İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANARAK HAVA KİRLİLİĞİ TAHMİNİ

YÜKSEK LİSANS TEZİ

Kıymet KAYA

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ

HAZİRAN 2019

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANARAK HAVA KİRLİLİĞİ TAHMİNİ

YÜKSEK LİSANS TEZİ

Kıymet KAYA
(504151552)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ

HAZİRAN 2019

İTÜ, Fen Bilimleri Enstitüsü'nün 504151552 numaralı Yüksek Lisans Öğrencisi Kıymet KAYA, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANARAK HAVA KİRLİLİĞİ TAHMİNİ" başlıklı tezini aşağıdaki imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Prof. Dr. Alper ÜNAL** İstanbul Teknik Üniversitesi	.....
	Dr. Öğr. Üyesi Reyhan AYDOĞAN Özyeğin Üniversitesi	.....
		.....

Teslim Tarihi : **3 Mayıs 2019**
Savunma Tarihi : **12 Haziran 2019**

Aileme,

ÖNSÖZ

Öncelikle danışmanım Prof. Dr. Şule GÜNDÜZ ÖĞÜDÜCÜ'ye tez süresince gösterdiği ilgi, destek ve emek için çok teşekkür ederim. Ailem gibi gördüğüm Filiz AYAZ, Onur IŞIK, Kübra YAŞAR ve Saime GÜMÜŞTAŞ'a, sevgili iş arkadaşım ve kırk yıllık hatırı kat be kat biriktirdiğimiz kahve dostum Tuğba PAMAY'a, ve beni bugünlere getiren canım Aileme destekleri için minnettarım.

Haziran 2019

Kıymet KAYA
Bilgisayar Mühendisi

İÇİNDEKİLER

Sayfa

KISALTMALAR

EAS : Ekstra Ağaç Sınıflandırıcı
ROS : Rasgele Orman Sınıflandırıcı
GAS : Gradyan Arttırma Sınıflandırıcı
DVMS : Destek Vektör Makinesi Sınıflandırıcı
EAR : Ekstra Ağaç Regresörü
ROR : Rasgele Orman Regresörü
GAR : Gradyan Arttırma Regresörü
DVMR : Destek Vektör Makinesi Regresörü
EPA : Environmental Protection Agency

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANARAK HAVA KİRLİLİĞİ TAHMİNİ

ÖZET

Hava kirliliği büyük şehirlerdeki çevre koşulları üzerinde önemli derecede etkilidir. Kirlilik risklerine karşı korunmak için geliştirilen hava kirliliği kontrol teknolojilerinin varlığı hava kirliliğinin doğru tahminine bağlıdır. Hava kirliliğini belirlemek için gösterge kirleticilerden faydalanılır. Partikül boyutu 10 μm altındaki maddeler (PM_{10}), partikül boyutu 2.5 μm altındaki maddeler ($PM_{2.5}$), azotoksitler (NO, NO_2, NO_2), ozon (O_3), kükürtoksitler (SO_2) ve karbonoksitler (CO) başlıca gösterge kirleticilerdir.

Hava kirliliğinin şehir sakinleri üzerinde, özellikle de çocuklar ve kalp/solunum yetmezliği olan insanlar gibi hassas grupların üyeleri üzerinde önemli olumsuz etkileri bulunmaktadır. Yüksek konsantrasyonlarda PM_{10}'a uzun süre maruz kalınması erken ölümlere, bozulmuş kardiyovasküler sisteme ve solunum yolu enfeksiyonlarına neden olabilmektedir. Bu zamana kadar İstanbul için yapılan hava kirliliği tahminleme çalışmalarınn hiçbiri dengesiz sınıf dağılımına sahip veri kümeleri ile değildir. Literatürdeki bu eksikliği gidermek için dengesiz veri dağılımı problemi ile başa çıkabilen ve PM_{10} kirleticisinin yoğunluğu aracılığıyla İstanbul için hava kirliliği tahmininde bulunan İki Katmanlı Hava Kirliliği Tahminleme Modeli'ni öneriyoruz.

Önerdiğimiz, hava kirliliği tahminleme modeli iki katmandan oluşmaktadır. İlk katmanında PM_{10} sınıflandırma problemi, zararsız sınıf (1) ve tehlikeli sınıf (0) olarak kodlanan dengesiz dağılan veriden ikili sınıflandırma problemi olarak değerlendirilmektedir. Dengesiz veri problemine çözüm olarak örnekleme yaklaşımları ve algoritmaların parametrelerinin dengesiz veriye göre ayarlanması üzerinde durulmuştur. Çözümün örnekleme bölümünde, verilerin Aşağı Örnekleme yöntemlerinden Rastgele Örnekleme ve Near-Miss örneklemesinin versiyonları ile oluşturulan veri dağılımı dengelenmiş versiyonları tatmin edici sonuçlar üretmemiştir. Algoritmik kısımda, dengesiz öğrenme problemleri üzerindeki olumlu etkileriyle öne çıkan topluluk modelleri Rastgele Orman Sınıflandırıcısı (ROS), Ekstra Ağaç Sınıflandırıcısı (EAS), Gradyan Arttırma Sınıflandırıcısı (GAS) ve çekirdek tabanlı algoritmalar çok terimli Destek Vektör Makinesi (poli-DVM), rbf Destek Vektör Makinesi (rbf-DVM) modellerinin performansları AUROC açısından karşılaştırılmıştır. İkili sınıflandırma için önerilen model, tüm eğitim kümesi örneklerini kullanır ve ROS ile tahmin yapar.

İkinci katman için başlangıçtaki ikili etiketlenmiş PM_{10} veri kümesi; gerçek etiketlerine göre tehlikeli sınıf örnekleri (0 etiketliler) bir veri grubu, zararsız sınıf örnekleri (1 etiketliler) bir veri grubu oluşturacak şekilde ikiye ayrılır. Bu veri gruplarında PM_{10} yoğunluğunu tahminleyen bağımsız regresyon modelleri eğitilir. İki Katmanlı Hava Kirliliği Tahminleme Modeli ile tahminleme aşamasında, ilk aşamada öne çıkan ikili sınıflandırıcı ile öncelikle örneğin hangi sınıfa ait olduğuna karar verilir. Sonrasında ise ait olduğu sınıfın regresyon modeli kullanılarak PM_{10} kirleticisinin yoğunluğu bulunur.

İki Katmanlı Hava Kirliliği Tahminleme Modeli'nin performansı, İstanbul'daki dokuz ölçüm noktasına ait veriler kullanılarak Ortalama Mutlak Hata (OMH) ve Ortalama Kare Hata (OKH) hata metriklerine göre saf regresyon modelleri ile kıyaslanmıştır. Aksaray, Alibeyköy, Beşiktaş, Esenler, Kartal, Sarıyer, Silivri, Üsküdar ve Yenibosna ölçüm noktalarındaki, Ağustos 2011 - Şubat 2018 aralığını kapsayan saatlik meteorolojik ve kirlilik verisi üzerinde, önerilen modelin dengesiz veri problemi ile daha iyi başa çıkabildiği görülmüştür.

PREDICTION OF AIR POLLUTION
USING
DATA MINING METHODS

SUMMARY

Air pollution has a significant effect on environmental conditions in many large cities. Presence of air pollution control technologies developed to protect against the risks of pollution depends on the accurate estimation of air pollution. Accurate air pollution prediction is particularly helpful in ensuring economic and social development in developing countries. Indicator pollutants are used for risk assessment and epidemiological analysis for air pollution studies. Particulate matter under 10 μm (PM_{10}), particulate matter under 2.5 μm ($PM_{2.5}$), sulphur oxides (SO_2), nitrogen oxides (NO, NO_2, NO_X), carbon oxides (CO) and ozone (O_3) are the indicator pollutants frequently seen in this area.

Air pollution has a significant impact on inhabitants of the cities, particularly members of vulnerable groups such as children and people with heart failure and respiratory failure. Prolonged exposure to high concentrations of PM_{10} may cause premature deaths, impaired cardiovascular system, and respiratory tract infections. Considering the threats posed to human health by particulate matter, we focus on PM_{10} density estimation in this study.

In order to guarantee the quality of life in urban and metropolitan centers, it is necessary to estimate the change of air pollution concentrations. In line with this need, to estimate the time at which the air quality is low and the pollution rate will be high at the regional and local scales before pollution occurs; Air quality estimation models have been developed by taking into account the characteristics of atmospheric pollution and the negative effects of air pollution on the standard of living.

When the data sets of the current studies in this area are examined, it is seen that meteorological data is used predominantly. Meteorological conditions are critical in determining the concentrations of pollutants in the air. Lower than normal ambient temperature and incoming solar radiation slow down photo-chemical reactions and cause secondary air pollutants, such as O_3, to be found in smaller amounts of air. Increased wind speed can increase or decrease air pollutant concentrations. Strong wind speeds can create dust storms by removing particles from the ground. High humidity often affects pollutants (PM, CO and SO_2) in the air with high concentrations, but may also result in low concentrations of some contaminants (such as NO_2 and O_3). One reason for this is that high humidity is an indicator of rainfall events.

In addition to using only meteorological data while performing air quality estimations, there are studies that use pollutant data or involve both meteorological and pollution data. The reason why the data set selection is limited in terms of pollution data; the installation and operation of pollutant measuring stations is more difficult and expensive than meteorological stations, the pollutant measuring stations are located in a small number of areas and are difficult to obtain data from pollutant measuring stations.

The aim of this thesis is to estimate the intensity of air pollution for İstanbul through PM_{10} indicator pollutant. For the study which includes meteorological and pollution data covering the period between August 2011 and February 2018, the pollutant measurement stations in İstanbul were examined and Aksaray, Alibeyköy, Beşiktaş, Esenler, Kartal, Sarıyer, Silivri, Üsküdar and Yenibosna stations which has the most data in the past were selected for use in the study. The meteorological station data, which is the closest to the pollution measurement stations, is taken from the Turkish State Meteorological Service.

The fact that the data to be used in the estimation of air pollution has some special characteristics may cause difficulties for urban air quality estimation. First, building a station in the city and operating the station requires a high cost, so there is a limited number of measuring stations. Accordingly, it is also difficult to obtain labeled data in this field. Secondly, data loss may occur in the event of a technical failure in stations. Generally, there is only one measuring device at each station, so the data in that time range is lost when the device is calibrated or maintained, not only when there is a problem with the device. Another problem is that urban air pollution data vary depending on the technology used in measuring stations. For example, the number of stations that can measure $PM_{2.5}$ in İstanbul is considerably less than the number of stations that can measure CO.

An accurate regression can replace air quality monitoring stations with a limited number and distribution in a city. A sensitive classification can provide valuable information to protect people from damage due to air pollution. Generally, both classification and regression provide solutions to support air pollution control, and in this way can have both social and scientific effects.

In the air pollution estimation problem for İstanbul, when the density information of the target pollutant PM_{10} is classified by using the EPA limit values, six classes are obtained. The fact that some classes have only a few examples after this transformation and that these few samples are not sufficient for the prediction model to learn the relevant class have shown that the problem cannot be treated as a sixth classification problem. Alternatively, when the problem is transformed into a binary classification problem and the distributions of the classes in the data set are examined, it is seen that the negative cases in the data (samples of minority class) are quite low compared to positive cases (samples of dominant class). In the classification problem, imbalanced distribution of the samples belonging to the classes is an important problem which makes the learning of the prediction models difficult.

In this thesis, we propose a Two Layer Air Pollution Estimation Model which predicts PM_{10} density by using machine learning algorithms and can cope with imbalanced distribution of data. In the first layer, the PM_{10} classification problem is considered to be the problem of binary classification on imbalanced data set coded as harmless class (1) and dangerous class (0). Sampling approaches and algorithmic approaches are addressed as a solution to imbalanced binary classification problem. In sampling part, the balanced versions of the data generated by Random Sampling and Near-Miss (three different versions) sampling from the Down Sampling approaches did not yield satisfactory results. In algorithmic part of the solution, ensemble models that stand out with their positive effects on imbalanced learning problems are Random Forest Classifier (RFC), Extra Tree Classifier (ETC), Gradient Boosting Classifier (GBC) and kernel based algorithms polynomial Support Vector Machine (poly-SVM), rbf-SVM

performances compared through Area Under ROC Curve (AUROC). The proposed model for binary classification uses all instances of the training set and predicts via RFC.

The initial binary-labeled PM_{10} data set are divided into two groups for the second layer: dangerous class samples (0 labeled) and harmless class instances (1 labeled). In these data groups, independent regression models that predict PM_{10} density are trained. In the first layer, the leading binary classifier determines which class the test sample belongs to. Then using the regression model of the class to which it belongs, density of the PM_{10} pollutant is found.

The performance of the Two Layer Air Pollution Estimation Model was compared with pure regression models according to the Mean Absolute Error (MAE) and Mean Square Error (MSE) error metrics using data from nine measurement points in İstanbul. It has been seen that the proposed model can better handle the imbalanced data problem on hourly meteorological and pollution data at the measurement points of Aksaray, Alibeyköy, Beşiktaş, Esenler, Kartal, Sarıyer, Silivri, Üsküdar and Yenibosna, covering the period of August 2011 - February 2018.

1. GİRİŞ

Hava kirliliği birçok büyük şehirde, yaşam koşullarını önemli derecede etkilemektedir. Kirliliğin oluşturduğu risklerden korunmak amacıyla geliştirilen hava kirliliği kontrol teknolojilerinin varlığı hava kirliliğinin doğru bir şekilde tahmin edilmesine bağlıdır. Doğru hava kirliliği tahmini özellikle gelişmekte olan ülkelerde ekonomik ve sosyal kalkınmanın sağlanmasına önemli derecede yardımcı olur.

Bireylerin maruz kaldığı hava kirliliği çok yönlüdür; yüzlerce gaz bileşiklerini ve kompleks fizikokimyasal bileşimlerin parçacıklarını içeren özel kirletici karışımlarını karakterize etmek, tanımlamak için standartlaşmış yaklaşımlar bulunmamaktadır. Bu tür karışımlar; farklı kirletici maddelerin değişen oranlarla birleşmesinden oluşur ve bölgenin sosyal, ekonomik, teknolojik faaliyetlerinden etkilenir. Bu nedenle hava kirliliği çalışmalarında, risk değerlendirmesi ve epidemiyolojik analiz için gösterge kirleticiler kullanılılır. Bilinen başlıca gösterge gaz kirleticileri partikül boyutu 10 μm altındaki maddeler (PM_{10}), partikül boyutu 2.5 μm altındaki maddeler ($PM_{2.5}$), azotoksitler (NO, NO_2, NO_2), ozon (O_3), kükürtoksitler (SO_2) ve karbonoksitlerdir (CO).

Hava kirliliği; kentin sakinleri, özellikle de çocuklar ve kalp/solunum yetmezliği olan kişiler gibi hassas gruplara üye kişiler üzerinde önemli derecede etkilidir. Artan mortalite (ölüm) ve morbidite (hastalık) oranlarının havadaki kirleticilerin (PM ve SO_2 gibi) yoğunluğunun artışı ile ilişkili olduğu ortaya çıkmıştır [1–3].

Havadaki partikül maddeler insan sağlığı üzerinde ciddi etkileri olan kirleticiler arasında yer alır. Önemli sağlık tehditlerine yol açan civa, kurşun, kadmiyum gibi ağır metaller ve kanserojen kimyasallar bu partikül maddeler içerisinde bulunur. Benzin ve dizel araç egzoz partikülleri benzo(a)pyrene gibi kansere neden olan maddeler içerir ve uzun süre solunduğunda kansere neden olabilir [4, 5]. Yüksek konsantrasyonlarda PM_{10}'a uzun süre maruz kalmak ayrıca erken ölümlere, kardiyovasküler sistemde bozukluklara, iç hastalıklara ve solunum yolu enfeksiyonlarına neden olabilir. Partikül maddelerin insan sağlığına oluşturduğu tehditleri göz önünde bulundurarak bu çalışmada PM_{10} yoğunluğu tahminine odaklanıyoruz.

Kent ve metropol merkezlerinde yaşam kalitesini garanti altına almak için, hava kirliliği konsantrasyonlarının değişimini tahmin etmek gereklidir. Bu ihtiyaç doğrultusunda, bölgesel ve yerel ölçeklerde hava kalitesinin düşük olacağı, başka bir deyişle kirlilik oranının yüksek olacağı zamanları kirlilik yaşanmadan önce tahmin etmek için; atmosferik kirliliğin özellikleri ve hava kirliliğinin yaşam kalitesi üzerindeki olumsuz etkileri göz önünde bulundurularak, hava kalitesi tahminleme modelleri geliştirilmiştir [6, 7].

