Int. J. Adv. Eng. Pure Sci. 2023, 35(1): 39-53
DOI: 10.7240/jeps.1175507

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Makine Öğrenmesi Yöntemleri ile Şehirlerin Hava Kalitesi Tahmini

Air Quality Prediction of Cities with Machine Learning Techniques

Mehtap ÖKLÜ [1], Pelin CANBAY [1]

[1]Kahramanmaraş Sütçü İmam Üniversitesi (KSÜ), Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Kahramanmaraş, Türkiye

Öz

Hava Kalite İndeksi (HKİ), Avrupa standartları çerçevesinde yer alan beş temel kirletici unsur (CO, SO2, NO2, O3 ve PM10) göz önünde bulundurularak değerlendirilen bir endekstir. Bu endeks ile şehirlerdeki kirlilik miktarları hakkında bilgi elde edilebilmekte ve şehirlerin daha temiz şehirlere dönüşmesi için çalışmalar yapılabilmektedir. Günümüzde bu ölçümlere gerekli önem verilmemekte, yeterli miktarda ve doğrulukta bu ölçümler yapılamamaktadır. Çalışmamızda, şehirlerin kirlilik oranına göre sınıflandırılabilmesi ve böylece kirlilik durumu kritik seviyede olan şehirlerin kısa sürede belirlenebilmesi amaçlanmıştır. Bu amaç doğrultusunda City-Data platformundaki 6000 üzeri nüfusa sahip Amerika Birleşik Devletleri (ABD) şehirlerinin hava kalitesi belirleyicileri olarak değerlendirilebilecek, şehirlerin hava kalitesine etkisi olan farklı parametreleri toplanarak bir araya getirilmiş, HKİ verileri ile birlikte veri seti olarak kullanılmıştır. Şehrin nüfusu, betonarme yapı sayısı, yeşil alan ve kullanılan ulaşım araç oranlarının da belirleyici olarak kullanıldığı çalışmamızda hava kalitesi 3 ve 5 sınıflı sınıflandırma problemi olarak ayrı ayrı ele alınmıştır. Çalışmamızda, HKİ değerinin insan sağlığına etki oranları hesaplanarak sınıf atamaları yapılmıştır. Makine öğrenmesi yöntemlerini kullanarak sunduğumuz çözümlerde hava kalitesi tahmini 3 sınıflı modellerde %87 oranında, 5 sınıflı modellerde ise %82 oranında başarılı sonuçlar üretmiştir.

Anahtar Kelimeler: Hava Kalitesi, Hava Kalite İndeksi (HKİ), Makine Öğrenmesi, Sınıflandırma

Abstract

The Air Quality Index (AQI) is evaluated by considering the five main pollutants (CO, SO2, NO2, O3 and PM10) within the framework of European standards. With this index, information about the amount of pollution in cities can be obtained and studies can be carried out to transform cities into cleaner cities. Today, these measurements are not given the necessary importance, and sufficient and accurate measurements cannot be made. Our study, it is aimed to classify the cities according to the pollution rate, and thus to determine the cities with critical pollution status in a short time. For this purpose, different parameters that can be considered as air quality determinants of the United States (USA) cities with a population of over 6000 from the City-Data platform, which have an impact on the air quality of the cities, were collected and used as a data set together with the AQI data. In our study, in which the population of the city, the number of reinforced concrete structures, the green area and the ratio of the transportation vehicles used are used as determinants, the air quality is handled as a classification problem separately with 3 and 5 classes. In our study, class assignments were made by calculating the effects of AQI value on human health. In the solutions we offer using machine learning methods, air quality prediction has produced 87% successful results in 3-class models and 82% in 5-class models.

Keywords: Air Quality, Air Quality Index (AQI), Machine Learning, Classification

I. GİRİŞ

Hava kirliliği, canlıların sağlığını ve yaşam kalitesini büyük oranda tehdit eden bir unsurdur [1] ve birçok zararlı etkiye neden olduğu için sürekli olarak izlenmelidir. İzlemenin en etkili yolu kirliliğin kaynağını, kökenini iyi bilmektir. Hava kirliliğinin seviyesi ölçülürken havayı kirleten bazı hava kirletici gazlar göz önünde bulundurulur. Başlıca hava kirleticileri; PM2.5, PM10, NO2, O3, CO, SO2'dir. Hava kalitesini etkileyen daha birçok verinin bulunduğu City-Data [2] platformu, HKİ değerini takip etmek için önemli kaynaklardan biridir. 18. yüzyılın sonlarından itibaren, sanayi devriminin de yaygınlaşmasıyla beraber yakıt ihtiyacı yüksek oranda artmıştır. Dönemin şartlarından dolayı en çok tercih edilen yakıt türü fosil yakıtlar olmuştur. Yüksek oranda fosil yakıtların (kömür, petrol ve doğal gaz) kullanımı sera gazlarının artmasına [3] ve dolayısıyla: hava kirliliği, buzulların erimesi, kuraklık, aşırı hava olayları, ekosistemde tahribat, vektörel hastalıklarda artış ve bazı canlıların neslinin tükenmesi gibi daha birçok soruna yol açmıştır [4]. Sera gazları dünyayı çevreleyerek dünyadan dışarı gitmesi gereken kızılötesi radyasyonu tutup, bu radyasyonun dışarı çıkmasını engelleyerek, dünyanın gitgide daha fazla ısınmasına da yol açmaktadır [5]. Bu sorunlar gibi birçok unsur günümüze kadar artış göstermeye devam etmiştir. Bu durumun en büyük sebepleri ise; fosil yakıtların

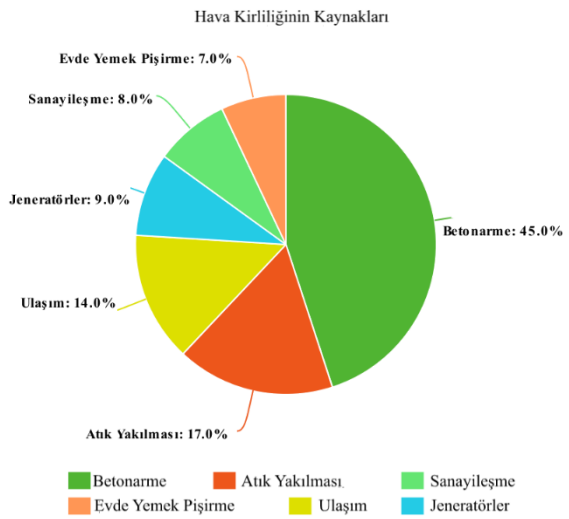
Sorumlu yazar: PELIN CANBAY, Tel: 0344 300 17 01-dahili, e-posta: pelincanbay@ksu.edu.tr

Gönderilme: 15.09.2022, **Düzenleme:** 28.12.2022, **Kabul:** 08.01.2023

kontrolsüz kullanımı, sanayileşmenin artması ve betonarme yapıların devamlı olarak artmasıdır [6]. Hava kirliliği sorunu küresel iklim sorununun temelidir. Küresel iklim sorunu 20. yüzyılın son çeyreğinden itibaren artarak ivmeli bir şekilde dünya gündeminde yer almaya başlamıştır [3]. Bu tehlikeyi önlemek için dünya çapında bir araya gelinip acil durum planları geliştirilmeli ve uygulanmalıdır. Aksi takdirde küresel iklim krizi daha tehlikeli bir hal almaya devam edecektir [5].

Yukarıda belirtilen etmenler her ne kadar insanların yaşam standardını yükselten etmenler olsa da çevreye ve insanlara çok büyük oranda zarar vermiştir, vermeye de devam etmektedir. Hava kirliliğindeki artış sonucunda insanlarda; karaciğer, böbrek ve beyinde uzun süreli hasar, akciğer kanseri ve doğum kusurları gibi yan etkiler yüksek oranda meydana gelmektedir [7]. Bu durum dünya, halk ve hükümetler için büyük bir problem olmaya devam etmektedir. Bu yüzden hava kirliliği daha da tehlikeli bir hal almadan etkisinin en aza indirilmesi acil bir ihtiyaçtır [8].

Dünya Sağlık Örgütü (DSÖ) verilerine göre [9], hava kirliliği sebebiyle yılda yaklaşık 7 milyon prematüre ölümü gerçekleşmektedir. Bu oran geçmiş yıllara göre daha yüksek olup her geçen yıl daha da artmaktadır. Hava kirliliğinin azaltılması hayati önem taşımakla birlikte toplumun tüm kesiminin bu amaç doğrultusunda bilinçlendirilmesi gerekmektedir. Hava kirliliğini en aza indirmek için öncelikle hava kalitesinin belirlenebilmesi gerekir. Şekil 1'de hava kalitesini en çok olumsuz etkileyen kaynakların dağılımı [9], [10] sunulmuştur.



Şekil 1. Hava kalitesini etkileyen faktörler

Özellikle büyük şehirlerdeki betonarme yapıların, sanayileşmenin, taşıt kullanımının ve atık yakılmasının sürekli artışı bu şehirlerin kontrol edilemez bir kirlilik artışına ortam sağladığı anlamına gelmektedir. Bu sürece müdahale edebilmek ve mevcut durum genelinde farkındalık oluşturabilmek adına bu çalışmada, şehirlerin hava kalitesinin tespit edilebileceği uygun belirleyicilerin seçimi ve bu belirleyicilerin yapay zekâ modelleri ile kullanımı önerilmiştir. Çalışmamızda, hava kalitesinin hangi öznitelikler ile belirlenebileceği ve hangi makine öğrenmesi yönteminin belirleyiciler üzerinde daha başarılı sonuçlar verdiği deneysel uygulamalar ile belirlenmiştir.

Bu çalışma kapsamında bir şehrin hava kalitesi belirleyicileri olarak; şehrin nüfusu, mil başına düşen insan sayısı, mil başına düşen betonarme yapı (ev) sayısı, evlerin kullanım oranı, şehir yüz ölçümünde yer alan verimli (yeşil) toprak alanı, insanların işe giderken tercih ettiği ulaşım şekli (yürümek, bisiklete binmek, özel araç kullanmak, toplu taşıma kullanmak) ve bahsi geçen kirletici unsurların (CO, SO_2, NO_2, O_3 ve PM_{10}) ölçüm değerleri göz önünde bulundurulmuştur. Çalışmamızda veri kümesi olarak nüfusu 6000 üzeri olan bazı ABD şehirlerinin HKİ değerleri ve bu şehirlerin hava kalitesini etkileyebilecek, yukarıda bahsedilen diğer bazı toplumsal verileri seçilerek kullanılmıştır. Veri kümesinde bulunan HKİ değerinin sürekli veri olması sebebiyle öncelikle bu değerin kategorileştirilmesi gerçekleştirilmiştir. Veri kümesindeki değerlerin dağılımı ve insan sağlığına etkisi göz önüne alınarak gerçekleştirilen kategorileştirme işlemi ile veriler 3 sınıflı (iyi, sağlıksız ve tehlikeli) ve 5 sınıflı (iyi, orta, sağlıksız, çok sağlıksız ve tehlikeli) olmak üzere 2 farklı yapıda ele alınmıştır.

