Academic Platform Journal of Engineering and Science

journal homepage: http://apjes.com/

Hava Kalite İndeksinin Tahmin Başarısının Artırılması için Topluluk Regresyon Algoritmalarının Kullanılması

*[1]Muhammet Emre Irmak, [2]İbrahim Berkan Aydilek
[1]Harran Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik Elektronik Mühendisliği Bölümü, Şanlıurfa, Türkiye
memreirmak@hotmail.com,
[2]Harran Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Şanlıurfa, Türkiye
berkanaydilek@harran.edu.tr,

Araştırma Makalesi Geliş Tarihi: 02.11.2018 Kabul Tarihi: 21.03.2019

Öz

Şehirlerdeki hava kalitesi seviyesinin düzenli aralıklarla ölçülmesi ve ölçüm sonuçlarının incelenerek gerekli önlemlerin alınması bu şehirlerde yaşayan insanların ve diğer canlıların sağlıkları için oldukça önemlidir. Ülkemizde bu amaçla ilgili bakanlık tarafından pek çok şehre hava kalitesi ölçüm istasyonları kurulmuştur. Bu çalışmada bu istasyonlardan biri olan Adana ili valilik istasyonuna ait ölçüm verileri kullanıldı. Kullanılan veriler kükürt dioksit (SO_2), azot dioksit (NO_2), ozon (O_3), karbon monoksit (CO) ve toz parçacıkları (PM10) gibi hava kirletici gazların ölçüm değerlerdir. Bu verilere farklı makine öğrenme algoritmaları uygulanarak hava kalite indeksi tespit edildi. Kullanılan makine öğrenmesi regresyon algoritmaları; rastgele orman, karar ağacı, destek vektör, k-en yakın komşu, doğrusal, yapay sinir ağı, yığın, uyumlu artırıcı, eğimli artırıcı ve örneklemeli toplam regresyonudur. Bu algoritmaların hata oranları ve çalışma süreleri bakımından başarı değerleri kıyaslanarak elde edilen sonuçlar değerlendirilmiştir.

Anahtar Kelimeler: Hava Kalite İndeksi, Makine Öğrenmesi, Regresyon Algoritmaları, Topluluk Öğrenme

Using Ensemble Regression Algorithms for Improving the Prediction Success of Air Quality Index

*[1]Muhammet Emre Irmak, [2]Ibrahim Berkan Aydilek
[1] Harran University, Graduate School of Natural and Applied Sciences, Department of Electrical and Electronics Engineering, Sanliurfa, Turkey,
memreirmak@hotmail.com
[2] Harran University, Faculty of Engineering, Department of Computer Engineering, Sanliurfa, Turkey,
berkanaydilek@harran.edu.tr

Abstract

Measuring the air quality level in the city at regular intervals and taking the necessary measures by examining the results of the measurement is very important for the health of the people and other living things in these cities. For this purpose, air quality measurement stations have been established in many cities by the relevant ministry. In this study, one of these stations, Adana province provincial station measurement data was used. The data used are the measured values of air pollutant gases such as sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO) and dust particles (PM10). The air quality index was determined by applying different machine learning algorithms to these data. Machine learning regression algorithms used; random forest, decision tree, support vector, k-nearest neighbor, linear, artificial neural network, stacking, adaboost, gradient boosting and bagging regression. The results obtained by comparing the success rates of these algorithms in terms of error rates and run times were evaluated.

Keywords: Air Quality Index, Machine Learning, Regression Algorithms, Ensemble Learning

1. GİRİŞ

Bulunduğumuz yüzyılda insanların büyük çoğunluğu yüksek nüfuslu şehirlerde yaşamaktadır. Yüksek nüfuslu şehirler pek çok sorunu içerisinde barındırabilmektedir. Yüksek nüfuslu bu şehirlerin sorunlarından biri de hava kirliliğidir. Hava kirliliği ulaşım, ısınma ve sanayi gibi insan kaynaklı olabileceği gibi çöl tozları, yanardağ faaliyetleri ve orman yangınları gibi doğal

*[1]Sorumlu Yazar: Harran Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik Elektronik Mühendisliği Bölümü, Şanlıurfa, Türkiye, memreirmak@hotmail.com, +905069273026

Doi: 10.21541/apjes.478038

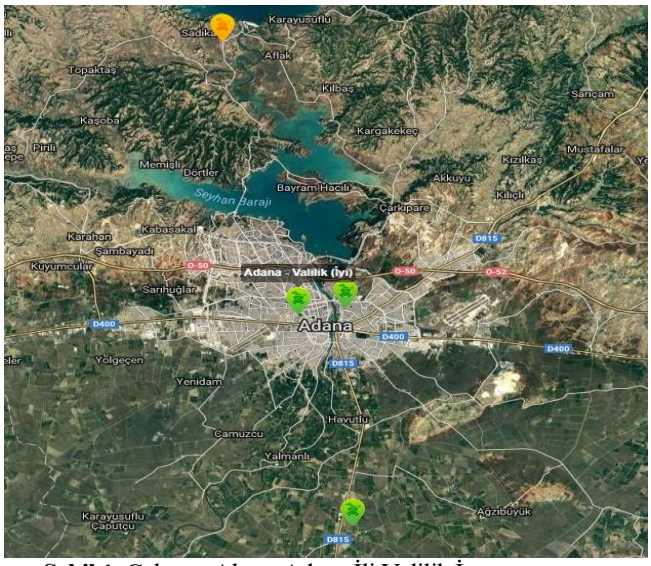
kaynaklı da olabilmektedir. Şehirlerdeki bu hava kirliliğinin önlenebilmesi için hava kirliliğinin belirli noktalardan ölçülmesi ve bu ölçüm sonuçlarına göre gerekli tedbirlerin alınması gerekmektedir [1].

