımsız değişkenlere dayanan bir hedef tahmin değerini modeller. Çoğunlukla değişkenler ve tahmin arasındaki ilişkiyi bulmak için kullanılır. Farklı regresyon modelleri, bağımlı ve bağımsız değişkenler, kullanılan bağımsız değişkenlerin sayısı arasındaki ilişkiye göre farklılık gösterir.
- **Basit Bayes:** Bayes teoremine dayanan bir denetimli bir makine öğrenmesi yöntemidir. Bu algoritma, koşullu olasılıklara dayalı olarak hedef sınıfta belirli bir değerin olasılığını inceler ve buna dayanarak, hedef sınıfın değerini tahmin eder.

Normalizasyon işlemi, makine öğrenmesi için genellikle veri hazırlamanın bir parçası olarak uygulanan bir tekniktir. Normalleştirmenin amacı, veri kümesindeki sayısal sütunların değerlerini, değerler aralığındaki farklılıkları bozmadan ortak bir ölçeğe uygun biçimde değiştirmektir. Makine öğrenmesi için, her veri kümesini normalleştirme gerekmeyebilir. Normalleştirme işlemi, verilerin boyutunu azaltmak veya işlemleri normalleştirilmiş değerlerle uygun aralıklarla gerçekleştirmek ve daha anlamlı ve kolayca yorumlanabilir sonuçlar elde etmek için kullanılabilir. (Dondurmacı & Çınar, 2014). Literatürde birçok veri normalizasyon çeşidi bulunmaktadır. Bunlar minimum-maksimum (min-max), ondalık ölçeklendirme, z-skor ve sigmoid gibi sıralanabilir (Jayalakshmi & Santhakumaran, 2011).

Bu çalışmada min-max normalizasyon metodu uygulanmış ve veri seti değerleri 0 ile 1 arasında normalize edilmiştir. Min-Max normalizasyon yöntemi için Denklem 1 kullanılır (Yapraklı ve Erdal, 2016).

$$x_{yeni} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Bu eşitlikte;

x_{yeni}: x değişkeni için yeni sayıyı,

x: x değişkeni için geçerli sayıyı,

x_{min}: veri setindeki bulunan en küçük sayı,

x_{max}: veri setinde bulunan en büyük sayıyı,

ifade etmektedir.

Tahmin etmeye ilişkin iki önemli adım bulunmaktadır; birincisi veriyi tahmin etmek için hazırlamadır. İkincisi ise farkı tahmin edici modellerin karşılaştırılmasıdır. Modelleri karşılaştırma ölçütleri; doğruluk, hız, sağlamlık, ölçeklenebilirlik, yorumlanabilirliktir. Yapay Sinir Ağları ve makine öğrenmesi yöntemlerinin performans değerlendirmelerinde kullanılan temel performans göstergeleri arasında R^2, MSE, RMSE ve MAE sayılabilir (Karasu ve ark., 2018).

Bu çalışmada Tablo 3'de verilen doğruluk ölçütleri kullanılmaktadır. Bu performans ölçüleri içerisinde R^2, modelin doğruluk oranı karar verme katsayısıdır. Bu katsayı değerinin yüksek olması tahmin ilişkisinin iyi olduğu gösterir. MSE, RMSE ve MAE ise birer hata ölçüsü olması nedeniyle düşük sonuçlar, performans ile ters orantılı olarak yüksek performansı gösteren ölçülerdir (Wang ve Xu, 2004). Örneğin RMSE sıfıra eşit olması durumunda iyi bir performans göstermektedir (Çınaroğlu, 2017). Belirli t zamanı

göstermek üzere r_t zaman aralığında gözlemlenen ve p_t tahmin edilen zaman serisi olmak üzere hata, Denklem 2'de e_t formülü ile ifade edilmektedir.

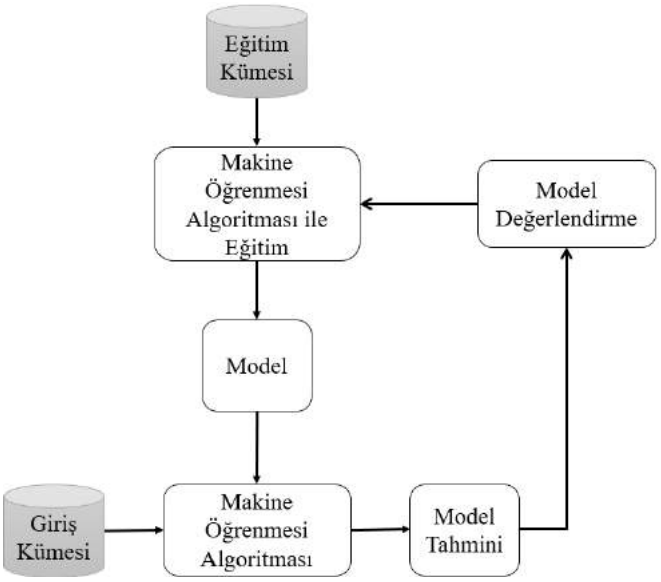
$$e_t = r_t - p_t \qquad (2)$$

Tablo 3. Hata Ölçütleri

Açıklaması	Hata Ölçüt Formülü
MSE	$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n} e_t^2$
RMSE	$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_t^2}$
MAE	$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n} \lvert e_t \rvert$