Bu çalışmada, makine öğrenmesi yöntemleri kullanılarak veri kümelerine sınıflandırma uygulanmıştır. K-En Yakın Komşuluk, Destek Vektör Makineleri, Lojistik Regresyon, Naive Bayes, Karar Ağaçları, Torbalama, Rastgele Orman ve Yapay Sinir Ağları algoritmaları sınıflandırıcı modeller oluşturmak için kullanılmıştır.

Çalışmanın bir sonraki alt bölümünde küresel iklim krizi ve hava kirliliği üzerine yapılan önemli çalışmaların bir kısmı özetlenmiştir. İkinci bölümde çalışmada kullanılan materyal ve metotların tanımı ve açıklamasına yer verilmiş, üçüncü bölümde ise deneysel uygulamalardan elde edilen sonuçlar sunulmuştur. Son bölümde ise çalışmanın sonucu, tartışma ve ileriki çalışmalar ele alınmıştır.

1.1. İlgili Çalışmalar

Küresel iklim krizi ve beraberinde karşılaşılan sorunların çözülmesine yönelik birçok bilimsel çalışma yürütülmüştür. Yürütülen çalışmaların büyük bir kısmı hava kirliliği ile alakalıdır. Çalışmamız kapsamında ele aldığımız, hava kirliliği kaynaklı

küresel iklim krizinin etkilerinin azaltılmasını hedefleyen, ön plana çıkmış bazı çalışmaların içeriği ve çıktıları aşağıda özetlenmiştir.

Janarthanan, R ve arkadaşları tarafından yapılan çalışmada [11], Hindistan'ın Chennai şehrinde devreye alınan Ulusal Hava İzleme Programı kapsamında 240 şehirde bulunan 342 ölçüm noktası ile gerekli veriler toplanmış, toplanan veriler Destek Vektör Makineleri ve LSTM (Uzun-Kısa Süreli Bellek) modelleri ile eğitilmiştir. Çalışmada elde edilen değerler; güvenli, orta, hassas, sağlıksız, çok sağlıksız ve tehlikeli olarak derecelendirilmiştir. Yapılan çalışmadan çıkan sonuçlar neticesinde, şehrin olumsuz hava kalitesine sahip olması durumu; insanları toplu taşıma kullanmaya, ağaçlandırma çalışmaları yapmaya ve sanayi atık filtrelerini düzenli olarak değiştirmeye teşvik etmiştir.

Hindistan'daki bir diğer çalışma [8] ise Mahalingam, U. ve arkadaşları tarafından Delhi şehri baz alınarak gerçekleştirilmiştir. Çalışmada kullanılması istenen veriler Merkezi Kirlilik Kontrolü Kurulu, Çevre Bakanlığı, Orman ve İklim Değişikliği ve Hindistan Hükümeti tarafından sağlanmıştır. Veri setindeki öznitelikler kullanılarak optimize edilen Yapay Sinir Ağları (YSA) ve Orta Gauss Destek Vektör Makineleri (Medium Gaussian Support Vector Machines) modelleri, diğer şehirler için de kullanılmaya hazır hale getirilmiş, modeller optimize edildikten sonra elde edilen en yüksek doğruluk oranları YSA için %91,62, Orta Gauss Destek Vektör Makineleri için %97,3 olarak elde edilmiştir.

Bir diğer Hindistan konumlu proje ise Bhalgat, P. ve arkadaşları tarafından geliştirilen, ortamdaki SO_2 konsantrasyonu, makine öğrenmesi algoritmaları ile tahmin edilmeye yönelik çalışmadır [12]. Çalışmada ayrıca zaman serisi algoritmaları ile yıllık ve aylık SO_2 değerleri de tahmin edilmeye çalışılmıştır. Çalışmada kullanılan AR (AutoRegressive) modelinin MSE (Ortalama Hata Karesi) değeri 166,358 olarak elde edilmiştir.

Pasupuleti, V. R. ve arkadaşlarının gerçekleştirdiği çalışmada [7], şehirlerin HKİ değeri tahmin edilmeye çalışılmıştır. Bu çalışmada Doğrusal Regresyon, Karar Ağacı ve Rastgele Orman gibi makine öğrenmesi teknikleri kullanılarak hava kalite endeksi tahmin edilmeye çalışılmıştır. Tahmin sonucunda en yüksek başarı değerini Rastgele Orman algoritması vermiştir. Nandigala Venkat Anurag, Y. ve arkadaşları ABD ve Moğolistan hükümeti tarafından başlatılan proje dâhilinde HKİ değerini tahmin etmeye yönelik bir çalışma [13] gerçekleştirmiştir. Bu çalışmada XGBoost Regressor, Karar Ağacı ve Çoklu Lineer Regresyon algoritmaları kullanılmış en başarılı sonucu 15,97 RMSE (Ortalama Hata Karesi Karekökü) değeri ile XGBoost algoritması vermiştir. Karar Ağacı 16,84, Çoklu Lineer Regresyon algoritması ise 18,72 RMSE değerini üretmiştir.

Tayvan'ın Zhongli, Changhua ve Fengshan bölgelerinden sırası ile 91672, 94453, 94145 adet veri temin edilen çalışmada [14] HKİ değeri tahmin edilmeye çalışılmıştır. Temin edilen veriler ile 1, 8 ve 24 saat önceden akut hava kirliliği olaylarında HKİ değerini tahmin etmek için Rastgele Orman, Adaboost, Destek Vektör Makineleri (DVM) ve YSA modelleri eğitilmiştir. En yüksek R^2 doğruluk skorunu ve en düşük RMSE ve MAE (Mean Absolute Error) hata skorlarını Adaboost algoritması vermiştir. Çalışma sonucunda en düşük doğruluk oranlarını çıkartan algoritma ise DVM olmuştur.

Kamal, M. M. ve arkadaşlarının geliştirdiği projede [15], Malezya'nın bazı eyaletlerinde Kasım 2004'ten beri toplanan veriler, YSA modellerinin eğitiminde kullanılmıştır. Bu çalışmada 2~12 arasında gizli katman ve 1000~11000 arasında iterasyon tercih edilmiş olup Jalan Tasek şehrinin Aralık ayı verisinde yüzde 64.23~99.99 arası bir başarı elde edilirken Nilai şehrinin Aralık ayı verisinden yüzde 97,19~99,99 arası bir performans alınmıştır.

ABD özelinde yapılan hava kalitesi tahmini ve değerlendirmesi çalışmaları da çoğunlukla en kalabalık şehirlerden biri olan Chicago şehrinin verileri üzerinden yapılmıştır [16], [17].

Çalışmamız kapsamında yapmış olduğumuz literatür araştırması sonucunda, hava kalitesi tespitine yönelik standart bir veri kümesi ve çözüm yönteminin ele alınmadığı, genellikle çalışmalarda yerel veriler üzerinden farklı algoritmalar ile çözümler sunulduğu gözlemlenmiştir.

Çalışmamızda, açık kaynak olarak elde edebileceğimiz en büyük hava kirliliği verisinin ve şehirlere özgü başka hava kirliliği belirleyici unsurlarının da erişilebilir olduğu veri tabanlarından faydalanılmış, başka şehirlerde de kullanılabilecek yapay zekâ modelleri üzerine yoğunlaşılmıştır. Çalışmamızda kullanılan verinin özellikleri ve yöntemler takip eden bölümlerde ayrıntılı olarak açıklanmıştır.

II. MATERYAL VE METOD

2.1. Veri Seti

Çalışma kapsamında kullandığımız veri seti City Data [2] web sitesi üzerinden toplanmıştır. Belirtilen kaynak hükümet ve özel kaynaklardan çeşitli verileri toplayıp analiz ederek ABD'deki her şehir için ayrıntılı ve bilgilendirici profiller oluşturmaktadır.

Şehirlerin suç oranlarından hava durumu akışına, iç ve dış göç oranlarından ekonomi politikalarına kadar çok çeşitli bilgiler barındıran kaynaktan, şehirlerdeki hava kirliliğini etkileyebilecek bazı parametrelerin [16] her biri özellikle seçilerek çalışmamız kapsamında bir araya getirilmiştir. Toplanan veriler ABD'nin 6000 ve üzeri nüfusa sahip olan bazı şehirlerini kapsamaktadır. Belirtilen kaynaktaki veriler, hava kalitesi endeksini şehirlere göre değerlendirebilmek için iki farklı belirleyici kümesi ile ele alınmıştır; (1) hava kirliliğine neden olan gazların oranlarına göre ve (2) şehirdeki nüfusun yaşam tercihleri ve özelliklerine göre. Veri kümesinde bulunan şehirlerdeki hava kirliliğine neden olan gazların hava kalitesi endeksine karşın değerlerinin örnek gösterimi Tablo 1'deki gibidir.

Tablo 1'de bulunan, hava kirletici gazlar olarak değerlendirilen $PM_{2.5}$, PM_{10}, NO_2, O_3, CO, SO_2 gazlarının karşılık geldiği HKİ değeri kullanılarak çalışmada ele alınacak veri setinin bir kısmı elde edilmiştir. Çalışmamızın özetinde de belirtildiği gibi bu değerlerin şehirler tarafından yeterli miktarda ve doğrulukta ele alınması zordur, bu zorluk veri kümemizdeki eksik değerlerden de görülebilmektedir. Çalışmamız kapsamında ele almak istediğimiz ikinci veri kümesi olan hava kalitesi belirleyici kümesi, şehirdeki nüfusun yaşam tercihleri ve bu tercihlerin özelliklerinin bir örnek gösterimi Tablo 2'de sunulmuştur.

Tablo 2'de örnek gösterimi bulunan, şehirlerdeki HKİ oranına karşın, nüfusun miktarı, nüfusun yoğunluğu, evleşme yoğunluğu, evleşme oranı, toprak alan, nüfusun ulaşım tercihleri çalışmamızda kullanılacak veri setinin bir parçası olarak ele alınmaktadır. Tablo 1 ve Tablo 2'de yer alan farklı öznitelikler (belirleyiciler) çalışmanın veri seti olarak birlikte kullanılacağından tek bir tabloda birleştirilmiştir. Şekillerde görüleceği üzere her bir satır bir şehir değerlerini barındırmaktadır, dolayısı ile birleştirme işlemi her iki veri kümesindeki şehirlerin doğru eşleştirilmesi ile yapılmıştır. Sonuç olarak iki farklı veri kümesindeki şehir öznitelikleri tek tabloda şehir ve HKİ değerlerinin eşleşmeleri kontrol edilerek birleştirilmiştir. Elde edilen veri setinin öznitelikleri ve bu özniteliklerin açıklamaları Tablo 3'te sunulmuştur.