Hava kalitesinin belirlenmesi için ölçülecek olan değerler uluslararası ve ulusal değerlendirmeler sonucunda belirlenmiştir. Buna göre hava kirliliği kükürt dioksit (SO_2), azot dioksit (NO_2), ozon (O_3), karbon monoksit (CO) gazları ve toz parçacıklarının (PM10) miktarlarına göre belirlenmektedir. Bu ölçümlerden istenen sonuçların çıkarılmasında geleneksel matematiksel modeller kullanıldığı gibi son zamanlarda yapay zekâ modelleri de kullanılmaktadır. Veljanovska ve Dimoski (2018) yaptıkları çalışmada Üsküp şehrinin hava kalite indeksini belirlemede k en yakın komşu, karar ağacı, yapay sinir ağları ve destek vektör makineleri sınıflandırma algoritmalarının başarı sonuçlarını kıyaslamışlardır [2]. Dragomir (2010) yaptığı çalışmada Romanya'daki hava kalite indeksini hesaplamak için k en yakın komşu sınıflandırmasını kullanmıştır [3]. Adams ve ark. (2013) yaptıkları çalışmada Kanada Ontaryo gölü bölgesinde havadaki PM2.5 ve NO_2 değerlerini yapay sinir ağları ile hesaplamışlar ve sonuçları geleneksel yöntemle karşılaştırmışlardır [4]. Raturi ve Prasad (2018) hava kalite indeksini hesaplamada yapay sinir ağlarını kullanmışlardır [5]. Zhai ve Chen (2017) yaptıkları çalışmada Çin'in başkenti Pekin'in hava kalite seviyesini tahminde genetik algoritmaları ve yapay sinir ağlarını kullanmışlar ve başarım sonuçlarını kıyaslamışlardır [6]. Wang ve ark. (2016) yaptıkları çalışmada Çin'in Nanjing şehrinin hava kalite indeksini hesaplamada otoregresif entegre hareketli ortalamayı ve bulanık zaman serilerini kullanmışlardır [7].

Bu çalışmada makine öğrenmesi regresyon algoritmalarının başarı değerleri kıyaslanmıştır. Kullanılan regresyon yöntemleri; rastgele orman regresyonu, karar ağacı regresyonu, destek vektör regresyonu, k-en yakın komşu regresyonu, doğrusal regresyon ve yapay sinir ağı regresyonu yığın regresyonu, uyumlu artırıcı regresyonu, eğimli artırıcı regresyonu ve örneklemeli toplam regresyonudur.

2. ULUSAL HAVA KALİTESİ İNDEKSİ VE ADANA

Bu çalışmada kullanılan veri seti, Adana ilinin valilik istasyonuna ait 2013-2017 yılları arasındaki PM10, SO_2, NO_2, O_3 ve CO hava kirleticilerinin saatlik ölçüm değerlerinden oluşmaktadır. Veri seti ilgili bakanlığa ait internet adresinden [8] elde edilmiştir. Şekil 1.'de Adana ilindeki 4 adet hava kalite izleme istasyonunun ve çalışma alanı olan valilik istasyonunun konumu görülmektedir. Hava kalite indeksi, hava kalitesinin sağlık açısından hangi seviyede olduğunu ifade etmektedir. Birleşik devletler çevre koruma ajansının (Unites States Environmental Protection Agency- EPA) yayımladığı değerleri Tablo 1.'de verilmiştir [9], [10]. Ulusal hava kalite indeks değerleri, EPA değerlerinin ulusal mevzuata ve sınır değerlerine uyarlanması ile oluşturulmuştur [10]. Kirleticilerin değerleri bir metre küp havada kaç mikrogram bulunduklarını ifade eden $\mu g/m^3$ cinsinden ölçülmektedir. Deneysel çalışmada; SO_2 ve NO_2 gazlarının saatlik ortalama değerleri, CO ve O_3 gazlarının son 8 saatlik ortalama değerleri ve PM10 toz değerinin son 24 saatlik ortalama değerleri hesaplanarak kullanılmıştır.



Şekil 1. Çalışma Alanı- Adana İli Valilik İstasyonu

Tablo 1. EPA Hava Kalite İndeksi Değerleri

Hava Kalitesi İndeksi (AQI) EPA Değerleri	Sağlık Endişe Seviyeleri	Renkler	Anlamı
0-50	İyi	Yeşil	Hava kalitesi memnun edici ve risk teşkil etmiyor
51-100	Orta	Sarı	Hava kalitesi uygun fakat hava kirliliğine hassas olan az sayıdaki insanlar için bazı kirleticiler açısından orta düzeyde sağlık endişesi oluşabilir
101-150	Hassas	Turuncu	Hassas gruplar için sağlık etkileri oluşabilir. Genel olarak kamunun etkilenmesi olası değildir.
151-200	Sağlıksız	Kırmızı	Herkes sağlık sorunları yaşamaya başlayabilir. Hassa gruplar için ciddi sağlık etkileri söz konusu olabilir
201-300	Kötü	Mor	Sağlık açısından acil durum oluşturabilir. Nüfusun tamamının etkilenme olasılığı yüksektir
301-500	Tehlikeli	Kahverengi	Sağlık Alarmı: Herkes çok ciddi sağlık sorunlarıyla karşılaşır

Tablo 2.'de verilen ulusal hava kalitesi indeks kesme noktalarına göre her bir hava kirleticisi için hava kalite indeks değeri iyi, orta, hassas, sağlıksız, kötü ve tehlikeli olarak hesaplanmıştır. Sonuç olarak hava kalite indeksi en kirli gazın indeks değerine eşit olmaktadır. Hesaplama için geliştirilen programda indeks sayısal karşılıkları kullanılmıştır. Bu amaçla, iyi indeks değeri için 1, orta indeks değeri için 2, hassas indeks değeri için 3, sağlıksız indeks değeri için 4, kötü indeks değeri için 5, tehlikeli indeks değeri için 6 değerleri kullanılmıştır.