Bu alandaki mevcut çalışmalara ait veri kümeleri incelendiğinde; ağırlıklı olarak meteorolojik verilerin kullanıldığı görülür. Meteorolojik koşullar, havadaki kirletici konsantrasyonlarının belirlenmesinde kritik öneme sahiptir [5, 8–12]. Normalden düşük ortam sıcaklığı ve gelen güneş radyasyonu, fotokimyasal reaksiyonları yavaşlatır ve O_3 gibi ikincil hava kirleticilerin daha az miktarda havada bulunmasına yol açar [11]. Artan rüzgar hızı hava kirletici konsantrasyonlarını artırabilir ya da azaltabilir [13]. Güçlü rüzgar hızları, zemindeki parçacıkları havaya uçurarak toz fırtınaları oluşturabilir [14]. Yüksek nem genellikle kirleticilerin (PM, CO ve SO_2 gibi) yüksek konsantrasyonlarla havada bulunmasına etki ederken, bazı kirleticilerin (NO_2 ve O_3 gibi) düşük konsantrasyonlarda bulunmasına da sebep olabilir [13] Bunun bir sebebi yüksek nemin yağış olaylarının göstergesi olmasıdır [15].

Hava kalitesi tahmini yaparken sadece meteorolojik veri kullanma yaklaşımının yanında, meteorolojik veriler olmaksızın ölçüm noktasındaki diğer kirleticilerin yoğunluklarını kullanarak tahmin yapan çalışmalar ile hem meteorolojik veri hem de kirletici verilerini içeren çalışmalara da rastlanır. Veri kümesi seçiminin kirlilik verisi açısından kısıtlı olmasının sebebi; kirletici ölçüm istasyonları kurulumunun ve işletilmesinin meteorolojik istasyonlara göre daha zor ve pahalı olması, kirletici ölçüm istasyonlarının az sayıda, belirli bölgelerde bulunması ve kirletici ölçüm istasyonlarından veri elde etmenin zorluğudur.

Bu tezin amacı PM_{10} gösterge kirleticisi aracılığıyla İstanbul için saatlik hava kirliliği yoğunluğu tahminlemektir. Meteorolojik veriler ile kirlilik verilerinin bir arada kullanıldığı, Ağustos 2011 – Şubat 2018 tarih aralığını kapsayan çalışmamız için; İstanbul'daki kirletici ölçüm istasyonları incelenmiş, geçmişe yönelik en fazla veriye sahip olan Aksaray, Alibeyköy, Beşiktaş, Esenler, Kartal, Sarıyer, Silivri, Üsküdar ve Yenibosna istasyonları çalışmada kullanılmak üzere seçilmiştir. Kirlilik ölçüm

istasyonlarına en yakın konumdaki meteorolojik istasyonlara ait veriler Meteoroloji Genel Müdürlüğü'nden [16] alınmıştır.

Hava kirliliği tahmininde kullanılacak verilerin bazı özel karakteristiklere sahip olması, kentsel hava kalitesi tahmini için zorluk oluşturabilmektedir. Birincisi şehir içerisinde bir istasyon inşa etmek ve istasyonu işletmek yüksek maliyet gerektirir, bu nedenle sınırlı sayıda ölçüm istasyonu bulunmaktadır. Buna bağlı olarak bu alanda etiketli veri elde etmek de zorlaşmaktadır. İkincisi, istasyonlarda meydana gelen teknik aksaklık durumunda veri kaybı yaşanabilmektedir. Genel olarak her istasyonda sadece bir ölçüm cihazı olduğundan; sadece cihazla alakalı aksaklık yaşandığında değil, cihaz kalibre edildiğinde ya da bakımı yapıldığında da o zaman aralığındaki veri kaybolmaktadır. Bir diğer sorun ise kentsel hava kirliliği ile ilgili veriler, ölçüm istasyonlarında kullanılan teknolojiye bağlı değişmektedir. Örneğin İstanbul'da $PM_{2.5}$ ölçümü yapabilen istasyon sayısı CO ölçümü yapabilen istasyon sayısına oranla oldukça azdır [4].

İyi bir regresyon modeli, bir şehirde sınırlı sayıda ve dağıtık halde bulunan hava kalitesi izleme istasyonlarının yerini tutabilir. Hassas bir sınıflandırma, insanları hava kirliliği nedeniyle zarar görmekten korumak için değerli bilgiler sağlayabilir. Genel olarak hem sınıflandırma hem regresyon yöntemleri kullanılarak, hava kirliliği kontrolünü desteklemek için etkin çözümler oluşturulabilir ve bu yolla hem toplumsal hem bilimsel büyük etkiler yaratılabilir.

İstanbul için hava kirliliği tahminleme probleminde hedef kirletici PM_{10}'un yoğunluk bilgisi EPA sınır değerlerine göre sınıflandırıldığında altı sınıf elde edilir. Bu dönüşümden sonra bazı sınıfların sadece birkaç örneğe sahip olması, tahmin modelinin ilgili sınıfı öğrenebilmesi için bu birkaç örneğin yeterli olmaması, problemin altılı sınıflandırma problemi olarak ele alınamayacağını göstermiştir. Alternatif olarak problem ikili sınıflandırma problemine dönüştürülüp; sınıfların veri kümesindeki dağılımları incelendiğinde, verideki negatif vakaların (azınlık sınıf örnekleri), pozitif vakalara oranla (baskın sınıf örnekleri) oldukça az olduğu görülmüştür. Buradaki negatif vakalar PM_{10} yoğunluğunun lokal eşik değerinin üstünde olduğu, hava kirliliğinin insan sağlığı için tehlike arz ettiği örnekleri temsil ederken, pozitif vakalar PM_{10} yoğunluğu bu eşik değerinin altında kalan, hava kirliliğinin kabul edilebilir seviyelerde olduğu zararsız örnekleri temsil etmektedir. Sınıflandırma

probleminde, sınıflara ait örneklerin veri kümesinde dengesiz dağılması tahmin modellerinin öğrenmesini zorlaştıran önemli bir sorundur. Dengesiz veri sorunu, makine öğrenmesi yöntemleriyle sınıflandırma yapan çalışmalarda sıkça karşılaşılan bir problemdir [17–19].

Biz bu tez kapsamında makine öğrenmesi algoritmalarından faydalanarak PM_{10} yoğunluğunu tahminleyen ve verideki dengesiz dağılım problemi ile başa çıkabilen İki Katmanlı Hava Kirliliği Tahminleme Modeli öneriyoruz. Modelimiz ilk aşamada problemi ikili sınıflandırma problemine dönüştürmekte ve Rastgele Orman Sınıflandırıcısı ile hava kirliliğinin yaşanıp/yaşanmadığına (1/0) karar vermektedir. İkinci aşamada ise örnekler hedef PM_{10} yoğunluğunu tahminlemek için, karar verilen sınıfa (1/0) göre regresyon modellerinden uygun olanına girdi olarak verilmektedir.

İkinci aşamadaki regresyon modelleri ikili etiketlenmiş verinin iki ayrı veri grubu olarak ele alınması, 0 ve 1 veri grupları üzerinde birer regresyon modeli eğitilmesi ile elde edilir. Buradaki regresyon modelleri birbirinden bağımsızdır. İkinci aşamada, örneklerin birinci aşamada tahminlenen sınıf değerine göre, uygun regresyon modelinde tekrar testi yapılmakta ve tahminlenen yoğunluk değeri çıktı olarak sunulmaktadır. Regresyon problemi öncesi yapılan ikili sınıflandırma ile tahminin doğruluğu için ön kontrol yapıldığı söylenebilir.

Tezin geri kalanı şu şekilde düzenlenmiştir. Bölüm 2, hava kirliliği/kalitesi tahminleme problemi ile ilgili literatürde öne çıkan çalışmaları sunmaktadır. 3. Bölüm önerilen tahminleme modeli ile birlikte, algoritmik yaklaşımlar ve örnekleme yaklaşımlarının ayrıntılarını açıklar. Bölüm 4; veri toplama ve veri kümesi üzerinde yapılan işlemleri ayrıntılı olarak anlatır. Modellerin değerlendirme ölçütleri, çalışmanın deneysel sonuçları yine bu bölümde sunulmaktadır. Tezden elde edilen sonuçlar ve yapılan çıkarımlar Bölüm 5'tedir.

2. LİTERATÜR TARAMASI

Hava kirliliği tahminleme yaklaşımları, uygulanan tekniklere dayalı olarak iki ana gruba ayrılmaktadır: deterministik modeller ve istatistiksel modeller.

Deterministik modeller, geçmişteki kirletici yoğunlukları ve gelecekteki kirletici yoğunluğuna dair öne sürülen senaryolar da dahil olmak üzere emisyon kaynakları, meteorolojik süreçler, fiziko-kimyasal değişimler ve kirletici yoğunlukları arasındaki deterministik ilişkiyi ölçen yöntemlerdir. Ayrıca kirleticilerin etkisini azaltma stratejileri de deterministik modelin bir parçasıdır. Diğer yandan, doğrusal ve doğrusal olmayan denetimli öğrenme yöntemlerini içeren istatistiksel modeller, rasgele olma özellikleriyle deterministik yöntemlerden kolayca ayırt edilmektedir. İstatistiksel modellerden makine öğrenmesi yaklaşımları, birçok hava kirliliği tahmini çalışmasında deterministik modellere üstünlüklerini kanıtlamıştır.

PM dışındaki hedef kirleticiler ile yapılan çalışmalar: Meteorolojik parametreler kullanılarak SO_2 ve NO_2 tahmini yapılan [20]'deki çalışmada, doğrusal model (Partial Least Square Regression (PLSR)) ve doğrusal olmayan modeller (Multivariate Polynomial Regression (MPR), Artificial Neural Networks (ANN)) kıyaslanmış, doğruluk değeri en yüksek olan sonuçlar ANN ile elde edilmiştir. ANN'in farklı yaklaşımları (Multilayer Perceptron Network (MLP), Radial-basis function network (RBFN), Generalized Regression Neural Network (GRNN)) karşılaştırıldığında ise öne çıkan GRNN olmuştur.

Özellik mühendisliğini vurgulayan çalışma [21]; sıcaklık, bağıl nem, yağış birikimi, dünya üzerindeki konum, ölçüm yapılan hafta ve konum numarası kullanarak Kasım 2009 ile Mart 2011 arasında Rawalpindi ve Islamabad bölgelerinde NO_2 tahmininde bulunur. Burada kullanılan konum numarası; bölgenin yakınında çift yönlü taşıma yolları, ana yollar, alt yollar, küçük yollar, devlet hastanesi, özel hastane, modern konut, ticaret alanı, dinlenme yerleri, otobüs durakları, okul, göl, orman vs. var ise (1) yoksa (0) şeklinde ikilik sayı sisteminde ifade edilmiş ve oluşturulmuştur. Model olarak ANN kullanılmış, en iyi ANN ağ yapısına evrimsel algoritma ile karar verilmiş, geri yayılımlı öğrenme ile de sonuç iyileştirilmiştir.

Dhirendra Mishra [22]'deki çalışmasında, Tac Mahal, Hindistan'da saatlik NO_2 tahmini için Multiple Linear Regression (MLR) ve Principle Component Analysis (PCA) ile desteklenmiş ANN modelini karşılaştırmış; PCA-ANN modelinin daha iyi performans gösterdiğini ve Tac Mahal, Agra'da hava kirliliğini tahmin etmek için kullanılabileceğini öne sürmüştür. Bir yıl sonra Hindistan'ın ikinci büyük metropolü Delhi'de NO_2, O_3 ve SO_2 tahmini için, öne çıkan model yine temeli yapay sinir ağlarına dayanan MLP olmuştur [23]. Boyut indirgemenin önemine vurgu yapan bir diğer çalışmada [24] Ana Russo ve arkadaşları sıcaklık, bağıl nem, yağış birikimi, sınır tabakası yüksekliği, basınç ve parlaklık meteorolojik parametrelerini kullanarak; Lizbon ve Portekiz için NO_2, NO ve CO tahmininde bulunmuşlardır.

Hedef kirleticinin PM olduğu çalışmalar: Milan'da hava kalitesi için kritik olarak görülen O_3 ve PM_{10}'un tahmininin yapıldığı [25]'de Feedforward Artificial Neural Network (FFANN) Based on Back Propagation yönteminin veri kümesindeki nadir durumları doğrusal modellere göre daha iyi öğrenmesinden bahsedilmiş, dezavantaj olarak ise FFANN ile oluşturulan modellerin uzun eğitim sürelerine sahip olmaları ve aşırı öğrenmeye sebep olabilmeleri gösterilmiştir.

Varşova için SVM, RBF ve MLP bireysel modelleri ile bu modelleri içeren topluluk modeli karşılaştırmış, topluluk modeli günlük ortalama PM_{10} tahmini için öne çıkan model olmuştur [26]. Kominski ve arkadaşları [27], Lodz için hava kalitesi sınıflandırma problemini ele alırken, farklı yapay sinir ağı modellerini denemiş, günlük ortalama PM_{10} tahmini için MLP ve RBF'in tatmin edici sonuçlar ürettiğini görmüşlerdir. Aynı yazarlar, [28]'de Lodz için günlük maksimum PM_{10} yoğunluğu tahminini iyi, kabul edilebilir, kötü sınıflarını tahmin etmeye yönelik sınıflandırma problemine dönüştürerek ele almışlar, yöntem olarak ise yine MLP ve RBF tercih etmişlerdir. Yapılan iki çalışma da [27,28] yapay sinir ağlarının tahmin sistemlerindeki önemini vurgulamaktadır.

Meteorolojik parametre ve kirlilik parametrelerinin iki ayrı veri kümesiymiş gibi ele alındığı [29]'da Extreme Learning Machine (ELM) ve SVM performansları eğitim süresi ve doğruluk değerleri yönünden karşılaştırılmıştır. ELM'in özellikle sınıflandırmada azınlıkta bulunun sınıf için SVM'e göre daha doğru sonuçlar ürettiği, genel değerlendirmede öne çıkan modelin de yine ELM olduğu görülmüştür. Hong Kong bölgesinde sıcaklık, rüzgar, bağıl nem ve mevsim bilgileri kullanılarak 2010

- 2015 yılları arasında NO_2, O_3, $PM_{2.5}$ ve SO_2 yoğunlukları tahmin edilmiştir [30]. Model olarak Multiple Linear Regression (MLR), FFANN ve ELM karşılaştırılmış; iki katmanlı yapay sinir ağının özelleştirilmesi ile elde edilen ELM en iyi sonucu üretmiştir.

ANN kullanılan [29]'da Barselona ve Montseny için PM_{10} yoğunluğu tahmini yapılmış, duyarlılık analizi ile meteorolojik özniteliklerin tahminlemeye katkısı incelenmiştir. İspanya'daki bir diğer çalışmada [31], Ovieodo için aylık ortalama PM_{10} yoğunluğu tahmininde; Vector Autoregressive Moving-Average (VARMA), Autoregressive Integrated Moving-Average (ARIMA), MLP ve SVM karşılaştırdığında; SVM'in diğer yöntemleri geride bıraktığı görülmüştür. Mekke'de saatlik PM_{10} tahmini için, MLR ve ANN performansları karşılaştırılmış, ANN tahmin için tercih edilen model olmuştur [32].

Ekvador'da Cotocollao ve Belisario için $PM_{2.5}$ ile hava kirliliği tahmini yapılmıştır [33]. Çalışmadaki ölçüm noktalarındaki veri kümeleri; yağış miktarı (mm), rüzgar yönü (0–360), rüzgar hızı (m/s) ve gözlenen partikül maddenin yoğunluğu (g/m^3) değerlerini gösteren 4 parametreye ait günlük periyottaki verileri içermektedir. Burada $PM_{2.5}$ için sınıflandırma problemi, ikili sınıflandırma ve üçlü sınıflandırma problemi şeklinde ele alınarak; farklı sınıflandırma yaklaşımları için doğrusal SVM ile topluluk modellerinden ADABoost'un performansları karşılaştırılmıştır.

2017'de Chicago'da Absip Village bölgesinde O_3 ve $PM_{2.5}$; Lemont Village bölgesinde ise O_3 ve SO_2 yoğunlukları aracılığıyla hava kirliliği tahmini yapılmıştır [34]. Kirleticilerin yoğunlukları tahminlenirken sıcaklık, rüzgar, bağıl nem, basınç, görünürlük, yağış birikimi ve çiğ noktası meteorolojik parametrelerinden faydalanılmıştır. Farklı genelleştirme ve optimizasyon yöntemlerinin kıyaslandığı araştırmada Consecutive Close (CC) genelleştirmesi ve Stochastic Alternating Direction Method of Multipliers (LA-SADMM) optimizasyon yöntemi kullanılan model ön plana çıkmıştır.

$PM_{2.5}$ kirleticisinin yoğunluğu tahminlenirken, uydu görüntüleri aracılığıyle elde edilmiş Aerosol Optical Depth (AOD) bilgisinden ve trafik yoğunluk verilerinden faydalanılmıştır [35]. Görüntü işlemeye dayalı bir diğer çalışmada [36], MODerate

resolution Imaging Spectroradiometer (MODIS) ve AOD uydu görüntülerine ek olarak uydu bazlı gece ışıkları ortalamaları PM_{10} tahmini için kullanılmıştır.