Tablo 3'te yer alan özniteliklerden ilki şehir isim bilgisi olup hesaplamalarda kullanılmamıştır. Tabloda yer alan HKİ değeri ise, diğer öznitelikleri kullanarak doğru bir şekilde tahmin etmeyi amaçladığımız hedef özniteliğimizdir. Veri setinin makine öğrenmesi yöntemleri üzerinde uygulanabilmesi için bir takım ön işlemlerden geçmesi gerekmektedir. Bu aşamada yapılan ön işlemler ve gereklilikleri alt başlıklarda ele alınmıştır.

Tablo 1. Hava kirliliğine sebep olan gazların hava kalitesine karşı değerleri

City (Population)	Air Pollution - Carbon (CO) [ppm] level	Air Pollution - Lead (Pb) [µq/m3] level	Air Pollution - Nitrogen Dioxide (NO_2) [ppb] level	Air Pollution - Ozone [ppb] level	Air Pollution - Particulate Matter ($PM_{2.5}$) [µq/m3] level	Air Pollution - Particulate Matter (PM_{10}) [µq/m3]level	Air Pollution - Sulfur Dioxide (SO_2) [ppb] level	Air Pollution - Air Quality Index (AQI) level
Middle Island, NY (10483)	0,36	-	8,78	27,1	9,22	17	1,03	87,1
Ridge, NY (13336)	0,36	-	8,78	27,1	9,22	17	1,03	87,1
Prichard, AL (22399)	0,32	0	-	27,2	8,77	19,7	0,94	68,9
Fairhope, AL (17386)	0,32	0,01	-	27,2	8,77	19,7	0,94	68,9
Tanner-Williams, AL (60211)	0,32	0	-	27,2	8,77	19,7	0,94	68,9
Spanish Fort, AL (7629)	0,32	0	-	27,2	8,77	19,7	0,94	68,9
Point Clear, AL (2125)	0,32	0,01	-	27,2	8,77	19,7	0,94	68,9
Mobile, AL (194889)	0,32	0	-	27,2	8,77	19,7	0,94	68,9
Daphne, AL (23633)	0,32	0	-	27,2	8,77	19,7	0,94	68,9
Belle Fontaine, AL (608)	0,32	0	-	27,2	8,77	19,7	0,94	68,9
Chickasaw, Al (6010)	0,32	0	-	27,2	8,77	19,7	0,94	68,9
Loxley, AL (1703)	0,32	0,01	-	25,1	8,77	19,7	-	68,6
Mount Pleasant, NC (1740)	0,31	-	8,67	28,1	9,22	15,5	0,28	87,3
Enochville, NC (2925)	0,31	-	8,67	28,1	9,22	15,5	0,28	87,4
Township 6, Rimertown, NC (2747)	0,31	-	8,67	28,1	9,22	15,5	0,28	87,3
Concord, NC (83506)	0,31	-	8,67	28,1	9,22	16,7	0,28	89,8
Mooresville, NC (34887)	0,31	-	8,67	28,1	9,22	15,5	0,28	87,4

Tablo 2. Şehirdeki nüfusun yaşam tercihleri ve özelliklerinin hava kalitesine karşı değerleri

City	Population	Population Density (people per square mile)	Housing Density (houses/condos per square mile)	Houses Occupied (%)	Land Area (square miles)	Means of Transportation to work - Walked (%)	Means of Transportation to work - Bicycle (%)	Means of Transportation to work - Car Alone (%)	Means of Transportation to work - Bus (%)	Air pollution - Air Quality Index (AQI) level
Cornwall, NY (12646)	12646	471	181	95,3%	26,8	2,46%	0%	84,6%	0,75%	81
West Lake Stevens, WA (20049)	20049	2071	665	96,2%	9,68	0,8%	0,29%	80,4%	1,53%	81,3
Arlington, WA (18664)	18664	2467	600	95%	7,57	0,88%	0,22%	85,8%	1,02%	81,3
Fairview, GA (6769)	6769	902	380	93,3%	7,5	0,27%	0%	83,7%	0%	81,3
Boynton Ridge, GA (13326)	13326	425	144	94,4%	31,3	0,74%	0%	87%	0,25%	81,3
Amherst, OH (12112)	12112	1690	643	96,9%	7,17	0,41%	0,11%	84,8%	0,69%	81,3
Nolensville, TN (6213)	6213	655	107	100%	9,49	0,92%	0%	82,9%	0%	81,5
Smyrna, TN (43060)	43060	1886	438	96,1%	22,8	0,21%	0,24%	86,6%	0,03%	81,5
Brentwood, TN (40021)	40021	1155	229	98,2%	34,6	0,33%	0%	84,7%	0,09%	81,5
La Vergne, TN (34077)	34077	1374	282	93,9%	24,8	0,33%	0%	87,8%	0,24%	81,5
Almaville, TN (20899)	20899	282	74,3	96,5%	74	0,36%	0%	84%	0%	81,5

2.2. Ön İşleme

Kaynağından topladığımız ve uygun eşleştirmeler ile birleştirdiğimiz veri setinde, Tablo 1'de de görülebileceği üzere bazı eksik değerler bulunmaktadır. Bu eksik değerlerin özniteliğine göre yönetilmesi aşağıdaki adımlar ile sağlanmıştır:

- Toplanan veri setinde AQI değeri eksik olan 494 örnek veri setinden çıkarılmıştır. Böylece ilk aşamada 6875 olan örnek sayısı 6381 olmuştur.
- Kullanılacak diğer özniteliklerdeki eksik değerler gözlemlenmiştir. Çalışmanın veri toplama aşamasında, eksik değer sayısı toplam örnek sayısının yarısından fazla olan öznitelikler veri seti oluşturmada kullanılmamıştır.
- Çalışmada kullanılan veri setinde bulunan her öznitelikte eksik değer sayısı toplam örnek sayısının yarısından azdır. Bu aşamada veri setinde bulunan her özniteliğin eksik değerleri söz konusu özniteliğin mevcut olduğu örneklerdeki değerlerinin ortalaması ile doldurulmuştur. Bu öznitelikler hava kirliliği belirleyicilerinden olan gazların değerlerinin bulunduğu özniteliklerdir. Şehirdeki nüfusun yaşam tercihleri ve bu tercihlerin özelliklerinin değerlendirildiği ikinci veri kümesinde eksik veri bulunmamaktadır.

Tablo 3. Çalışmada kullanılan veri setinde bulunan öznitelikler ve açıklamaları

Öznitelik	Açıklama
City	Şehir adı
Population	Şehir nüfusu
Population Density	Şehirdeki mil kare başına düşen birey sayısı
Housing Density	Şehirdeki mil kare başına düşen ev/ apartman sayısı
Houses Occupied(%)	Şehirde bulunan evlerin kullanım oranı
Land Area(square miles)	Şehirde işlenmemiş, ellenmemiş topraklar
Means of Transportation to Work	Şehirde işe giderken tercih edilen ulaşım aracı
Means of Transportation to Work	*Walked (%)* Yürüyerek işe gidenlerin oranı
Means of Transportation to Work	*Bicycle (%)* Bisiklet sürerek işe gidenlerin oranı
Means of Transportation to Work	*Car alone (%)* Özel araç kullanarak işe gidenlerin oranı
Means of Transportation to Work	*Bus (%)* Toplu taşıma kullanarak işe gidenlerin oranı
Air Pollution	Hava kirliliğine neden olan gazlar
Air Pollution	Carbon Monoxide(CO) [ppm]
Air Pollution	Lead(Pb) [µg/m3]
Air Pollution	Nitrogen Dioxide(NO_2) [ppb]
Air Pollution	Ozone [ppb]
Air Pollution	Particulate Matter ($PM_{2.5}$) [µg/m3]
Air Pollution	Particulate Matter (PM_{10}) [µg/m3]
Air Pollution	Sulfur Dioxide (SO_2) [ppb]
Air Quality Index(AQI)	Hava kirliliğinin düzeyini belirten ölçü birimi (HKİ)

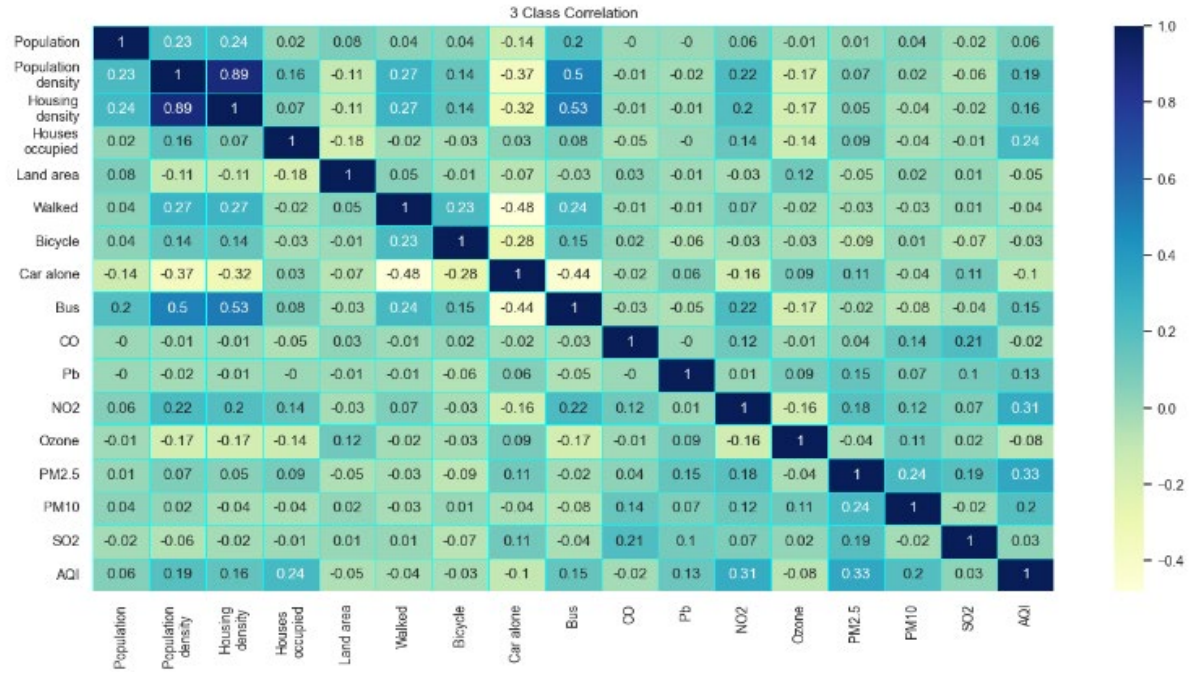
Veri setimizde bulunan özniteliklerin makine öğrenmesi yöntemleri ile kullanılabilecek yapıya getirilmesinin ardından hedef özniteliğin kategorileştirilmesi gerçekleştirilmiştir. Kategorileştirme işlemi şehirlerin hava kalite sınıflarını bilme ve karşılaştırma yapabilmesi için önemlidir. Böylece daha temiz bir şehir olmak ve daha kalitesi yüksek sınıflara geçebilmek için ayırt edici bir unsur olarak kategoriler ele alınabilecektir. Hedef özniteliğimiz olan AQI değeri sürekli veridir. Çalışmamızın amacı şehirlerin hava kalitesini sınıflandırmak ve hava kalitesi sınıflarını tahminlemek olduğundan, veri setindeki AQI değeri hem veri setindeki dağılımı hem de canlı sağlığına etkileri [19]-[21] göz önüne alınarak kategorileştirilmiştir. AQI değerlerini kategorilere ayırma işleminde 3 sınıflı ve 5 sınıflı olmak üzere iki farklı veri seti elde edilmiştir. Sınıflara ayırma işleminde kullanılan AQI aralığı, her aralığın sınıf etiketi ve her sınıfta bulunan örnek sayıları 3 sınıflı ve 5 sınıflı veri setleri için Tablo 4'te sunulmuştur.

