



**MAKİNE ÖĞRENİMİ TEKNİKLERİ İLE VERİYE DAYALI KISA VADELİ
DOĞAL GAZ TALEP TAHMİNİ**

Mustafa ÇELEBİ

**YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

MAYIS 2024

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Mustafa ÇELEBİ

07/05/2024

MAKİNE ÖĞRENİMİ TEKNİKLERİ İLE VERİYE DAYALI KISA VADELİ DOĞAL GAZ TALEP TAHMİNİ

(Yüksek Lisans Tezi)

Mustafa ÇELEBİ

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Mayıs 2024

ÖZET

Doğal gazın arz ve talebinin doğru bir şekilde planlanması ve yönetilmesi sürecinde gaz talebinin yüksek doğruluk ile öngörülmesi kritik öneme sahiptir. Doğal gaz talep tahminlerinde yapılacak büyük hatalar hem doğal gaz dağıtım şirketlerine hem de ülke ekonomisine milyonlarca dolar zarara uğratmaktadır. Bu kayıpların en aza indirilmesi için, yüksek doğrulukta tahmin modelleri geliştirmek elzemdir. Bu amaçla ilk olarak Türkiye'nin dört farklı bölgesinin dokuz yıl boyunca günlük doğalgaz tüketim verisi temin edilmiştir. Elde edilen zaman serileri kullanılarak gecikmeli, pencere ve zamansal özellikler başlıkları altında toplamda 33 adet bağımsız değişken oluşturulmuştur. Oluşturulacak tahmin modellerine en yüksek katkıyı sağlayacak değişkenleri belirlemek için LASSO regresyonu kullanılmıştır. Ardından uzun kısa süreli bellek, topluluk öğrenme stratejileri ve çok değişkenli regresyon modeli uygulanarak bu şehirlerin günlük doğalgaz tüketimi tahmin edilmeye çalışılmıştır. Oluşturulan modellerin performansı OMÖH, R^2 , OMYH ve SOMYH yöntemleri ile değerlendirilmiştir. Oluşturulan farklı tahmin modellerin test kümesindeki tahminleri arasındaki farklılıklarının gözlenmesi amacı ile Friedman sıralama toplam testi uygulanmıştır. Çok değişkenli regresyon ve ağaç tabanlı yöntemler tahmin doğruluğu, uyarlanabilirlik ve genelleme yetenekleri ile değerlendirildiğinde iyi performans gösterdiği için doğal gaz dağıtım ve kontrol sistemlerinde başarılı bir şekilde kullanılabileceği görülmüştür. Çalışmanın son aşamasında, makine öğrenimi modellerine ait tahminlerin tüm kombinasyonları, maksimum, minimum, ortalama ve medyan metrikleri ile birleştirilerek doğruluk performansının artırılması amaçlanmıştır. Elde edilen sonuçlar, kombinasyon modellerinin kısmen daha iyi sonuçlar verdiğini göstermiştir.

Bilim Kodu : 90607

Anahtar Kelimeler : Günlük doğal gaz talep tahmini, topluluk öğrenmesi, UKSB, çok değişkenli regresyon

Sayfa Adedi : 64

Danışman : Doç. Dr. Hakan ÇERÇİOĞLU

DATA-BASED SHORT-TERM NATURAL GAS DEMAND FORECASTING WITH MACHINE LEARNING TECHNIQUES

(M. Sc. Thesis)

Mustafa ÇELEBİ

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

May 2024

ABSTRACT

Predicting daily natural gas demand accurately is crucial for effective management of supply and demand of natural gas consumption. Significant errors in natural gas demand forecasts can result in millions of dollars in losses for natural gas distribution companies and the national economy. Developing prediction models with high accuracy is essential to minimize these losses. To this end, daily natural gas consumption data for four different regions of Turkey over nine years was initially obtained. For time series modeling, 33 independent variables were created under the headings of lagged, window, and time features. LASSO regression was employed to identify the variables that would contribute most to the prediction models to be constructed. Subsequently, long short-term memory, ensemble learning strategies, and multivariate regression models were applied to predict these cities' daily natural gas consumption. The performance of the constructed models was evaluated using metrics such as MAPE, R^2 , MASE, and sMAPE. The Friedman rank sum test was applied to observe differences in predictions among various forecast models in the test dataset. Multivariate regression and tree-based methods demonstrated remarkable performance in terms of prediction accuracy, adaptability, and generalization ability, indicating their potential successful utilization in natural gas distribution and control systems. In the study's final stage, combinations of machine learning models' predictions were merged using maximum, minimum, mean, and median metrics to enhance accuracy performance. The obtained results indicated that combination models partially provided better results.

Science Code : 90607

Key Words : Daily natural gas demand forecasting, ensemble learning, LSTM, multiple linear regression

Page Number : 64

Supervisor : Assoc. Prof. Dr. Hakan ÇERÇİOĞLU

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Doç. Dr. Hakan ÇERÇİOĞLU'na, veri temini konusunda gösterdiği kolaylık için AKMERCAN ENERJİ Genel Müdürü Yunus KESGİN'e, teknik bilgilendirmeleri ile beni sürekli destekleyen Ticari Hizmetler Müdürü Hakan KIZILELMA'ya ve son olarak manevi destekleriyle bana her zaman destek olan çok değerli aileme ve arkadaşlarıma sonsuz teşekkürlerimi sunarım.



İÇİNDEKİLER

	Sayfa
ÖZET	v
ABSTRACT.....	vi
TEŞEKKÜR.....	vii
İÇİNDEKİLER	viii
ÇİZELGELER LİSTESİ.....	x
ŞEKİLLER LİSTESİ	xi
SİMGELER VE KISALTMALAR.....	xii
1. GİRİŞ.....	1
2. LİTERATÜR ARAŞTIRMASI	11
3. MATERYAL METOD	17
3.1. Kullanılacak Makine Öğrenmesi Algoritmalar	17
3.2. Karar Ağaçları (KA)	17
3.3. Torbalama	19
3.4. Güçlendirme	20
3.5. Rastgele Ormanlar (RO) ve Aşırı Rastgeleleştirilmiş Ağaçlar (ARA)	22
3.6. Çok Değişkenli Regresyon (ÇDR).....	23
3.7. LASSO Regresyon	24
3.8. Aşırı Gradyan Artırma (XGBoost).....	25
3.9. Hafif Gradyan Artırma Makinesi (LightGBM).....	26
3.10. Uzun Kısa Süreli Bellek (UKSB).....	28
3.11. Değerlendirme Ölçütleri.....	29
3.12. İstatistiksel Test.....	30
4. UYGULAMA.....	33
4.1. Çalışma Alanı ve Verinin Tanımlanması	35
4.2. Eksik Değer Analizi	39

	Sayfa
4.3. Bağımsız Değişken Oluşturma ve Seçme	39
4.4. Modellerin Uygulanması.....	40
5. SONUÇ VE DEĞERLENDİRME.....	53
KAYNAKLAR	52
EK-1. Bölgelerin tüketim mutlak hata grafikleri	56
EK-2. Bölgelerin tüketim ısı grafikleri.....	59
ÖZGEÇMİŞ	60



ÇİZELGELER LİSTESİ

Çizelge	Sayfa
Çizelge 1.1. Dünya’da birincil enerji kaynakları 2021 dağılımı.....	1
Çizelge 1.2. Dünya’da birincil enerji tüketim bölgeleri 2021 dağılımı.....	2
Çizelge 2.1. Türkiye'deki doğal gaz tüketim tahmini üzerine yapılan bazı çalışmalar...	13
Çizelge 3.1. Model değerlendirme ölçütleri.....	30
Çizelge 3.2. Fridmen testi için veri tablosu.....	31
Çizelge 4.1. Çalışmaya konu edinilen şehirlere ait bazı tanımlayıcı bilgiler.....	35
Çizelge 4.2. Veri setlerinin istatistiksel özellikleri (Sm^3).....	37
Çizelge 4.3. LASSO regresyonu tarafından seçilen bağımsız değişkenler.....	40
Çizelge 4.4. Seçilen modeller ve tahmin hataları.....	42
Çizelge 4.5. Eğitim verisi için modellerin tahmin hatası.....	42
Çizelge 4.6. Test verisi için modellerin tahmin hatası.....	43
Çizelge 4.7. İstatistiksel test sonuçları (p değeri).....	44

ŞEKİLLER LİSTESİ

Şekil	Sayfa
Şekil 1.1. Yıllara göre Dünya’da kullanılan enerji miktarı.....	1
Şekil 1.2. Bölgesel doğal gaz tüketimi.....	3
Şekil 1.3. Yıllara göre Türkiye’nin doğal gaz tüketimi	4
Şekil 1.4. Doğalgaz ithal edilen ülkeler	5
Şekil 1.5. Kullanım türüne göre doğal gaz tüketimi	6
Şekil 1.6. Doğal gazın değer akışı	6
Şekil 3.1. Karar ağacı.....	17
Şekil 3.2. Sınıflandırma ağaçlarının algoritması.....	19
Şekil 3.3. Regresyon ağaçlarının algoritması	19
Şekil 3.4. Torbalama süreci	20
Şekil 3.5. Karar ağacı ve rastgele ormanlar	22
Şekil 3.6. UKSB hücresinin yapısı	29
Şekil 4.1. Tahmin sürecine ait akış diyagramı	34
Şekil 4.2. Tahmin işlemi gerçekleştirilecek dört bölge.....	36
Şekil 4.3. Batıkar bölgesinin tüketim ısı grafiği	38
Şekil 4.4. Model eğitimi için çapraz doğrulama şeması	41
Şekil 4.5. Bölgelere göre modellerin ürettiği performans ölçütleri	46
Şekil 4.6. Dört farklı şehrin gerçek ve tahmin edilen tüketim değerleri.....	47
Şekil 4.7. Tahmin modellerinin MAE değerleri	48
Şekil 4.8. Yirmi dört model (6 yöntem × 4 bölge) için ekstrem veri noktalarının sayısı	49
Şekil 4.9. Kombinasyon modellerin SOMYH değerleri.....	50
Şekil 4.10. Dağıtım bölgesi kırılımında modellerin eğitim süreleri	51

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklamalar
τ	Sürekli değişkenin bölüneceği eşik değeri
β_p	Çok değişkenli regresyonun katsayıları
ε_t	Çok değişkenli regresyonun hata terimi
λ	LASSO regresyonun düzensileştirme katsayısı
Ω	Regresyon ağacının karmaşıklığı
L	Kayıp fonksiyonu
T	Regresyon ağacındaki yaprak sayısı
γ	Düzensileştirme katsayısı
km	Kilometre
x_t	t zamanındaki girdi verisi
h_t	UKSB biriminin çıktısı (t zamanındaki kısa dönemli hafıza)
i_t	Girdi kapısı
f_t	Unutma kapısı
o_t	Çıktı kapısı
C_t	Uzun dönemli hafıza
$\hat{C}_t$	Güncellenmiş uzun dönemli hafıza
W	UKSB birimindeki ağırlık değerleri
b	Bias değeri
σ	Sigmoid aktivasyon fonksiyonu
y_t	t zamanındaki doğalgaz tüketim miktarı
$\hat{y}_t$	t zamanı için tahmin edilen doğalgaz tüketim miktarı
Sm^3	Standart metreküp
e_t	Hata terimi
$\text{Gain}(X_i)$	i bağımsız değişken için kazanım değeri