Tablo 2. Ulusal Hava Kalitesi İndeksi Kesme Noktaları

Hava Kalite İndeksi	**SO_2 [µg/m^3]**	**NO_2 [µg/m^3]**	**CO [µg/m^3]**	**O_3 [µg/m^3]**	**PM10 [µg/m^3]**
	1 Sa. Ort.	1 Sa. Ort.	8 Sa. Ort.	8 Sa. Ort.	24 Sa. Ort.
İyi	0-100	0-100	0-5500	0-120	0-50
Orta	101-250	101-200	5501-10000	121-160	51-100
Hassas	251-500	201-500	10001-16000	161-180	101-260
Sağlıksız	501-850	501-1000	16001-24000	181-240	261-400
Kötü	851-1100	1001-2000	24001-32000	241-700	401-520
Tehlikeli	>1100	>2000	>32000	>700	>520

Veri seti üzerinde, hava kalitesi ölçüm istasyonundaki bakım çalışmaları ve donanım hatalarından oluştuğu düşünülen eksik veriler mevcuttur. Eksik veriler her bir gazın 2013-2017 yılı ortalama değeri ile tamamlanmıştır. Veri setinde yer alan nitelikler arasında çok büyük farklar olduğunda bir nitelik diğerini baskılayabilmektedir. Bu gibi baskın niteliklere ait verileri tek bir düzen içinde ele almak ve aynı algoritmada kullanılabilir hale getirmek için normalleştirme işlemi uygulanmaktadır. Örneğin CO niteliği 0 ile 32000 arasında değer alırken, PM10 niteliği 0 ile 520 arasında değer almaktadır. Burada, normalleştirme yapılmazsa yapay sinir ağı için CO niteliği daha baskın hale gelebilecektir. Bu sebeple, mevcut diğer niteliklerin sonuca olan etkileri baskılanabileceğinden dolayı doğru sonuca ulaşılamayacaktır. Bu çalışmada min-maks (asgari-azami) normalleştirme kullanılarak veriler 0 ve 1 arasında olacak şekilde normalleştirilmiştir. Hesaplama işlemleri, veri madenciliği ve makine öğrenmesi için faydalı kütüphaneleri içeren Python programlama dili derleyicisi Spyder [11] ile yapılmıştır. Hesaplama işlemi bittikten sonra 2013-2017 yılları için 43838 kayıt içeren bir veri seti elde edilmiştir. Bu veri setinin içerisinden rastgele seçilerek elde edilmek şartıyla %75 oranında bir eğitim veri seti ve %25 oranında bir test veri seti oluşturulmuştur. Spyder derleyicisi üzerinde Pandas ve Sklearn kütüphanesi yardımıyla makine öğrenmesi algoritmaları çalıştırılmıştır. Regresyon algoritmaları önce eğitim veri seti ile eğitilmiştir. Sonrasında aynı algoritmaya test verisi uygulanmış ve hata değerleri kaydedilmiştir. Her bir algoritmaya aynı işlemler uygulanmış ve değerleri kaydedilmiştir. Sonuç kısmında algoritmalar, Eş. (20) ile hesaplanan belirlilik katsayısı (r^2), Eş. (21) ile hesaplanan ortalama mutlak hata (OMH), Eş. (22) ile hesaplanan ortalama karesel hata (OKH) ve işlem süreleri açısından kıyaslanmıştır.

3. KULLANILAN MAKİNE ÖĞRENMESİ REGRESYON ALGORİTMALARI

3.1. Temel Regresyon Yöntemleri

3.1.1. Doğrusal Regresyon (Linear Regression-LR)

Bağımsız değişkenler olan girdilerin ve bunlara bağlı olarak hesaplanan çıktının arasındaki ilişkinin doğrusal olarak belirlendiği bir algoritmadır [12]. Çalışmada 5 tane bağımsız değişken olduğu için doğrusal regresyon Eş. (1) ile ifade edilecektir.

$$Yi = \beta 0 + \beta 1\, Xi1 + \beta 2\, Xi2 + \cdots + \beta k\, Xik + \varepsilon i \quad i = 1, 2, \ldots, n \qquad (1)$$

Doğrusal regresyon algoritması girdiler için çıktıları tahmin etmeye çalışmaktadır. Her denemede hata değerini hesaplayarak bir sonraki denemede bu değeri düşürmeye çalışmaktadır. Algoritma β0 sabitini -0.00514807 olarak β1, β2, β3, β4 ve β5 ağırlıklarını sırasıyla 0.6885489, 0.6766748, 0.08699996, 1.32685881, 2.17248983 olarak hesaplamıştır.

3.1.2. Karar Ağacı Regresyon (Decision Tree Regression-DTR)

Karar ağacı yapısı ilk olarak 1986 yılında Quinlan tarafından yayınlanmıştır [13]. Karar ağacı karar düğümlerinden ve yaprak düğümlerinden oluşmaktadır. Düğüm hesaplamalarında regresyon işlemi yapılacaksa bilgi kazanımı yerine standart sapma (Eş. (2)) kullanılmaktadır. Öncelikle hedef kümesinin standart sapması hesaplanmaktadır. Sonra diğer kümelerle hedef kümeler arasında ikili standart sapma değerleri (Eş. (3)) hesaplanmaktadır ve her birisinin sonucu hedef kümesinin standart sapma değerinden (Eş. (4)) çıkarılmaktadır. SDR değeri en büyük olan küme kök olarak belirlenmektedir. Bu adımlar her bir düğüm için devam ettirilerek ağaç yapısı oluşturulmaktadır.

$$S = \sqrt{\frac{\sum(x-\mu)^2}{n}} \qquad (2)$$

$$S(T, X) = \sum_{c \in X} P(c) S(c) \qquad (3)$$

$$SDR(T, X) = S(T) - S(T, X) \qquad (4)$$

3.1.3. K-En Yakın Komşu Regresyonu (K-Nearest Neighbor Regression- k-NNR)

Algoritma her bir test verisi için eğitim verileriyle olan uzaklık mesafesini hesaplamaktadır. Uzaklık mesafesi aşağıda verilen öklid uzaklık fonksiyonu (Eş. (5)) manhatın uzaklık fonksiyonu (Eş. (6)) veya minkowski uzaklık fonksiyonundan (Eş. (7)) istenen birisi ile hesaplanabilmektedir [14]. Sonra belirlenen k tane en yakın komşunun hedef verisi test verisinin sonucu olmaktadır.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (5)$$

$$\sum_{i=1}^{k}|x_i - y_i| \quad (6)$$

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{\frac{1}{q}} \quad (7)$$

3.1.4. Destek Vektör Regresyon (Support Vector Regression- SVR)

Destek vektör makineleri ilk olarak Vapnik tarafından ortaya atılmıştır [15]. Destek vektör regresyon ise Smola tarafından geliştirilmiştir [16]. Destek vektör regresyonunda amaç birbirleriyle aynı özellikler taşıyan özniteliklerin, özniteliklere en yakından geçen bir doğru yardımıyla ayrılmasıdır (Eş. (8), Eş. (9), Eş. (10), Eş. (11)). Doğrusal olarak ayrılamayan öznitelikler olduğu durumlarda farklı çekirdek (kernel) fonksiyonları yardımıyla ayrılabilmesi sağlanabilmektedir (Eş. (12), Eş. (13), Eş. (14), Eş. (15)).