Tablo 4. Veri setinin 3 sınıflı ve 5 sınıflı dağılım özellikleri

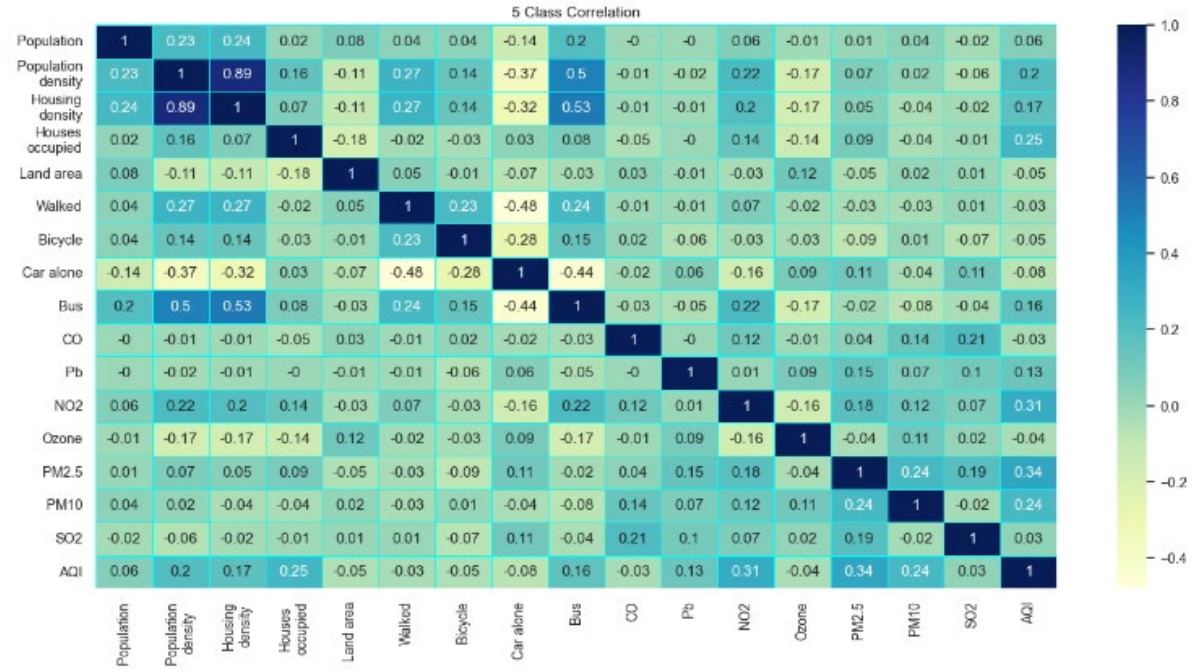
Sınıf Bilgisi	Aralık	Sınıf Etiketi	Örnek Sayısı
3 sınıflı dağılım	AQI < 82	İyi	2193
	83 < AQI < 103	Sağlıksız	2081
	104 < AQI	Tehlikeli	2107
5 sınıflı dağılım	AQI < 69	İyi	1232
	70 < AQI < 86	Orta	1226
	87 < AQI < 100	Sağlıksız	1378
	101 < AQI < 115	Çok Sağlıksız	1320
	116 < AQI	Tehlikeli	1225

Çalışmanın amacına uygun seçilen ve makine öğrenmesi yöntemleri ile veri kümesinin sınıflara ayrılma işleminde kullanılabilecek en dengeli aralıklar Tablo 4'teki gibi tercih edilmiştir. Mevcut iki veri setinin sınıflandırma işlemine tabi tutulmadan önce kontrol edilmesi gereken bir diğer özelliği de kullanılan özniteliklerin birbirleri ile ne kadar ilişkili olduğudur. Veri setindeki öznitelikler hava kalitesi tahmininde etkili olabilecek öznitelikler olarak belirlendiğinden beklenti, bu özniteliklerin birbirleri arasında ve hedef değişken arasında belli bir korelasyonun olması gerektiği yönündedir. Bu durumun analiz edilebilmesi için veri setinde bulunan özniteliklerin korelasyon haritası çıkarılmış, Şekil 2 ve Şekil 3'te sunulmuştur.

Şekil 2 ve Şekil 3'te görüleceği üzere öznitelikler arası birtakım ilişkiler bulunmaktadır. Korelasyon haritasında her bir özniteliğin, veri setindeki diğer özniteliklerle arasındaki lineer ilişki yer alır. Bu ilişki değeri 1'e ne kadar yakınsa, öznitelikler arası o kadar yüksek bir lineer ilişki olduğu anlaşılmaktadır. Benzer şekilde korelasyon haritasındaki iki öznitelik arasındaki ilişki değeri -1'e ne kadar yakınsa, o kadar lineer olmayan bir ilişkisi olduğu anlaşılmış olur. İki öznitelik arasındaki ilişki değerinin 0 olması bu iki özniteliğin birbirlerinin değişimlerinden etkilenmediğini göstermektedir. Söz konusu haritalarda çalışmamız kapsamında en önemli nokta; belirleyici öznitelikler ile hedef öznitelik arasındaki ilişkinin varlığıdır. Hem 3 sınıflı veri seti için hem de 5 sınıflı veri seti için belirleyici öznitelikler ile hedef öznitelik arasında belirli oranlarda korelasyon değeri mevcut olup, 0 değerinde öznitelik ilişkisi bulunmamaktadır. Şekil 2 ve Şekil 3'teki korelasyon tabloları incelendiği zaman AQI değerinin; havada bulunan gazlar ve nüfus yerleşimi ile alakalı parametreler ile yüksek korelasyona sahip olduğu, ulaşım sağlamak için kullanılan araçların oranı ile düşük korelasyona sahip olduğu görülmektedir.



Şekil 2. Veri seti öznitelikler arası korelasyon haritası (3 sınıflı)



Şekil 3. Veri seti öznitelikler arası korelasyon haritası (5 sınıflı)

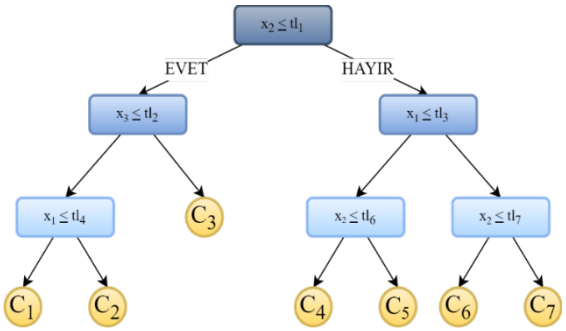
2.3. Makine Öğrenmesi Algoritmaları

Çalışma kapsamında belirlenen 7 adet geleneksel makine öğrenmesi yöntemi, Python programlama dilinde kullanılan Scikit Learn kütüphanesindeki varsayılan standart parametreleri ile uygulanmıştır. Fakat Yapay Sinir Ağları (YSA) mimari bir tasarım gerektirdiğinden diğer modellerden ayrı olarak ele alınmıştır. Çalışmada kullanılacak YSA modelinin katmanlarını ve yoğunluklarını belirlemek için kaba kuvvet (brute force) yaklaşımı kullanılmıştır.

2.3.1. Karar Ağacı

Karar Ağaçları (Decision Trees), ağacın oluşturulmasında kullanılan veri setini yinelemeli olarak daha küçük alt bölümlere ayıran bir sınıflandırma prosedürü olarak tanımlanır. Ağaç; kök düğüm, dizi dâhili düğüm ve dizi uç düğümden (yaprak) oluşur [22]. Veri setinde bulunan öznitelikleri kullanarak mantıksal sıraya uygun dalları olan bir ağaç yapısı oluşturur. Ağaçtaki

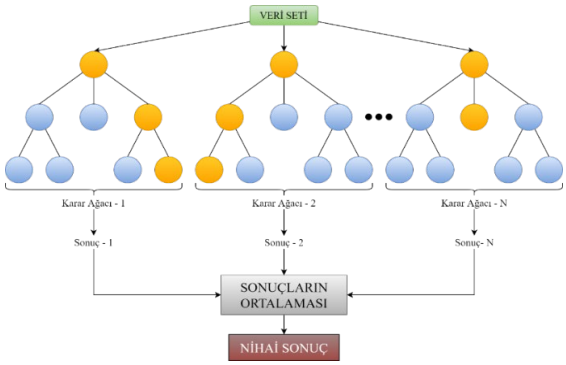
seviyeleri belirlemek için özniteliklerin değerleri ve sınıflandırma başarısındaki ağırlıkları kullanılır. Karar Ağacı'nı diğer sınıflandırıcılardan ayıran en önemli özelliği, ele aldığı problemi böl ve yönet tekniğiyle çözmeye çalışmasıdır. Oluşturulan ağaç yapısında; veri setindeki öznitelikler ağacın düğümlerini, tahmin edilecek olan sonuç sütunundaki değerler ise ağacın yapraklarını oluşturur. Veri setine dahil edilen her yeni örnek, mevcut özniteliklerin değerleri baz alınarak ağaçta uygun olan sınıfa yerleştirilir. Bu yerleştirme işlemini yapmak için, dâhil edilecek olan örnek, ağaç üzerinde hareket ettirilerek sınıfı tespit edilir. Karar Ağaçları sınıflandırma algoritmasının çalışma prensibini açıklayıcı örnek bir görsel Şekil 4'te sunulmaktadır.