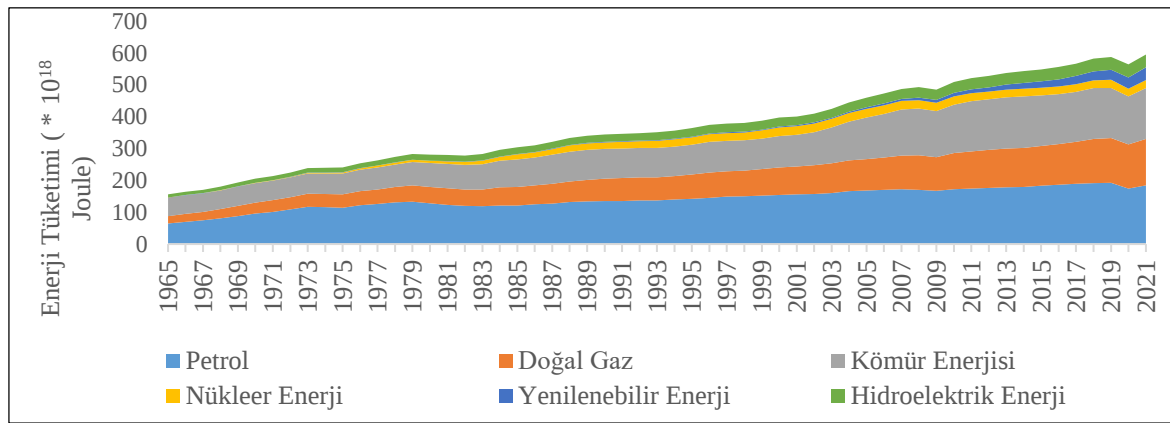
Simgeler	Açıklamalar
$H(Y)$	Bağımlı değişkene ait entropi değeri
k_i	i bağımsız değişkenin kategori sayısı
$P(X_{ij})$	i bağımsız değişkenin j düzeyinin ortaya çıkma olasılığı
$H(Y / X_{ij})$	i bağımsız değişkenin j düzeyi için bağımlı değişkene ait entropi değeri
m	Bağımsız değişken sayısı
$S(T)$	Sürekli bağımlı değişkenin standart sapması
$S(T, X)$	X bağımsız değişkeni kırılımında sürekli bağımlı değişkenin ağırlıklı standart sapması
$P(c)$	X bağımsız değişkeninin alt düzeyinin ortaya çıkma olasılığı
$S(c)$	X bağımsız değişkeni kırılımında sürekli bağımlı değişkenin standart sapması

Kısaltmalar	Açıklamalar
AMD	Ampirik mod ayrıştırması
ARA	Aşırı Rastgeleleştirilmiş Ağaçlar
BDÖİ	Basıncı Düzenleme ve Ölçüm İstasyonu
BOTAŞ	Türkiye Petrol Boru Hatları Anonim Şirketi
BP	İngiliz Petrolleri
BT	Benzetilmiş Tavlama
ÇDR	Çok Değişkenli Regresyon
ÇDURE	Çok Değişkenli Uyarlamalı Regresyon eğrileri
DD	Dalgacık Dönüşümü
DR	Doğrusal Regresyon
DVM	Destek Vektör Regresyonu
ESA	Evrişimli sinir ağları

Kısaltmalar	Açıklamalar
GA	Genetik Algoritma
GAA	Gradyan Arttırma Ağacı
GTSA	Geliştirilmiş tekil spektrum analizi
KA	Karar Ağaçları
KKP	Kurumsal Kaynak Planlaması
KP	Karmaşıklık Parametresi
LightGBM	Hafif Gradyan Artırma
MKVTS	Gözetleyici Kontrol ve Veri Toplama Sistemi
OKHK	Ortalama Kare Hatasının Karekökü
OMANH	Ortalama Mutlak Aralık Normalleştirme Hatası
OMÖH	Ortalama mutlak ölçekli hata
OMH	Ortalama Mutlak Hata
OMYH	Ortalama Mutlak Hata Yüzdesi
R²	R-Kare
RO	Rastgele Ormanlar
SD	Sıvılaştırılmış Doğalgaz
SDG	Sıkıştırılmış Doğal gaz
SKH	Sürdürülebilir Kalkınma Hedefleri
SOMYH	Simetrik ortalama mutlak yüzde hata
TANAP	Trans Anadolu Doğal Gaz Boru Hattı
TPAO	Türkiye Petrolleri Anonim Ortaklığı
TANAP	Trans Anadolu Doğal Gaz Boru Hattı
TSA	Tekrarlayan Sinir Ağı
UKSB	Uzun Kısa Süreli Bellek
USBÇ	Uyarlamalı Sinirsel Bulanık Çıkarım
YSA	Yapay Sinir Ağları
XGBoost	Aşırı Gradyan Artırma

1. GİRİŞ

Hızla artan nüfus, sanayileşme ve kentleşme ile dünya çapında enerji tüketimi son yıllarda önemli ölçüde artmıştır. British Petroleum (BP) istatistiksel incelemelerine göre, Dünya enerji kullanımı özellikle 1990'lı yılların sonu itibariyle büyük bir hızla artmaktadır [1]. Şekil 1.1'de Dünya'da yıllara göre tüketilen enerjinin joule cinsinden değerleri enerji üretim çeşidine göre gösterilmiştir.



Şekil 1.1. Yıllara göre Dünya'da kullanılan enerji miktarı [1]

Hidroelektrik, nükleer ve yenilenebilir enerji kaynakları başta olmak üzere alternatif enerji kaynaklarındaki ilerlemelere rağmen Dünyamız enerji tüketiminin büyük bölümünü hala fosil kaynaklar ile karşılamaktadır. Dünya enerji tüketiminin 2021 yılında birincil enerji kaynaklarına göre dağılımı Çizelge 1.1'de verilmektedir [1]. Çizelge incelendiğinde Dünya enerji kullanımının %87'lik kısmının hala fosil kaynaklardan karşılandığı görülmektedir.

Çizelge 1.1. Dünya'da birincil enerji kaynakları 2021 dağılımı

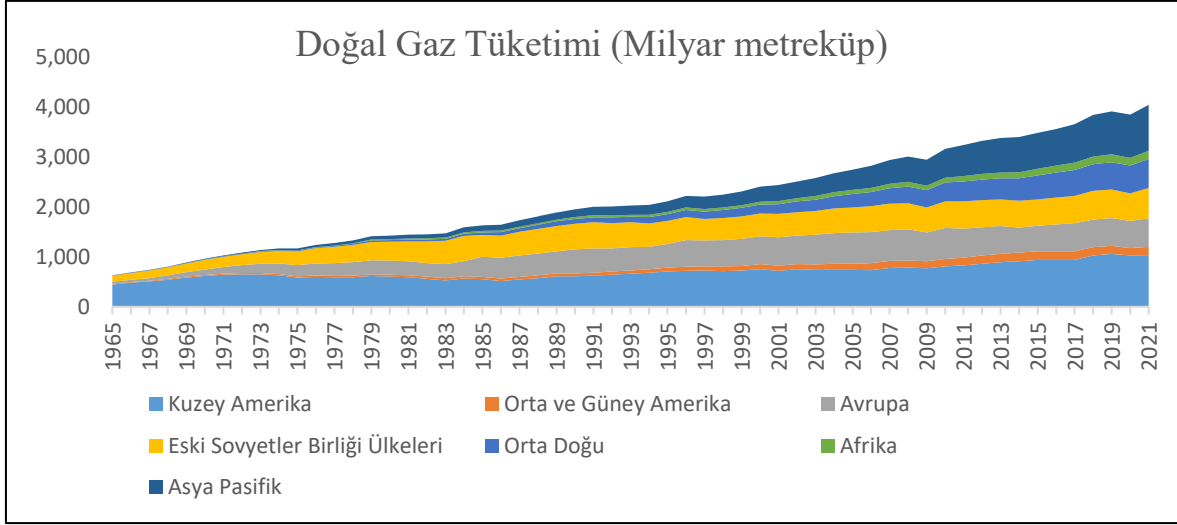
Tür	Oran
Petrol	%38
Doğal Gaz	%21
Kömür Enerjisi	%28
Nükleer Enerji	%5
Yenilenebilir Enerji	%2
Hidroelektrik Enerji	%7

Dünya'daki enerji tüketimi coğrafi olarak incelendiğinde Asya Pasifik, Kuzey Amerika ve Avrupa en çok enerji tüketen üç bölge olarak göze çarpmaktadır. Bu üç bölge, Dünya enerji tüketiminin %78'ini oluşturmaktadır. BP istatistiklerine göre Dünya enerji tüketiminin 2021 yılında bölgelere göre dağılımı Çizelge 1.2'de verilmektedir.

Çizelge 1.2. Dünya'da birincil enerji tüketim bölgeleri 2021 dağılımı

Tür	Oran
Kuzey Amerika	%27
Orta ve Güney Amerika	%5
Avrupa	%21
Eski Sovyetler Birliği Ülkeleri	%10
Orta Doğu	%4
Afrika	%3
Asya Pasifik	%30

Enerji tüketiminin dünya ölçeğinden böylesine yayılması ve yaygınlaşması, farklı enerji üretim alternatiflerinin değerlendirilmesine yol açmıştır. Bu noktada; depolanabilirliği, boru hatları veya gemiler ile kolay taşınabilmesi, yandığı zaman diğer fosil yakıtlardan çok daha az CO₂ ve diğer kirletici gaz salınımı yapması, temiz ve çok yönlü kullanımı, doğal gazı bir enerji üretim seçeneği olarak öne çıkarmıştır [2]. Doğal gaz, enerji güvenliği açısından da önemli bir rol oynar. Dünya genelinde enerji çeşitliliğini artırarak, enerji arzının daha dirençli ve istikrarlı hale gelmesine katkıda bulunur. Bu, enerji talebinin dalgalı olduğu ve çeşitli faktörlere bağlı olarak değişebilen bir dünyada stratejik bir avantaj sağlar. Tüm bu avantajları doğal gazı dünya genelinde çokça talep edilen bir enerji kaynağı haline getirmiştir. Küresel enerji kullanımında doğal gaz tüketimi her geçen yıl artmaktadır. İstatistikler Dünya genelinde son 20 yıl içinde doğal gaza olan talebin 2,4 trilyon m³'ten 4 trilyon m³'e arttığını ortaya koymaktadır [1]. Sadece 2021 yılında Dünya çapında doğalgaz tüketiminde 191,9 milyar m³ artış gerçekleşmiştir. Doğal gaz üretiminde ise yaklaşık %5'lik bir artış meydana gelmiştir [1]. BP istatistiklerine göre küresel doğal gaz tüketiminin bölgesel dağılımı Şekil 1.2'de gösterilmiştir.