2.3. Makine Öğrenmesi Adımları

Makine öğrenmesi algoritmaları ile hava kirliliği tahmini, iki temel süreçten oluşmaktadır. Birincisi, makine öğrenmesi için temel gereksinimlerden olan veriyi toplamaktır. KastamonuDataSet'te tanımlanan parametreler, T.C. Çevre ve Şehircilik Bakanlığı Hava Kalitesi İzleme İstasyonları Web Sitesi'nden Kastamonu'da 4 Ekim 2015 ile 30 Ekim 2018 tarihleri arasındaki PM10, SO_2, hava sıcaklığı, hava basıncı, nem, rüzgar yönü ve rüzgar hızı ölçüm sonuçlarından elde edilen verilerdir. İkincisi ise veri seti üzerinde makine öğrenme modelleri oluşturularak kirleticilerin ve konsantrasyonlarının tahmin edilmesidir. Kirleticiler olarak özniteliklerin doğru seçilmesi ve doğru şekilde temsil edilmesi çok önemlidir. Temel makine öğrenmesi süreci, Şekil 4'de gösterilmiştir (Bilbro, 2016). Model(ler) oluşturulur ve bu model(ler)in amacı tahmin edilmesi istenilen şeyin, en verimli, en yüksek olasılıkla en hızlı şekilde tahmin edilmesidir.



Şekil 4. Temel makine öğrenmesi süreci

KastamonuDataSet için hangi algoritmanın başarılı olacağını önceden tahmin etmek zor bir iştir. Bu çalışmanın sonuçları incelendiğinde hangi algoritmanın hangi doğruluk derecesiyle en iyi sonucu ürettiği söylenebilecektir. Ayrıca gerçekleştirilecek algoritmalar arasında da algoritma başarısına göre de bir sıralama işlemi uygulanacaktır. YSA uygulaması Matlab ile diğer makine öğrenmesi yöntemleri ise Pyhton ortamında kodlanmıştır. Python dili, yapay zeka ve makine öğrenmesi uygulamalarında oldukça popüler ve kullanışlı bir dildir.

3. Araştırma Sonuçları ve Tartışma

YSA ve makine öğrenmesi ile yapılan testlerde, modellerin performansının değerlendirilmesinde R^2, MSE, RMSE ve MAE esas alınmıştır. Hava kirliliği tahmini için yapılan analiz ortalamaları, Tablo 4'de gösterilmiştir. Rastgele Orman ve Karar Ağacı yöntemleri en yüksek R^2, en düşük MSE ve MAE oranlarına sahiptir. R^2 değerinin 1'e yaklaşırken MSE değerinin 0'a yaklaştığı göz önüne alınan modeller arasında en başarılı model parametreleri gösterilmektedir.

Deneysel sonuçlar, YSA modeli için doğru tahmin oranı %87 ve diğer makine öğrenmesi modellerinden Rastgele Orman %99 ve Karar Ağacı %99 değerleri ile en iyi sonuçları vermişlerdir. Lineer Regresyon yöntemi %30'lık doğruluk oranı ile oldukça kötü performans sergilemektedir. Ayrıca deneysel sonuçlar, makine öğrenme yöntemlerinin (Rastgele Orman (%99) ve Karar Ağacı (%99)), YSA'nın %87 doğruluk oranına göre daha yüksek doğruluk elde edildiğini göstermiştir.

Tablo 4. KastamonuDataSet Üzerinde Kullanılan YSA ve Makine Öğrenmesi Yöntemlerin Sonuçlarının Karşılaştırılması

Değerlendirme Kriterleri	**Doğru Tahmin Oranı**	**R^2**	**MSE**	**RMSE**	**MAE**
YSA	%87	0.76	0.19	0.04	0.08
Rastgele Orman	%99	0.97	0.01	0.11	0.01
K-En Yakın Komşu	%97	0.98	0.03	0.17	0.03
Lojistik Regresyon	%87	0.86	0.15	0.38	0.13
Karar Ağacı	%99	0.97	0.01	0.10	0.01
Lineer Regresyon	%30	0.29	0.27	0.52	0.44
Basit Bayes	%94	0.93	0.06	0.25	0.06

4. Sonuç

Bu çalışmada çeşitli meteorolojik parametreler ve hava kirliliği arasındaki ilişki incelenmiştir. Bu amaçla, YSA modeli ve beş farklı makine öğrenmesi modeli kullanılarak hava kirliliği tahmin edilmiş ve model sonuçları karşılaştırılmıştır. Minimum-Maksimum (Min-Max) normalizasyon tekniği, öğrenme yöntemleri ile birlikte kullanılmıştır. Analizde kullanılan YSA ve makine öğrenmesi yöntemleri için 2015-2018 yılları arasındaki toplam 2500 verinin %70'lik kısmına karşılık gelen 1750 veri ile eğitim aşaması; geri kalan %30'lik kısmına karşılık gelene 750 veri ile test aşaması gerçekleştirilmiştir. Elde edilen sonuçlara göre hava kirletici tahmini için Rastgele Orman ve Karar Ağacı makine öğrenme yöntemleri en yüksek performansı göstererek ön plana çıkmıştır. En kötü performansı ise Lineer Regresyon yöntemi sergilemiştir.