Şekil 4. Karar Ağacı çalışma prensibi

2.3.2. *Rastgele Orman*

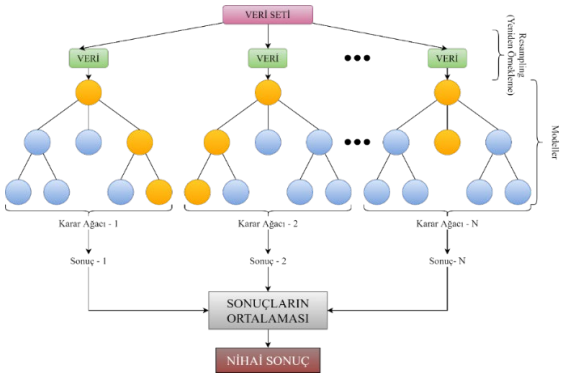
Topluluk (ensemble) öğrenme yöntemini kullanan Rastgele Orman (Random Forest) sınıflandırıcı algoritması, birden fazla modelin oluşturulması ve eğitilmesini üzerinden sınıflandırma gerçekleştirir [23]. Rastgele Orman sınıflandırıcı modeli, eğitim verisini kullanarak rastgele karar ağacı algoritmaları oluşturup sınıflandırma işlemleri yapar. Parametrik olmayan Rastgele Orman sınıflandırıcı modeli, bir model eğitmek yerine birden çok modelin sonucunu göz önünde bulundurarak ortak bir sonuç elde eder. Öznitelikler arasındaki bağlantıları hesaplayabilir. Bu yüzden veri sayısının düşük olduğu durumlarda, çok sayıda öznitelikli veri setleri için, diğer algoritmalara göre daha ideal ve tercih edilebilirdir [24]. Rastgele Orman sınıflandırma algoritmasının çalışma prensibini açıklayıcı örnek bir görsel Şekil 5'te sunulmaktadır.



Şekil 5. Rastgele Orman çalışma prensibi

2.3.3. *Torbalama*

Torbalama (Bagging) sınıflandırma algoritması, eğitim veri setinin N defa rastgele örneklenmesi sonucunda orijinal eğitim setiyle eşit boyutta N adet eğitim verisi üretilmesini sağlar. Elde edilen veri kümelerini N adet sınıflandırıcı üzerinde eğitim ve tahmin işlemlerine tabi tutar [25], [26]. Eğitilen tüm modellerin sonuçlarını birleştirerek ana sonucu elde eder [27]. Rastgele Orman'dan farklı olarak, veri setini bölerek birden çok Karar Ağacı modeli oluşturur. Rastgele Orman ise tek bir veri seti kullanarak birden çok karar modeli oluşturmaktadır [27]. Şekil 6'da Torbalama algoritmasının örnek bir sınıflandırma prensibi sunulmaktadır.

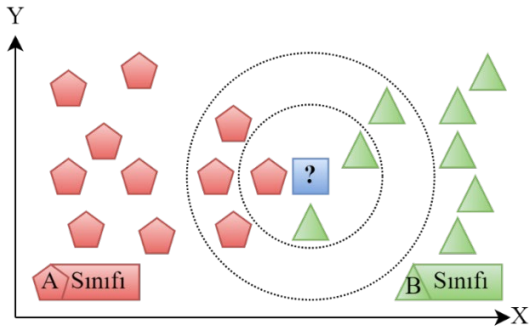


Şekil 6. Torbalama çalışma prensibi

2.3.4. *K-En Yakın Komşu*

K-En Yakın Komşuluk (K-Nearest Neighbor) sınıflandırma algoritması, örüntü tanıma alanında yaygın olarak kullanılan parametresiz öğrenme algoritmasıdır [28]. Sınıflandırma kuralları, eğitim örnekleri tarafından oluşturulur. Yeni örneklerin sınıflarına karar verirken eğitim kümesindeki örneklerin konumunu göz önünde bulundurur. Yeni gelen örneğin sınıfı tahmin edilirken, kendisine en yakın K adet komşusunun sınıfları dikkate alınır. Dikkate alınan sınıflardan en fazla sayıda bulunan sınıf, yeni örneğin sınıfı olarak tanımlanır. Bu yüzden K değeri olarak tek (3, 5, 7…) sayılar tercih edilir. Çünkü komşular arasında en fazla bulunan

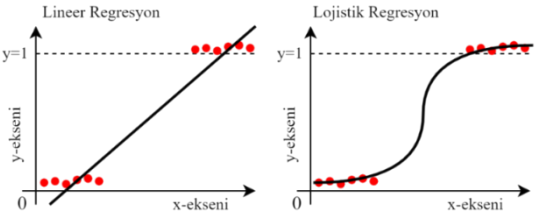
sınıf incelenirken, eşit sayıda çıkması durumu, istenmeyen bir durumdur. Örneklerin birbirlerine olan yakınlıkları; Öklid, Manhattan, Mamming ve Minkowski gibi uzaklık hesabı yapan fonksiyonlarla [29] hesaplanır. K-En Yakın Komşu sınıflandırma algoritmasının çalışma prensibini açıklayıcı, örnek bir görsel Şekil 7'de sunulmaktadır.



Şekil 7. K-En Yakın Komşu çalışma prensibi

2.3.5. *Lojistik Regresyon*

En temel istatistiksel sınıflandırma yöntemlerinden biri olan Lojistik Regresyon (Logistic Regression), parametrik bir sınıflandırma yöntemidir. En çok kullanılan sınıflandırma yöntemlerinden biri olan Lojistik Regresyon, bir veya daha fazla değişkene bağımlı olan değişkenin modellenmesinde kullanılır [30]. Hedef değişkenin kategorik olmadığı, sürekli değer analizi için benzer çalışma prensibine sahip Lineer Regresyon kullanılabilir. Lojistik Regresyon ile Lineer Regresyon'u birbirinden ayıran en temel özellik Lojistik Regresyon'un çıktısının sürekli değer içermemesidir [31]. Lojistik Regresyon sınıflandırma algoritmasının çalışma prensibini ve Lineer Regresyondan farkını açıklayıcı, örnek bir görsel Şekil 8'de sunulmaktadır.



Şekil 8. Lojistik Regresyon çalışma prensibi ve Lineer Regresyon'dan farkı

2.3.6. *Gaussian Naive Bayes*

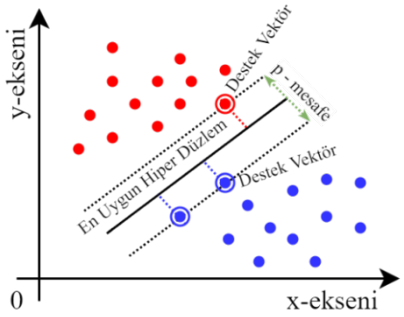
Naive Bayes, Bayes teoreminin uygulanmasına bağlı, olasılık tabanlı istatistiksel bir sınıflandırma algoritmasıdır [32]. Navie Bayes, denetimli öğrenme problemlerinde etkili bir şekilde eğitilebilir ve gerçek dünya durumlarında kullanılabilir [32]. Naive Bayes algoritmasının çalışma prensibi Eşitlik (1)'de sunulmuştur.

$$p(sınıf|veri) = \frac{p(veri|sınıf)\ x\ p(sınıf)}{p(veri)} \quad (1)$$

Naive Bayes algoritması, probleme yaklaşırken öznitelikler arasında şartlı bağımsız, sınıflar arasında da şartlı bağımlı bir yaklaşım izler [33]. Üretici bir sınıflama algoritması olup sınıfı bilinmeyen bir verinin sorgulanan sınıfta olma olasılığını üretir. Gaussian Naive Bayes, veri kümesinde bulunan her sınıfın bir gauss dağılımını takip ettiğini varsayar. Bu yaklaşımı sergilemesindeki amaç, hatalı tahmin etme oranını olabildiğince düşük tutmaktır [33].

2.3.7. *Destek Vektör Makineleri*

Destek Vektör Makineleri (Support Vector Machines, DVM) sınıflandırmada sıklıkla kullanılan denetimli makine öğrenmesi algoritmalarından birisidir [34], [35]. Destek Vektör Makinelerinin amacı, sınıflar arasındaki en yüksek mesafeli hiper düzlemi bulmaktır [34]. Veri setindeki sınıflara ait olan örneklerin arasındaki mesafelere göre sınıfları ayırt etmeye çalışır. Bunu yaparken hiper düzlem(ler) oluşturur. DVM algoritması, sınıflandırmadaki hata oranının en düşük olduğu senaryoyu elde edebilmek için hiper düzlemler arasındaki uzaklığı olabildiğince uzak tutmaya çalışır. DVM sınıflandırma algoritmasının çalışma prensibini açıklayıcı, örnek bir görsel Şekil 9'da sunulmaktadır.

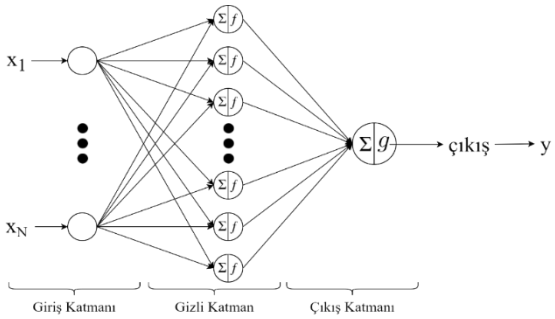


Şekil 9. Destek Vektör Makineleri çalışma prensibi

2.3.8. *Yapay Sinir Ağları*

Son zamanlarda dünya genelinde, beynin çalışma prensibinin bilgisayar ortamında da taklit edilmesini sağlayacak olan Yapay Sinir Ağları (Artificial Neural Networks) üzerine yapılan çalışmalar yüksek bir hızla artış göstermiştir. Yapay Sinir Ağları insan beynindeki nöronların işleyiş şekli ve bağlantıları örnek alınarak tasarlanan bir algoritmadır [36]. Veri setindeki özniteliklere ağırlıklar atanır ve sinir ağlarıyla bağlanmış olan katmanlar arasında hesaplamalar gerçekleştirilerek

kategori veya değer tahminleme işlemi yapılır. Yapay sinir ağlarında ilk katman, veri setindeki öznitelikleri girdi olarak alan girdi katmanı; son katman ise çıktıyı sınıf veya sürekli değer tahmini olarak veren çıktı katmanıdır. N adet öznitelikten oluşan bir veri seti, yapay sinir ağı kullanılarak sınıflandırma işlemine tâbi tutulacak olursa; girdi katmanında N adet nöron, çıktı katmanında ise sınıf sayısı kadar nöron bulunan bir model geliştirilmelidir. Nöronlar; girdileri ve bu girdilerin ağırlıklarını, model tasarlanırken belirlenmiş olan bir f fonksiyonuna parametre olarak verir ve çıktı değeri hesaplanır. Giriş-çıkış katmanları arasına uygun sayıda ve yoğunlukta gizli katman eklenebilir. Gizli katmanlar, nöronların arasında kalan sinir ağlarının alacağı ağırlık değerinin hesaplanması işlemini daha detaylı hale getirir. Şekil 10'da, yapay sinir ağları yaklaşımıyla görselleştirilmiş; N adet girişe, tek çıkışa, opsiyonel sayıda nörona ve tek gizli katmana sahip, ikili sınıflandırıcı örneği yer almaktadır.