Şekil 1.2. Bölgesel doğal gaz tüketimi [1]

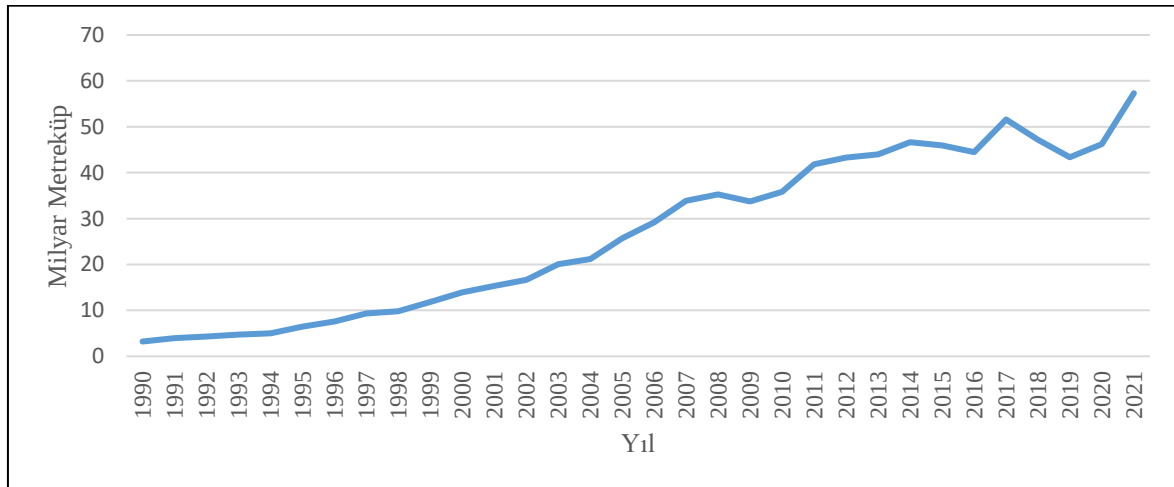
Buna göre küresel doğalgaz tüketimi esas olarak Asya Pasifik, Kuzey Amerika ve Avrupada yoğunlaşmaktadır. Asya Pasifik ve Orta Doğu'da ise tüketimin küresel payı hızla artmaktadır. Gelecekte küresel doğalgaz tüketimindeki artışın üçte ikisi başta Asya ve Orta Doğu bölgesi olmak üzere gelişmekte olan ülkelerde meydana gelecektir ve özellikle yükselen bir piyasa olarak Asya bölgesi talebin ana motoru olacaktır. 2040 yılına kadar Asya'nın gelişmekte olan ülkelerindeki doğal gaz talebindeki artışın, toplam küresel talep büyümesinin yarısını oluşturacağı tahmin edilmektedir [3].

Günümüzde enerji, sürdürülebilir kalkınmanın temel taşlarından biri olarak kabul edilmektedir. Bu bağlamda, doğal gaz tüketiminin sürdürülebilir kalkınma hedefleri (SKH) ile ilişkisi incelenmesi gereken önemli bir konudur [4]. SKH #7, "temiz, uygun fiyatlı, güvenilir, sürdürülebilir ve modern enerjiye erişimi" hedefler. Elektrik üretiminden ısınma sistemlerine kadar geniş bir yelpazede kullanılabilen doğalgaz, enerji erişimini artırabilir ve aynı zamanda temiz enerji hedeflerine ulaşılmasını sağlayabilir. SKH #11, "dayanıklı, kapsayıcı, güvenli ve sürdürülebilir şehirlerin ve yerleşim yerlerinin oluşturulması"nı hedefler. Bu bağlamda, şehirlerin enerji ihtiyaçlarını karşılamak için doğalgaz gibi temiz enerji kaynaklarına geçiş, hava kalitesini artırarak şehirlerde yaşayan insanların yaşam kalitesini iyileştirebilir. Enerji altyapısının güçlendirilmesi ve modernize edilmesi, SKH #9'un odak noktalarından biridir. Doğal gaz, enerji sektöründeki altyapıyı güçlendirmek ve modernize etmek için önemli bir araçtır. Doğalgazın kullanımı, enerji üretiminde verimliliği artırabilir, enerji iletim ve dağıtım sistemlerini iyileştirebilir ve sürdürülebilir enerji

altyapısının inşasına katkıda bulunabilir. Bu da endüstrinin daha sürdürülebilir ve yenilikçi hale gelmesine olanak tanır. Tüm bu hedefler göz önüne alındığında, doğal gaz, sürdürülebilir kalkınma hedefleri doğrultusunda önemli bir enerji kaynağı olarak değerlendirilmelidir. Temiz enerjiye geçiş, sanayide inovasyon ve enerji altyapısının güçlendirilmesi gibi hedeflere ulaşmak için doğalgazın etkili bir şekilde kullanılması, sürdürülebilir bir enerji geleceğinin anahtarı olabilir.

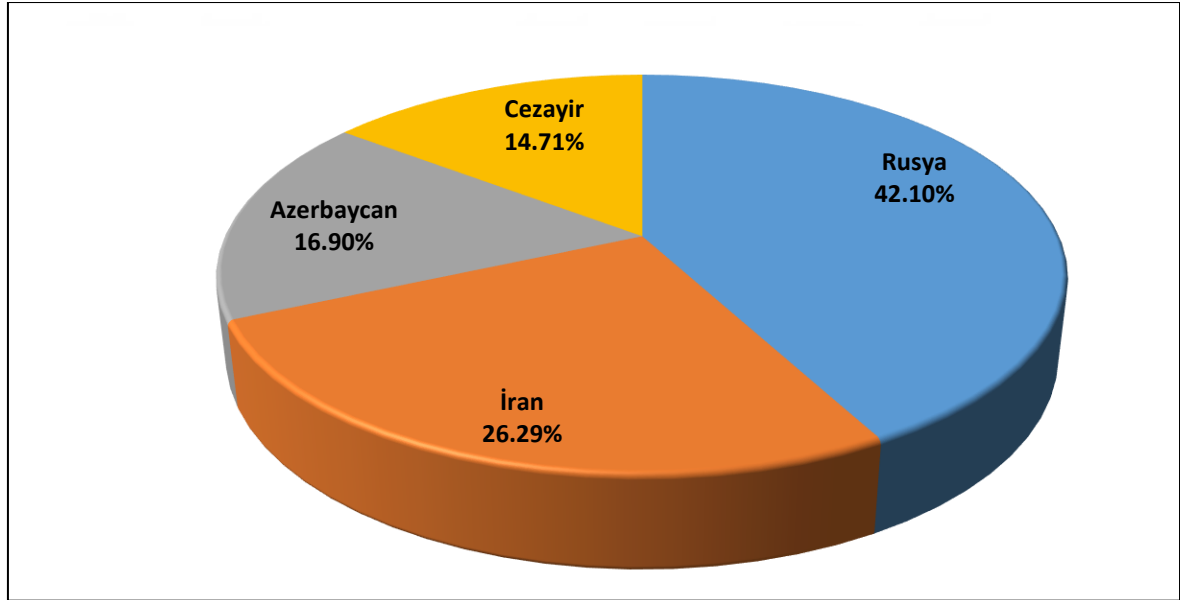
Türkiye’de doğal gaz ilk kez 1970 yılında Türkiye Petrolleri Anonim Ortaklığı (TPAO) tarafından Hamitabat ve Kumrular’da aranmıştır ve ilk kez 1976 yılında Pınarhisar çimento fabrikasında kullanılmaya başlanmıştır. Türkiye’de doğal gaza olan talep 1980’li yılların sonundan itibaren katlanarak artmıştır. Türkiye, artan doğal gaz enerji talebini karşılamak amacıyla 1984 yılında eski Sovyetler Birliği ile doğal gaz dağıtımına ilişkin ilk anlaşmayı imzalayarak, artan çevresel kaygıların dikkate alındığını ve aynı zamanda enerji kaynaklarının çeşitlendirilmesi politikasını yerine getirdiğini gösterdi. O tarihten bu yana doğal gaz ihtiyacını karşılamak amacıyla ek satın alma anlaşmaları hızla yürütüldü [5].

Diğer birçok ülke ile kıyaslandığında, doğal gaz Türkiye’nin toplam enerji tüketiminde çok daha büyük bir paya sahiptir. Örneğin Türkiye’nin 2000 yılında 13,9 milyar m³ olan doğalgaz tüketimi 2021 yılı sonu itibarıyla 57,3 milyar m³’e ulaşmıştır. Türkiye, bu tüketimi ile Avrupa’nın en yüksek gaz tüketimine sahip 5. ülkesi konumundadır [1]. Şekil 1.3’te Türkiye’nin 1990-2021 yılları arasında doğal gaz tüketimi gösterilmektedir.



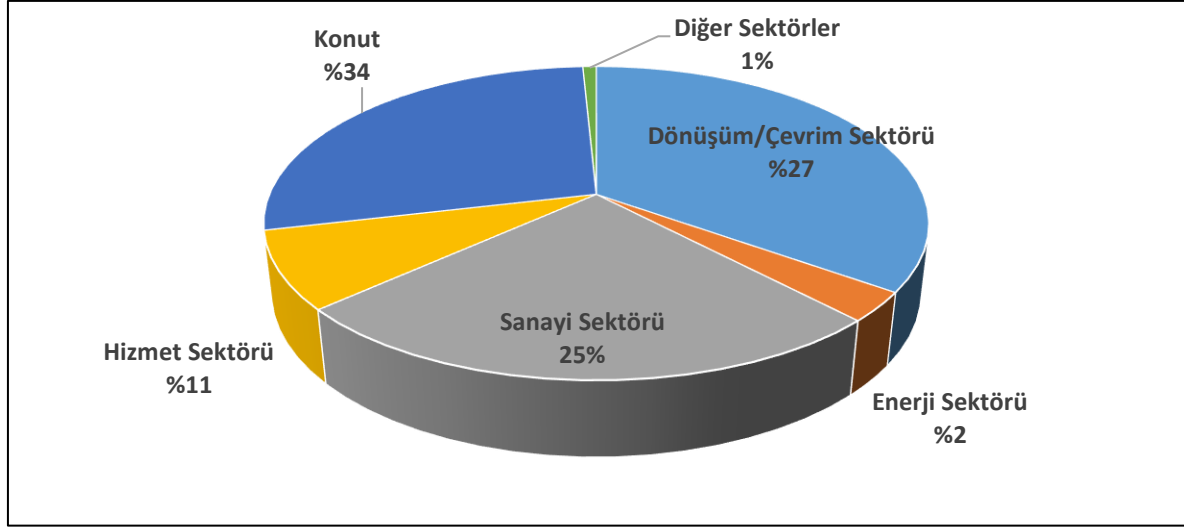
Şekil 1.3. Yıllara göre Türkiye’nin doğal gaz tüketimi [1]

Türkiye, yerli doğal gaz rezervlerinin kıtlığı nedeniyle ihtiyacı olan doğalgazı tamamen yabancı doğalgaz kaynaklarından (%99,33) karşılamaktadır. Şekil 1.4'te 2022 yılında Türkiye'nin doğal gaz ithal ettiği ülkelerin paylarını göstermektedir. Türkiye gaz ithalatının büyük kısmını (%93,42) Türkiye Petrol Boru Hatları Anonim Şirketi (BOTAŞ) tarafından gerçekleştirmektedir. BOTAŞ ülke genelinde ithalat, iletim, dağıtım ve satış faaliyetlerinde münhasır haklara sahip, Türkiye'nin devlete ait gaz ticaret ve iletim şirkettir. BOTAŞ, 2022 yılında yaptığı toplam 53,5 milyar metreküp doğal gaz ithalatının büyük bölümünü Rusya Federasyonu (%42,10), İran (%26,29), Azerbaycan (%16,90) ve Cezayir'den (%14,1) gerçekleştirmiştir [6].