Çin'in Chongqing şehrinin işlek caddelerinden Zhongshan caddesinin, iki yanındaki lazer toz monitörlerinden elde edilen $PM_{2.5}$, PM_5 ve PM_{10} yoğunluk verileri hava kalitesi tahmini için kullanılmıştır [37]. Caddenin iki tarafındaki bina yüksekliği, sokak genişliği, trafik yoğunluğu, caddedeki araç tipleri gibi özellikler kullanılarak yapılan tahminlemede, yöntem olarak RBF ve geri beslemeli yapay sinir ağı seçilmiştir.

PM_{10}, SO_2, NO_2, sıcaklık, basınç, nem, rüzgar yönü, rüzgar hızı parametreleri ile $PM_{2.5}$ konsantrasyonunu tahmin etmek için yapılan [38]'deki çalışmada, Gauss transfer fonksiyonu kullanan RBF'in geri beslemeli yapay sinir ağlarına göre yüksek doğruluk değerleri ürettiği görülmüştür.

PM_{10} ile zaman serileri analizi için Gri Tahmin Modeli (GTM) önerilmiştir [39]. Sistem davranışını tanımlamak ve sistemde sürekli değişen süreci ortaya koymak için sadece küçük miktarlarda veriye ihtiyaç duyulan gri tahmin teorisi, sosyal bilimler araştırmalarında da yaygın olarak kullanılmaktadır. PM_{10} ve $PM_{2.5}$ yoğunluk tahmini için [40]'da yapay sinir ağlarının döngü içeren modeli, Recurrent Neural Network (RNN) kullanılmıştır. RNN'in performansı, Güney Kore'nin başkenti Seul'daki metro istasyonlarından alınan veri üzerinde; FFANN ve MLR ile karşılaştırılmıştır. Kirletici yoğunluklarının PM_{10} ve $PM_{2.5}$ tahminine etkisinin de incelendiği araştırmada, yapısında azot bulunan bileşiklerin, yapısında karbon bulunan bileşiklere göre partikül maddelerin tahmininde daha etkili olduğu görülmüştür.

İstanbul için hava kirliliği tahminleme problemini EPA [4] sınır değerlerinden faydalanarak ikili sınıflandırma problemi olarak ele aldığımızda, veri kümesinin dengesiz veri dağılımı probleminden muzdarip olduğu görülmüştür. Dengesiz dağılım sorunu sınıflandırma problemlerinde sınıflara ait örneklerin veri kümesinde aynı ya da birbirlerine yakın oranlarda bulunmamasından kaynaklanmaktadır. Literatürde dengesiz veriden öğrenme olarak da yer bulan bu problem için önerilen çözümler temelde ikiye ayrılır: Örnekleme Yaklaşımları ve Algoritmik Yaklaşımlar [18]. Örnekleme yaklaşımlarına bakıldığında, yukarı örnekleme işlem süresini uzatması ve yapay veri kullanması, aşağı örnekleme ise veriyi temsil eden alt kümeyi belirlemenin

zorluğu nedeniyle algoritmik yaklaşımların gerisinde kalmaktadır [41]. Algoritmik yaklaşımlarda öne çıkan modeller ise hiyerarşik yapısının sağladığı avantaj ile ağaç temelli topluluk modelleridir.

Daha önce İstanbul'da Beşiktaş bölgesi için araştırma yapan Kurt ve Oktay; günlük periyottaki kirlilik verilerini, meteorolojik verileri ve mekansal bilgileri kullanarak SO_2, CO ve PM_{10} seviyelerinin tahmini için coğrafik sınıflandırma modeli oluşturmuşlardır [42]. Burada verinin dengesiz dağılma problemine çözüm sunulmamış, tehlikeli sınıfa ait örnekler göz ardı edilmiştir.

Ortamdaki hava kirliliğinin sonuçları yerel ve global sonuçlar olarak ikiye ayrılabilir. Yerel sonuçlar insan sağlığı, bitki örtüsü, ham madde ve kültürel ürünler üzerinde etkiliyken, küresel sonuçlar sera etkisine, iklim değişikliğine ve troposferik/stratosferik ozon etkisine neden olabilir.

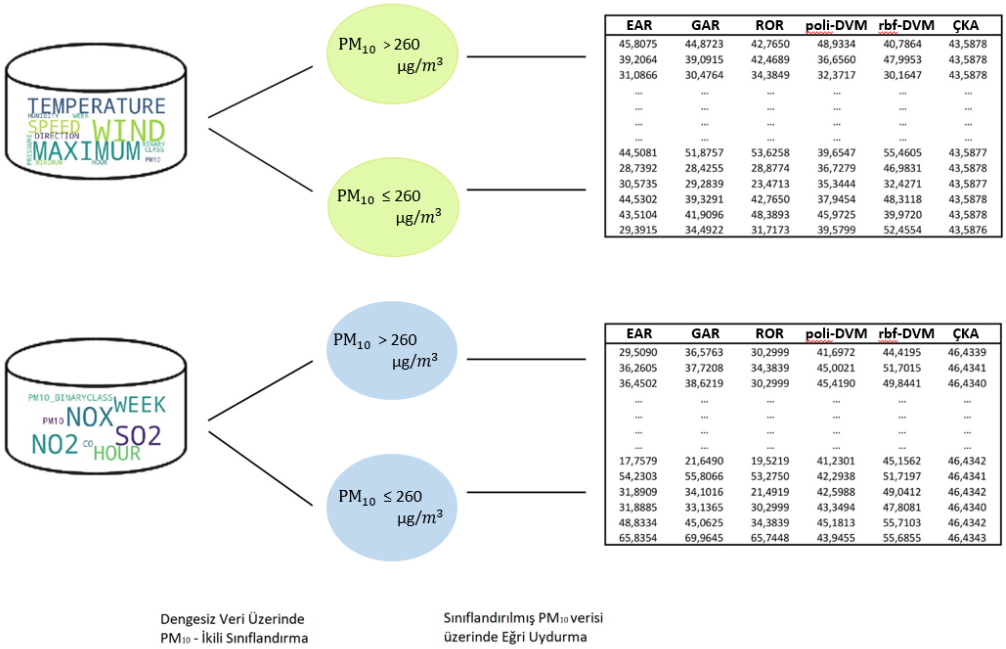
Yapılan çalışmalar incelendiğinde, Türkiye'de hava kirliliği tahmini yapan çalışma sayısı oldukça azdır. Mevcut çalışmalarda, hava kirliliğinin tehlikeli seviyelerde olduğu ölçümlere ait örnek sayısının az olması sebebiyle bu ölçümler göz ardı edilmiş, dengesiz dağılım sorununa yönelik çözüm üretilmemiştir. Toplululuk modellerinden bagging, boosting gibi yöntemleri kullanan çalışma ise bulunmamaktadır. Sürekli değişen mevsimsel olaylar (küresel ısınma-mevsimsel farklılıklar), günden güne artan kirlilik oranları düşünüldüğünde eski çalışmaların gerçekçi sonuçlar üretmesi mümkün değildir.

Bu tez kapsamında İki Katmanlı Hava Kirliliği Tahminleme Modeli kullanılarak PM_{10} aracılığıyla İstanbul için hava kirliliğinin yoğunluğu tahminlenmektedir. Önerilen İki Katmanlı Hava Kirliliği Tahminleme Modeli şu ana kadar önerilen modellerden farklı, iki katmanlı bütünleşik yapısı sayesinde dengesiz veriden öğrenme problemi ile başa çıkabilen bir modeldir. Modeldeki yenilik regresyon modelinin ön aşamasında uygulanan ikili sınıflandırma tekniğinden kaynaklanmaktadır. Modelimiz benzer veri dağılımına sahip, İstanbul dışındaki diğer bölgeler için hava kirliliği tahmini çalışmalarında da kullanılabilir.

3. HAVA KİRLİLİĞİ TAHMİNLEME MODELİ

PM_{10} kirleticisi aracılığıyla İstanbul için hava kirliliği tahminleme problemi dengesiz dağılan veri kümesine sahiptir. Yüksek oranlarla dengesiz dağılan veri kümelerinin sıkça görüldüğü hava kirliliği tahminleme çalışmalarının çok az bir bölümü [29] dengesiz dağılım problemine dikkat çekmektedir.

Farklı çalışma alanlarında da rastlanan bu probleme çözüm olarak temelde iki yaklaşım bulunmaktadır: Algoritmik Yaklaşımlar ve Örnekleme Yaklaşımları. Bunların içinde Algoritmik Yaklaşımlar, dengesiz dağılıma uygun tahminleme yöntemlerini ve uygun yöntem parametrelerini seçerek sorunu çözmeye çalışır. Bu yaklaşım ile veri kümesinde sınıfların dağılım oranları korunur. Öte yandan, Örnekleme Yaklaşımları, örnek ekleyerek veya çıkararak veri kümesindeki sınıfların dağılımını değiştirir. Buradaki amaç sınıflara ait örnek sayılarının eşit olduğu, dengeli veri kümesini elde etmektir.



Şekil 3.1 : Meteorolojik Veri ve Kirlilik Verisi ile İki Katmanlı Hava Kirliliği Tahminleme Modeli.

Bu tez kapsamında, dengesiz veri dağılımı ile başa çıkabilen İki Katmanlı Hava Kirliliği Tahminleme Modeli önerilmektedir. Şekil 3.1'de tasarımı verilen modelin ilk aşamasında EPA'dan hedef kirletici için yoğunluk eşik değeri elde edilir ve problem ikili sınıflandırma problemi olarak ele alınır. İkinci aşamada ise kirleticinin yoğunluk değeri bulunur. PM_{10} kirleticisi için özelleştirilmiş haliyle önerilen modelin aşamaları adım adım şu şekildedir:

1. EPA'dan elde edilen bilgiye göre PM_{10} için lokal eşik değeri 260 $\mu g/m^3$'tür. Bu eşik değerine göre PM_{10} yoğunluğu 260 $\mu g/m^3$'ten küçük veya eşit olan örnekler 1, eşik değerinin üstündeki örnekler ise 0 olarak işaretlenir.

2. 0 ve 1 olarak işaretlenmiş veri, iki veri grubu olarak ele alınır. Her bir veri grubu için iki ayrı regresyon modeli eğitilir. Bu yolla modelin ikinci katmanında kullanılmak üzere pozitif örnekler için bir, negatif örnekler için bir regresyon modeli elde edilir.

3. Veri kümesi sınıf dağılımları korunacak şekilde eğitim ve doğrulama kümelerine ayrılır.

4. Modelin ilk aşamasında, ikili sınıflandırma problemi için eğitim kümesi üzerinde Rastgele Orman Sınıflandırıcısı eğitilir ve eğitilen modelle test kümesindeki örneklerin sınıf bilgisi bulunur.

5. Test örnekleri, 4. adımda bulunan sınıf bilgisine göre, 2. adımdaki regressörlerden uygun olanına verilir. Bu regressörün çıktısı hedef değişkenimiz PM_{10}'un tahminlenen yoğunluk bilgisini sunar.

Modelin oluşturulmasında faydalanılan ve tez içerisinde kullanılan yöntemler algoritmik düzeyde ve örnekleme düzeyinde Bölüm 3.1 ve Bölüm 3.2'de detaylı olarak açıklanmaktadır.

3.1 Örnekleme Yaklaşımları

Dengesiz dağılan veri kümesi üzerinde ikili sınıflandırma için örnekleme yaklaşımları, pozitif ve negatif sınıflardaki örnek sayılarını eşitlemeyi amaçlamaktadır. Seçilen yönteme bağlı olarak, ya Yukarı Örnekleme ile tehlikeli/azınlık sınıf örneklerinin sayısı arttırılır ya da Aşağı Örnekleme ile zararsız/baskın sınıf örneklerinin sayısı azaltılır.

Tehlikeli sınıftaki örneklerin sayısı zararsız sınıftaki örneklerin sayısına eşit olacak şekilde veri kümesinin dağılımı dengelenir.

3.1.1 Aşağı örnekleme

Aşağı Örnekleme, baskın sınıfın örnekleri arasından azınlık sınıfın örnek sayısı kadar örnek seçerek veri kümesini dengeler. Bu tür örnekleme, rastgele seçimle veya sezgisel yöntemler ile yapılabilir. Farklı dengeli veri kümeleri elde etmek için aşağı örnekleme yaklaşımlarından rastgele örnekleme ve NearMiss örneklemesinin üç farklı versiyonu kullanılmıştır. Aşağı örnekleme ile amacımız, baskın sınıfı en iyi temsil eden daha küçük ve azınlık sınıfla dengeli dağılan veri kümesini bulmaktır.

- Rastgele Örnekleme: Tehlikeli sınıfa ait örnek sayısı kadar örnek, zararsız sınıftan rastgele seçilir.
- Near Miss Örneklemesi : Near Miss en yakın komşu algoritmasına dayanan sezgisel örnekleme yöntemidir [41]. Farklı versiyonları olan bu yöntem, temelde azınlık sınıftan N kadar komşu seçilimi ile başlar. Bu çalışmada denemeler sonucu N üç olarak seçilmiştir. Tezde yer bulan Near Miss'in üç farklı versiyonu:
  - **NearMiss-1** öncelikle zararsız sınıfa en yakın üç tehlikeli sınıf örneği tespit edilir. Bu üç örneğe ortalama uzaklığı en az olan zararsız sınıf örneklerinden, tehlikeli sınıf örnek sayısı kadar örnek uzaklıkla ters orantılı olarak sırayla seçilir.
  - **NearMiss-2**, tehlikeli sınıfın en uzak üç örneğine ortalama uzaklığı en az olan zararsız sınıf örnekleri arasından tehlikeli sınıf örnek sayısı kadar örnek seçer.
  - **NearMiss-3** iki adımlı bir algoritmadır. İlk olarak, her negatif sınıf örneği için M en yakın komşuları tutulur. Daha sonra, seçilen pozitif sınıf örnekleri, en yakın üç komşuya olan ortalama mesafenin en fazla olduğu örneklerdir.

3.1.2 Yukarı örnekleme

Yukarı Örnekleme, azınlık sınıfın etkisini güçlendirmek için bu sınıfa ait örneklerin sayısını arttırarak veri kümesindeki sınıfların oranlarını dengeleme sürecidir. Yukarı örnekleme için, veri kümesindeki azınlık sınıf örnekleri rastgele çoğaltılabilir ya

da SMOTE [43] gibi algoritmalar kullanarak sentetik veriler üretilebilir. Yukarı Örnekleme, veri kümemizin boyutunu neredeyse iki katına çıkararak, eğitim süresini kayda değer ölçüde artırdığından bu çalışmada kullanılmamıştır.

3.2 Tahminleme Algoritmaları

Dengesiz veri üzerinde yapılan çalışmalarda, problem ister sınıflandırma problemi olsun ister regresyon problemi olsun makine öğrenmesi yöntemlerinden topluluk modelleri öne çıkmaktadır [26]. Topluluk modelleri, tek bir sınıflandırıcı yerine birden fazla zayıf sınıflandırıcı kullanır. Bu modellerin kendi içlerindeki farklılıkları tahmin sürecinde izlenen yoldan kaynaklanmaktadır. Boosting yöntemleri zayıf sınıflandırıcıların tahminlerinin ağırlıklı ortalamasını alırken, bagging yöntemleri zayıf sınıflandırıcılardan gelen tahminlerin aritmetik ortalamasını almaktadır. Tek bir karar ağacı yerine karar ağaçlarından oluşan topluluk modelini tercih etmek, tahmin modelini genelleştirir ve yapılan tahminin doğruluğunu arttırır.

Bu çalışmada topluluk modellerinden üçü Ekstra Ağaç, Rastgele Orman ve Gradyan Arttırma kullanılmıştır. Dengesiz veriden öğrenmede popülerleşmiş topluluk modellerine ek olarak veri madenciliği ile yapılan çalışmalarda öne çıkan DVM'e de yer verilmiştir. Algoritmalar ile tahminleme aşamasında Python Scikit Learn kütüphanesinden [44] faydalanılmıştır.

Tahminleme algoritmaları tez içerisinde regresyon veya sınıflandırma için kullanılmasına bağlı olarak farklı şekilde anılmaktadır. Örneğin DVM algoritması regresyon için kullanılıyorsa DVMR, sınıflandırma için kullanılıyorsa DVMS ile ifade edilmektedir. Bu yolla problemin sınıflandırma aşamasında mı yoksa regresyon aşamasında mı olduğu belirtilmektedir.