$$y = wx + b \quad (8)$$

$$\min \frac{1}{2}\|w\|^2 \quad (9)$$

$$y_i - wx_i - b \leq \varepsilon \quad (10)$$

$$wx_i + b - y_i \leq \varepsilon \quad (11)$$

$$y = \sum_{i=1}^{N}(a_i - a_i^*)\,K(x_i, x) + b \quad (12)$$

$$k(x, x') = \langle x, x'\rangle^P \quad (13)$$

$$k(x, x') = (\langle x, x'\rangle + c)^P \quad (14)$$

$$k(x, x') = \tanh(\vartheta + \kappa\langle x, x'\rangle) \quad (15)$$

3.1.5. Yapay Sinir Ağı Regresyon (Neural Network Regression- NNR)

Giriş katmanından, n sayıda nöron içerebilen bir veya daha fazla gizli katmandan ve çıkış katmanından oluşan bir yapıdır. Katmandaki nöronlar birbirlerine bağlıdırlar. Katmanlardaki nöronlar bir ağırlık katsayısına ve aktarım fonksiyonuna sahiptir.

3.2.3. Uyumlu Artırıcı Regresyon (Adaboost Regression- ADBR)

Temel regresyon algoritması olarak kullanılan karar ağacı regresyonu veri seti ile eğitilmektedir. Bir sonraki regresyon işlemi yapılırken ilk regresyonda yanlış tahmin edilen verilere ait eğitim verilerinin göreceli ağırlığı artırılarak eğitim işlemine devam edilmektedir [20]. Ağırlıklar güncellenerek durdurma şartı oluşana kadar regresyon işlemine devam edilmektedir.

3.2.4. Eğimli Artırıcı Regresyon (Gradient Boosting Regression- GRBR)

Temel regresyon modeli veri seti ile eğitilmektedir. Tahmin hata değerleri bir sonraki tahmin değerlerine eklenmektedir. Gradyan iniş kullanarak ve tahminleri öğrenme oranına göre güncelleyerek, hata değerinin en az olduğu model bulunmaya çalışılmaktadır [21].

3.2.5. Örneklemeli Toplam Regresyon (Bagging Regression- BAGR)

Bu yöntemde ise aynı algoritma verinin farklı altkümeleri üzerinde çalıştırılmaktadır [22]. Altkümeleri oluştururken, örnekleme işlemi yerine koyma ile yapılırsa bu yöntem bagging olarak adlandırılmaktadır. Tüm tahmin ediciler eğitildikten sonra, tahminleri birleştirmek için regresyon probleminde tahminlerin ortalaması kullanılmaktadır.

Giriş katmanından gelen bilgi çıkışa doğru ilerlerken nöronlardaki ağırlıklar ile çarpılarak aktarım fonksiyonundan Eş. (17), Eş. (18), Eş. (19) geçirilir ve diğer nöronlara aktarılır. Ağ yapısı geri yayılımlı ise her eğitim döngüsünde ağdaki nöronların ağırlık değerleri güncellenir [17]. Algoritmada giriş katmanında 5 nöron, gizli katmanda 100 nöron ve çıkış katmanında 1 nöron bulunmaktadır. Doğrusal aktarım fonksiyonu kullanılmıştır.

$$y = f(w1.X1 + w2.X2 + b) \quad (16)$$

$$y = x \quad (17)$$

$$y = \frac{1}{1+e^{-x}} \quad (18)$$

$$y = \frac{1-e^{-2x}}{1+e^{2x}} \quad (19)$$

3.2. Topluluk Tabanlı Öte (Meta) Algoritmalar

3.2.1. Rastgele Orman Regresyon (Random Forest Regression- RFR)

Birden fazla sayıda karar ağacı oluşturularak bunların içerisinden en iyi sonucu verenin seçilmesidir [18]. Çalışmada 30 adet karar ağacı oluşturularak bunlar içerisinden en iyi olan seçilmiştir.

3.2.2. Yığın Regresyon (Stacking Regression- STCKR)

Birden fazla regresyon modelinin birlikte kullanılarak tahmin sonucunun iyileştirilmesinde kullanılan bir regresyon algoritmasıdır [19]. Öncelikle birden fazla regresyon algoritması tüm veri seti kullanılarak eğitilmekte ve tahminler üretilmektedir. Tahminler ile önceki veri seti kullanılarak yeni bir veri seti üretilmektedir. Son olarak bir regresyon algoritması yeni veri seti ile son tahminleri üretmede kullanılmaktadır.

4. DOĞRULUK BAŞARI DEĞERLERİNİ HESAPLAMA YÖNTEMLERİ

Çalışmada kullanılan regresyon algoritmaların başarı sonuçları değerlendirilirken belirlilik katsayısı (r^2), ortalama mutlak hata (OMH- mean absolute error (MAE)) ve ortalama karesel hata (OKH – mean squared error (MSE)) ölçütleri kullanılmıştır. Eş. (20) ile ifade edilen belirlilik katsayısı, regresyon kareler toplamının genel kareler toplamına bölümüyle elde edilen istatistiksel bir sonuçtur. Eğitim verileri ile eğitilmiş algoritmaların test verileri için ürettikleri tahminlerin gerçek değerlere ne ölçüde yakın olduğunun bir göstergesidir. Değer ne kadar 1'e yakın ise regresyon algoritması o derece başarılıdır.

$$r^2 = \frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2} = 1 - \frac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2} \quad (20)$$

Ortalama mutlak hata, test verisindeki gerçek değerler ile tahmin değerleri arasındaki farkların mutlak değerinin ortalamasıdır (Eş. (21)). Değerin 0'a yakın olması hatanın düşük olduğunu ifade etmektedir.

$$\frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \quad (21)$$

Ortalama karesel hata, test verisindeki gerçek değerler ile tahmin değerleri arasındaki farkların karelerinin ortalamasıdır (Eş. (22)).