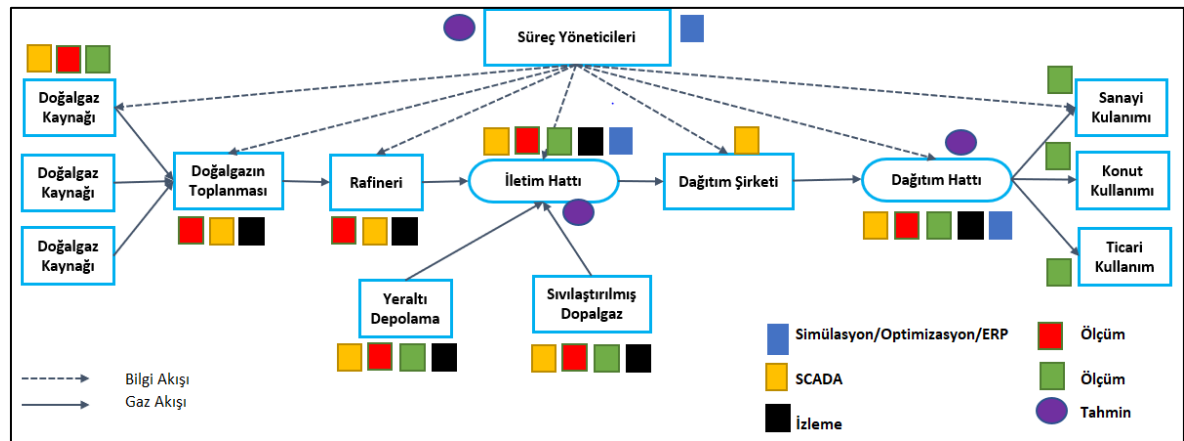
Şekil 1.4. Doğalgaz ithal edilen ülkeler [6]

Şekil 1.5'te 2022 yılında doğal gaz tüketiminin kullanım amacına göre dağılımını göstermektedir [7]. Doğal gaz yoğun olarak elektrik üretimi (%27), sanayi (%25) ve konut tüketimi (%34) alanlarında kullanılmaktadır.



Şekil 1.5. Kullanım türüne göre doğal gaz tüketimi [7]

Şekil 1.6’da doğal gazın kaynağından son tüketiciye ulaşana kadar tüm bilgi ve değer akışı zinciri gösterilmiştir. Doğal gaz birincil enerji kaynağı olarak doğadan çıkarıldıktan sonra toplanır ve rafinerilerde bir dizi kimyasal işlemden geçerek saflaştırılır. Ardından saflaştırılan doğalgaz iletim hatları ile taşınarak doğal gaz dağıtım şirketlerine teslim edilir. Dağıtım şirketleri sorumluluk alanlarındaki nihai tüketicilere inşa ettikleri dağıtım hatları ile doğalgaz tedarik etmek ile sorumludurlar.



Şekil 1.6. Doğal gazın değer akışı [8]

Doğal gazın fiziksel akışı kaynaktan tüketicilere olmak üzere tek yönlüdür. Tüm bu akış, süreç yöneticileri tarafından önceden yapılan sözleşmeler doğrultusunda yönetilir. Her yılın başında, BOTAŞ ile ilgili şehirde doğalgaz dağıtım lisansına sahip firma arasında doğalgaz tedarikiyle ilgili bir sözleşme imzalanmaktadır. Bu sözleşme ile dağıtım şirketi, aylar

bazında dağıtımını gerçekleştireceği doğalgaz miktarını BOTAŞ'a bildirir. Bildirilen gaz miktarının %80' ine tekabül eden miktar, dağıtım şirketinin sorumluluğundadır. Yani, fiili çekim gerçekleştirilse de gerçekleştirilmese de bu miktarın bedelini BOTAŞ'a ödemekle yükümlüdür. Bu durum, dağıtım şirketinin BOTAŞ'a karşı olan "Al ya da Öde" yükümlülüğünü ifade eder.

Doğal gaz akış sürecinin etkin bir şekilde yönetilebilmesi için Merkezi Kontrol ve Veri Toplama Sistemi (MKVTS) ve Kurumsal Kaynak Planlama (KKP) gibi gelişmiş bilgi teknolojileri ve tüketim tahminleri yoğun bir şekilde kullanılır. Bir sözleşme yılında BOTAŞ ile iki farklı tüketim tahmin verisi paylaşılır. Birincisi, sözleşme yılı başında bir sonraki yılın aylarına ilişkin doğalgaz tüketim tahminidir. Bu tahminler, uzman görüşleriyle desteklenir ve ilgili şehirdeki konut ve büyük tüketici abone sayısındaki artış dikkate alınarak oluşturulur. Abone sayısındaki artış ile ortalama abone tüketim miktarı çarpılarak hesaplanan gaz miktarı bir önceki yılın tüketim verisine eklenir. Elde edilen nihai tüketim miktarı aylara dağıtılır. Uzman görüşleri doğrultusunda bu verilerin yenilenmesi mümkündür ve revizyonlar yapıldıktan sonra BOTAŞ'a bildirilir. İkinci tahmin verisi ise günlük doğalgaz tüketim tahmin verisidir. Buna göre dağıtım şirketi, söz konusu günlük bildirimin yapıldığı günden sonraki gün için talep ettiği doğalgaz miktarını BOTAŞ'a bildirmekle yükümlüdür. Çalışmada konu alınan dağıtım şirketi, bu tahmini, bir gün öncesinin gerçekleşen değerini bir gün sonrasının tahmini olarak BOTAŞ'a bildirmektedir.

Doğal gaz Türkiye'deki her bir şehrin giriş noktasına iletim hatları ile BOTAŞ tarafından getirilir. İlgili şehrin dağıtım lisansına sahip şirket inşa ettiği basınç düşürme ve ölçüm istasyonları (BDÖİ) ile doğal gazı teslim alır. Teslim noktaları olarak ifade edilen bu noktalarda gazın mülkiyeti ve riski dağıtım şirketine geçer. Her şehirdeki gaz tedarikinden sorumlu olan dağıtım şirketi, şehir sakinlerinin ihtiyacı olan doğal gazı arz etmekle yükümlüdür. Dağıtım şirketleri gün öncesi kentsel doğal gaz tüketim tahmini ile kentsel doğal gaz tedarik sistemlerinin operasyonel verimliliğini artırabilir ve tüketim planlarının oluşturulmasına rehberlik edebilir. Tahmin sonuçlarına göre, bölgesel doğal gaz boru hattı ağının koordineli çalışması sağlanır. Böylece doğal gaz sevkiyatı iyileştirilebilir ve optimize edilebilir. Bu durum doğal gaz şirketlerinin ekonomik faydalarını da önemli ölçüde etkileyecektir. Doğal gaz yükü değişikliklerini önceden anlamak, şehir sakinlerinin ve endüstriyel gaz talebinin garanti altına alınmasını sağlayabileceği gibi, gecikmeli sevkiyattan kaynaklanan ekstra maliyetlerin de önlenmesine olanak tanır.

Doğal gazın arz ve talebi arasında kritik bir denge vardır. Doğal gaza olan talebin fazla tahmin edilmesi kullanılmayan kaynaklar için enerji maliyetlerinin artmasına sebep olur. Doğal gaz talebinin gereğinden az tahmin edilmesi durumunda ise dağıtım şirketleri endüstriyel tesislere doğalgaz arzını keser. Bu kesinti, ülke ekonomisi için milyonlarca dolarlık zarara sebep olur ve nihayetinde ülkede ciddi krizlere sebep olabilecektir. Ayrıca dağıtım şirketleri her gün bir sonraki gün için kullanacakları doğal gaz miktarını BOTAŞ'a bildirmek zorundadır. Bildirilen tahmini tüketim ile fiili tüketim arasında oluşacak yüksek sapmalar ceza ödemelerine yol açmaktadır. Etkili bir gaz tedarik yönetimi; temel olarak, ülkelerin ve işletmelerin; makul gaz tedarik planları oluşturmasını, tedarik sözleşmelerini yönetmesini, operasyonel verimliliği artırmasını sağlayacaktır.

Doğal gaz altyapı inşaatları önemli miktarda sermaye yatırımı gerektirmektedir. Kesin doğal gaz talebinin yüksek doğruluk ile belirlenmesinin bir diğer önemli tarafı da yatırım planlarının geniş bir bakış açısı ile formüle edilmesine yardımcı olmasıdır. Aynı zamanda, bu tür tahminler şehir gaz iletim boru hattı ağının, sıvılaştırılmış doğalgaz (SD) ve sıkıştırılmış doğalgaz (SDG) alım istasyonlarının, gaz depolama tesislerinin planlanması ve inşaat yatırımı hesaplamalarının oluşturulması için de faydalıdır. Bu nedenle, doğal gaz talebinin yüksek doğruluk ile belirlenmesi; gaz boru hatlarının inşasının doğal gaz talebiyle eşleştirilmesine yardımcı olacak ve böylece inşaat maliyetlerinin azalmasına etki edecektir [9].

Hükümet açısından, doğru doğal gaz yükü tahmini ve eksiksiz bir doğal gaz tedarik sistemi, kamu kurumlarının plan ve politikaları formüle etmesi ve yayınlaması için ön koşullardır. Kentlerde yaşayan vatandaşların yaşamları, şehirlerdeki doğal gaz arz ve talep dengesinden doğrudan etkilenmektedir. Özellikle, doğru doğal gaz talep tahmini, hükümetin şehrin doğal gaz tüketim örüntüsünü anlamasına ve buna bağlı olarak doğal gaz depolama tesislerinin inşası, ek tedarikçilerin belirlenmesi gibi planların oluşturulmasına yardımcı olabilir [9].

Tüm bu nedenler göz önünde bulundurulduğunda doğal gaz tüketiminin gün öncesinde doğru tahmini, doğal gaz endüstrisi için önemli bir zorluktur ve doğal gaz talebini yüksek doğruluk seviyesi ile tahmin eden bir sistemin kurulması yöneticiler için de önem arz etmektedir. Bu çalışmanın amacı; Türkiye'nin farklı bölgelerinde doğalgaz dağıtım faaliyetleri gerçekleştiren bir firmanın günlük doğalgaz talebini yüksek doğruluk ile tahmin edecek bir karar destek aracı oluşturmaktır.

Çalışmanın geri kalan kısmı şöyle organize edilmiştir: İkinci bölümde, doğalgaz tüketim tahmin yöntemlerine ilişkin kapsamlı bir literatür çalışması sunulmuştur. Üçüncü bölümde, çalışmada kullanılacak veri ön işleme ve makine öğrenmesi tekniklerinin matematiksel temeli oluşturulmuştur. Dördüncü bölümde, çalışmada önerilen tahmin süreci kısaca açıklanmıştır. Veri toplama ve ön işleme adımları gerçekleştirilmiştir. Önceki bölümde bahsedilen makine öğrenmesi teknikleri veri setlerine uygulanmıştır. Oluşturulan modellerin performansları değerlendirilip sonuçlar bölge bazında sunulmuştur. Beşinci bölümde ise tez çalışması kapsamında elde edilen sonuçlar ve öneriler verilmiştir.