3.2.1 Topluluk modelleri

Alt başlıklarda çalışma prensiplerinin detayları verilen Ekstra Ağaç, Rastgele Orman ve Gradyan Arttırma algoritmaları için model eğitimleri sırasında optimize edilen hiper parametreler:

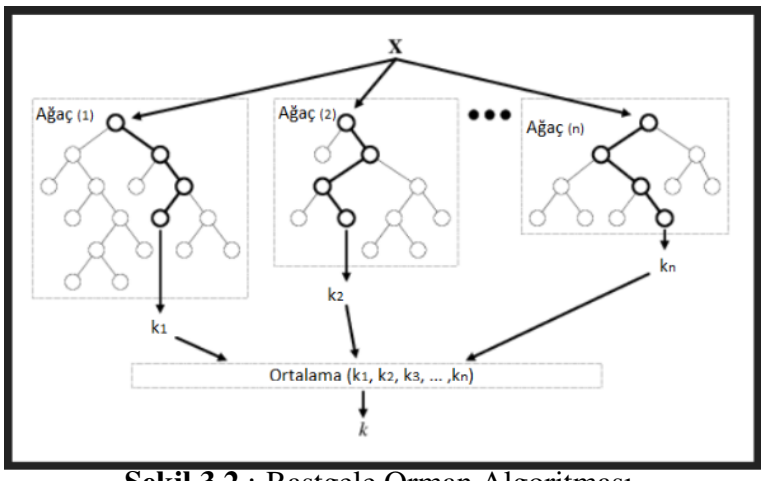
- **min_samples_leaf** : Bir yaprak düğümde olması gereken minimum örnek sayısı. Tamsayı bir değer ise, yaprak düğümdeki minimum örnek sayısıdır, eğer ondalıklı

değer ise örnek sayısı üzerindeki oran üzerinden hesaplanır. Bu hiper parametre, özellikle regresyonda modeli yumuşatma konusunda işe yarayabilir.

- **min_samples_split** : Ağaca ait bir düğümü ayırmak için gereken minimum örnek sayısını ifade eder. Tamsayı bir değer ise, düğümü ayırmak için kullanılacak minimum örnek sayısıdır, eğer ondalıklı değer ise örnek sayısı üzerindeki oran üzerinden hesaplanır.
- **max_depth** : Bireysel karar ağacı tahminleyicilerinin maksimum derinliği. Maksimum derinlik, ağaçtaki düğüm sayısını sınırlar.
- **max_features** : En iyi bölünmeyi ararken dikkate alınacak özellik sayısı. Toplam özellik sayısından az seçilmesi varyansın azalmasına ve hatanın artmasına yol açabilir.

3.2.1.1 Rastgele Orman

Topluluk algoritmalarından Rastgele Orman, *bootstrap* [45] metodunu kullanarak eğitim kümesinin alt kümelerini elde eder. Bu alt kümeler üzerinde bir dizi karar ağacını çalıştırır ve rastgele orman oluşturur.



Şekil 3.2 : Rastgele Orman Algoritması.

Son tahmin için karar ağacı modellerinin (Şekil 3.2) ürettiği sonuçların ortalamasını alır. Tek bir ağaç modeli ile çalışmaktansa, birden çok ağaç modeli kullanıp bu modellerin ürettiği sonuçların ortalamasının alınması, başarımı arttırır ve aşırı uyum problemini engeller.

3.2.1.2 Ekstra Ağaç

Ekstra Ağaç [46] algoritmasının iki özelliği onu diğer topluluk modellerinden farklı kılmaktadır. Birincisi, kesme noktalarını rastgele seçerek ağaç düğümlerini ayırmasıdır. İkincisi ise, *bootstrapping* ile veri kümesini çoğaltma yerine tüm eğitim kümesini kullanarak ağaçları büyütmesidir.

Ağaç oluşumunu ekstra rasgeleleştirmeye çalışan Ekstra Ağaç modelini eğitmek genellikle Rastgele Ormana göre daha kolaydır. Ancak genelleştirme konusunda bazı nadir uygulamalar dışında Rastgele Orman yöntemi ön plana çıkar.

3.2.1.3 Gradyan Arttırma

Boosting bir algoritma olan Gradyan Arttırma [47] öncelikle sınıflandırma problemlerinde kullanılmak üzere önerilmiş, zamanla hem regresyon hem sınıflandırma problemlerinde tercih edilen bir yönteme dönüşmüştür. Gradyan Arttırma'da tekli karar ağacı sınıflandırıcıları çoğunlukla sırayla eğitilmektedir.

Örneklerin sınıflarını tahmin etmek için, öncelikle zayıf bir sınıflandırıcı üretilir. Daha sonra, bu sınıflandırıcının çıktısı, kayıp fonksiyonunun değerini hesaplamak için hedef ile karşılaştırılır. Bu çalışmada *"lojistik regresyon"* olarak seçtiğimiz kayıp fonksiyonu, Gradyan Arttırma için değiştirilebilen bir parametredir. İkinci adımda daha güçlü bir sınıflandırıcı elde etmek için, ilk adımdaki kayıp fonksiyonu kullanılır.

Her yinelemede, kayıp fonksiyonunun kalıntısı, *Gradient Descent* [48] yöntemi kullanılarak hesaplanır ve bir sonraki iterasyon için hedef değer haline gelir. Özetle Gradyan Arttırma, *Gradient Descent* yöntemini kullanarak her bir zayıf sınıflandırıcının türevlenebilir kayıp fonksiyon değerini minimuma indirmeye çalışır.

3.2.2 DVM

Destek Vektör Makinesi (DVM), ilk olarak ikili sınıflandırma problemine çözüm olarak sunulmuş çekirdek tabanlı makine öğrenmesi algoritmasıdır. Şimdilerde ise sınıflandırma, regresyon ve aykırı durum saptaması alanlarında kullanılan gözetimli öğrenme metodu olarak çalışmalarda yer bulmaktadır.

DVM'in ikili sınıflandırmadaki temel amacı sınıfları ayıran hiper düzlemler içerisinden en uygun hiper düzlemi seçmektir. Bunun için farklı sınıflara ait destek vektörleri arasındaki uzaklığı maksimize ederek ayıran hiper düzlemi bulmaya çalışır. DVM, çekirdek parametresinin iki farklı seçimi, rbf-DVM ve poli-DVM versiyonları ile bu çalışmada yer bulmuştur. DVM için optimize edilen hiper parametreler ise şunlardır:

- **C**: Hata teriminin penaltı parametresi.
- **degree**: Polinom çekirdek fonksiyonunun derecesi. Diğer çekirdekler tarafından kullanılmaz.

4. DENEYSEL ÇALIŞMALAR

Bu tez kapsamında İstanbul için PM_{10} kirleticisi aracılığıyla hava kirliliği yoğunluğu tahmini yapılmıştır. Avrupa ve Anadolu yakasında hava kirliliğine oldukça maruz kalan toplamda dokuz bölgeye ait Ağustos 2011 - Şubat 2018 tarih aralığını kapsayan gerçek dünya verileri tahminlemede kullanılmıştır. Bu bölgeler Aksaray, Alibeyköy, Beşiktaş, Esenler, Kartal, Sarıyer, Silivri, Üsküdar ve Yenibosna bölgeleridir.

Bir bölgedeki hava kirliliği öncelikle bölgenin meteorolojik özelliklerinden etkilenir. Komşu bölgelerdeki kirlilik; rüzgar, yağış, nem gibi meteorolojik etkenler ile taşınarak, kirliliğin kaynağı olmayan bölge için bile tehdit oluşturabilir. Meteorolojik koşulların havadaki kirletici konsantrasyonlarının ölçümündeki etkileri ve önemi [24, 25] göz önünde bulundurularak; çalışmada kullanılmak üzere saatlik sıcaklık, rüzgar yönü, rüzgar hızı, aktüel basınç, nispi nem, minimum sıcaklık, maksimum sıcaklık, maksimum rüzgar yönü ve maksimum rüzgar hızı parametrelerini içeren meteorolojik veri Meteoroloji Genel Müdürlüğü'nden alınmıştır. Elde edilen meteorolojik verideki özniteliklerin birimleri ve veri tipleri "Çizelge 4.1" 'deki gibidir.

Çizelge 4.1 : Meteorolojik Veri Kümesi Öznitelikleri.

	METEOROLOJİK PARAMETRE	BİRİM	VERİ TİPİ
1	Sıcaklık	°C	Tamsayı
2	Rüzgar Yönü	°	Tamsayı
3	Rüzgar Hızı	m ÷ s	Ondalıklı Sayı
4	Aktüel Basınç	hPa : hektopaskal	Ondalıklı Sayı
5	Nispi Nem	%	Ondalıklı Sayı
6	Minimum Sıcaklık	°C	Ondalıklı Sayı
7	Maksimum Sıcaklık	°C	Ondalıklı Sayı
8	Maksimum Rüzgar Yönü	°	Tamsayı
9	Maksimum Rüzgar Hızı	m ÷ s	Ondalıklı Sayı

Tahminlemede kullanılan meteoroloji veri kümesindeki öznitelikler, ölçüm istasyonlarına göre farklılık gösterebilir. Yeni kurulan istasyonlarda rüzgar, sıcaklık, basınç gibi temel meteorolojik parametrelere ek olarak minimum sıcaklık, maksimum rüzgar hızı gibi parametreler de bulunabilir. Meteorolojik parametrelerden sıcaklık, rüzgar yönü ve rüzgar hızı parametrelerinin ölçümlerinin güvenilirliğini garantilemek için ölçüm saati içinde bir dizi ölçüm yapılmaktadır. "Çizelge 4.1"de 'maksimum' ile

belirtilen sıcaklık, rüzgar hızı veya rüzgar yönü parametre değerleri, verilen parametre için o saat içinde görülen en yüksek değeri ifade etmektedir. 'minimum' ise benzer şekilde sıcaklık parametresi için aynı saat içindeki en düşük değeri ifade etmektedir.

Kirleticilerin meteorolojik parametrelerden etkilendiği ve bu bağlamda birbirlerini de etkiledikleri göz önünde bulundurulduğunda, hava kirliliği tahmini yaparken meteorolojik veri yanında kirlilik verisini de kullanmak faydalıdır. Bu amaçla CO, NO, NO_2, NO_X, O_3 ve SO_2 kirleticilerinin saatlik yoğunluk verileri meteoroloji istasyonlarına en yakın kirletici istasyonlarından toplanmıştır.

Elde edilen meteorolojik veri ve kirlilik verisi PM_{10} yoğunluğunu tahmin etmek için ayrı ayrı kullanılmıştır. Üzerinde çalıştığımız kirlilik ölçüm istasyonları ve meteoroloji istasyonlarının eşleşmesi "Çizelge 4.2"de verilmiştir.

Çizelge 4.2 : Kirlilik Ölçüm İstasyonu – Meteoroloji İstasyonu Eşleşmesi.

KİRLİLİK ÖLÇÜM İSTASYONU	***METEOROLOJİ İSTASYONU***		
	ID	***İlçesi***	***Adı***
Aksaray	17603	FATİH	İst.Den.Bil.Ens.
Alibeyköy	18101	EYÜP	Eyüp
Beşiktaş	18401	ŞİŞLİ	Şişli
Esenler	17814	GÜNGÖREN	Davutpaşa Marmara
Kartal	17064	KARTAL	İstanbul Bölge
Sarıyer	17061	SARIYER	Sarıyer
Silivri	18400	SİLİVRİ	Silivri
Üsküdar	18404	ÜSKÜDAR	Üsküdar
Yenibosna	17060	BAKIRKÖY	Atatürk Havaalanı

Havadaki kirleticilerin yoğunluklarının pozitif değerlere sahip olması gerekmektedir [4]. Bir kirleticinin yoğunluğunun sıfır olması ya da negatif bir değere sahip olması, ölçüm yapılırken ya da ölçüm değerinin kaydedilmesi sırasında (cihaz ya da insan kaynaklı) hata yapıldığını gösterir. Dolayısıyla kirletici yoğunluğunun sıfır olduğu ya da negatif bir değere sahip olduğu saatler, o saatte ölçüm yapılmaması ile eş değerdir. Kirletici yoğunluğunun sıfır olması, negatif olması ya da o saatte ölçümünün yapılmamış olması (NAN) problemleri seyreklik yaratmaktadır. "Çizelge 4.3"de kirletici ölçüm istasyonlarına ait doluluk oranları verilmiştir. Burada kalın yazı tipiyle belirtilen saatlik ölçüm sayıları her bir istasyondaki toplam örnek sayısını gösterir. Veri kümelerindeki eksik verilerin sayısı göz ardı edilebilecek miktarda olduğundan, eksik veri barındıran örneklerde veri doldurma yöntemlerinden birinin uygulanması

Çizelge 4.3 : Kirlilik Veri Kümeleri Doluluk Oranları.

		KİRLETİCİLER								
	Kirletici Ölçüm İstasyonu Adı		**CO**	**NO**	**NO_2**	**NO_x**	**O3**	**PM_{10}**	**$PM_{2.5}$**	**SO_2**
1	**Aksaray**	**NAN**	746	780	809	780	4849	675	7857	479
		=0	2312	1437	1360	1332	9097	1208	16435	203
		<0	4	43	56	1	0	2	1	158
		51413 saatlik ölçüm								
2	**Alibeyköy**	**NAN**	1523	873	872	850	1421	246	13386	5569
		=0	1733	6109	5418	5301	1039	6248	5569	2349
		<0	9	1015	84	39	128	9	3	1542
		54704 saatlik ölçüm								
3	**Beşiktaş**	**NAN**	122	721	717	734	9494	219	42079	181
		=0	1936	1625	1624	1630	653	1151	1377	817
		<0	159	40	221	14	83	36	26	523
		52203 saatlik ölçüm								
4	**Esenler**	**NAN**	122	1130	1128	1109		74	37854	260
		=0	897	713	665	287		340	1874	197
		<0	800	80	47	22		0	0	2657
		52245 saatlik ölçüm								
5	**Kartal**	**NAN**						229		177
		=0						292		16159
		<0						0		180
		56080 saatlik ölçüm								
6	**Sarıyer**	**NAN**						369		464
		=0						2733		2733
		<0						0		2
		55832 saatlik ölçüm								
7	**Silivri**	**NAN**		3789	687	4076	718	701	179	
		=0		31	30	30	40	148	154	
		<0		1389	0	0	320	37	6	
		44132 saatlik ölçüm								
8	**Üsküdar**	**NAN**	196	4032	175	4299		375		
		=0	12	0	0	0		84		
		<0	207	355	2	1		1		
		44048 saatlik ölçüm								
9	**Yenibosna**	**NAN**	16736					300		242
		=0	5232					392		4086
		<0	6					0		437
		55442 saatlik ölçüm								

yerine bu örneklerin veri kümelerinden çıkarımı tercih edilmiştir. Tabloda kırmızı ile belirtilen kirleticiler ise eksiksiz veri kümesi oluşturulurken veri kümesinin boyutunu oldukça küçülteceği için veri kümesine dahil edilmemiştir.

Meteorolojik verideki rüzgar yönü ve maksimum rüzgar yönü parametrelerinin değerleri, 0 ile 360 arasında değişen tamsayılardır. Bu öznitelikler kategorik değişkenlerdir ve diğer özniteliklerden farklı olarak model eğitiminden önce ön işleme uygulanması zorunludur. Kategorik öznitelikler dörtlü, sekizli ve onaltılı etiketleme yoluyla ifade edilmiş ve tahminlemeye etkileri test edilmiştir. Bu dönüşüm ile örneğin dörtlü etiketlemede; rüzgar yönü için 0-90, 91-180, 181-270 ve 271-360 aralıkları sırasıyla 1, 2, 3 ve 4 ile temsil edilmiştir. Öznitelikler gösterimleri değiştirildikten sonra, *'One Hot Encoding'* uygulanmış ve bu gösterimlerin tahmin modelleri üzerindeki etkileri karşılaştırılmıştır. Dörtlü etiketleme gösterimininin diğer etiketlemelerden ve özgün gösterimden daha iyi sonuçlar ürettiği görülmüştür. Bu yüzden rüzgar yönü ve maksimum rüzgar yönü öznitelikleri için dörtlü etiketleme dönüşümü uygulanmıştır.

Kirleticilerin yoğunluğu, saatlik, günlük, yıllık bazda belirli özellikler gösteren meteorolojik parametrelerden etkilenir. İstanbul için havadaki kirletici maddelerin yoğunluklarının yaz aylarında yüksek sıcaklıklara ve buharlaşmaya, kış aylarında ise yüksek düzeyde yakıt tüketimine bağlı olarak en yüksek seviyelerde görülmesi; diğer mevsimlerin ise ortalama karakteristiğe sahip olması sürpriz değildir. Veri kümesindeki zamansal bağımlılığı tahmin modellerinde ifade etmek için ölçüm tarihinden saat (0-23) ve hafta (1-52) bilgisi çekilmiş, bu iki yeni öznitelik veri kümelerine eklenmiştir.

4.1 Öznitelik Seçimi

Öznitelik seçimi, veri kümesinde hedef değeri öğrenmeye daha çok katkısı olan özniteliklerin tutulması, sınırlı katkısı olan ya da öğrenmeye yardımcı olmayan özniteliklerin ise veri kümesinden çıkarılması ilkesine dayanır. Doğru öznitelik seçimi ile sınıflandırma modelinin öğrenme hızı artmakta, hesaplama maliyeti ve hafıza kullanımı azalmaktadır. Öznitelik seçimi ayrıca aşırı öğrenme probleminden kaçınmak için de kullanılabilmektedir. Veri kümelerinde öznitelik seçimi için 'Tek Değişkenli

Öznitelik Seçimi' ve 'Ağaç Temelli Topluluk Modeli Kullanarak Öznitelik Seçimi' yöntemleri uygulanmıştır.