Değerin 0'a yakın olması hatanın düşük olduğunu ifade etmektedir.

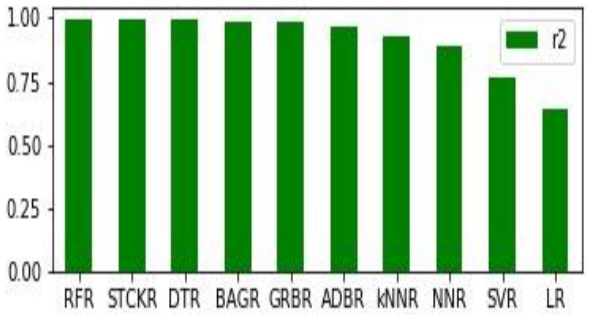
$$\frac{1}{n}\sum_{j=1}^{n}\left(y_j-\hat{y}_i\right)^2 \quad (22)$$

5. BULGULAR VE TARTIŞMA

Çalışmada kullanılan regresyon algoritmalarının belirlilik katsayısı (r^2), ortalama mutlak hata (OMH- mean absolute error (MAE)), ortalama karesel hata (OKH – mean squared error (MSE)) ve saniye cinsinden işlem süresi değerleri büyükten küçüğe sıralı olarak Tablo 3.'de verilmiştir.

Tablo 3. Regresyon Algoritmalarının Başarı Değerleri

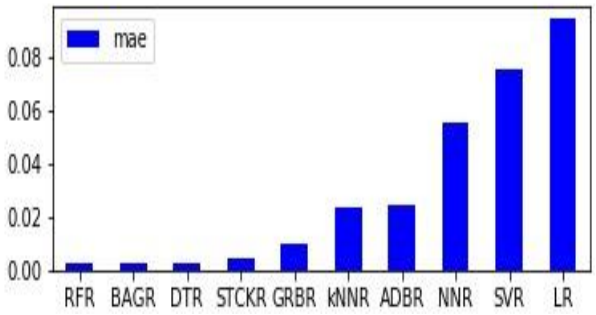
Regresyon Algoritması	r^2	mae	mse	süre
Rastgele Orman (RFR)	**0.99049**	**0.00275**	**0.00037**	1.12502
Yığın (STCKR)	0.99047	0.00442	**0.00037**	21.42568
Karar Ağacı (DTR)	0.99036	0.00293	0.00038	0.04088
Örneklemeli Toplam (BAGR)	0.99006	0.00283	0.00039	1.01229
Eğimli Artırıcı (GRBR)	0.98704	0.00967	0.00051	0.33809
Uyumlu Artırıcı (ADBR)	0.96364	0.02447	0.00142	0.57346
K-En Yakın Komşu (KNNR)	0.92685	0.02331	0.00286	0.29321
Yapay Sinir Ağı (NNR)	0.88783	0.05604	0.00439	1.08011
Destek Vektör (SVR)	0.76300	0.07609	0.00927	13.88685
Doğrusal (LR)	0.64496	0.09443	0.01389	**0.00699**



Şekil 2. Regresyon Algoritmalarının Belirlilik Katsayıları Kıyaslaması

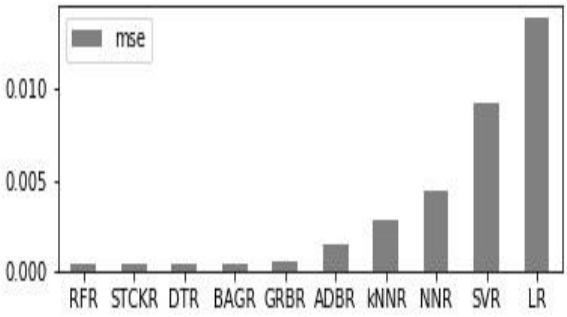
Şekil 2.'de regresyon algoritmalarının belirlilik katsayılarının (r^2) kıyaslama grafiği verilmiştir. Belirlilik katsayı değerinin 1'e yakın olması algoritmanın başarısının yüksek olduğunu ifade etmektedir. Çalışma sonucunda en iyi regresyon algoritması 0.99049 r^2 değeriyle rastgele orman regresyonu olmuştur.

Şekil 3.'de regresyon algoritmalarının ortalama mutlak hata (OMH) değerlerinin kıyaslama grafiği verilmiştir. Ortalama mutlak hata değerinin 0'a yakın olması algoritmanın başarısının yüksek olduğunu ifade etmektedir.



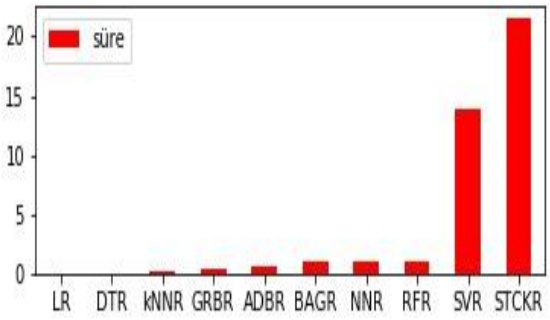
Şekil 3. Regresyon Algoritmalarının Ortalama Mutlak Hata Değerleri Kıyaslaması

Şekil 4.'de regresyon algoritmalarının ortalama karesel hata (OKH) değerlerinin kıyaslama grafiği verilmiştir. Ortalama karesel hata değerinin 0'a yakın olması algoritmanın başarısının yüksek olduğunu ifade etmektedir.