2. LİTERATÜR ARAŞTIRMASI

Doğal gaz talebini tahmin etme konusunda istatistiksel ve makine öğrenimi modelleri literatürde yaygın bir şekilde kullanılmıştır. Liu ve diğerleri [10], 1950-2019 yılları arasında yayımlanan toplamda 72 makaleyi başlangıç aşaması, geleneksel aşama, yapay zekâ aşaması ve çok yönlü aşama olacak şekilde 4 farklı dönemde sınıflayıp; tahmin ufku, tüketimi etkileyen faktörler, değerlendirme teknikleri ve tahmin performansı açısından değerlendirmiştir. Beyca, Ervural, Tatoğlu, Ozuyar ve Zaim [5], yaptıkları çalışmada çok değişkenli regresyon (ÇDR), destek vektör makinesi (DVM) ve yapay sinir ağları (YSA) modellerini İstanbul'un aylık doğalgaz tüketimini belirlemek için kullanmıştır. Kullanılan yöntemler içerisinde DVM (OMYH: %5,53) ile en iyi tahmin sonucu elde edilmiştir. Çalışmada ayrıca İstanbul'un gelecek 12 ay içinde tüketeceği doğalgaz miktarı tahmin edilmiştir. Özmen [11], Ankara'nın günlük doğal gaz talebini tahmin etmek için doğrusal regresyon (DR), en küçük mutlak küçülme ve seçme operatörü (LASSO) ve çok değişkenli uyarlamalı regresyon eğrileri (ÇDURE) gibi doğrusal modeller kullanmışlardır. Kullanılan yöntemler içerisinde ÇDURE en iyi tahmin performansını göstermiştir (OMYH: %4,8; R^2 : 0,994). Taşpınar, Çelebi ve Tutkun [12], Sakarya'nın günlük doğal gaz talebini belirlemek için meteorolojik verileri bağımsız değişken olarak kullanan bir YSA modeli geliştirmiştir. Özdemir, Aydemir, Olgun ve Mulbay [13], Türkiye'nin yıllık doğal gaz talebini tahmin etmek için DR'ye dayalı hibrit Genetik Algoritma (GA)-Benzetilmiş Tavlama (BT) algoritması uygulamıştır. Nüfus, gayri safi yurt içi hâsıla (GSYİH) ve büyüme oranı bağımsız değişken, yıllık doğal gaz tüketimi bağımlı değişken olarak kullanılmıştır. Burada GA-BT ile regresyon katsayıları belirlenmiştir ve hibrit algoritma (OMYH: %4,97) en iyi tahmin sonucunu elde etmiştir. Es [14], Türkiye'nin aylık doğal gaz tüketimini düzeltilmiş mevsimsel gri model ve SARIMA modellerini karşılaştırarak belirlemeye çalışmıştır. Karşılaştırma sonucunda düzeltilmiş mevsimsel gri model en üstün performans göstermiştir (OMYH: %8,67). Bilici, Özdemir ve Temurtaş [15], Türkiye'nin aylık doğalgaz talebini tahmin etmek için dört farklı metasezgisel yöntem ve üç farklı matematiksel model kullanmışlardır. Bağımsız değişken olarak beş farklı meteorolojik değişkenin kullanıldığı çalışmada parçacık sürü optimizasyonu (PSO) ile optimize edilen kuadratik model tüm değerlendirme ölçütlerinde en iyi performansı göstermiştir. Ceylan [16], COVID-19 döneminde Türkiye'nin haftalık doğalgaz tüketimini belirlemek için uzun kısa süreli bellek (UKSB), ARIMA, DVM ve doğrusal olmayan otoregresif sinir ağı (DOOSA) modellerini

kullanmıştır. UKSB modelinin üstünlüğü, Diebol-Mariona (DM) istatistik testiyle doğrulanmıştır. Arık [17], Türkiye'nin yıllık doğal gaz tüketimini belirlemek amacıyla yapay arı kolonisi (YAK) algoritmasını uygulamıştır. Çalışmada, nüfus, ithalat, ihracat ve GSYİH gibi bağımsız değişkenler kullanılmıştır. Elde edilen tahminler DR yöntemi ile karşılaştırılmıştır. Yapılan karşılaştırmada, YAK algoritmasının daha düşük OMYH değerleri sergilediği gözlemlenmiştir. Sen, Günay ve Tunç [18], nüfus, GSYİH, enflasyon, doğalgaz birim fiyatı, ortalama kış mevsim sıcaklığı gibi beş farklı bağımsız değişken kullanarak Türkiye'nin yıllık doğalgaz tüketimini belirlemeye çalışmışlardır. Bu değişkenler arasından GSYİH ve enflasyon tahmin için anlamlı olan tek göstergeler olarak bulunmuştur. Ardında çok değişkenli logaritmik, doğrusal ve doğrusal olmayan regresyon modelleri kullanılarak Türkiye'nin 2015-2025 yıllık doğalgaz tüketimi tahmin edilmiştir. Akpınar, Adak ve Yumuşak [19], Sakarya ilinin günlük doğalgaz tüketimini tahmin etmek için geçmiş yedi günlük doğalgaz tüketimini bağımsız değişken olarak kullanan bir YSA modeli önermiştir. Önerilen YSA modeli YAK algoritması ile eğitilmiştir. %14,9 OMYH ile günlük doğalgaz tüketim tahmini gerçekleştirilmiştir. Yücesan, Pekel, Çelik, Gül ve Serin [20], Türkiye'nin günlük doğalgaz tüketimini tahmin etmek için istatistiksel, metasezgisel ve derin öğrenme tekniklerini kullanan kapsamlı bir çerçeve kullanmışlardır. ARIMA-YSA modeli %0,35 OMYH ile en iyi performansı göstermiştir. Szoplik [21], Polonya'nın Szczecin şehrinin saatlik ve günlük doğal gaz tüketimini tahmin etmek için sıcaklık ve zamansal değişkenleri girdi olarak kullanan bir YSA modeli önermiştir. Rehman, Zhu, Segovia ve Anderson [22], İspanya'nın beş farklı bölgesinin saatlik doğal gaz talebini belirlemek istatistiksel ve derin öğrenme modellerini kullanan kapsamlı bir tahmin çerçevesi sunmuştur ve çoklu doğrusal regresyon modeli en iyi tahmin doğruluğunu elde etmiştir. Sharma ve diğerleri [8], gradyan artırma ağaçları (GAA) ve YSA modellerini kullanarak altışar saatlik doğal gaz tüketim miktarını tahmin etmiştir. Modellerden elde edilen sonuçlar birleştirildiğinde tahmin performansında anlamlı bir artış meydana gelmiştir. Hribar, Potocnik, Silc ve Papa [23], Slovenya'nın Ljubljana şehrinde konutların doğal gaz tüketimini belirlemek için ÇDR ve tekrarlayan sinir ağı (TSA) modellerini kullanmışlardır. Sıcaklık, geçmiş doğal gaz tüketimi ve tatil günlerini bağımsız değişken olarak kullanan bu çalışmada TSA en iyi performansı göstermiştir (OMYH: %6,8). UKSB ve evrişimli sinir ağı (ESA) gibi derin öğrenme teknikleri de doğalgaz tüketim öngörü çalışmalarında kendilerine uygulama alanı bulmuşlardır [24, 25]. Laib, Khadir ve Mihaylova [26], Cezayir'in saatlik doğal gaz tüketim verilerini k-ortalamalar kümeleme tekniğini uygulayarak beş farklı tüketim grubu oluşturmuştur. Tüketim gruplarını doğru tahmin etmek için 98,43'lük

doğruluğa sahip YSA sınıflama modeli geliştirilmiştir. Her bir tüketim grubu için oluşturulan UKSB modeli ortalama %4,73 OMYH değeri ile bir sonraki günün doğalgaz talebini belirleyebilmektedir. Doğal gaz tüketim verilerindeki gürültüler tahmin performansını düşürebilmektedir. Wei ve diğerleri [27], KP (Karmaşıklık Parametresi) ile ifade edilen ve verilerdeki karmaşıklığı belirleyen bir metrik tanımlamışlardır. KP değerindeki artış tahmin performansına olumsuz yansıdığı deneysel bir çalışma ile ortaya konmuştur. Tahmin performansını arttırmak için literatürde dalgacık dönüşümü (DD), ampirik mod ayrıştırması (AMD) ve geliştirilmiş tekil spektrum analizi (GTSA) gibi farklı veri ön işleme algoritmaları kullanılmıştır. Wei, Li, Peng ve Zeng [28], üç farklı iklim bölgesinde bulunan dört farklı şehrin günlük doğal gaz tüketim verilerindeki gürültüyü gidermek adına GTSA yöntemini önermiştir. GTSA ile temizlenmiş veri üzerinden UKSB ile yapılan tahmin sonucu ortalama mutlak aralık normalleştirme hatası (OMANH) [%4,68 - %14,1] elde edilmiştir. Panapakidis ve Dagoumas [29], Yunanistan'ın tüm doğalgaz dağıtım noktalarının günlük gaz tüketimini tahmin etmek için DD - Uyarlamalı Sinirsel Bulanık Çıkarım (USBC)/GA ve YSA birlikte kullanan kapsamlı bir metodoloji önermiştir. Dağıtım noktaları için elde edilen OMANH değerleri [0,0187-0,1212] arasındadır. Su ve diğerleri [30], saatlik doğalgaz tüketim verisini DD ile üç alt bileşene ayırmış ve her bir alt bileşenin gelecek değerlerini tahmin etmek için GA ile optimize edilmiş TSA modeli kullanmışlardır. Elde edilen tahmin sonuçları ters DD tekniği ile birleştirilerek saatlik doğalgaz tüketim tahminleri elde edilmiştir. Li, Zheng, Li ve Yang [9], ayrıştırma-füzyon tekniği, özellik seçimi ve istifleme topluluk öğrenme stratejisini kullanarak Çin'in 10 farklı şehrinin günlük doğalgaz tüketimini belirlemeye çalışmışlardır.

Çizelge 2.1. Türkiye'deki doğal gaz tüketim tahmini üzerine yapılan bazı çalışmalar

Yazar	Tahmin Modeli	Tahmin Ufku	Veri Boyutu	Bağımsız Değişkenler	Değerlendirme Kriteri
Taşpınar ve diğerleri (2013)	YSA	Günlük	2007-2011	Yağış, basınç, rüzgâr hızı, sıcaklık	R^2
Özdemir ve diğerleri (2016)	GA-BT ile optimize edilmiş DR	Yıllık	1985-2009	Nüfus, GSYİH,	OMYH
Akpınar ve diğerleri (2017)	YAK-YSA	Günlük	2011-2014	Geçmiş yedi günlük tüketim verisi	OHK OMYH R^2

Çizelge 2.2. (devam) Türkiye'deki doğal gaz tüketim tahmini üzerine yapılan bazı çalışmalar

Beyca ve diğerleri (2019)	DVM	Aylık	2004-2015	Sıcaklık, mevsimsel indeks, nüfus, doğalgaz birim fiyatı	OMYH OHK
Arık (2019)	YAK-DR	Yıllık	1998-2017	GSYİH, nüfus, ithalat, ihracat	OMYH
Sen ve diğerleri (2019)	Doğrusal olmayan çok değişkenli regresyon	Yıllık	1986-2014	GSYİH, enflasyon	R ² OKHK OMH
Özmen (2020)	DR, LASSO, ÇDURE	Günlük	2004-2013	Meteorolojik değişkenler, gecikmeli tüketim, zamansal değişkenler	OMYH MMYH R ² OMH OKHK
Es (2021)	Mevsimsel Gri Model	Aylık	2000-2018	Geçmiş Tüketim Verisi	OMH OKHK OMYH
Yücesan ve diğerleri (2021)	SARIMAX-YSA	Günlük	2017-2019	Zamansal değişkenler, dolar kuru	R ² OKHK OMYH OHK
Bilici ve diğerleri (2023)	Hibrit PSO-Kuadratik Model	Aylık	2010-2020	Sıcaklık, basınç, nem, rüzgâr, yağış	OMH R ² OMYH OKHK
Ceylan (2023)	UKSB	Haftalık	2020-2022	Geçmiş tüketim verisi	OKHK

Bu çalışma literatürdeki doğalgaz tüketim tahmin çalışmalarına üç temel katkıda bulunmaktadır. İlk olarak topluluk öğrenme stratejilerini, derin öğrenme ve doğrusal regresyon modellerini kullanan ve bunların sonuçlarını karşılaştıran kapsamlı bir tahmin metodolojisi sunulmuştur. İkinci olarak mevcut çalışmaların büyük çoğunluğu yalnızca belirli bir bölgedeki doğalgaz tüketim tahminine odaklanırken bu çalışmada dört farklı bölgenin günlük doğalgaz tüketim tahmini gerçekleştirilmiş ve farklı bölgelerdeki tahmin değişkenlikleri karşılaştırılmıştır. Son olarak önceki çalışmalarda doğalgaz tüketimini yüksek doğruluk ile belirleyebilmek için meteorolojik ve ekonomik değişkenler kullanılırken bu çalışmada tüketim verisi kullanılarak oluşturulan gecikmeli tüketim

değişkenleri, pencere değişkenleri ve zamansal değişkenler dikkate alınarak tahmin işlemi gerçekleştirilmiştir.