4.1.1 Tek değişkenli öznitelik seçimi

Tek değişkenli öznitelik seçiminde; veri kümesindeki öznitelikler, tahminlemeye sağladıkları fayda açısından değerlendirilirken, özniteliklerin kendi aralarındaki korelasyon ilişkileri göz ardı edilir. Öznitelikleri bireysel değerlendirmeye dayanan bu yönteme puanlama yöntemi de denilmektedir. Bir öznitelik için puan değeri ne kadar yüksekse, o öznitelik hedef değeriyle o kadar yüksek bağıntılı demektir.

Özniteliklerin değerlendirilmesi için f-testine dayanan *f_regression* yöntemi kullanılmıştır. f-testi temelde iki özniteliğin varyanslarının ya da standart sapmalarının birbirine eşit olup olmadığını test eder ve bu hipotezin gerçekliğini sınar. Sadece f-istatistiklerini hesaplar ve en önemli özniteliği seçerse, doğrusal regresyon analizine benzer.

Çizelge 4.4 : Kirlilik Verisi Tek Değişkenli Öznitelik Seçimi Puan Değerleri.

	Aksaray		**Alibeyköy**		**Beşiktaş**		**Esenler**	
	Öznitelik	**Puan**	**Öznitelik**	**Puan**	**Öznitelik**	**Puan**	**Öznitelik**	**Puan**
1.	NO	7962.33	NO	11682.91	NO	2582.44	NOx	35585.31
2.	NOx	7006.23	NOx	11475.22	SO2	1607.93	NO	30943.29
3.	CO	4203.73	CO	6376.54	NO2	1536.05	NO2	19580.17
4.	NO2	1520.62	O3	3744.20	O3	402.88	CO	2441.68
5.	SO2	1087.93	NO2	3480.86	CO	322.75	SO2	2093.71
6.	saat	76.85	saat	350.02	saat	106.20	saat	7.89
7.	hafta	70.38	hafta	141.27	NOx	39.4	hafta	0.18
8.	O3	47.64	SO2	89.61	hafta	1.73		

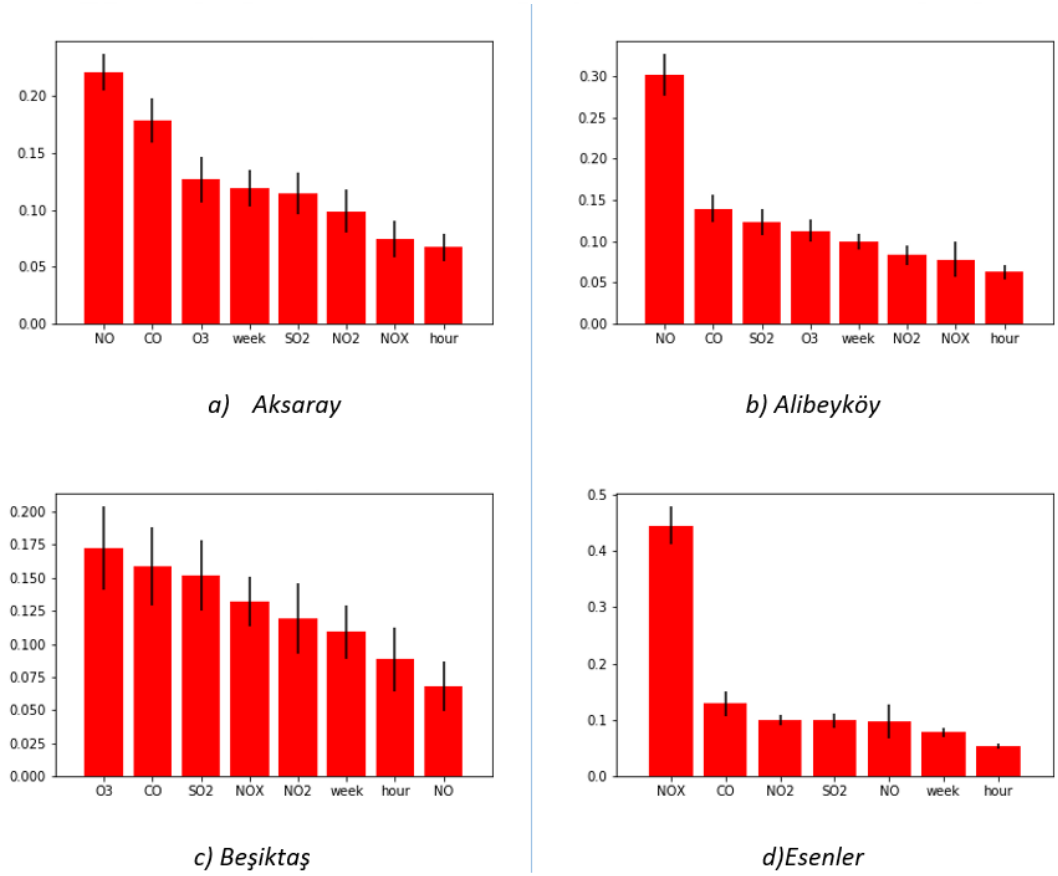
Daha iyi açıklamak için bir örnek verecek olursak; veri kümemizde 'x1, x2, x3' özniteliklerimizin olduğunu ve hedef değişkenimiz 'y' için belirli öznitelikleri seçmek istediğimizi varsayalım. Bu özniteliklerden x1 ve x2, y ile oldukça koreledir, ancak aynı zamanda kendi aralarında da yüksek korelasyon değerine sahiplerdir. x3'ün ise sadece y ile ilişkisi vardır. *f_regression*, en yüksek puanları x1 ve x2'ye atayacaktır.

Aksaray, Alibeyköy, Beşiktaş ve Esenler bölgelerine ait kirlilik verisindeki özniteliklerin bireysel önemlerini belirlemek için yapılan *f_regression* sonuçları özniteliklerin önem sırasına göre sıralanmış haliyle "Çizelge 4.4"te verilmiştir. "Çizelge 4.4"teki puanlara göre, PM_{10} tahmini için 'NO, NO_2, NO_X, CO' öznitelikleri

öne çıkarken, ‘hafta, saat’ öznitelikleri önem sıralamasında diğer özniteliklere kıyasla sonlarda yer almıştır.

4.1.2 Ağaç temelli topluluk modeli kullanarak öznitelik seçimi

Ağaç temelli topluluk modellerinden Rastgele Orman yöntemi, veri kümesindeki öznitelikler ormandaki ağaçlarla ilişkilendirilerek, önemli özniteliklerin seçiminde kullanılmaktadır. Rastgele Orman yöntemi öznitelikleri birbirinden bağımsızmış gibi ele almaz, özniteliklerin birbiri ile ilişkisini de hesaba katar. Puanlama yöntemi değil, sıralı öznitelik seçme yöntemidir.



Şekil 4.1 : Rastgele Orman Yöntemine Göre Özniteliklerin Önem Sıralaması.

f_regression’ın x1 ve x2’yi seçtiği bir önceki bölümdeki örnekte, Rastgele Orman yöntemi için durum değişebilir. Rastgele Orman yöntemi ilk turda x1’i seçer. Daha sonra hem x2 hem x3 özniteliklerini değerlendirir. x2 zaten seçilmiş bir öznitelik ile yüksek derecede korele olduğundan, içerdiği bilgilerin çoğu modele dahildir. Bu nedenle sıralı model sonraki aşamada, x2 yerine x3’ü seçebilir. x3 , y ile daha az ilişkili

bir öznitelik olsa da; x1'in açıklayamadığı kısımları x2'ye göre daha iyi açıkladığı için öne çıkan öznitelikler 'x1, x3' olmuştur.

Şekil 4.1'de Aksaray, Alibeyköy, Beşiktaş ve Esenler bölgelerine ait kirlilik veri kümelerindeki özniteliklerin Rastgele Orman yöntemine göre önem sıralaması verilmiştir. Alibeyköy için tek değişkenli öznitelik seçimine göre NO ilk sırada, NO_X ise ikinci sırada yer alırken, Rastgele Orman yöntemi kullanılarak seçim yapıldığında puanlama yöntemine göre daha önemli olan NO yine ilk sırada olmakla birlikte, NO_X son sıralardadır.

Öznitelikler arası korelasyon değerleri incelendiğinde; Aksaray, Alibeyköy ve Esenler için NO ve NO_X özniteliklerinin yaklaşık %97 korele olduğu görülmüştür. Rastgele Orman, öznitelikler arası korelasyonu da dikkate aldığından *f_regression*'dan farklı sonuçlar üretmiştir. Bu durumda, NO_X veya NO örnek sayısını arttırmak, tahmin modelinin eğitim süresini azaltmak ve genelleştirme performansını arttırmak için veri kümesinden çıkarılabilir. Veri kümesinden NO_X'i çıkarmak "Çizelge 4.5"de görüldüğü gibi dört bölge için toplamda sadece 13 örnek artışına sebep olurken, NO çıkarıldığında veri kümelerinde toplam 1998 örnek artışı olmaktadır. Bu sebeple kirlilik veri kümelerinden NO özniteliği çıkarılmıştır.

Çizelge 4.5 : NO ve NO_X Özniteliklerine Göre Kirlilik Veri Kümelerindeki Örnek Sayısı.

	sadece NO_X	sadece NO	NO ve NO_X
Aksaray	32128 (+140)	31988 (=)	31988
Alibeyköy	39993 (+1385)	38608 (=)	38608
Beşiktaş	38992 (+34)	38971 (+13)	38958
Esenler	49086 (+439)	48647 (=)	48647

Meteorolojik veri için Rastgele Orman ile aynı şekilde özniteliklerin önemleri incelendiğinde ise, öznitelikler arası %90'lara yaklaşan korelasyon değerleri görülmemiş, öznitelik sayısının az olduğu da göz önünde bulundurularak veri kümesi özgün haliyle korunmuştur.

İstanbul için yapılan önceki çalışmaya [42] ait veri kümesinde, PM_{10} [9 $\mu g/m^3$ - 206 $\mu g/m^3$] aralığındadır. Ortalaması 53 $\mu g/m^3$, standart sapması ise 32 $\mu g/m^3$ olarak verilmiştir. Zaman serileri analizi yapılan [49]'daki çalışmada, PM_{10} [0 kg/m^3 - 300 kg/m^3] aralığında değerlere sahiptir. Ortalaması 33.87 kg/m^3, standart sapması ise

21.18 kg/m^3'tür. 2004-2007 tarih aralığındaki veri kümesi üzerinde saatlik PM_{10} yoğunluğunun hesaplandığı [27]'de, saatlik PM_{10} değerleri 60 µg/m^3 ile 80 µg/m^3 arasında değişirken; [22]'de PM_{10} [0 µg/m^3 ile 1000 µg/m^3] aralığında değişmektedir. Önceki çalışmalardan hareketle PM_{10} değeri [0 µg/m^3 - 1000 µg/m^3] aralığında olmayan örnekler veri kümelerinden çıkarılmıştır.

Ulusal Hava Kalitesi İndeksi, EPA Hava Kalitesi İndeksinin ulusal mevzuatımız ve sınır değerlerimize uyarlanmasıyla oluşturulmaktadır [4]. Hava kirliliği tahmininde kullandığımız hedef değişkenimiz PM_{10} için hava kalitesi indeks aralıkları ile Aksaray, Alibeyköy, Beşiktaş, Esenler, Kartal, Sarıyer, Silivri, Üsküdar ve Yenibosna ölçüm istasyonlarında bu aralıklara denk gelen örnek sayıları "Çizelge 4.6"daki gibidir. Örneklere ait PM_{10} değerlerinin tüm ölçüm istasyonları için [0 µg/m^3 - 260 µg/m^3] aralığında yoğunlaştığı görülmektedir.

"Çizelge 4.6"dan görüldüğü üzere PM_{10} için Ulusal Hava Kalitesi İndeksi kesme noktalarına göre oluşturulan sınıflandırma problemi altı sınıftan oluşmaktadır: "İyi", "Orta", "Hassas", "Sağlıksız", "Kötü" ve "Tehlikeli". Bu sınıflar içinde hava durumunun insan sağlığı açısından tehlikeli olduğu, PM_{10} yoğunluğunun 260 µg/m^3 'ten fazla olduğu örnek sayısı diğer sınıflardaki örnek miktarlarıyla kıyaslandığında oldukça azdır. Örneğin Üsküdar istasyonundaki 38710 örnek içerisinde 'kötü' sınıfına ait örnek sayısı sadece 1'dir. Bu örneği test kümesine dahil ettiğimizi varsayalım. Bu durumda bu sınıfı öğrenmek için eğitim kümesine eklenecek örnek elimizde yoktur.

Çizelge 4.6 : Ulusal Hava Kalitesi İndeksi Kesme Noktalarına Göre Sınıf Dağılımları.

PM_{10} [µg/m^3]	***0-50*** **İ**[a]	***51-100*** **O**[a]	***101-260*** **H**[a]	***261-400*** **S**[b]	***401-520*** **K**[b]	***>521*** **T**[b]	***Örnek Sayısı***
Aksaray	14713	13425	3761	163	35	31	32128
Alibeyköy	26095	10323	3350	170	29	26	39993
Beşiktaş	26521	10818	1557	35	21	40	38992
Esenler	28040	16183	4529	256	30	30	49068
Kartal	19467	12812	6358	442	105	44	39228
Sarıyer	40132	9786	2476	30	7	10	52441
Silivri	30815	5683	1172	26	3	19	37718
Üsküdar	29828	7393	1464	24	1	0	38710
Yenibosna	18609	9941	2716	261	82	32	31641

[a]İ: İyi, O: Orta, H: Hassas - "Pozitif Sınıf"

[b]S: Sağlıksız, K: Kötü, T: Tehlikeli - "Negatif Sınıf"

Farklı yoğunluklardaki PM_{10} değerlerini doğru temsil edebilmek ve tahminleyebilmek amacıyla PM_{10} için sınıflandırma problemi altı sınıftan iki sınıfa dönüştürülmüştür: PM_{10} yoğunluğunun 260 $\mu g/m^3$'e eşit veya daha az olduğu örnekleri içeren pozitif sınıf ve PM_{10}'un yoğunluğunun 260 $\mu g/m^3$'ten fazla olduğu örnekleri içeren negatif sınıf. Sınıflar açısından bakıldığındaysa, PM_{10} ikili sınıflandırma problemi için "İyi", "Orta", "Hassas" sınıflara ait örnekler pozitif sınıfı temsil ederken, "Sağlıksız", "Kötü" ve "Tehlikeli" sınıf örnekleri ise negatif sınıfı temsil etmiştir.

"Çizelge 4.3"de göze çarpan bir başka problem Kartal, Sarıyer ve Yenibosna istasyonları gibi bazı istasyonlarda ölçümü yapılabilen kirletici türlerinin sınırlı sayıda olmasıdır. Bu gibi durumlarda tahminleme modeli sınırlı sayıdaki öznitelik sebebiyle yüksek başarım elde edemeyebilir. Bunun önüne geçmek için o bölgedeki meteorolojik veri ve kirlilik verisi ayrı ayrı kullanılarak iki farklı model oluşturulabilir ya da bu iki veri kümesi örnek sayısının azalması önemsiz boyuttaysa birleştirilebilir.

Veriler model eğitimi için uygun hale getirildikten sonra, meteorolojik verideki örnek sayısının, kirlilik verisindeki örnek sayısından daha az sayıda olduğu görülmüştür. Çoğunlukla tercih edilen veri kümelerinin kesişimlerini alma yaklaşımı yerine, eğitim kümelerindeki örneklerin sayısını maksimize etmek amacıyla bölge başına iki veri grubu ayrı ayrı kullanılmıştır. Bu yaklaşımın örnek sayısından optimum şekilde yararlanma dışındaki bir diğer avantajı, veri gruplarından birinin belirli bir bölge için mevcut olmaması durumunda halen hava kirliliği tahmini yapmamıza olanak sağlamasıdır.

Sınıflandırma probleminde, veri kümesindeki sınıflara ait örneklerin farklı oranlarla, dengesiz dağılması makine öğrenmesi ile tahminleme çalışmalarında oldukça sık karşılaşılan bir sorundur [18, 41, 50, 51]. 260 $\mu g/m^3$ yoğunluğunu eşik değeri olarak kabul edip, problemi ikili sınıflandırma problemine dönüştürdükten sonra bile "Çizelge 4.6"dan da açıkça anlaşılacağı üzere veri kümesi halen dengesiz veri dağılmı probleminin etkisindedir. Veri kümesinin yaklaşık %92'sini baskın sınıfa ait pozitif örnekler , %8'ini ise azınlık sınıfa ait negatif örnekler oluşturmaktadır.