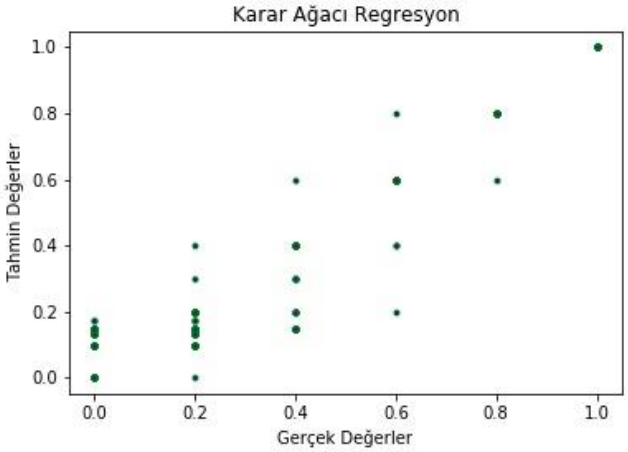
Şekil 4. Regresyon Algoritmalarının Ortalama Karesel Hata Değerleri Kıyasları

Şekil 5.'de regresyon algoritmalarının saniye cinsinden hesaplama süreleri kıyaslama grafiği verilmiştir. Destek vektör ve yığın regresyon dışındaki algoritmalar oldukça hızlı bir şekilde sonuç üretmişlerdir.



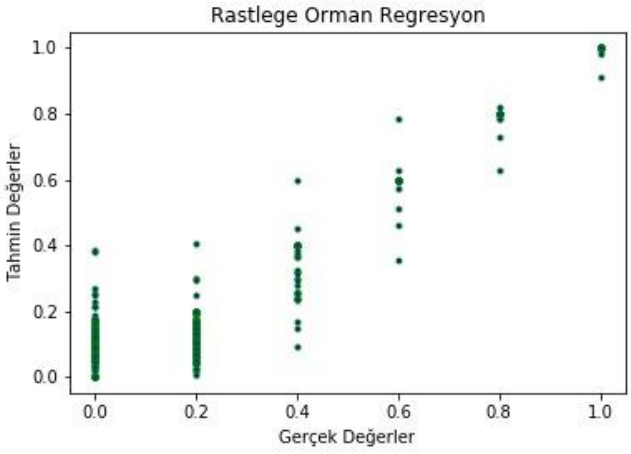
Şekil 5. Regresyon Algoritmalarının Hesaplama Süreleri Bakımından Kıyasları

Eğitim verileri ile eğitilmiş regresyon algoritmalarına test verileri uygulanmış, algoritmaların tahmin ettiği hava kalitesi indeks değerleri ile test verisindeki gerçek indeks değerlerini bir arada gösteren grafikler sırasıyla verilmiştir (Şekil (6.a., 6.b., 6.c., 6.d., 6.e., 6.f., 6.g., 6.h., 6.ı, 6.i)).

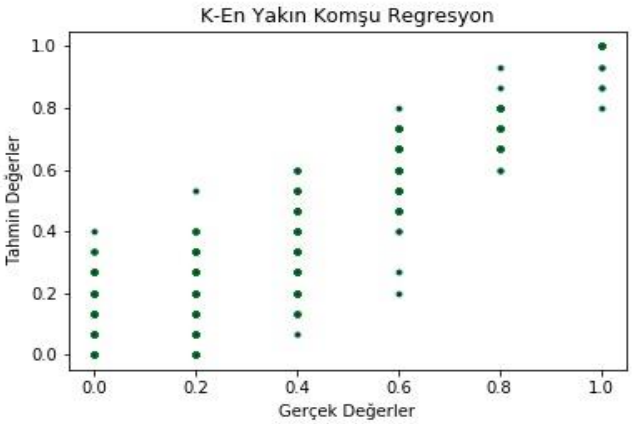


Şekil 6.a. Karar Ağacı Regresyon Değerleri

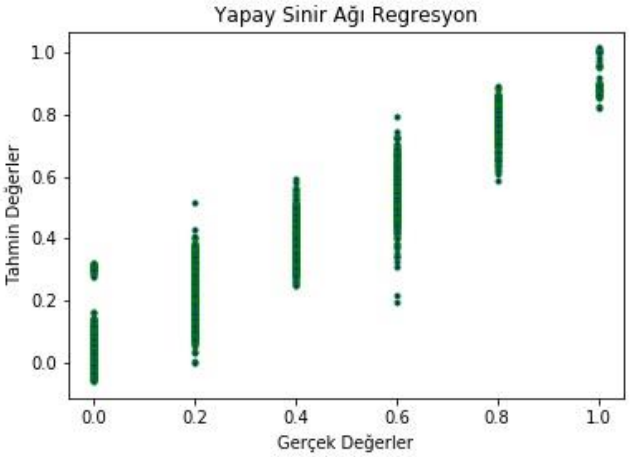
Bu grafiklerde gerçek değerler 0, 0.2, 0.4, 0.6, 0.8 ve 1 olarak görülmektedir. Bunun sebebi hava kalite indeks değerleri olan 1, 2, 3, 4, 5 ve 6 değerlerinin 0-1 arasında normalleştirilmiş olmasındandır.