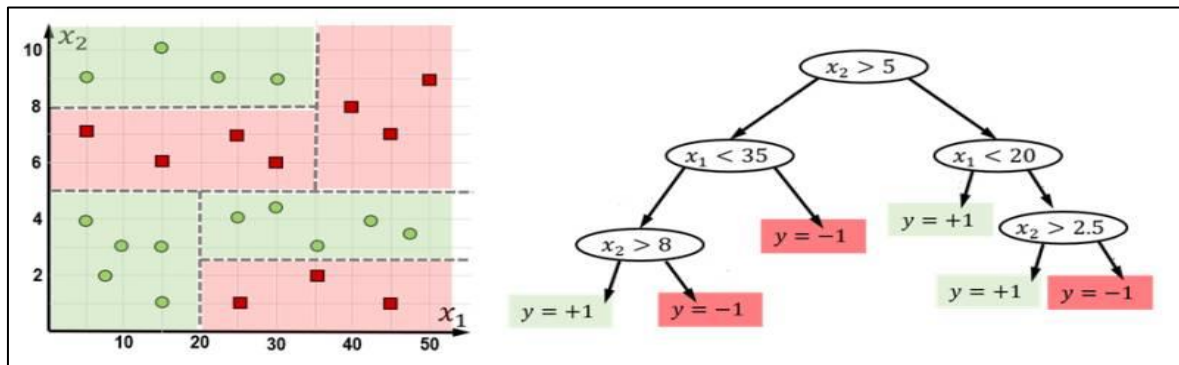
3. MATERYAL METOD

3.1. Kullanılacak Makine Öğrenmesi Algoritmalar

Literatür incelendiğinde doğalgaz talep tahmin problemi UKSB, YSA, istatistiksel modeller (ARIMA-SARIMA, ÇDR) ve metasezgisel yöntemler gibi çok farklı makine öğrenmesi teknikleri ile ele alınmıştır. Ancak, XGBoost, LightGBM ve RO gibi ağaç tabanlı algoritmaların bu alandaki uygulamaları sınırlıdır. Bu gelişmiş ağaç tabanlı algoritmaların performansının istatistiksel ve derin öğrenme modelleriyle karşılaştırılması, bu çalışmanın amaçlarından biridir. Bu algoritmaların seçilmesindeki bir diğer neden ise modelin performansını artıracak ve genelleştirebilme yeteneği yüksek bir model oluşturulmasını sağlayacak çok sayıda etkili hiperparametreler içermektedir. Bu hiperparametrelerinin dikkatlice ayarlanmasıyla, modelin performansını optimize etmek ve genelleme yeteneğini artırmak amaçlanmaktadır. Bu çalışmada UKSB, ÇDR, RO, ARA, XGBoost ve LightGBM modelleri kullanılmak üzere belirlenen makine öğrenmesi algoritmalarıdır.

3.2. Karar Ağaçları (KA)

En popüler makine öğrenimi modellerinden biri olan karar ağaçları (KA) sınıflandırma ve regresyon görevleri için sıklıkla kullanılır. Bir KA, karar düğümleri ve yaprak düğümlerden oluşur. Her karar düğümü mevcut örneği belirli bir durum için test eder (örneğin, yaş > 50 mi?) ve cevabı temel alarak sağ yola veya sol yola yönlendirir. Sınıflandırma için yaprak değeri bir sınıf etiketidir, regresyon için ise yaprak, bir reel sayı değeri döndürür. Şekil 3.1’de örnek bir KA gösterilmiştir.



Şekil 3.1. Karar ağacı [31]

Derinliği 1 olan karar ağacına karar kütüğü denir ve en basit ağaçtır. Bir karar kütüğü tek bir karar düğümü ve iki yaprak düğümünden oluşur. Sığ bir karar ağacı (örneğin derinlik 2 veya 3) az sayıda karar düğümüne ve yaprak düğümüne sahip olan basit bir modeldir. Sonuç olarak, yalnızca basit işlevleri temsil edebilir. Öte yandan, daha derin bir karar ağacı çok daha fazla karar düğümüne ve yaprak düğümüne sahip olacaktır. Böylece daha karmaşık fonksiyonları temsil edebilecektir [31].

Karar ağacı algoritması, dallanma noktasında hangi bağımsız değişkenin seçileceğine karar verirken kazanç (gain) değerini kullanır. Kazanç değeri Eş. 3.1 ile hesaplanır. Kazanç değeri yüksek olan bağımsız değişken, bağımlı değişken üzerinde en fazla etkiye sahip olan değişken olarak seçilir ve dallanma için tercih edilir. Eğer bir düğümde bağımlı değişken değerleri tek bir kategoriye toplanmışsa, bu düğüme saf düğüm denir ve bu düğümde dallanma işlemi durdurulur [32].

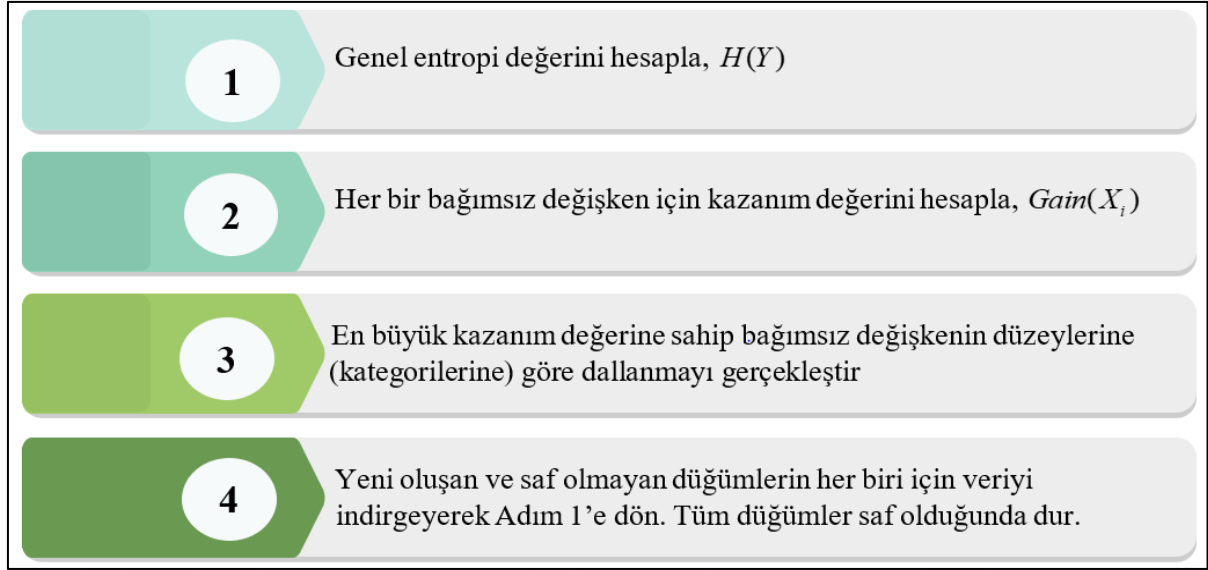
$$Gain(X_i) = H(Y) - \sum_{j=1}^{k_i} P(X_{ij})H(Y/X_{ij}) \quad i = 1, 2, \dots, m \quad (3.1)$$

Entropi bir değişkendeki belirsizliği ölçmek için kullanılır. Değişkenin bütün değerleri birbirine eşit olduğunda belirsizlik yoktur ve entropi değeri 0 olur ancak değişken tamamen birbirinden farklı değerler alıyor ise entropi maksimum değerine ulaşır.

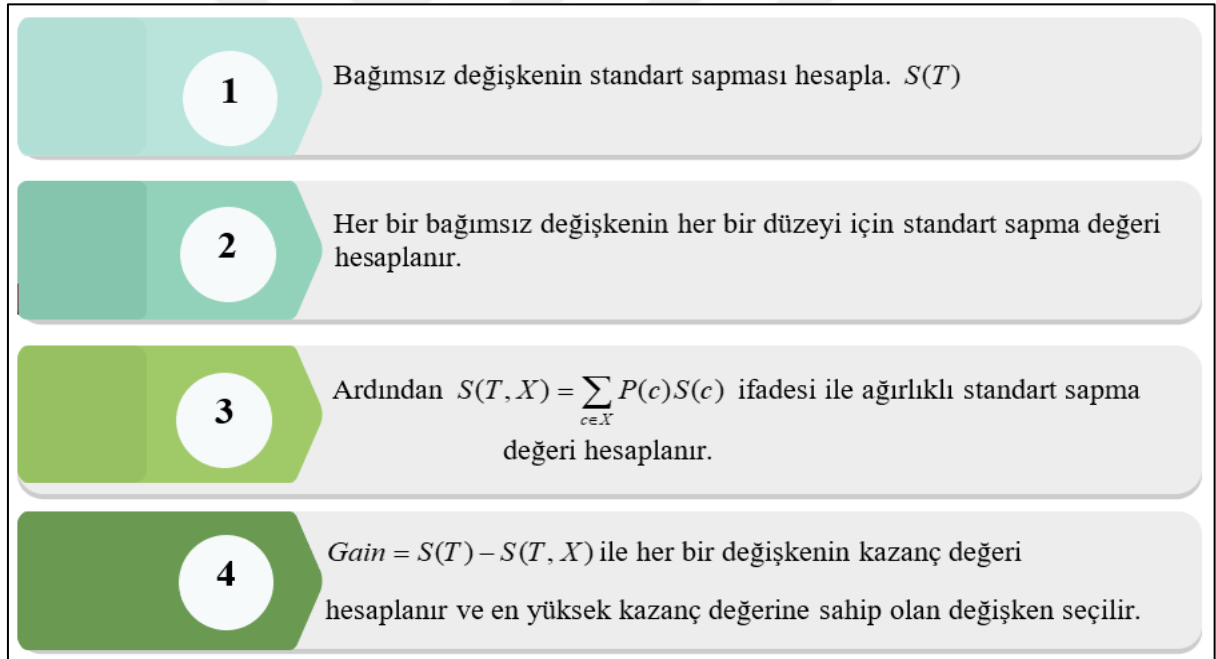
Y değişkeninin k tane düzeyi olsun bu durumda entropi Eş. 3.2 ile gösterilen eşitlik ile hesaplanır. Burada p_j notasyonu j. düzeyin ortaya çıkma olasılığını göstermektedir.

$$H(Y) = H(p_1, p_2, \dots, p_k) = - \sum_{j=1}^k (p_j \log_2(\frac{1}{p_j})) \quad (3.2)$$

Karar ağacı algoritması regresyon problemlerine uygulanır iken entropi yerine standart sapma azalması göz önünde bulundurulur. Şekil 3.2’de sınıflama, Şekil 3.3’te ise regresyon problemleri için karar ağaçlarının algoritmik akışı açıklanmıştır.