Hava kirliliği tahmini yapan mevcut çalışmaların çok az bir bölümü, problemi dengesiz dağılma sorunuyla birlikte ele almaktadır. Bu alandaki çalışmaların bir çoğu negatif örnekleri veri kümesinden çıkararak problemin dengesiz dağılımını göz ardı eder ve

baskın sınıfları kullanarak baskın sınıfları tahminler. Ancak asıl önemli olan kirlilik seviyesinin tehlikeli seviyelerde olduğu zamanları önceden tahmin etmektir. Bir çeşit uyarı sisteminin temelini oluşturan hava kirliliği tahmin modeli ile kirliliğin olumsuz etkilerini azaltacak önlemler alınarak, insan ve çevre sağlığına katkı sağlanabilir.

İki Katmanlı Hava Kirliliği Tahminleme Modeli, hava kalitesi tahminleme çalışmalarında karşılaşılan dengesiz dağılan veri problemine çözüm sunar. Modelimizin ilk aşamasında ikili sınıflandırıcı ile örneğin sınıfına karar verilmekte, ikinci aşamadaki regresyon modeli ile de sınıfı belirlenen örneğin PM_{10} yoğunluk bilgisi tahminlenmektedir. Deneysel çalışmalar öncelikle Aksaray, Alibeyköy, Beşiktaş ve Esenler bölgeleri üzerinde yapılmış, öne çıkan modeller ile çalışma dokuz bölge için genişletilmiştir.

4.2 Değerlendirme Ölçütleri

Hava kirliliği tahminleme modelinin ikili sınıflandırma aşamasında modeller Doğruluk ve AUROC metrikleri aracılığıyla değerlendirilmiştir. Regresyon aşamasında bu metriklerin yerini Ortalama Mutlak Hata (OMH) ve Ortalama Kare Hata (OKH) metrikleri almıştır.

Sınıflandırma problemlerinde en çok kullanılan değerlendirme ölçütü Doğruluk, dengesiz veri problemlerinde aldatıcı sonuçlar doğurabilir. Örneğin üzerinde çalıştığımız veri kümesinde basit bir doğrusal ikili sınıflandırıcı ile PM_{10}'un sınıfını tahminledğimizde, %90 civarında doğruluk sonuçları elde etmek mümkündür. Dengesiz veri dağılımına sahip karmaşık problemler için doğrusal ve basit sınıflandırıcılarla elde edilen yüksek başarım değerleri aslında sadece baskın sınıfı doğru tahminlemekten kaynaklanır ve farklı metrikler kullanmadan sadece Doğruluk ile sınıflandırıcıları değerlendirmek dengesiz veri dağılımından muzdarip problemlerde aldatıcıdır. Bu nedenle Doğruluk'a ek olarak AUROC aracılığıyla da sınıflandırıcıların performansları karşılaştırılmıştır.

"Çizelge 4.7" ikili sınıflandırmadaki karışıklık matrisini ifade etmektedir. Karışıklık matrisinde kullanılan terimlerin anlamları:

- Doğru Pozitif (DP): '1' sınıfında, '1' olarak tahmin ettiğimiz örnek sayısı.

Çizelge 4.7 : İkili Sınıflandırma Karışıklık Matrisi.

		Öngörülen Sınıf	
		Sınıf 1[a]	***Sınıf 0***[b]
Gerçek Sınıf	***Sınıf 1***[a]	DP	YN
	Sınıf 0 [b]	YP	DN

[a] Sınıf 1: Pozitif sınıf (baskın sınıf).
[b] Sınıf 0: Negatif sınıf (azınlık sınıf).

- Doğru Negatif (DN): '0' sınıfında, '0' olarak tahmin ettiğimiz örnek sayısı.
- Yanlış Pozitif (YP): '0' sınıfında, '1' olarak tahmin ettiğimiz örnek sayısı.
- Yanlış Negatif (YN): '1' sınıfında, '0' olarak tahmin ettiğimiz örnek sayısı.

Karışıklık matrisindeki terimler yardımıyla, "Doğruluk" ve "AUROC" performans metrikleri formülleri ile birlikte Bölüm 4.2.1 ve Bölüm 4.2.2'de açıklanmıştır.

4.2.1 Doğruluk

Doğruluk (4.1), doğru tahmin edilen örnek sayısının test kümesindeki toplam örnek sayısına oranıdır.

$$\text{Doğruluk} = \frac{DP + DN}{DP + DN + YP + YN} \tag{4.1}$$

4.2.2 AUROC

Reciever Operator Characteristics (ROC) Curve eğrisi, Doğru Pozitif Oranı (DPO)'nun (4.2) Yanlış Pozitif Oranı (YPO)'ya (4.3) oranını verir. Area Under ROC (AUROC) ise ROC eğrisi altındaki alanı hesaplar.

$$DPO = \frac{DP}{DP + YN} \tag{4.2}$$

$$YPO = \frac{YP}{YP + DN} \tag{4.3}$$

Dengesiz dağılan veride sınıflandırma probleminde AUROC değeri, hem baskın hem de azınlık sınıf için yapılan doğru tahminlerin sayısı arttıkça artmaktadır. Böylelikle

hem pozitif hem negatif sınıfın doğru tahminlenip tahminlenemediğinin kontrolü yapılır.

Hava kirliliği tahminleme modelinin ikinci aşamasındaki aday regresyon modellerinin başarımının değerlendirilmesinde ise regresyon problemlerinde öne çıkan OMH ve OKH kullanılmıştır.

4.2.3 OMH

OMH iki sürekli değişken arasındaki farkın ölçüsüdür. OMH (4.4), her gerçek değer ile veriye en iyi uyan çizgi arasındaki ortalama dikey mesafedir. OMH değeri 0'dan sonsuza kadar değişebilir. Negatif yönelimli puanlar yani daha düşük değerlere sahip tahminleyiciler daha iyi performans gösterir.

$$OMH = \frac{1}{n}\sum_{1}^{n}|e_t| \tag{4.4}$$

4.2.4 OKH

Basitçe, OKH bir regresyon eğrisinin bir dizi noktaya ne kadar yakın olduğunu söyler. OKH (4.5), bir makine öğrenmesi modelinin, tahminleyicinin performansını ölçer, her zaman pozitif değerlidir ve OKH değeri sıfıra yakın olan tahminleyicilerin daha iyi bir performans gösterdiği söylenebilir.

$$OKH = \frac{1}{n}\sum_{1}^{n}e_t^2 \tag{4.5}$$

4.3 Hava Kirliliği Tahminleme Modeli Sonuçları

Çalışmanın ilk aşamasında üzerinde çalışılan dokuz bölge arasından dört tanesi pilot bölge olarak seçilmiş ve hava kirliliği tahminleme modeli oluşturmak için öncelikle bu bölgelere ait veri kümeleri üzerinde modeller eğitilmiştir. Bu bölgeler Aksaray, Alibeyköy, Beşiktaş ve Esenler'dir. Pilot bölgeler seçilirken; yoğun nüfusa sahip olmaları ve yoğun trafiğe maruz kalmaları göz önünde bulundurulmuştur. Aksaray Ölçüm İstasyonu; Yenikapı Feribot İskelesi, İstanbul Üniversitesi Tıp Fakültesi ile İstanbul Eğitim ve Araştırma Hastanesi'nin kesişiminde yer alır. Yenikapı İstasyonu ile marmaray ve metro için aktarma noktasıdır. Portekiz'de başlayıp, İran Gürbulak

Sınır Kapısı'nda sona eren Avrupa E-Yolu (TEM-E80) Alibeyköy'den geçer. Ulaşımın kalbi niteliğindeki Büyük İstanbul Otogarı Esenler'dedir. İstanbul'un en önemli stadyumlarından Vodafone Park Stadyumu Beşiktaş'ta yer alır. Bu bölgeler için öne çıkan tahminleme modelleri ile, çalışmanın son hali dokuz bölge için genişletilmiştir.

İki ayrı ölçüm istasyonunundan (Meteoroloji ve Kirlilik) elde edilen veriler kullanılarak oluşturulan tahminleme modelleri ile elde edilen ilk sonuçlar [52]'de yayımlanmış, Bölüm 4.2.1'de sunulmuştur. Yeni yöntemler eklenmiş ve dokuz bölgeye genişletilmiş haliyle ikili sınıflandırma sonuçları Bölüm 4.2.2'de yer almaktadır. Önerilen İki Katmanlı Hava Kirliliği Tahminleme Modeli ile geleneksel regresyon yöntemlerinin karşılaştırılması Bölüm 4.3'tedir.

- M : Meteorolojik Veri
- K : Kirlilik Verisi

Her bir bölgeye ait modellerin performansını gözlemlemek için eğitim kümesi ve test kümesi aşağıdaki şekilde oluşturulmuştur: her ölçüm istasyonu için, meteorolojik verinin ve kirlilik verisinin ortak kayıt içerdiği zamanlara ait verinin %30'u test verisi olarak ayrılmıştır. Meteorolojik eğitim kümesi ve kirlilik eğitim kümesi ise, test kümelerinin başlangıç veri kümelerinden çıkarılmasından sonra kalan veri kümeleridir. Modeller, hiper parametrelerin farklı kombinasyonları denenerek ve 10 kat çapraz doğrulama kullanılarak optimize edilmiştir.

Ekstra Ağaç, Rastgele Orman ve Gradyan Arttırma algoritmalarının eğitimi sırasında kullanılan parametre değerleri "Çizelge 4.8"de gösterilmiştir.

Çizelge 4.8 : Parametreler & Parametre Değerleri.

Model	***Parametre***	***Parametre Değerleri***
GA	min_samples_leaf	[40, 50, 80]
	min_samples_split	[200, 300, 400]
	max_depth	[5, 8]
EA & RO	max_features	['auto', 'sqrt']
	min_samples_leaf	[1, 2, 4]
	min_samples_split	[2, 5, 10]
	max_depth	[10 - 110]

DVM için optimize edilen hiperparametreler ise şunlardır:

- **C**: poli-DVM için "1" seçilmiş, rbf-DVM içinse "0.1, 1, 10, 100" sayısal değerleri kullanılmıştır.
- **degree**: poli-DVM için "2, 3, 4" değerleri en iyi hiper parametreleri bulmak için kullanılmıştır.

4.3.1 İkili sınıflandırma sonuçları - I

İkili sınıflandırmada örnekleme yöntemlerinin performansı öncelikle Aksaray, Alibeyköy, Beşiktaş ve Esenler bölgelerine ait veri kümeleri üzerinde kıyaslanmıştır. Dengesiz dağılan veri üzerinde ikili sınıflandırma için düşündüğümüz ilk çözüm, tüm veri kümesinde topluluk modellerini kullanmaktı. Her ölçüm bölgesi için Meteoroloji (M) ve Kirlilik (K) verileri üzerinde ayrı ayrı eğitilen ve tüm veriyi kulanan topluluk modellerinin ürettiği Doğruluk ve AUROC sonuçları "Çizelge 4.9" daki gibidir. Önde gelen model ROS; ***max_depth*** 10 ve 40 aralığında seçildiğinde, ***min_samples_split*** 2, ***min_samples_leaf*** 4 ve ***max_features*** "auto" olduğunda, hem meteorolojik hem de kirlilik verileri için en doğru sonuçları üretmiştir.

Diğer yandan, EAS'daki en iyi sonuçlar, meteorolojik veri için ***max_depth*** 30, ***min_samples_split*** 5, ***min_samples_leaf*** 1 ve ***max_features*** "sqrt" seçildiğinde, kirlilik verisi için ise ***max_depth*** 10, ***min_samples_split*** 2, ***min_samples_leaf*** 4 ve ***max_features*** "sqrt" seçildiğinde elde edilmiştir.

Son olarak, GAS için ***min_samples_leaf*** 50 veya 80, ***min_samples_split*** 400 ve ***max_depth*** 8 seçilerek eğitilen model en iyi sonuçları üretmiştir.

Özgün veri kümesine alternatif olarak, aşağı örnekleme yoluyla dört farklı veri kümesi oluşturulmuştur. "Çizelge 4.10", rastgele örnekleme uygulandıktan sonra elde edilen dengeli veri kümesi üzerinde topluluk modellerinin performanslarını göstermektedir.

Öte yandan, "Çizelge 4.11", "Çizelge 4.12" ve "Çizelge 4.13", Near-Miss aşağı örnekleme yaklaşımının üç farklı versiyonuna dair sonuçları sunar.

Sonuçlara genel olarak bakıldığında, aşağı örnekleme yaklaşımlarının, veri kaynağından (meteorolojik veya kirlilik) bağımsız olarak, tüm veri kümesini kullanan tahmin modellerinden daha iyi performans göstermediği ortaya çıkmıştır.

Çizelge 4.9 : Tüm Veri Kümesi Kullanılarak Elde Edilen Sonuçlar.

		DRS		*ROS*		*EAS*		*GAS*	
		Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
M	***Aksaray***	0.9945	0.5755	0.9947	0.9590	0.9951	0.9796	0.9982	0.9695
K		0.9901	0.6060	0.9902	0.9030	0.9902	0.8647	0.9902	0.8612
M	***Alibeyköy***	0.9944	0.7684	0.9962	0.9919	0.9945	0.9915	0.9955	0.9893
K		0.9910	0.8869	0.9911	0.9399	0.9911	0.8998	0.9912	0.9257
M	***Beşiktaş***	0.9971	0.6809	0.9996	0.9996	0.9975	0.9994	0.9997	0.9981
K		0.9973	0.6145	0.9974	0.8387	0.9974	0.7642	0.9974	0.8360
M	***Esenler***	0.9931	0.7442	0.9943	0.9689	0.9931	0.9640	0.9959	0.9674
K		0.9877	0.7256	0.9906	0.8659	0.9903	0.8424	0.9899	0.8444

Çizelge 4.10 : Rastgele Aşağı Örnekleme ile Elde Edilen Sonuçlar.

		DRS		*ROS*		*EAS*		*GAS*	
		Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
M	***Aksaray***	0.5949	0.6058	0.7509	0.9443	0.6833	0.8406	0.6094	0.6775
K		0.8697	0.7081	0.8649	0.8561	0.8984	0.8473	0.8303	0.7190
M	***Alibeyköy***	0.6183	0.7644	0.7609	0.9373	0.7356	0.9511	0.7079	0.8577
K		0.8699	0.8719	0.9011	0.9240	0.9024	0.9225	0.9002	0.8807
M	***Beşiktaş***	0.6690	0.7071	0.7864	0.9771	0.8028	0.9971	0.5029	0.5600
K		0.6353	0.6272	0.6764	0.6998	0.6820	0.7800	0.5026	0.5553
M	***Esenler***	0.5616	0.7327	0.7599	0.9521	0.7719	0.9543	0.7152	0.8913
K		0.9149	0.8021	0.9284	0.8485	0.9463	0.8246	0.9039	0.7716

Çizelge 4.11 : NearMiss-1 Örnekleme ile Elde Edilen Sonuçlar.

		DRS		*ROS*		*EAS*		*GAS*	
		Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
M	***Aksaray***	0.2707	0.5291	0.1211	0.6964	0.1601	0.7484	0.1865	0.4839
K		0.6426	0.7198	0.2507	0.6856	0.4313	0.7274	0.4127	0.6885
M	***Alibeyköy***	0.2520	0.4709	0.1022	0.7057	0.0825	0.6690	0.1482	0.5550
K		0.6445	0.8858	0.2335	0.8616	0.4932	0.8931	0.6096	0.8542
M	***Beşiktaş***	0.2922	0.4255	0.1830	0.8790	0.1202	0.7669	0.9971	0.5000
K		0.2691	0.5016	0.1052	0.6337	0.0336	0.6034	0.9974	0.5000
M	***Esenler***	0.2904	0.4826	0.2092	0.8525	0.2224	0.8158	0.3485	0.6995
K		0.4812	0.6963	0.0663	0.7228	0.1479	0.7126	0.2611	0.6688

Çizelge 4.12 : NearMiss-2 Örnekleme ile Elde Edilen Sonuçlar.

		DRS		*ROS*		*EAS*		*GAS*	
		Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
M	***Aksaray***	0.4221	0.5704	0.0233	0.7255	0.0710	0.8283	0.0333	0.5181
K		0.1318	0.3211	0.0157	0.5065	0.0270	0.5328	0.0184	0.4418
M	***Alibeyköy***	0.5885	0.6509	0.0249	0.8046	0.0741	0.8291	0.0287	0.6580
K		0.0855	0.2479	0.0107	0.4663	0.0117	0.2123	0.0105	0.2468
M	***Beşiktaş***	0.6149	0.7321	0.1546	0.9201	0.1725	0.9459	0.9971	0.5000
K		0.3032	0.5769	0.0788	0.6198	0.1643	0.6594	0.9974	0.5000
M	***Esenler***	0.3687	0.5825	0.0456	0.7736	0.1341	0.8661	0.1605	0.6036
K		0.0260	0.2729	0.0164	0.6543	0.0178	0.4765	0.0163	0.6491

Çizelge 4.13 : NearMiss-3 Örnekleme ile Elde Edilen Sonuçlar.