Çalışma sonucunda, hava kirliliği tahmininde YSA, lojistik regresyon, lineer regresyon makine öğrenmesi yöntemlerine göre daha başarılı olmuştur. Ayrıca, önceki yıllardaki hava kirliliğine ait kirletici ve konsantrasyon verilerini kullanarak gelecek yıllar için uygun önlemlerin alınmasının mümkün olacağı sonucuna varılmıştır.

Kaynakça

Alimissis, A., Philippopoulos, K., Tzanis, C.G., and Deligiorgi, D. (2018). Spatial estimation of urban air pollution with the use of artificial neural network models, Atmospheric Environment, 191, 205-213, 2018.

Bilbro, R. (2016). An Introduction to Machine Learning with Python. Erişim adresi: https://districtdatalabs.silvrback.com/an-introduction-to-machine-learning-with-python.

Aydınlar, B., Güveni H. ve Kırksekiz, S., 2009. Hava Kirliliği ve Modellenmesi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Çevre Mühendisliği Bölümü Yüksek Lisans Rapor.

Çınaroğlu, S. (2017). Sağlık Harcamasının Tahmininde Makine Öğrenmesi Regresyon Yöntemlerinin Karşılaştırılması, Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, 22(2): 179-200.

Demirarslan, O. Çetin, Ş., & Ayberk, S. (2008). Hava Kirliliği Belirlemelerinde Modelleme Yaklaşımı ve Modelleme Aşamasında Karşılaşılabilecek Sorunlar, Environmental Problems Symposium, Kocaeli 2008.

Dey, A. (2016). Machine Learning Algorithms: A Review, International Journal of Computer Science and Information Technologies, 7(3): 1174-1179.

Dondurmacı, G.A. & Çınar, A. (2014). Finans Sektöründe Veir Madenciliği Uygulaması, Akademik Sosyal Araştırmalar Dergisi, 2(1): 258-271.

Hu, K. & Rahman, A. (2017). HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation From Fixed and Mobile Sensors, IEE Sensors, 17(11): 3571-3525.

Huang, C-J., & Kuo, P-H. (2018). A Deep CNN-LSTM Model for Particulate Matter ($PM_{2.5}$) Forecasting in Smart Cities, Sensors 2018.

Jayalakshmi, T. & Santhakumaran A. (2011). Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering, 3(1): 89-93

Karasu, S., Hacıoğlu, R. & Altan, A. (2018). Prediction of Bitcoin Prices with Machine Learning Methods using Time Series Data, 26th signal Processing and Communications Applications Conference.

Kunt, F. (2007). Hava Kirliliğinin Yapay Sinir Ağları Yöntemiyle Modellenmesi ve Tahmini, Selçuk University Graduate School of Natural and Applied Sciences, M.Sc. Thesis, Environmental Engineering Department, Konya, 2007.

Martínez-Espaňa, R., Bueno-Crespo, A., Timón, I., Soto, J., Muňoz, A. & Cecilia, J.M. (2018). Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain.

Peng, H. (2013). *Air Quality Prediction by Machine Learning Methods*. The Degree of Master of Science, the University of British Columbia.

Saygın, H., Eren, Ö., & Oral, H.V. (2018). Peaks Over Threshold Method Application on Airborne Particulate Matter (PM_{10}) and Sulphur Dioxide (SO_2) Pollution Detection in Specified Regions of İstanbul, Avrupa Bilim ve Teknoloji Dergisi, 14:399-407.

Tamas, W., Notton, G., Paoli, C., Nivet, M-L. & Voyant, C. (2016). Hybridization of Air Quality Frecasting Models Using Machine Learning and Clustering: An Orginal Approach to Detect Pollutant Peaks, Aerosol and Air Qaulity Research, 16: 405-416.

Yapraklı, T.Ş. & Erdal, H. (2016). Firma Başarısızlığı Tahminlemesi: Makine Öğrenmesie Dayalı Bir Uygulama, Bilişim Teknolojileri Dergisi, 9(1): 21-31.

Zaree, T. & Honarvar, A.R. (2018). Improvement of Air Pollution Prediction in a Smart City and its Correction with Weather Conditions using Metrological Big Data, Turkish Journal of Electrical Engineering & Computer Sciences, 26: 1302-1313.

Zhao, Y. & Hasan, Y.A. (2013). Machine Learning Algorithms for Predicting Roadside Fine Particulate Matter Concentration Level in Hong Kong Central, Computation Ecology and Software, 3(3): 61-73.

Wang, W. & Xu, Z. (2004). A Heuristic Training for Support Vector Regression, Neurocomputing, 61: 259-275.