Şekil 3.2. Sınıflandırma ağaçlarının algoritması

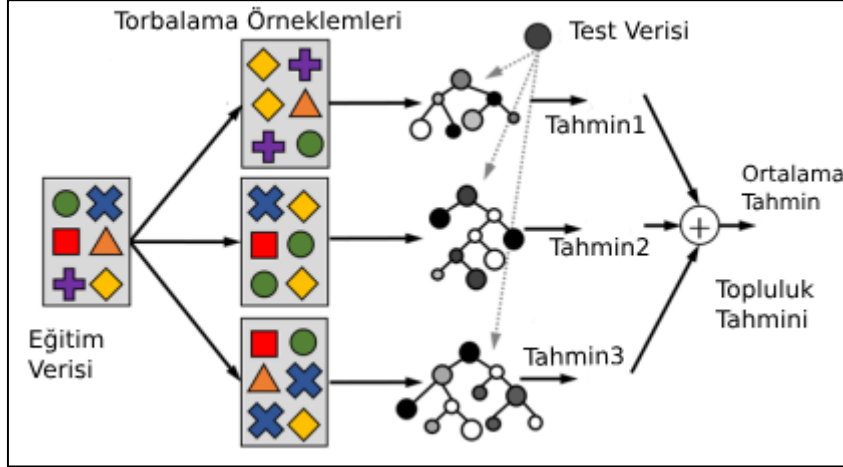


Şekil 3.3. Regresyon ağaçlarının algoritması

3.3. Torbalama

Torbalama, en temel homojen paralel topluluk öğrenme yöntemidir. Torbalama, tek bir veri kümesinden benzer ancak tam olarak aynı olmayan alt kümeler oluşturmak için torbalama örneklemesini kullanır. Bu alt kümeler oluşturulur iken eğitim verisindeki her bir verinin eşit olasılıkla seçilme ve birden fazla kez seçilme şansı vardır. Oluşturulan bu veri setleri

üzerinde modeller eğitilir ve sonuçta benzer ancak tam olarak aynı olmayan temel tahmin edici modeller oluşturulmuş olur. Bu süreç Şekil 3.4'te gösterilmiştir. Tahmin sırasında, bireysel temel modellerin tahminlerini tek bir topluluk tahmininde birleştirmek için çeşitli birleştirme operatörleri kullanılır. Genellikle, sınıflandırma için nihai topluluk tahmini çoğunluk oyu, regresyon için ise aritmetik ortalamadır.



Şekil 3.4. Torbalama süreci [31]

3.4. Güçlendirme

Bir sıralı topluluk yöntemi olan güçlendirme, birkaç zayıf öğrenciyi tek bir güçlü öğrenciye dönüştürmeyi amaçlar. Başka bir ifade ile, güçlendirme, kelimenin tam anlamıyla bir grup zayıf öğrencinin performansını artırmayı amaçlamaktadır. Güçlendirmenin uygulaması oldukça basittir, hesaplama açısından verimlidir ve çok çeşitli temel öğrenme algoritmalarıyla birlikte kullanılabilir. Regresyon problemleri için geliştirilmiş güçlendirme algoritması aşağıda gösterilmiştir [33].

İlk olarak eğitim verisinde bulunan tüm veri noktalarına Eş. 3.3 kullanılarak ağırlıklar atanır. Burada w_i eğitim verisindeki her bir noktanın ağırlığını göstermektedir. N_i ise veri sayısını ifade eder. $i = 1, \dots, N_i$.

$$w_i = \frac{1}{N_i} \quad (3.3)$$

İkinci aşamada eğitim verisi kullanılarak bir regresyon modeli (zayıf öğrenici) oluşturulur. Oluşturulan modelin eğitim verisindeki tahminleri $y_i^{(p)}$ ile gösterilir.

Her bir eğitim veri noktası için mutlak hata değerleri Eş. 3.4 ile hesaplanır. Eş. 3.5 ve Eş. 3.6 kullanılarak mutlak hatalar 0-1 arasında normalize edilir.

$$L_i = L[|y_i^{(p)} - y_i|] \quad (3.4)$$

$$D = \text{enbüyük} |y_i^{(p)} - y_i| \quad (3.5)$$

$$L_i = \frac{|y_i^{(p)} - y_i|}{D} \quad (3.6)$$

Oluşturulan tahmin modeli için Eş. 3.7 ile ortalama hata, Eş. 3.8 ile oransal ortalama hata değerleri hesaplanır.

$$\bar{L} = \sum_{i=1}^{N_i} L_i \cdot w_i \quad (3.7)$$

$$\beta = \frac{\bar{L}}{1 - \bar{L}} \quad (3.8)$$

Veri noktalarının mutlak hatalarına göre ağırlık güncellemesi Eş. 3.9 ile yapılır.

$$w_{\text{yeni}} = w_i \cdot \beta^{(1 - L_i)} \quad (3.9)$$

Ardından güncellenen ağırlıklar göz önünde bulundurularak yerine koyma tekniği ile N_i adet veri seçilir. Seçilen veriler ile yeniden regresyon modeli kurulur. Tüm bu süreç istenilen hata oranına ulaşılan kadar tekrarlanır.

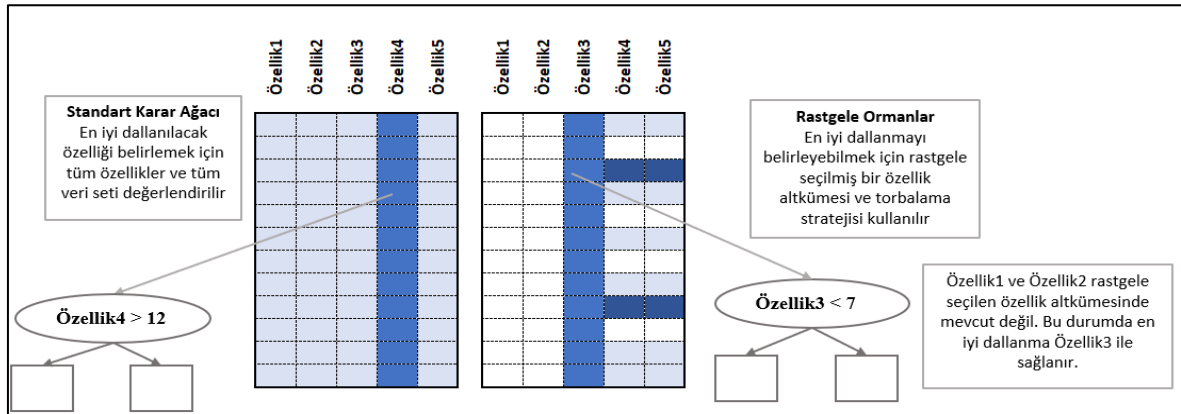
Eş. 3.10'de her bir regresyon modelinin normalize ağırlık değeri gösterilmiştir. Eş. 3.11 ile güçlendirme topluluğunun nihai tahmin değeri hesaplanır. Burada t regresyon algoritmasını göstermektedir.

$$w_t = \frac{\log\left(\frac{1}{\beta_t}\right)}{\sum_t \log\left(\frac{1}{\beta_t}\right)} \quad (3.10)$$

$$h_f = \sum_t \hat{y}_t \cdot w_t \quad (3.11)$$

3.5. Rastgele Ormanlar (RO) ve Aşırı Rastgeleleştirilmiş Ağaçlar (ARA)

Rastgele ormanlar (RO) torbalama tekniği kullanılarak oluşturulan rastgele karar ağaçlarından oluşan bir topluluğu ifade eder. Rastgele karar ağaçları, rassallığı sağlamak için değiştirilmiş bir karar ağacı öğrenme algoritması kullanılarak eğitilir. Bu ek rassallık oluşturma süreci ile topluluğun çeşitliliği artırılır ve genellikle daha iyi tahmin performansı gözlemlenir. Standart bir karar ağacı ile RO arasındaki temel fark, karar düğümlerinin nasıl oluşturulduğudur. Standart karar ağacında ilk dallanmanın belirleneceği özellik seçilirken, veri setinde bulunan mevcut tüm özellikler kapsamlı bir şekilde değerlendirilir. Burada karar ağacı öğrenimi açgözlü bir algoritma olduğundan, dallanmanın gerçekleşeceği özellik en yüksek kazanç değerini sağlayan özelliktir. Bu durum baskın özelliğin farklı ağaçlarda tekrar tekrar kullanılmasına ve topluluğun çeşitliliğinin daha az olmasına sebep olur. RO ise en iyi bölünmeyi belirlemek için tüm özellikleri kullanmak yerine, rastgele bir özellik alt kümesi oluşturur ve bu özellik altkümesini değerlendirerek en iyi dallanma özelliğini belirler. RO bir karar düğümü oluşturmadan önce ilk olarak özellikleri rastgele örnekleyen değiştirilmiş bir ağaç öğrenme algoritması kullanır. Ortaya çıkan ağaç ise rastgele bir karar ağacıdır [31]. Şekil 3.5'te KA ve RO algoritmasının çalışma prensibi karşılaştırmalı olarak gösterilmiştir.



Şekil 3.5. Karar ağacı ve rastgele ormanlar

Aşırı rassallaştırılmış ağaçlar (ARA), rastgele bir özellik alt kümesinden yalnızca bölme değişkenini değil, aynı zamanda bölme eşiğini de rastgele seçer. Tipik karar ağacı öğreniminde, seçilen değişkenin bölüneceği en iyi eşiği (τ) ararız. Bunun yerine, aşırı rassallaştırılmış ağaçları eğitirken, eşik τ da rastgele seçilir. Uygulamada, ARA toplulukları, çok sayıda sürekli özelliğe sahip yüksek boyutlu veri kümeleri için oldukça uygundur [31].

3.6. Çok Değişkenli Regresyon (ÇDR)

Çok değişkenli regresyon (ÇDR) modeli geleneksel olarak yaygın bir şekilde kullanılan bir tahmin tekniğidir. Burada birincil amacımız bağımlı değişkeni tahmin etmek iken, ikincil amacımız ise her bağımsız değişkenin marjinal etkisini tahmin etmektir. ÇDR modelinde bağımlı değişkendeki değişkenliği açıklayacak bir model kurarız. P adet bağımsız değişkenden ve bir bağımlı değişkenden oluşan ÇDR modeli Eş. 3.12’de gösterilmiştir [34].

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i \quad (3.12)$$

Burada ε_i , ortalaması 0 ve varyansı σ^2 olan rassal hata terimini göstermektedir. β_j terimleri $\beta^T = [\beta_0 \ \beta_1 \ \dots \ \beta_p]$ $j = 1, \dots, p$ için bağımsız değişkenlerin katsayılarını göstermektedir. i gözlem verilerini gösterir $i = 1, \dots, n$. x_{ij} gibi küçük harfler, X_j değişkeninin i gözlemindeki belli değerleri göstermektedir.