		DRS		*ROS*		*EAS*		*GAS*	
		Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
M	***Aksaray***	0.5130	0.4807	0.6121	0.8897	0.6539	0.8966	0.6720	0.5485
K		0.0430	0.3004	0.1087	0.6110	0.0826	0.5311	0.1314	0.4847
M	***Alibeyköy***	0.5654	0.6553	0.5946	0.9500	0.6457	0.8309	0.5689	0.6888
K		0.0436	0.1350	0.0895	0.4320	0.0679	0.2516	0.0370	0.2232
M	***Beşiktaş***	0.5389	0.6840	0.6839	0.9524	0.6760	0.9325	0.9971	0.5000
K		0.4284	0.5255	0.5722	0.6608	0.5729	0.5994	0.9974	0.5000
M	***Esenler***	0.3798	0.5484	0.7305	0.9504	0.7680	0.8996	0.6715	0.7600
K		0.4284	0.5255	0.5722	0.6608	0.5729	0.5994	0.9974	0.5000

Dört örnekleme yaklaşımı kıyaslandığında, en yüksek AUROC değerlerine sahip modelin rastgele örnekleme ile oluşturulan model olduğu görülmüştür. (M_{Aks}=0.9443, P_{Aks}=0.8561, M_{Alb}=0.9511, P_{Alb}=0.9240, $M_{\text{Bşk}}$=0.9971, $P_{\text{Bşk}}$=0.7800, M_{Esn}=0.9543, P_{Esn}=0.8485). Performans açısından rastgele örnekleme yöntemini sırasıyla NearMiss-3, NearMiss-2 ve NearMiss-1 yöntemleri takip etmektedir.

Veri kaynaklarına göre modellerin performanslarına bakıldığında, meteorolojik verilerin kirlilik verilerine göre her zaman daha doğru sonuçlar ürettiği açıktır. Örneklemeden ziyade, dengesiz dağılımı göz önüne alarak topluluk modellerinin parametrelerini ayarlamanın daha doğru bir yaklaşım olduğu ortaya çıkmıştır. Topluluk modellerinde, bagging yöntemleri boosting'e kıyasla daha iyi performans göstermiştir.

4.3.2 İkili sınıflandırma sonuçları - II

İkili sınıflandırmada Aksaray, Alibeyköy, Beşiktaş ve Esenler ölçüm istasyonları için modellerin performansları değerlendirildikten sonra Kartal, Sarıyer, Silivri, Üsküdar ve Yenibosna istasyonları da eklenerek çalışma genişletilmiştir. İkili sınıflandırmanın ikinci aşamasında artık temel yaklaşımımız örnekleme yöntemlerinden vazgeçerek, tüm veri kümesini kullanma ve algoritmaların parametrelerini dengesiz veri dağılımına uyum sağlayacak şekilde ayarlamaya dayanmaktadır.

Çizelge 4.14 : Meteorolojik Veri Üzerinde İkili Sınıflandırma Sonuçları.

	DRS		***ROS***		***EAS***		***GAS***	
	Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
Aksaray	0.9945	0.5557	0.9945	0.9332	0.9945	0.9032	0.9943	0.8903
Alibeyköy	0.9939	0.7456	0.9939	0.8937	0.9939	0.8714	0.9939	0.8753
Beşiktaş	0.9971	0.6057	0.9974	0.9213	0.9971	0.8914	0.9978	0.9349
Esenler	0.9937	0.6956	0.9946	0.9242	0.9937	0.9010	0.9942	0.8874
Kartal	0.9887	0.7914	0.9887	0.9282	0.9887	0.9504	0.9895	0.9340
Sarıyer	0.9992	0.7775	0.9993	0.9445	0.9992	0.9399	0.9960	0.6771
Silivri	0.9993	0.7469	0.9994	0.9952	0.9993	0.9935	0.9994	0.9865
Üsküdar	0.9997	0.9413	0.9999	0.9992	0.9997	0.9989	0.9997	0.9997
Yenibosna	0.9890	0.7742	0.9893	0.9092	0.9890	0.9057	0.9886	0.9063

Üzerinde çalıştığımız dokuz bölge için topluluk modelleriyle elde edilen sonuçlar, "Tablo 4.14" de meteorolojik veri üzerinde, "Tablo 4.15" de ise kirlilik verisi üzerinde kıyaslanmıştır.

Meteorolojik veri üzerinde elde edilen sonuçlara göre öne çıkan model, parametreleri ***max_depth*** 40, ***min_samples_split*** 2, ***min_samples_leaf*** 4 ve ***max_features*** "auto"

Çizelge 4.15 : Kirlilik Verisi Üzerinde İkili Sınıflandırma Sonuçları.

	DRS		*ROS*		*EAS*		*GAS*	
	Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
Aksaray	0.9946	0.6240	0.9948	0.9694	0.9945	0.9220	0.9947	0.9208
Alibeyköy	0.9939	0.8372	0.9939	0.8942	0.9939	0.8913	0.9942	0.8793
Beşiktaş	0.9970	0.6839	0.9973	0.9413	0.9971	0.9615	0.9971	0.8858
Esenler	0.9936	0.7633	0.9942	0.9361	0.9940	0.9358	0.9941	0.8760
Kartal	0.9885	0.5856	0.9887	0.8817	0.9887	0.8832	0.9887	0.8701
Sarıyer	0.9992	0.6870	0.9992	0.8319	0.9992	0.8306	0.9990	0.8194
Silivri	0.9993	0.8313	0.9997	0.9989	0.9994	0.9995	0.9993	0.9911
Üsküdar	0.9997	0.4291	0.9997	0.9991	0.9997	0.9991	0.9997	0.6197
Yenibosna	0.9900	0.8770	0.9920	0.9523	0.9917	0.9507	0.9914	0.9524

olarak seçilerek eğitilen ROS olmuştur. Kirlilik verisi üzerinde de en doğru sonuçlar yine aynı modelle elde edilmiştir.

Topluluk modellerine ek olarak, makine öğrenmesine dayalı çalışmalarda başarısı ispatlanmış DVM'nin iki versiyonu poli-DVM ve rbf-DVM de dengesiz dağılan eğitim kümesi üzerinde eğitilmiş ve test edilmiştir. Çekirdek tabanlı modellerle birlikte Doğrusal Regresyon Sınıflandırıcısından (DRS) dokuz bölge için elde edilen sonuçlar, "Tablo 4.16" da meteorolojik veri üzerinde, "Tablo 4.17" de ise kirlilik verisi üzerinde kıyaslanmıştır.

Çizelge 4.16 : Meteorolojik Veri Üzerinde İkili Sınıflandırma Sonuçları - 2.

	DRS		*poli-DVMS*		*rbf-DVMS*	
	Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
Aksaray	0.9945	0.5557	0.9945	0.6280	0.9945	0.6347
Alibeyköy	0.9939	0.7456	0.9939	0.6910	0.9939	0.6489
Beşiktaş	0.9971	0.6057	0.9971	0.6349	0.9971	0.7477
Esenler	0.9937	0.6956	0.9937	0.6292	0.9937	0.8330
Kartal	0.9887	0.7914	0.9887	0.6510	0.9887	0.7917
Sarıyer	0.9992	0.7775	0.9992	0.7981	0.9992	0.9493
Silivri	0.9993	0.7469	0.9993	0.5782	0.9993	0.7810
Üsküdar	0.9997	0.9413	0.9997	0.8469	0.9997	0.4913
Yenibosna	0.9890	0.7742	0.9890	0.4695	0.9890	0.8312

Topluluk modelleri EAS, GAS, ROS; çekirdek tabanlı poli-DVMS, rbf-DVMS ve DRS performansları test kümeleri üzerinde karşılaştırıldığında ROS, dokuz bölge için genişletilmiş testlerde de öndedir. DVMS modellerinden beklenen performanlar elde edilememiş, bu modellerin performansının ne kadar kötü olduğunu çıkarsamak amacıyla sonuçlar doğrusal bir sınıflandırıcı olan DRS ile de karşılaştırılmıştır. Çekirdek tabanlı modeller topluluk modellerindense DRS'e yakın sonuçlar üretmekte ve tahminlemede beklenen performansı sergileyememektedir.

Çizelge 4.17 : Kirlilik Verisi Üzerinde İkili Sınıflandırma Sonuçları - 2.

	DRS		***poli-DVMS***		***rbf-DVMS***	
	Doğruluk	***AUROC***	***Doğruluk***	***AUROC***	***Doğruluk***	***AUROC***
Aksaray	0.9946	0.6240	0.9945	0.7144	0.9945	0.7162
Alibeyköy	0.9939	0.8372	0.9939	0.7397	0.9939	0.2940
Beşiktaş	0.9970	0.6839	0.9971	0.4178	0.9971	0.3499
Esenler	0.9936	0.7633	0.9937	0.6509	0.9937	0.7327
Kartal	0.9885	0.5856	0.9887	0.5970	0.9887	0.2870
Sarıyer	0.9992	0.6870	0.9992	0.6851	0.9992	0.3405
Silivri	0.9993	0.8313	0.9993	0.7378	0.9993	0.4993
Üsküdar	0.9997	0.4291	0.9997	0.5430	0.9997	0.6461
Yenibosna	0.9900	0.8770	0.9890	0.7140	0.9890	0.8188

Bu bulgular doğrultusunda İki Katmanlı Hava Kirliliği Tahminleme Modeli'nin ilk katmanındaki ikili sınıflandırma için ROS kullanılması kesinleştirilmiştir.

4.3.3 İki Katmanlı Hava Kirliliği Tahminleme Modeli sonuçları

Önerilen yeni İki Katmanlı Hava Kirliliği Tahminleme Modeli ile saf regresyon modellerinin performansları OMH metriğine göre "Tablo 4.18"de meteorolojik veri için, "Tablo 4.19"da kirlilik verisi için karşılaştırılmıştır. OKH metriğine göre alınan sonuçlar ise aynı sırayla "Tablo 4.20" ve "Tablo 4.21"de sunulmuştur.

Burada her saf regresyon yöntemi; ilk katmanında ROS, ikinci katmanında ise yine aynı yöntemden üretilmiş regresyon modelleri kullanılan İki Katmanlı Hava Kirliliği Tahminleme Modeli ile kıyaslanmaktadır. Örneğin ROR ile karşılaştırmada iki katmanlı modelin birinci katmanında ROS, ikinci katmanında iki ayrı ROR kullanılırken; EAR için birinci katmanda ROS, ikinci katmanda iki ayrı EAR kullanılır. Regresyon modellerini saf halleriyle kullanmaktansa, öncesinde ikili sınıflandırıcı ile birlikte kullanmak, oldukça dengesiz dağılan veri üzerinde OKH ve OMH hata değerlerini önemli ölçüde azaltmıştır.

Sonuçlar göstermiştir ki DVM modelleri, topluluk modellerine kıyasla dengesiz dağılan veri üzerinde kötü performans sergilemektedir. Önerilen model ile de DVM için performans artışı kısmen sağlanabilirken, iyileşme yaşanmayan durumlar da görülebilir. Topluluk modellerinde ise birkaç ender durum dışında, iki katmanlı model öne çıkmaktadır. İki katmanlı modellerden ilk katmanda ROS; ikinci katmanda meteorolojik veri için ROR, kirlilik içinse EAR kullanan modeller öncüdür.

Çizelge 4.18 : Meteorolojik Veri Üzerinde OMH Sonuçları.

		EAR	***GAR***	***ROR***	***poli-DVM***	***rbf-DVM***
Aksaray	***Saf Regresyon***	23.48	23.52	23.68	26.51	26.10
	İki Katmanlı	21.99	22.35	22.85	25.45	24.05
Alibeyköy	***Saf Regresyon***	24.96	25.17	25.21	26.57	26.42
	İki Katmanlı	23.02	23.27	23.71	24.09	24.95
Beşiktaş	***Saf Regresyon***	13.68	15.70	16.25	19.10	18.55
	İki Katmanlı	12.53	13.45	14.18	18.30	17.74
Esenler	***Saf Regresyon***	16.62	20.45	20.40	23.44	23.17
	İki Katmanlı	13.73	18.43	19.35	23.10	22.83
Kartal	***Saf Regresyon***	25.06	26.69	24.77	35.10	33.67
	İki Katmanlı	24.47	25.75	22.83	34.19	33.46
Sarıyer	***Saf Regresyon***	16.69	16.43	17.41	21.61	21.44
	İki Katmanlı	15.92	15.42	16.76	19.61	19.27
Silivri	***Saf Regresyon***	13.14	13.47	12.49	10.77	10.32
	İki Katmanlı	11.80	12.26	11.39	10.07	10.25
Üsküdar	***Saf Regresyon***	12.06	12.34	11.73	15.08	18.29
	İki Katmanlı	11.55	11.41	10.79	15.02	18.23
Yenibosna	***Saf Regresyon***	21.60	22.85	21.33	26.33	26.51
	İki Katmanlı	20.09	21.08	20.26	25.23	25.02

Çizelge 4.19 : Kirlilik Verisi Üzerinde OMH Sonuçları.

		EAR	***GAR***	***ROR***	***poli-DVM***	***rbf-DVM***
Aksaray	***Saf Regresyon***	21.56	19.95	22.15	28.72	33.54
	İki Katmanlı	19.15	18.12	20.63	28.29	32.10
Alibeyköy	***Saf Regresyon***	24.47	23.19	24.96	28.55	26.10
	İki Katmanlı	23.26	22.17	23.42	27.48	25.22
Beşiktaş	***Saf Regresyon***	15.29	16.50	17.08	20.86	21.63
	İki Katmanlı	13.28	14.87	16.22	18.83	21.52
Esenler	***Saf Regresyon***	20.87	21.24	22.16	26.92	25.90
	İki Katmanlı	19.23	19.79	20.86	25.30	24.33
Kartal	***Saf Regresyon***	36.64	33.41	33.40	35.10	34.67
	İki Katmanlı	34.97	33.08	33.08	34.74	34.57
Sarıyer	***Saf Regresyon***	20.91	19.92	19.92	21.61	21.44
	İki Katmanlı	19.48	19.25	19.64	21.20	20.64
Silivri	***Saf Regresyon***	8.76	8.79	8.30	10.77	10.32
	İki Katmanlı	8.51	8.43	8.02	10.24	9.75
Üsküdar	***Saf Regresyon***	10.96	10.24	10.16	15.08	18.29
	İki Katmanlı	10.74	10.20	9.96	14.86	17.77
Yenibosna	***Saf Regresyon***	21.01	20.58	20.25	26.63	26.51
	İki Katmanlı	20.15	20.06	20.15	25.77	25.66

Çizelge 4.20 : Meteorolojik Veri Üzerinde OKH Sonuçları.

		EAR	*GAR*	*ROR*	*poli-DVM*	*rbf-DVM*
Aksaray	***Saf Regresyon***	1812.22	1714.01	1674.88	2095.98	2080.22
	İki Katmanlı	1725.25	1676.87	1571.92	2045.83	2063.32
Alibeyköy	***Saf Regresyon***	2045.21	1930.76	1910.49	2309.13	2293.61
	İki Katmanlı	1927.28	1908.85	1758.80	2301.95	2290.03
Beşiktaş	***Saf Regresyon***	1220.20	1239.75	1109.27	1556.18	1533.70
	İki Katmanlı	1210.23	1236.65	1070.17	1530.36	1507.25
Esenler	***Saf Regresyon***	1156.89	1338.12	1120.46	1872.55	1830.00
	İki Katmanlı	1121.91	1229.11	1063.72	1695.97	1737.66
Kartal	***Saf Regresyon***	2049.97	2206.60	2004.84	2963.46	2935.30
	İki Katmanlı	1954.27	1990.72	1937.88	2895.20	2930.20
Sarıyer	***Saf Regresyon***	645.76	624.57	690.19	953.93	926.66
	İki Katmanlı	643.77	621.00	621.00	911.86	938.79
Silivri	***Saf Regresyon***	532.74	559.41	514.71	760.23	739.77
	İki Katmanlı	510.56	518.27	506.67	748.96	739.77
Üsküdar	***Saf Regresyon***	316.54	331.43	306.43	494.07	462.98
	İki Katmanlı	312.76	327.28	302.27	450.91	462.98
Yenibosna	***Saf Regresyon***	1592.65	1715.39	1569.00	2477.56	2433.97
	İki Katmanlı	1320.23	1385.37	1299.75	2393.47	2434.91

Çizelge 4.21 : Kirlilik Verisi Üzerinde OKH Sonuçları.