Şekil 10. Yapay Sinir Ağları çalışma prensibi

2.4. Değerlendirme Yöntemleri

Çalışmamızda veri temini, ayrıntılı veri ön işleme adımları ve kullanılacak makine öğrenmesi algoritmalarının belirlenmesinden sonra yapılacak sınıflandırma işleminin sonuçlarının nasıl değerlendirileceği de önemli bir aşamadır. Bu aşamada kullanılacak yöntemler, bu yöntemlerin tanımları ve çalışmamızda kullanılma sebepleri alt başlıklarda ele alınmıştır.

2.4.1. Çapraz Doğrulama

Sınıflandırma çalışmalarında, sınıflandırıcı algoritma ile bir model eğitilirken, veri seti eğitim ve test verisi olarak ikiye (genelde %70-30 oranında) ayrılır. Ancak bu durumda dengesiz bir dağılım gerçekleşirse, modelin gerçek performansının görülmesi mümkün olmayabilir. Bu doğrultuda K-Kat Çapraz Doğrulama (K-Fold Cross Validation) yöntemi kullanılarak daha güvenilir sonuçlar elde edilebilir. Çapraz doğrulama, veri setini K eşit parçaya böler ve parçalardan birini test için saklar. Verinin geri kalanını ise modeli eğitmek için kullanır. Bu eğitim işlemini, her K parça için gerçekleştirir. Böylece veri setinin sunabileceği maksimum performans görülmüş olur. Çapraz Doğrulama işleminin örnek bir görüntüsü Şekil 11'de sunulmuştur.

5-Katmanlı Doğrulama (K=5)						
Katman 1	Eğitim verisi				Test verisi	ε1
Katman 2	Eğitim verisi			Test verisi	Eğitim verisi	ε2
Katman 3	Eğitim verisi		Test verisi	Eğitim verisi		ε3
Katman 4	Eğitim verisi	Test verisi	Eğitim verisi			ε4
Katman 5	Test verisi	Eğitim verisi				ε5
	20%	40%	60%	80%	100%	

Şekil 11. K=5 için K-Katlı Çapraz Doğrulama Dağılımı

2.4.2. Hata Matrisi

Sınıflandırma problemlerinde, eğitilen modelin performansını ölçmek için modelin tahmin sonuçları Hata Matrisi (Confusion Matrix) şekline getirilerek daha açıklamalı bir inceleme yapılması sağlanır. Hata matrisinde, True adı altındaki değerler doğru yapılan tahminleme sayısını, False adı altındaki değerler ise yanlış yapılan tahminleme sayısını belirtir. Her bir sütun tahminlenen değeri, her bir satır ise gerçek değeri içerir. Böylece sol üstten sağ alta doğru çapraz bir biçimde devam eden hücreler, tahmin sonucunun doğruluğunu, geri kalan hücreler ise hata oranını verir. Çalışmamızda kullandığımız modellerin performansını ölçmek için kullanılacak hata matrislerinin içerik tanımları 3 sınıflı sınıflandırma modelleri için Şekil 12'de, 5 sınıflı sınıflandırma modelleri için Şekil 13'te sunulmuştur.

	Tahmin edilen etiket (C1)	Tahmin edilen etiket (C2)	Tahmin edilen etiket (C3)
Gerçek etiket (C1)	True (C1-C1)	False (C1-C2)	False (C1-C3)
Gerçek etiket (C2)	False (C2-C1)	True (C2-C2)	False (C2-C3)
Gerçek etiket (C3)	False (C3-C1)	False (C3-C2)	True (C3-C3)

Şekil 12. 3 sınıflı veri setlerinin hata matrisi

	Tahmin edilen etiket (C1)	Tahmin edilen etiket (C2)	Tahmin edilen etiket (C3)	Tahmin edilen etiket (C4)	Tahmin edilen etiket (C5)
Gerçek etiket (C1)	True (C1-C1)	False (C1-C2)	False (C1-C3)	False (C1-C4)	False (C1-C5)
Gerçek etiket (C2)	False (C2-C1)	True (C2-C2)	False (C2-C3)	False (C2-C4)	False (C2-C5)
Gerçek etiket (C3)	False (C3-C1)	False (C3-C2)	True (C3-C3)	False (C3-C4)	False (C3-C5)
Gerçek etiket (C4)	False (C4-C1)	False (C4-C2)	False (C4-C3)	True (C4-C4)	False (C4-C5)
Gerçek etiket (C5)	False (C5-C1)	False (C5-C2)	False (C5-C3)	False (C5-C4)	True (C5-C5)

Şekil 13. 5 sınıflı veri setlerinin hata matrisi

Yukarıda sunulan hata matrislerinde, C1, C2, C3, C4 ve C5 ile belirtilen değerler sınıfları temsil

etmektedir. Matrislerde yer alan her bir satırda, ilgili sınıf belirtilmiş olup her sütunda ise kullanılan sınıflandırma modelinin ilgili sınıfın kaç elemanını hangi sınıfta tahminlediğinin bilgisi sunulur. Örnek olarak; 1. satır 1. sütunda yer alan değer, C1 sınıfına ait olan verilerin, C1 olarak tahmin edilme sayısını içerir. Bu senaryo, doğru tahmin yapılması durumunda gerçekleşir. Ancak 1. satır 2. sütunda yer alan değer, C1 sınıfına ait olan verilerin, C2 sınıf etiketine sahip olarak tahmin edilme sayısını içerir. Bu senaryoda ise yanlış tahminleme yapılmış demektir. Bu durumda doğru tahminleme işlemi yapılan senaryolar sarı, yanlış tahminleme işlemi yapılan senaryolar ise kırmızı ile renklendirilmiştir. Hata matrisleri kullanılarak, eğitilen modellerin doğruluk değeri elde edilebilir.

2.4.3. Performans Değerlendirmesi

Hata matrisi çıkarılarak sınıflandırma algoritmalarının örnekleri hangi sınıflarda tahmin ettiğini ve bu örneklerin gerçekte hangi sınıflarda olduğunu görebilmek bu algoritmaların başarısını kıyaslayabilmek için yeterli değildir. Sınıflandırma işlemlerinde elde edilen başarıyı değerlendirebilmek için genellikle dengesiz dağılıma sahip veri için f-skor hesaplanırken dengeli veri setleri için doğruluk (accuracy) ölçeği kullanılmaktadır. Çalışmamızda kullanılan doğruluk ölçeği Eşitlik (2)'deki gibi hesaplanmaktadır.

$$\textbf{Doğruluk} = \frac{\textbf{Doğru Sınıflandırılmış Örnek Sayısı}}{\textbf{Veri Setinde Bulunan Tüm Örneklerin Sayısı}} \quad \textbf{(2)}$$

Çalışmamızda hazırlamış olduğumuz 3 sınıflı ve 5 sınıflı veri setleri dengeli dağılıma sahip olduğundan her iki veri seti için de kullanılan algoritmaların başarısı Eşitlik (2)'deki doğruluk ölçeği üzerinden karşılaştırılmış ve yorumlanmıştır.

III. DENEYSEL SONUÇLAR

Bir önceki bölümde ayrıntıları verilen veri seti, belirtilen ön işleme adımlarından geçirilerek sınıflandırılmaya uygun hale getirilmiştir. Yine aynı bölümde tanımları ve özellikleri verilen, Torbalama, Karar Ağaçları, K-En Yakın Komşuluk, Lojistik Regresyon, Gaussian Naive Bayes, Rastgele Orman, Destek Vektör Makineleri ve Yapay Sinir Ağları olmak üzere 8 farklı sınıflandırma algoritması çalışmamızda uygulanmıştır. Uygulamada kullanılan algoritmaların hepsinde veri setleri 5-katlı çapraz doğrulamayla işleme alınmıştır. Sınıflandırma algoritmalarının eğitimi sonucunda, sınıflandırıcı modellerden elde edilen test seti başarı sonuçları 3 sınıflı veri seti için Tablo 5'te, 5 sınıflı veri seti için Tablo 6'da sunulmuştur. Kaba kuvvet yöntemi sonucunda elde edilen en başarılı YSA modelinin ayrıntıları Tablo 7'de sunulmuştur. Kullanılan YSA modelinde son Dense katmanının çıktısı 3 sınıflı veri seti için 3, 5 sınıflı veri seti için 5 çıktı verecek şekilde tasarlanmıştır. Modelimizdeki ayarlanabilir parametrelerde; optimizer=Adam, epoch sayısı=200, batch size=16 olarak seçilmiştir.

Tablo 5. Kullanılan sınıflama algoritmalarından elde edilen 3 sınıflı sınıflandırma başarısı

Algoritma	Çapraz Doğrulama Ortalama Doğruluk	Çapraz Doğrulama Standart Sapma
Rastgele Orman	***86,69%***	1,01%
Torbalama	84,86%	0,54%
Karar Ağaçları	81,93%	0,83%
K-En Yakın Komşuluk	81,69%	0,97%
Destek Vektör Makineleri	76,58%	0,11%
Lojistik Regresyon	57,46%	0,65%
Gaussian Naive Bayes	50,39%	0,67%

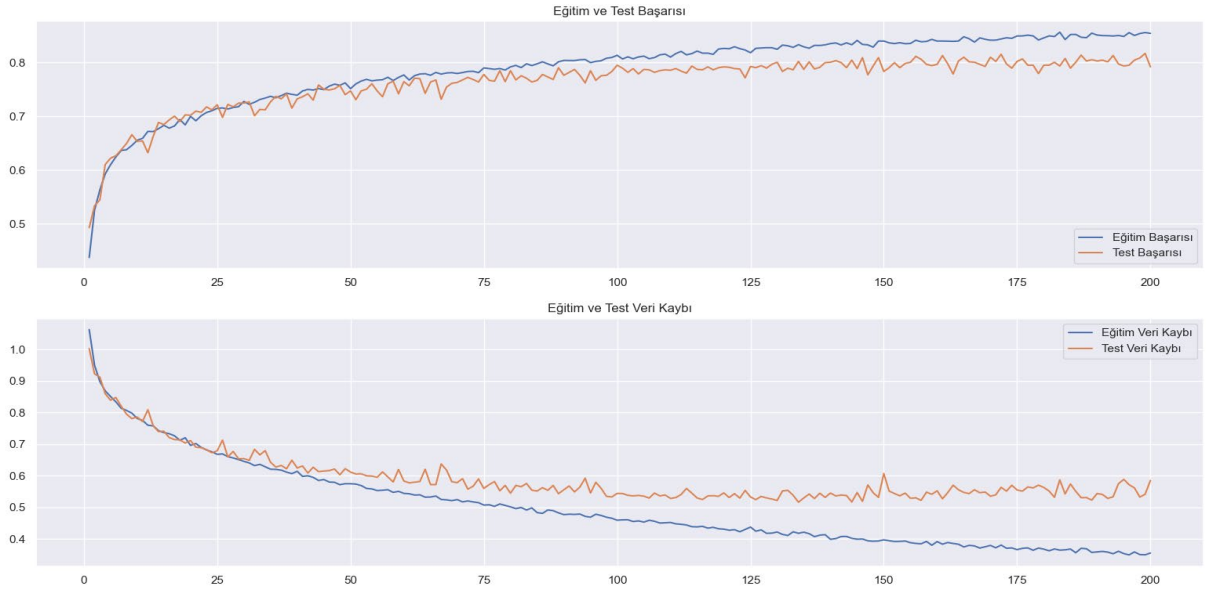
Tablo 6. Kullanılan sınıflama algoritmalarından elde edilen 5 sınıflı sınıflandırma başarısı