Şekil 6.b. Rastgele Orman Regresyonu Değerleri

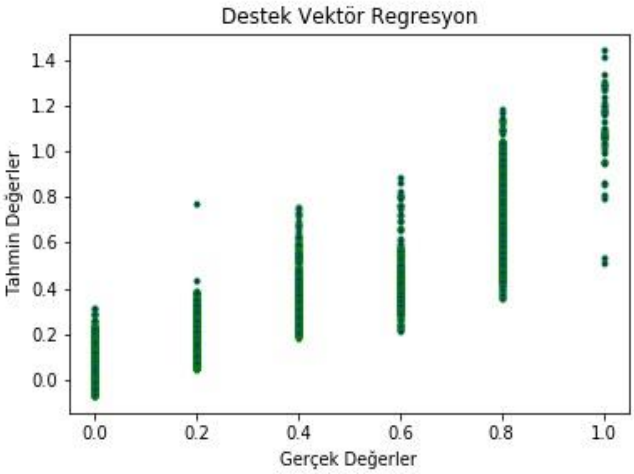


Şekil 6.c. K En Yakın Komşu Regresyonu Değerleri

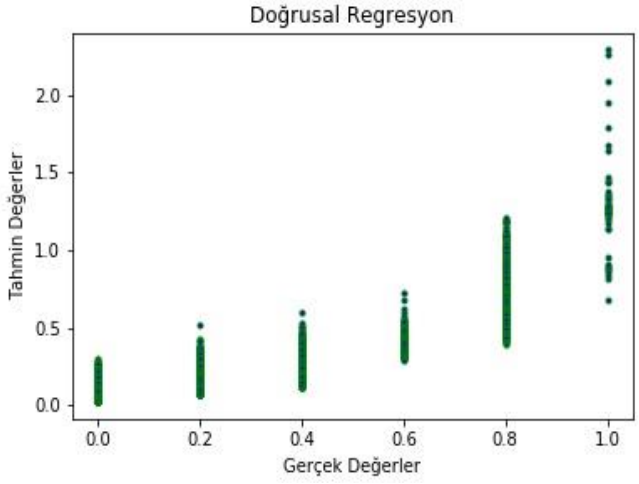


Şekil 6.d. Yapay Sinir Ağı Regresyonu Değerleri

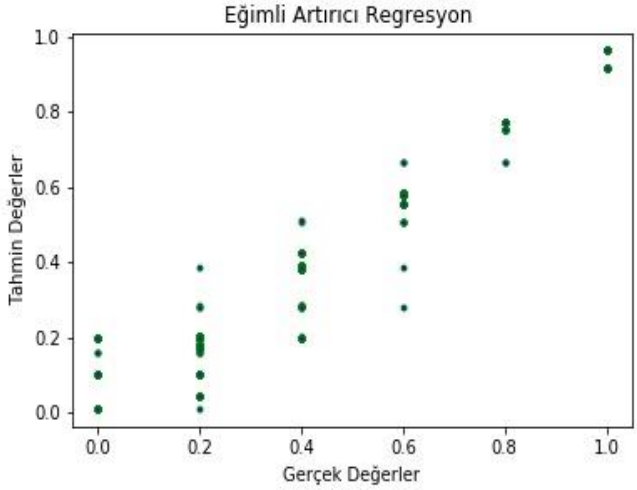
Şekillerde gerçek değerlere karşılık tahmin değerlerinin dağılımı görülmektedir. Örneğin gerçekte 0.4 olan değerlere ait tahminlerin 0.4 ve etrafında dağıldığı görülmektedir. Gerçek değerlere karşılık tahmin değerlerinin tek bir noktada toplanmadığı ve dağılım gösterdikleri şekillerde görülmektedir. Bu farklar hataları ifade etmektedir



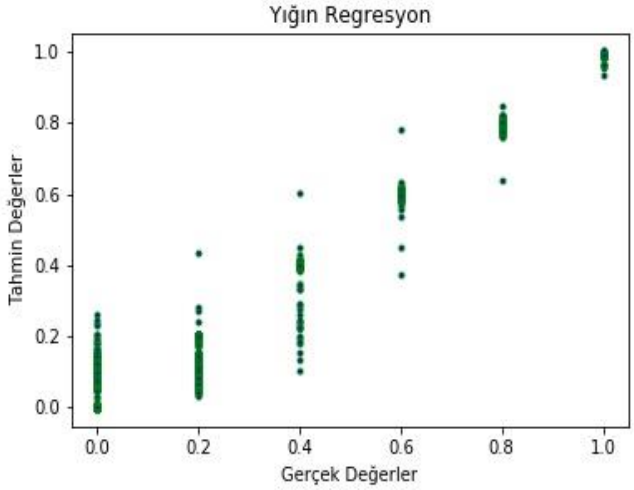
Şekil 6.e. Destek Vektör Regresyonu Değerleri



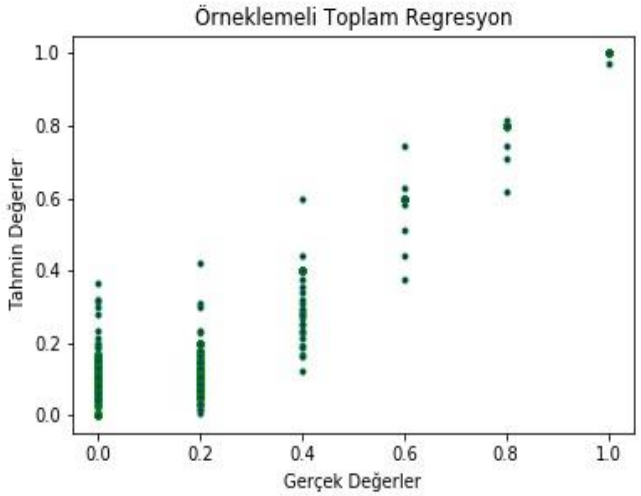
Şekil 6.f. Doğrusal Regresyon Değerleri



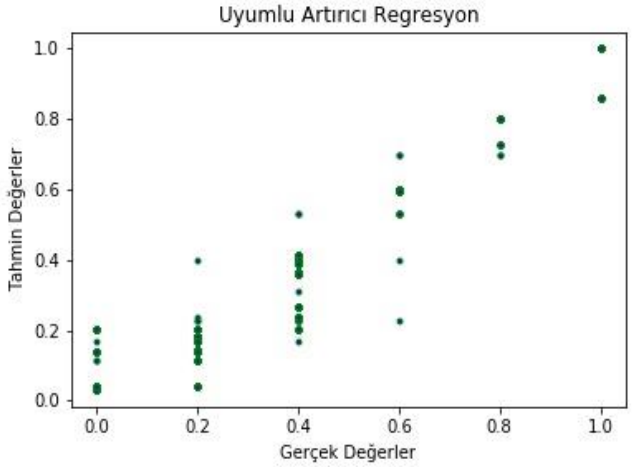
Şekil 6.ı. Eğimli Artırıcı Regresyonu Değerleri