		EAR	*GAR*	*ROR*	*poli-DVM*	*rbf-DVM*
Aksaray	***Saf Regresyon***	1356.65	1425.92	1386.42	2005.36	1748.73
	İki Katmanlı	1256.47	1384.96	1384.96	1970.09	1722.70
Alibeyköy	***Saf Regresyon***	1774.90	1600.01	1598.81	2184.51	1978.90
	İki Katmanlı	1694.81	1590.51	1580.51	2164.08	1965.07
Beşiktaş	***Saf Regresyon***	1171.37	1188.25	1257.64	1533.90	1476.59
	İki Katmanlı	1102.21	1159.79	1219.79	1500.25	1442.31
Esenler	***Saf Regresyon***	1110.00	1179.16	1178.01	1612.21	1439.22
	İki Katmanlı	1055.94	1176.11	1176.11	1563.68	1382.75
Kartal	***Saf Regresyon***	2952.03	2724.52	2794.46	3077.38	3023.33
	İki Katmanlı	2882.71	2720.67	2624.67	3067.74	3015.69
Sarıyer	***Saf Regresyon***	931.49	863.45	866.24	1113.64	1094.22
	İki Katmanlı	925.20	852.07	852.07	1105.27	1091.99
Silivri	***Saf Regresyon***	331.12	333.05	315.88	448.60	384.73
	İki Katmanlı	295.45	296.94	296.94	413.76	358.89
Üsküdar	***Saf Regresyon***	236.91	214.85	222.31	373.59	489.89
	İki Katmanlı	212.78	202.98	202.98	416.58	575.27
Yenibosna	***Saf Regresyon***	1312.49	1298.09	1346.63	2552.12	2509.38
	İki Katmanlı	1274.96	1230.26	1230.26	2125.76	2098.36

5. SONUÇLAR

Bu tez kapsamında İstanbul için hava kirliliği tahminleme problemine çözüm olarak İki Katmanlı Hava Kirliliği Tahminleme Modeli önerilmiştir. Modelimizin performansı PM_{10} kirleticisi tahmininde saf regresyon yöntemleri EAR, ROR, GAR, poly-DVMR, rbf-DVMR ile karşılaştırılmıştır. İki performans ölçütüne (OMH ve OKH) göre de İki Katmanlı Hava Kirliliği Tahminleme Modeli'nin dengesiz dağılan veri üzerinde üstün performans sergilediği ispatlanmıştır.

Üzerinde çalıştığımız veri kümelerinin oldukça dengesiz dağılıma sahip olması; topluluk modellerinin dengesiz dağılan veride hiyerarşik yapısı ile öne çıkması ve sınıflandırmada daha yüksek performans sergilemesi bizi bu tasarıma yönlendiren sebeplerdir.

Önerilen İki Katmanlı Hava Kirliliği Tahminleme Modeli sadece PM_{10} ile hava kirliliği tahmini için değil, benzer veri dağılımına sahip farklı uygulamalar için de kullanılabilir. Sadece ROR ya da EAR için değil, diğer saf topluluk modellerine de birinci katmanın eklenmesinin (Örneklerin regressörden önce ROS ile ikili sınıflandırma aşamasından geçirilmesi), bilinen tekli kullanıma göre modellerin performansını arttırdığı ortadadır.

Bu çalışmanın katkıları; (1) PM_{10} kirletici seviyesi açısından İstanbul için tehdit edici durumları öngörme kabiliyeti, (2) hava kirliliği tahmininde dengesiz dağılan veri sorununa odaklanması, (3) örnekleme yöntemlerini incelemesi ve topluluk modellerini uygulaması, (4) İki Katmanlı Hava Kirliliği Tahminleme Modeli ile dengesiz veriden doğru PM_{10} yoğunluk tahmini yapması.

Gelecekteki çalışmalar için iki hedefimiz var: lokal hedefimiz İstanbul'daki diğer ölçüm istasyonlarından veri toplamak ve tüm verileri birleştiren bir model tasarlamak, global hedefimiz ise İki Katmanlı Hava Kirliliği Tahminleme Modeli'nin performansını farklı hava kirliliği tahmini problemlerinde izlemektir.

KAYNAKLAR

[1] **Van Donkelaar, A.**, **Martin, R.V.**, **Brauer, M.**, **Kahn, R.**, **Levy, R.**, **Verduzco, C. ve Villeneuve, P.J.** (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application, *Environmental health perspectives*, *118*(6), 847.

[2] **Martin, R.V.** (2008). Satellite remote sensing of surface air quality, *Atmospheric environment*, *42*(34), 7823–7843.

[3] **Hoff, R.M. ve Christopher, S.A.** (2009). Remote sensing of particulate pollution from space: have we reached the promised land?, *Journal of the Air & Waste Management Association*, *59*(6), 645–675.

[4] **Çevre ve Şehircilik Bakanlığı**. `http://havaizleme.gov.tr/`.

[5] **Kalkstein, L.S. ve Corrigan, P.** (1986). A synoptic climatological approach for geographical analysis: assessment of sulfur dioxide concentrations, *Annals of the Association of American Geographers*, *76*(3), 381–395.

[6] **Lal, B. ve Tripathy, S.S.** (2012). Prediction of dust concentration in open cast coal mine using artificial neural network, *Atmospheric Pollution Research*, *3*(2), 211–218.

[7] **Raischel, F.**, **Russo, A.**, **Haase, M.**, **Kleinhans, D. ve Lind, P.G.** (2012). Searching for optimal variables in real multivariate stochastic data, *Physics Letters A*, *376*(30-31), 2081–2089.

[8] **Wehner, B.**, **Birmili, W.**, **Gnauk, T. ve Wiedensohler, A.** (2002). Particle number size distributions in a street canyon and their transformation into the urban-air background: measurements and a simple model study, *Atmospheric Environment*, *36*(13), 2215–2223.

[9] **Jacob, D.J. ve Winner, D.A.** (2009). Effect of climate change on air quality, *Atmospheric environment*, *43*(1), 51–63.

[10] **Fiore, A.M.**, **Naik, V.**, **Spracklen, D.V.**, **Steiner, A.**, **Unger, N.**, **Prather, M.**, **Bergmann, D.**, **Cameron-Smith, P.J.**, **Cionni, I.**, **Collins, W.J.** *ve diğerleri* (2012). Global air quality and climate, *Chemical Society Reviews*, *41*(19), 6663–6683.

[11] **Rasmussen, D.**, **Hu, J.**, **Mahmud, A. ve Kleeman, M.J.** (2013). The ozone–climate penalty: past, present, and future, *Environmental science & technology*, *47*(24), 14258–14266.

[12] **Seinfeld, J.H. ve Pandis, S.N.** (2012). *Atmospheric chemistry and physics: from air pollution to climate change*, John Wiley & Sons.

[13] **Elminir, H.K.** (2005). Dependence of urban air pollutants on meteorology, *Science of the Total Environment*, *350*(1-3), 225–237.

[14] **Hamidi, M.**, **Kavianpour, M.R. ve Shao, Y.** (2013). Synoptic analysis of dust storms in the Middle East, *Asia-Pacific Journal of Atmospheric Sciences*, *49*(3), 279–286.

[15] **Seinfeld, J.H. ve Pandis, S.N.** (2016). *Atmospheric chemistry and physics: from air pollution to climate change*, John Wiley & Sons.

[16] **Müdürlüğü, M.G.** `https://www.mgm.gov.tr/`.

[17] **Tang, Y.**, **Krasser, S.**, **Judge, P. ve Zhang, Y.Q.** (2006). Fast and effective spam sender detection with granular svm on highly imbalanced mail server behavior data, *2006 International Conference on Collaborative Computing: Networking, Applications and Worksharing*, IEEE, s.1–6.

[18] **Chawla, N.V.**, **Japkowicz, N. ve Kotcz, A.** (2004). Special issue on learning from imbalanced data sets, *ACM Sigkdd Explorations Newsletter*, *6*(1), 1–6.

[19] **Phua, C.**, **Alahakoon, D. ve Lee, V.** (2004). Minority report in fraud detection: classification of skewed data, *Acm sigkdd explorations newsletter*, *6*(1), 50–59.

[20] **Singh, K.P.**, **Gupta, S.**, **Kumar, A. ve Shukla, S.P.** (2012). Linear and nonlinear modeling approaches for urban air quality prediction, *Science of the Total Environment*, *426*, 244–255.

[21] **Sheikh Saeed Ahmad, R.U..M.N.** (2015). Air Pollution Monitoring and Prediction, *Intech Open*.

[22] **Taneja, S.**, **Sharma, N.**, **Oberoi, K. ve Navoria, Y.** (2016). Predicting trends in air pollution in Delhi using data mining, *Information Processing (IICIP), 2016 1st India International Conference on*, IEEE, s.1–6.

[23] **Mishra, D. ve Goyal, P.** (2015). Development of artificial intelligence based NO2 forecasting models at Taj Mahal, Agra, *Atmospheric Pollution Research*, *6*(1), 99–106.

[24] **Russo, A.**, **Raischel, F. ve Lind, P.G.** (2013). Air quality prediction using optimal neural networks with stochastic variables, *Atmospheric Environment*, *79*, 822–830.

[25] **Corani, G.** (2005). Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning, *Ecological Modelling*, *185*(2-4), 513–529.

[26] **Siwek, K.**, **Osowski, S.**, **Garanty, K. ve Sowinski, M.** (2009). Ensemble of predictors for forecasting the PM10 pollution, *Theoretical Engineering (ISTET), 2009 XV International Symposium on*, VDE, s.1–5.

[27] **Kaminski, W.**, **Skrzypski, J. ve Jach-Szakiel, E.** (2008). Application of artificial neural networks (ANNs) to predict air quality classes in big cities, *19th International Conference on Systems Engineering*, IEEE, s.135–140.

[28] **Skrzypski, J.**, **Jach-Szakiel, E. ve Kaminski, W.** (2008). Neural Models for Prediction of Maximum Daily Particulate Matter PM10 Concentration in the Air in Big Cities as Ecological Safety Management Tools, *Systems Engineering, 2008. ICSENG'08. 19th International Conference on*, IEEE, s.141–146.

[29] **Vong, C.M.**, **Ip, W.F.**, **Wong, P.K. ve Chiu, C.C.** (2014). Predicting minority class for suspended particulate matters level by extreme learning machine, *Neurocomputing*, *128*, 136–144.

[30] **Zhang, J. ve Ding, W.** (2017). Prediction of air pollutants concentration based on an Extreme Learning Machine: The case of Hong Kong, *International journal of environmental research and public health*, *14*(2), 114.

[31] **Nieto, P.G.**, **Lasheras, F.S.**, **García-Gonzalo, E. ve de Cos Juez, F.** (2018). PM 10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study, *Science of the Total Environment*, *621*, 753–761.

[32] **Sayegh, A.S.**, **Munir, S. ve Habeebullah, T.M.** (2014). Comparing the performance of statistical models for predicting PM10 concentrations, *Aerosol and Air Quality Research*, *14*(3), 653–65.

[33] **Kleine Deters, J.**, **Zalakeviciute, R.**, **Gonzalez, M. ve Rybarczyk, Y.** (2017). Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters, *Journal of Electrical and Computer Engineering*, *2017*.

[34] **Zhu, D.**, **Cai, C.**, **Yang, T. ve Zhou, X.** (2018). A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization, *Big Data and Cognitive Computing*, *2*(1), 5.

[35] **Tang, M.**, **Wu, X.**, **Agrawal, P.**, **Pongpaichet, S. ve Jain, R.** (2017). Integration of diverse data sources for spatial PM2. 5 data interpolation, *IEEE Transactions on Multimedia*, *19*(2), 408–417.

[36] **Campalani, P.**, **Nguyen, T.N.T.**, **Mantovani, S. ve Mazzini, G.** (2011). On the Automatic Prediction of PM 10 with in-situ measurements, satellite AOT retrievals and ancillary data, *Signal Processing and Information Technology (ISSPIT), 2011 IEEE International Symposium on*, IEEE, s.093–098.

[37] **Mingjian, F.**, **Guocheng, Z.**, **Xuxu, Z. ve Zhongyi, Y.** (2011). Study on air fine particles pollution prediction of main traffic route using artificial neural network, *Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM), 2011 International Conference on*, IEEE, s.1346–1349.

[38] **Haiming, Z. ve Xiaoxiao, S.** (2013). Study on prediction of atmospheric PM2. 5 based on RBF neural network, *Digital Manufacturing and Automation (ICDMA), 2013 Fourth International Conference on*, IEEE, s.1287–1289.

[39] **Fan, Q.**, **Li, Y. ve Ren, N.** (2009). Application of Grey Prediction Model to Forecast the Main Air Contaminant PM10 in Harbin City, *Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on*, IEEE, s.1–4.

[40] **Kim, M.**, **Kim, Y.**, **Sung, S. ve Yoo, C.** (2009). Data-driven prediction model of indoor air quality by the preprocessed recurrent neural networks, *ICCAS-SICE, 2009*, IEEE, s.1688–1692.

[41] **Galar, M.**, **Fernandez, A.**, **Barrenechea, E.**, **Bustince, H. ve Herrera, F.** (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463–484.

[42] **Kurt, A. ve Oktay, A.B.** (2010). Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks, *Expert Systems with Applications*, *37*(12), 7986–7992.

[43] **Chawla, N.V.**, **Bowyer, K.W.**, **Hall, L.O. ve Kegelmeyer, W.P.** (2002). SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, *16*, 321–357.

[44] **Pedregosa, F.**, **Varoquaux, G.**, **Gramfort, A.**, **Michel, V.**, **Thirion, B.**, **Grisel, O.**, **Blondel, M.**, **Prettenhofer, P.**, **Weiss, R.**, **Dubourg, V.** *ve diğerleri* (2011). Scikit-learn: Machine learning in Python, *Journal of machine learning research*, *12*(Oct), 2825–2830.

[45] **Felsenstein, J.** (1985). Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, *39*(4), 783–791.

[46] **Geurts, P.**, **Ernst, D. ve Wehenkel, L.** (2006). Extremely randomized trees, *Machine learning*, *63*(1), 3–42.

[47] **Friedman, J.H.** (2002). Stochastic gradient boosting, *Computational Statistics & Data Analysis*, *38*(4), 367–378.

[48] **Bengio, Y.**, **Simard, P. ve Frasconi, P.** (1994). Learning long-term dependencies with gradient descent is difficult, *IEEE transactions on neural networks*, *5*(2), 157–166.

[49] **Kim, D. ve Kim, C.** (1997). Forecasting time series with genetic fuzzy predictor ensemble, *IEEE Transactions on Fuzzy systems*, *5*(4), 523–535.

[50] **Estabrooks, A.**, **Jo, T. ve Japkowicz, N.** (2004). A multiple resampling method for learning from imbalanced data sets, *Computational intelligence*, *20*(1), 18–36.

[51] **Krawczyk, B.** (2016). Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence*, *5*(4), 221–232.

[52] **Kaya, K. ve Öğüdücü, Ş.G.** (2018). A Binary Classification Model for PM 10 Levels, *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, IEEE, s.361–366.

EKLER

EK A.1 : Terimler Sözlüğü

EK A.1: Terimler Sözlüğü

Bu bölümde tez içerisinde kullanılan Türkçe terimlerin literatürde yer alan İngilizce karşılıkları verilmektedir.

Destek Vektör Makinesi	Support Vector Machine
Rasgele Orman	Random Forest
Ekstra Ağaç	Extra Trees
Gradyan Arttırma	Gradient Boosting
Çok Katmanlı Algılayıcı	Multilayer Perceptron
Ortalama Kareli Hata (OKH)	Mean Squared Error (MSE)
Ortalama Mutlak Hata (OMH)	Mean Absolute Error (MAE)

ÖZGEÇMİŞ

Ad Soyad: Kıymet Kaya

Doğum Tarihi ve Yeri: 23.07.1991, Antakya

E-Posta: kayak16@itu.edu.tr

ÖĞRENİM DURUMU:

- **Lisans:** 2014, Ege Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü.
- **Y. Lisans:** 2018, İstanbul Teknik Üniversitesi, Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programı.

MESLEKİ DENEYİMLER VE ÖDÜLLER:

- Araştırma Görevlisi, 2016 - halen. İTÜ Bilgisayar Bilişim Fakültesi.
- Araştırma Görevlisi, 2015 - 2016. RTEÜ Mühendislik Fakültesi.

YÜKSEK LİSANS TEZİNDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- Kaya K., Öğüdücü S., 2018. A Binary Classification Model for PM10 Levels. *International Conference on Computer Science and Engineering (UBMK)* , September 20-23, 2018 Sarajevo, Bosnia & Herzegovina.
- Kaya K., Öğüdücü S., 2019. Deep Flexible Sequential (DFS) Model for Air Pollution Forecasting. *Environ Monit Assess*, (2019). [Değerlendirme aşamasında.]

DİĞER YAYINLAR, SUNUMLAR VE PATENTLER:

- M. F. Akay, E. Çetin, İ. Yarım, F. Abut and K. Kaya, "New Regression Equations for Estimating the Maximal Oxygen Uptake of College of Physical Education and Sports Students in Turkey", 5th Cyprus International Conference on Educational Research (CYICER-2016), North Cyprus, 1-2 April 2016, pp. 162.
- K. Kaya, M. F. Akay, E. Çetin and İ. Yarım, "Development of New Prediction Models for Maximal Oxygen Uptake Using Artificial Intelligence Methods", International Conference on Natural Science and Engineering (ICNASE'16), Kilis, Turkey, 19-20 March 2016, pp. 988-986.