Algoritma	Çapraz Doğrulama Ortalama Doğruluk	Çapraz Doğrulama Standart Sapma
Rastgele Orman	***81,61%***	0,62%
Torbalama	80,11%	1,22%
Karar Ağaçları	74,98%	0,79%
K-En Yakın Komşuluk	72,96%	1,20%
Destek Vektör Makineleri	66,80%	1,13%
Lojistik Regresyon	42,90%	1,43%
Gaussian Naive Bayes	37,78%	2,04%

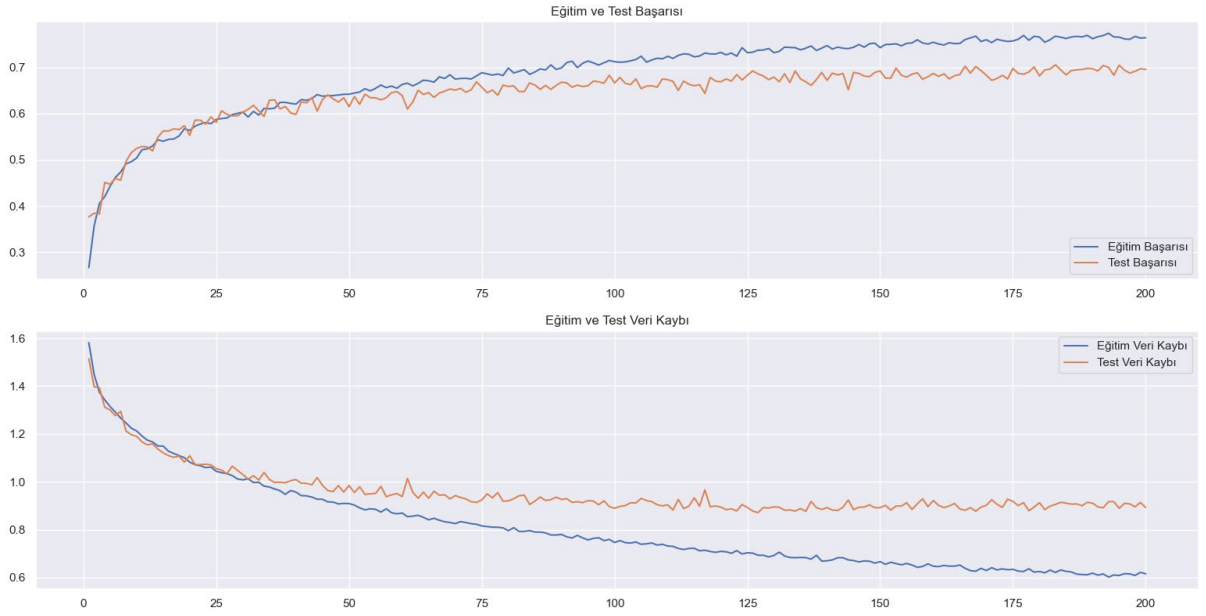
Tablo 7. Kullanılan YSA modelinin özellikleri

Katman	Çıktı	Aktivasyon
Dense	64	Relu
Dense	32	Relu
Dense	16	Relu
Dropout	0.1	
Dense	(3 veya 5)	Softmax

Yapay Sinir Ağları eğitimlerinin 200 epoch'ta kesilmesinin sebebi, 200. epoch'tan sonra meydana gelmeye başlayan aşırı öğrenme (over-fitting) durumudur. Şekil 14 ve Şekil 15'de YSA modelinin eğitim sonuçları (sırasıyla 3 sınıflı ve 5 sınıflı veri seti için) verilmiştir. Bu eğitimler sonucunda; 3 sınıflı veri seti %80,34, 5 sınıflı veri seti ise %70,56 doğruluk oranına ulaşmıştır.

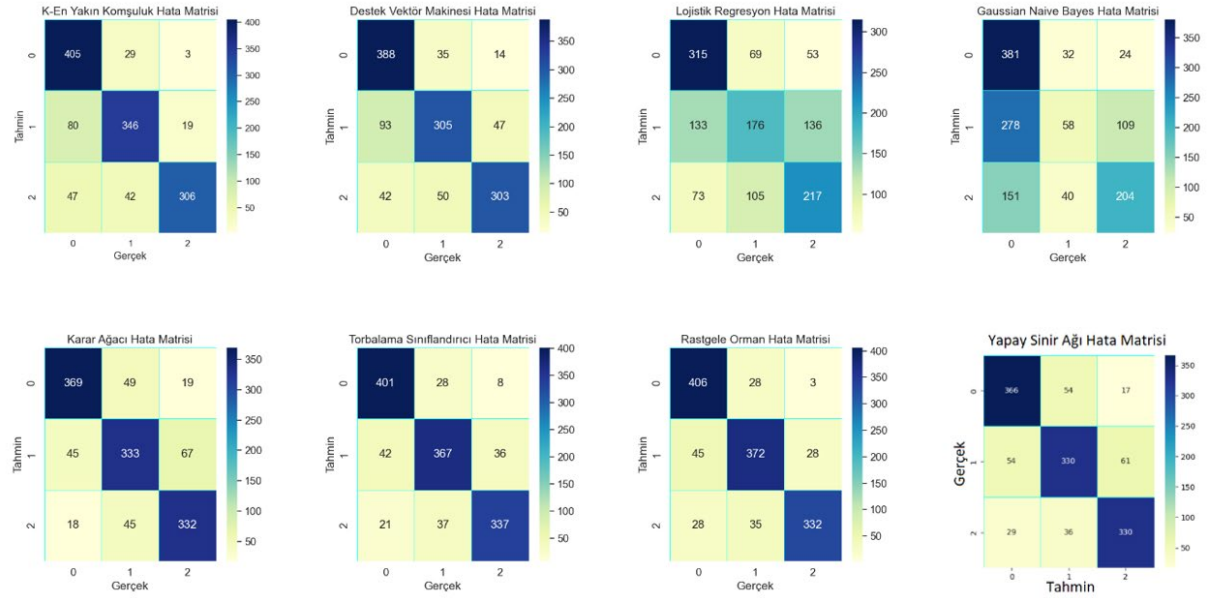


Şekil 14. 3 sınıflı veri setinin sınıflandırılmasından elde edilen YSA modeli sonuç grafikleri

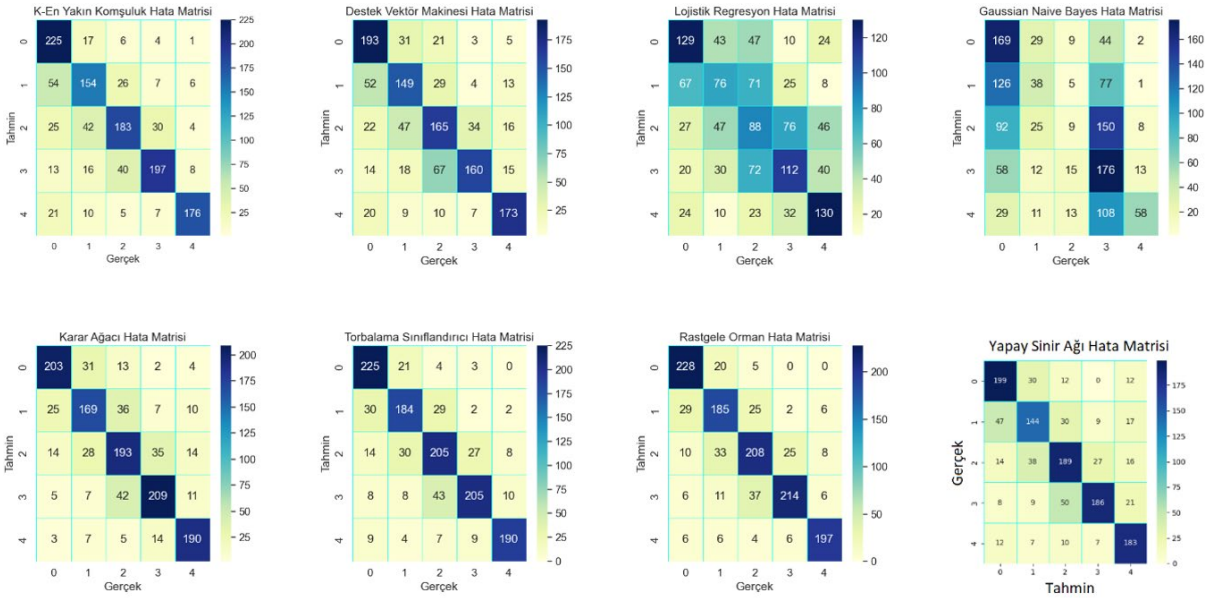


Şekil 15. 5 sınıflı veri setinin sınıflandırılmasından elde edilen YSA modeli sonuç grafikleri

Toplamda 2 veri seti için 8 farklı sınıflandırma işleminin gerçekleştirildiği çalışmamızda, yukarıda bahsedildiği üzere örneklerin sınıflara dağılımlarını inceleyebilmek için hata matrisi kullanılmıştır. Şekil 16'da 3 sınıflı sınıflandırma sonucu sınıflandırıcı modellerden elde edilen hata matrisleri, Şekil 17'de ise 5 sınıflı sınıflandırma sonucu sınıflandırıcı modellerden elde edilen hata matrisleri sunulmuştur. Şekil 16 ve Şekil 17 incelendiği zaman 3 sınıflı sınıflandırıcıların 5 sınıflı sınıflandırıcılara göre daha başarılı sonuçlar sağladığı görülmektedir. Bunun sebebi sınıf sayısı azaldıkça doğru tahmin etme olasılığının artmasıdır.