Şekil 6.g. Yığın Regresyonu Değerleri



Şekil 6.i. Örneklemeli Toplam Regresyon Değerleri



Şekil 6.h. Uyumlu Artırıcı Regresyonu Değerleri

Veljanovska ve Dimoski [2] yaptıkları çalışmada sınıflandırma algoritmaları kullanmışlardır. Destek vektör makineleri ile yapay sinir ağı sınıflandırmasının başarı sonuçları bu çalışmadaki sonuçlar ile birbirine yakın bulunmuştur. K- en yakın komşu ve karar ağaçları sınıflandırmasının başarı sonuçları ise bu çalışmada kullanılan k- en yakın komşu ve karar ağaçları regresyona göre %15 daha düşük bulunmuştur. Bu farkın veri setindeki örnek sayısından ziyade algoritma yapılarındaki farklılıklardan kaynaklandığı düşünülmektedir. Dragomir [3] yaptığı çalışmada k- en yakın komşu sınıflandırması ile 1 aylık veri seti üzerinde hava kalite indeksi tahmini yapmıştır. Bulduğu başarı sonucu, bu çalışmada kullanılan k- en yakın komşu regresyon başarı sonucundan %35 daha düşük bulunmuştur. Bu farkın oluşmasında kullanılan algoritma farklılığının yanı sıra veri setinin küçük boyutta olmasının da etkili olduğu düşünülmektedir. Zhai ve Chen [6] yaptıkları çalışmada hava kalite indeksini tahminde geri yayılımlı sinir ağını ve genetik algoritma ile sinir ağını birlikte kullanmışlardır. Belirlilik katsayılarını (r^2) sırasıyla 0.72 ve 0.75 olarak hesaplamışlardır. Bu çalışmada kullanılan yapay sinir ağı regresyon yönteminin belirlilik katsayısı 0.88 olarak

hesaplanmış ve daha iyi sonuç ürettiği gözlenmiştir. Wang ve ark. [7] yaptıkları çalışmada hava kalite indeksini hesaplamada otoregresif entegre hareketli ortalamayı ve bulanık zaman serilerini kullanmışlar ve ortalama mutlak hata değerlerini sırasıyla 14.43 ve 10.80 olarak hesaplamışlardır.

6. SONUÇ

Bu çalışmada hava kalite indeksinin doğru tahmin edilebilmesi için regresyon temelli algoritmalar kullanılmıştır. Adana ili örneği üzerinde yapılan çalışmalar, belirlilik katsayısı (r^2) bakımından kıyaslandığında hava kalite indeksini en iyi tahmin edebilen algoritma rastgele orman regresyonu olmuştur. Topluluk tabanlı regresyon algoritmaları diğer algoritmalara göre daha başarılı sonuçlar üretmişlerdir. Tablo 3.'den görülebileceği üzere belirlilik katsayısı yüksek olan regresyon algoritmalarının hata değerleri olan ortalama mutlak hata (OMH) ve ortalama karesel hata (OKH) değerleri 0'a daha yakın bulunmuştur. Süre bakımından yığın regresyonu 21.42568 saniye, destek vektör regresyonu ise 13.88685 saniye ile diğer algoritmalara göre daha uzun işlem sürelerine sahiptir. En hızlı çalışan algoritma ise 0.00699 saniye ile doğrusal regresyon olmuştur. Bu çalışmadaki regresyon algoritmalarının ortalama mutlak hata ve ortalama karesel hata değerleri literatürde hava indeksi hesaplamalarında kullanılan diğer yöntemlere oranla düşüktür ve bu regresyon algoritmaları ile daha iyi sonuçlar üretildiğini göstermektedir.

REFERANSLAR

[1] L.H. Tecer, "Hava Kirliliği ve Sağlığımız. Bilim ve Aklın Aydınlığında Eğitim", S. 135, ss. 15-29., Mayıs 2011.
[2] K. Veljanovska and A. Dimoski, "Air Quality Index Prediction Using Simple Machine Learning Algorithms", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 7, Issue 1, pp. 025-030, ISSN 2278-6856, January - February 2018.
[3] E.A. Dragomir, "Air Quality Index Prediction using K-Nearest Neighbor Technique", BULETINUL Universităţii Petrol – Gaze din Ploieşti, Volume 62, No 1, pp. 103 – 108, 2010.
[4] M.D. Adams et al., "Air Quality Health Index Mapping: A Data Driven Modelling Approach", Proceedings of the 13th International Conference on Environmental Science and Technology Athens, Greece, 5-7 September 2013.
[5] R. Raturi and J.R. Prasad, "Recognition of Future Air Quality Index Using Artificial Neural Network", International Research Journal of Engineering and Technology (IRJET), Volume: 05, Issue: 03, e-ISSN: 2395-0056, 2018.
[6] B. Zhai and J. Chen, "Research on the forecasting of Air Quality Index (AQI) based on FS-GABPNN: A case study of Beijing, China", Proceedings of the 14th ISCRAM Conference – Albi, France, May 2017.
[7] H. Wang et al, "Air Quality Index Forecast Based on Fuzzy Time Series Models", Journal of Residuals Science & Technology, Vol. 13, No. 5, doi:10.12783/issn.1544-8053/13/5/161, 2016.
[8] http://www.havaizleme.gov.tr (Nisan 2018'de erişildi)
[9] https://www3.epa.gov/airnow/aqi_brochure_02_14.pdf (Ekim 2018'de erişildi)
[10] http://www.havaizleme.gov.tr/home/HKI (Nisan 2018'de erişildi)
[11] https://www.spyder-ide.org (Nisan 2018'de erişildi)
[12] J.M. Stanton , "Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors", Journal of Statistics Education, 9:3, DOI: 10.1080/10691898.2001.11910537, 2017.
[13] J.R. QUINLAN, Machine Learning 1: 81-106, 1986
[14] K. Alkhatib et al., "Stock Price Prediction Using K-Nearest Neighbor (k-NN) Algorithm", Int. J. Bus. Humanit. Technol., vol. 3, no. 3, pp. 32–44, March., 2013.
[15] V. Vapnik, The nature of statistical learning theory, Springer-Verlag, New York, 2000.
[16] A.J. Smola and B. Schölkopf, A tutorial on support vector regression, Statistics and Computing, 14 (3), 199-222, 2004.
[17] F. Murtagh, "Multilayer perceptron for classification and regression", Neurocomputing, Volume 2, Issues 5-6, Pages 183-197, doi.org/10.1016/0925-2312(91)90023-5, 1991.
[18] L. Breiman, "Random forests". Machine Learning, 45 (1): s.5-32., 2001.
[19] L. Breiman, "Stacked regressions". Machine learning, 24.1. 49-64, 1996.
[20] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting", 1995.
[21] J.H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", 1999.
[22] L. Breiman, "Bagging predictors", Machine Learning, 24(2), 123-140, 1996.