Şekil 16. 3 sınıflı veri setinin sınıflandırılmasından elde edilen hata matrisleri

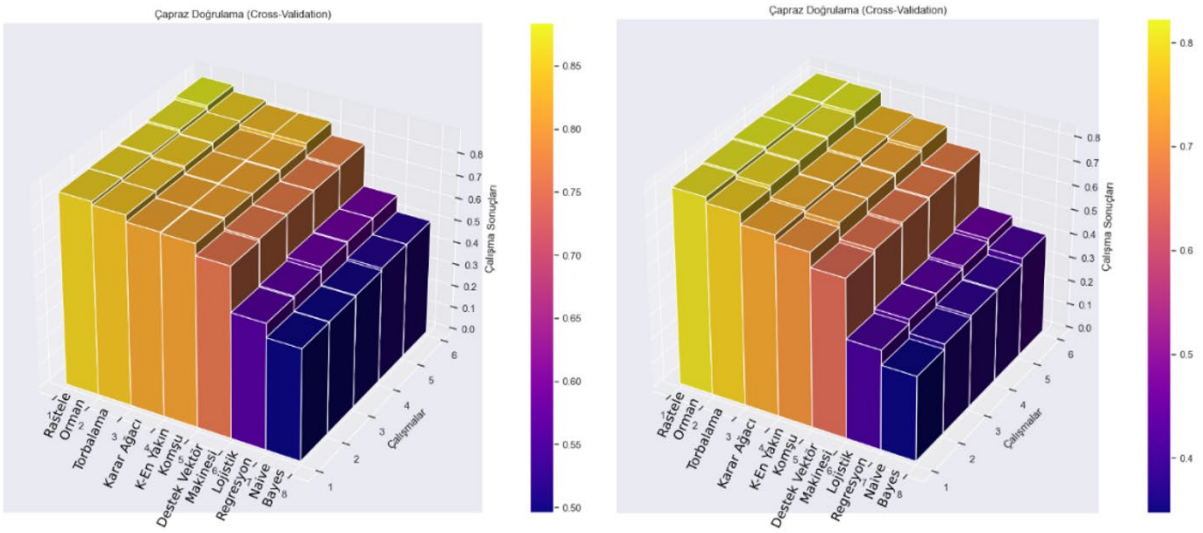


Şekil 17. 5 sınıflı veri setinin sınıflandırılmasından elde edilen hata matrisleri

Kullanılan sınıflandırıcı modellerden üretilen sonuçlar 5 Katlı Çapraz Doğrulama ile elde edildiğinden her kattan elde edilen sonuçların her model için tutarlı olması beklenir. Çalışmamızda kullandığımız modellerin tutarlılığı, modellerin her test seti için ürettiği sınıflandırma sonuçlarının karşılaştırması Şekil 18'de sunulmuştur.

Şekil 18'de X-ekseninde algoritmaların isimleri, Y-ekseninde ise bu algoritmaların çalışma sayıları görülmektedir. Eldeki X-Y şeklindeki 2D ortama 3. bir boyut (Z-ekseni) olarak, bu algoritmaların çalışma sonuçları da eklenmiştir. Çalışma sonuçları küçükten büyüğe sıralanarak bir rampa oluşması sağlanmıştır. Bu sayede rampanın yüksek kısmında kalan algoritmaların, yüksek verimliliğe sahip algoritmalar olduğu rahatlıkla görülebilmektedir. Oluşturulan 3D grafikler ve doğruluk sonuçları incelendiğinde, en yüksek doğruluk ortalamasına, yani en yüksek başarıya sahip modelin Rastgele Orman algoritması ile üretilen sınıflandırıcı model olduğu görülebilmektedir. Çalışmamızda hem 3 sınıflı veri setinde hem de 5 sınıflı veri setinde en iyi performansın Rastgele Orman sınıflandırma algoritmasından elde edilmesinin en belirgin özelliği, kullanılan belirleyicileri farklı dallanmalar ile kullanabilmesi olarak görülmektedir. Bu durum, çalışmamızda hava kirliliği tahmininde kullanılmasının anlamlı olduğu düşünülen özniteliklerin belirlenmesinde başarılı bir aşama gerçekleştirildiğinin göstergesi olmaktadır.

Lojistik Regresyon ve Naive Bayes algoritmalarının, kullanılan diğer algoritmalara göre daha başarısız sonuçlar sağladığı Şekil 18'de görülmektedir. Bu algoritmaların sunulan problem için yetersiz olması veya veri seti ile uyumsuz olması, düşük başarının elde edilmesinin olası sebeplerindendir.



Şekil 18. Makine öğrenmesi algoritmalarının 5-katlı çapraz doğrulama sonuçlarının 3 sınıflı ve 5 sınıflı veri setleri için karşılaştırılması

IV. SONUÇLAR ve TARTIŞMALAR

Gün geçtikçe daha da kirlenen havanın küresel boyutlarda iklim krizine yol açmış olması, bilim insanlarının bu konuda çözümler geliştirmesini gerekli kılmıştır. İleride karşılaşılacak olan küresel tehlike ve tehditlerle baş edilebilmesi için öncelikle mevcut durumun bilincinde olunması gerekmektedir. Günümüzde ne yazık ki şehirlerde hava kalitesi ölçümü yeterli miktarda ve doğrulukta yapılamamaktadır. Çalışmamızda, şehirlerdeki hava kalitesinin hangi belirleyiciler gözlenerek değerlendirilebileceği ve bu değerlendirmenin yapay zekâ yöntemleri kullanılarak nasıl sonuçlandırılabileceği ele alınmıştır. Hava kalitesi belirleyicilerinin makine öğrenmesi algoritmaları ile sınıflandırıldığı çalışmamızda, hava kalite endeksi değerinin şehirlerdeki dağılımı ve bu değerin insan sağlığına etkisi göz önüne alınarak, şehirler; iyi, sağlıksız ve tehlikeli olmak üzere 3 sınıflı, iyi, orta, sağlıksız, çok sağlıksız ve tehlikeli olmak üzere 5 sınıflı yapılarda ele alınmıştır. Sekiz farklı sınıflandırma algoritmasının kullanıldığı çalışmamızda, en yüksek sınıflandırma başarısı; 3 sınıflı sınıflandırmada %86,69 oranında, 5 sınıflı sınıflandırmada %81,61 oranında Rastgele Orman algoritmasından üretilen modelden elde edilmiştir. Çalışmamızda ele aldığımız belirleyiciler ve geliştirdiğimiz yöntemler ile şehirlerdeki hava kirliliği seviyesi kısa sürede belirlenebilmektedir. Bu sayede şehirlerdeki hava kirliliğine gerekli farkındalığın oluşması, önlem ve tedbirlerin erkenden alınması sağlanabilecektir. Sonraki çalışmalarda, sürekli elde edilecek hava kalitesi değeri ile şehirlerde kritik durumlar için uyarı verebilen modellerin geliştirilmesi öngörülmektedir.

KAYNAKLAR

[1] Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *The lancet*, *360*(9341), 1233-1242.

[2] Advameg Inc. City Data, Advanced U.S. City Search, http://www.city-data.com/advanced/search.php (February 2022).

[3] AKYÜZ, A. A. (2019). YAŞAMSAL BİLİNMEZLİK: İKLİM KRİZİ VE GIDA.

[4] TÜBİTAK Bilim ve Teknik, 632's Extra Article, https://bilimteknik.tubitak.gov.tr/pdf/temmuz-2020, (February 2022).

[5] Houghton, J. (2005). Global warming. *Reports on progress in physics*, *68*(6), 1343.

[6] Bayram, H. (2005). Türkiye'de hava kirliliği sorunu: Nedenleri, alınan önlemler ve mevcut durum. *Toraks Dergisi*, *6*(2), 159-165.

[7] Pasupuleti, V. R., Kalyan, P., & Reddy, H. K. (2020, March). Air quality prediction of data log by machine learning. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1395-1399). IEEE.

[8] Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019, March). A machine learning model for air quality prediction for smart cities. In *2019 International conference on wireless communications signal processing and networking (WiSPNET)* (pp. 452-457). IEEE.

[9] Birleşmiş Milletler, Ambient (outdoor) air pollution, https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (November 2021).

[10] Wikimedia Foundation, Sources of Air Pollution, https://commons.wikimedia.org/wiki/File:Sources_of_Air_Pollution.png (November 2021).

[11] Janarthanan, R., Partheeban, P., Somasundaram, K., & Elamparithi, P. N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. *Sustainable Cities and Society*, *67*, 102720.

[12] Bhalgat, P., Bhoite, S., & Pitare, S. (2019). Air quality prediction using machine learning algorithms. *International Journal of Computer Applications Technology and Research*, *8*(9), 367-390.

[13] NandigalaVenkatAnurag, Y., & Sharanya, S. (2019). Air Quality Index Prediction with Meteorological Data Using Feature Based Weighted Xgboost. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(1), 1355-1358.

[14] Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C. (2020). Machine learning-based prediction of air quality. *Applied Sciences*, *10*(24), 9151.

[15] Kamal, M. M., Jailani, R., & Shauri, R. L. A. (2006, June). Prediction of ambient air quality based on neural network technique. In *2006 4th Student Conference on Research and Development* (pp. 115-119). IEEE.

[16] Holloway, T., Spak, S. N., Barker, D., Bretl, M., Moberg, C., Hayhoe, K. & Wuebbles, D. (2008). Change in ozone air pollution over Chicago associated with global climate change. Journal of Geophysical Research: Atmospheres, 113(D22).

[17] Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. Big data and cognitive computing, 2(1), 5.

[18] Almetwally, A. A., Bin-Jumah, M., & Allam, A. A. (2020). Ambient air pollution and its influence on human health and welfare: an overview. *Environmental Science and Pollution Research*, *27*(20), 24815-24830.

[19] Wong, T. W., San Tam, W. W., Yu, I. T. S., Lau, A. K. H., Pang, S. W., & Wong, A. H. (2013). Developing a risk-based air quality health index. *Atmospheric environment*, *76*, 52-58.

[20] Rovira, J., Domingo, J. L., & Schuhmacher, M. (2020). Air quality, health impacts and burden of disease due to air pollution (PM10, PM2. 5, NO2 and O3): Application of AirQ+ model to the Camp de Tarragona County (Catalonia, Spain). *Science of The Total Environment*, *703*, 135538.

[21] Stripe Payments Europe Ltd., Paris Air Pollution Has Reached A Critical Level https://www.statista.com/chart/7152/paris-air-pollution-has-reached-a-critical-level/ (June 2021).

[22] Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, *61*(3), 399-409.

[23] Deus, D. (2018). Assessment of Supervised Classifiers for Land Cover Categorization Based on Integration of ALOS PALSAR and Landsat Data. *Advances in Remote Sensing*, *7*(2), 47-60.

[24] Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99*(6), 323-329.

[25] Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, *36*(6), 1291-1302.

[26] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

[27] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

[28] Suguna, N., & Thanushkodi, K. (2010). An improved k-nearest neighbor classification using genetic algorithm. *International Journal of Computer Science Issues*, *7*(2), 18-21.

[29] Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, *1*(2), 1.

[30] Hilbe, J. M. (2011). Logistic regression. *International encyclopedia of statistical science*, *1*, 15-32.

[31] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[32] Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019, June). Cancer classification using gaussian naive bayes algorithm. In *2019 International Engineering Conference (IEC)* (pp. 165-170). IEEE.

[33] ŞAHİN, A. R., DOĞAN, K., & SİVRİ, S. (Eds.). (2020). *Sağlık Bilimlerinde Yapay Zeka*. Akademisyen Kitabevi.

[34] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, 1578-1585.

[35] Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, *24*(12), 1565-1567.

[36] 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*, *4*(11), e00938.