ÇDR Eş. 3.13’te gösterilen kalıntı kareler toplamının en küçük değerine karşılık gelen katsayıları hesaplar.

$$\min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad (3.13)$$

Eş. 3.14 gözlemlere ait girdi matrisini ve Eş. 3.15 ise bağımlı değişkeni gösterir.

$$X = \begin{bmatrix} 1 & x_1^T \\ 1 & x_1^T \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n^T \end{bmatrix} \quad (3.14)$$

$$y^T = [y_1 \quad y_2 \quad \cdot \quad \cdot \quad y_n] \quad (3.15)$$

Eş. 3.13 ile gösterilen fonksiyon dışbükey bir fonksiyon olduğu için tek bir minimum noktası vardır. Bu minimum noktaya karşılık gelen bağımsız değişken katsayıları Eş. 3.16 ile elde edilir.

$$\beta = (X^T X)^{-1} X^T y \quad (3.16)$$

3.7. LASSO Regresyon

LASSO regresyonu, doğrusal regresyon türlerinden biri olup, en küçük kareler maliyet fonksiyonuna bir ceza terimi ekleyerek modelde bulunan ve tahminde sınırlı katkı sağlayan bağımsız değişkenlerin katsayılarını sıfıra indirme yeteneğine sahiptir. Dolayısıyla bu özellikler modelden çıkarılabilir [35]. Buna karşılık ridge regresyonu, model karmaşıklığını kontrol etmek için tüm bağımsız değişkenlerin katsayılarını birlikte küçültür, bu da onu ilişkili ve kovaryant özelliklerin ele alınması için çok uygun hale getirir.

Regresyon katsayıları için LASSO tahmincisi Eş. 3.17’de gösterilmiştir:

$$\beta = \min \left\{ \sum_{t=1}^T (y_t - \beta_1 * x_{1,t} - \beta_2 * x_{2,t} - \dots - \beta_p * x_{p,t})^2 + \lambda * \sum_{j=1}^p |\beta_j| \right\} \quad (3.17)$$

Düzenleme parametresi λ , LASSO tahmincisinin regresyon katsayılarını sıfıra doğru küçültme derecesini kontrol eder. λ değeri büyüdükçe katsayılar sıfıra doğru yaklaşır.

3.8. Aşırı Gradyan Artırma (XGBoost)

Gradyan artırma yöntemi olarak bilinen XGBoost birçok zayıf öğreniciyi (rastgele tahminden daha iyi tahmin üreten regresyon ağacını) birleştirerek güçlü bir öğrenici oluşturmayı amaçlayan bir makine öğrenme algoritmasıdır. XGBoost algoritması, bir kayıp fonksiyonunu minimuma indirmeyi hedefleyen bir karar ağacı topluluk modelidir [36]. Bu algoritmanın temel fikri, ilk ağacın yaptığı hataları düzeltmek için sürekli yeni ağaçlar oluşturmak ve nihai tahmin sonucunu belirlerken her bir ağacın ürettiği tahmini, öğrenme oranı ile çarpıp toplamını almaktır.

Eş. 3.18’de t ağacının için XGBoost algoritmasının minimize edilmesi hedeflenen kayıp fonksiyonu gösterilmiştir:

$$L^{(t)} = \sum_{i=1}^n \left(\frac{1}{2} \left[y_i - \left(\hat{y}_i^{(t-1)} + w_{x_i}^{(t)} \right) \right]^2 \right) + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^{(t)2} \quad (3.18)$$

Bu eşitlikte i veri noktasını, j ise yaprak düğüm numarasını, T ağaçta bulunan toplam yaprak düğüm sayısını göstermektedir. Bu amaç fonksiyonu Eş. 3.19’da gösterilen ortalama kare hata ve Eş. 3.20’de gösterilen düzenlileştirme terimi olmak üzere iki kısımdan oluşmaktadır.

$$\sum_{i=1}^n \left(\frac{1}{2} \left[y_i - \left(\hat{y}_i^{(t-1)} + w_{x_i}^{(t)} \right) \right]^2 \right) \quad (3.19)$$

$$\gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^{(t)2} \quad (3.20)$$

Eş. 3.21’de önceki ağaç topluluğunun tahmini gösterilmiştir. Burada η öğrenme oranını, $w_{x_i}^{(t)}$ t ağacının i. veri noktası için hesapladığı tahmini değeri göstermektedir. Öğrenme oranı ağacın çıktı değerini düşürerek öğrenme sürecini dengelemektir.

$$\hat{y}_i^{(t-1)} = \hat{y}_i^{(0)} + \eta w_{x_i}^{(1)} + \eta w_{x_i}^{(2)} + \eta w_{x_i}^{(3)} + \dots + \eta w_{x_i}^{(t-1)} \quad (3.21)$$

Eş. 3.20 ile yaprak düğüm sayısına ve t . ağacın üreteceği çıktı miktarına bir ceza uygulanır. Yaprak düğüm sayısı ve çıktı miktarının artması modelin karmaşıklığını arttıracaktır. Bu durumda daha fazla ceza uygulanarak modelin karmaşıklığı kontrol altında tutulur. Böylece aşırı uyumdan kaçınılarak genellenebilme yeteneği yüksek ağaçlar oluşturulur.

Eş. 3.22’de benzerlik değerinin hesaplanmasında kullanılan eşitlik gösterilmiştir. Burada r_i i . düğüm noktasında bulunan hata değerlerini, $|hata|$ hata sayısını adet olarak göstermektedir. λ ise düzenlileştirme sabiti olarak kullanılır.

$$\text{Benzerlik Puanı} = \frac{\left(\sum_{i \in k \ddot{o} k} r_i^{(1)} \right)^2}{|hata|_{k \ddot{o} k} + \lambda} \quad (3.22)$$

XGBoost en iyi bölünme noktalarını bulmak için benzerlik puanlarından hesaplanan kazanç değerini kullanır. Kazanç değeri hesaplanması Eş. 3.23 ile gerçekleştirilir. Kazanç değeri, düğüm bölünmesinden önceki ve sonraki benzerlik puanları arasındaki fark olarak hesaplanır. Burada γ bir budama sabitidir. Bu değer, kazanç değerinin belirli bir değerin altında olması durumunda oluşabilecek ağacı budar. Böylece modelin aşırı öğrenmesini engeller.

$$\text{Kazanç} (L_{\text{ayırma}}) = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (3.23)$$

XGBoost algoritması, büyük veri setlerindeki başarısının yanı sıra eğitim hızını artırmak ve aşırı uyum sorununu azaltmak için çeşitli rassallaştırma yöntemleri kullanır. XGBoost tüm bu özellikleri nedeniyle birçok uygulama için güçlü bir araçtır.

3.9. Hafif Gradyan Artırma Makinesi (LightGBM)

LightGBM, 2017 yılında Microsoft tarafından geliştirilen histogram tabanlı bir gradyan artırma yaklaşımıdır. Büyük miktarda veriyi XGBoost’tan daha verimli bir şekilde işlemek için önerilmiştir. LightGBM özünde her sürekli değişkeni histogram kullanarak ayrık hale

getirir. En iyi bölünmeyi hesaplamak için ise tüm veri noktalarını değerlendirmek yerine ayrıık hale getirilen veri üzerinden yinelemeler yapar Büyük miktarda veriyi işlemek için gradyan tabanlı tek taraflı örnekleme (GTTÖ) ve özel özellik paketleme (ÖÖP) olmak üzere iki farklı yenilik getirmiştir [37].

Çoğu ağaç tabanlı topluluk öğrenme tekniği eğitim verimliliğini arttırmak için verileri rastgele örnekler. LightGBM birinci dereceden gradyan olan hataların, mutlak değerini dikkate alarak örnekleme yapar. Bu yöntemin temel fikri küçük hata değerine sahip veri noktaları, tahminlerin zaten güvenilir olduğu için bir sonraki iterasyonda eğitimden çıkarılabilir. Bunun aksine, büyük hata değerlerine sahip veri noktaları bir sonraki iterasyonda eğitim için seçilmelidir [31]. Algoritma adımları aşağıda gösterilmiştir:

- Her bir veri noktası için mutlak hatalar hesaplanır ve veri noktaları mutlak hatalarına göre azalan sırada sıralanır.
- Üst oran (a) diğer oran (b) olmak üzere iki parametre tanımlanır. Bu değerler sıfır ile bir arasındadır. En büyük gradyanlara sahip örneklerin en üst a %'sı seçilir; bu alt küme 'en üst' olarak adlandırılır.
- Kalan örneklerin rastgele b % si örneklenir; bu alt küme 'rast küme' olarak adlandırılır.
- Örneklerin her iki alt kümesine de ağırlıklar atanır. Bu ağırlıklar Eş. 3.24 ve Eş. 3.25'te gösterilmiştir.

$$\text{üst ağırlık} = 1 \quad (3.24)$$

$$\text{rast ağırlık} = \frac{1 - a}{b} \quad (3.25)$$

- Son olarak örneklenen verileri, hataları ve ağırlıkları kullanan yeni bir model oluşturulur.

Bir veri setinde çok sayıda veri noktasının olmasının yanı sıra çok sayıda bağımsız değişkeninde olması da genel eğitim sürecini yavaşlatabilmektedir. Eğitim örneklerini küçültme gibi, özellikleri de küçültme imkanına sahipsek, eğitim süresinde iyileştirmeler meydana gelebilir. LightGBM modelinde bu durum için ÖÖP sürecini sunulmuştur. Özellik değerleri aynı anda sıfır olmayan durum 'çatışma' olarak adlandırılır. Çatışmaları veya her iki özelliğin de sıfır olmadığı zamanları ölçülür, böylece bir araya getirilebilecek özellikleri

tanımlanır. Bu durum, graf boyama problemi olarak ele alınabilir. Buradaki sezgi, eğer iki özellik sıklıkla aynı anda sıfırda, düşük çatışma gösterir ve bir araya getirilebilirler.

3.10. Uzun Kısa Süreli Bellek (UKSB)

UKSB çeşitli tahmin görevlerini çözmek için özelleştirilmiş bir derin öğrenme ağıdır. UKSB bir dizinin her ögesi için aynı işlemleri gerçekleştirir ve çıktısı bir önceki hesaplamalara bağlıdır. Bu yaklaşım zaman serisi problemlerinin çözümünde büyük fayda sağlar. UKSB, kapı denetçileri sayesinde, gradyanın yok olması probleminin üstesinden gelerek yüksek doğruluğa sahip sonuçlar üretebilmektedir [28]. UKSB hücresinin yapısı Şekil 3.6'da gösterilmiştir. Aşağıdaki denklemler tek bir UKSB hücresinin hesaplama sürecini göstermektedir. Burada x_t t anında modele verilen girdi vektörünü gösterir. C_t ve h_t uzun ve kısa dönemli hafıza değerleridir. UKSB ağında unutmaya kararlarını verme ve yeni girdi vektörleri oluşturmak için bazı kapılar bulunmaktadır. Bu kapılar f_t unutmaya kapısı, i_t girdi kapısı o_t ise çıktı kapısıdır. Bir UKSB hücresinde 4 adet sinir ağı katmanı bulunur. Bu katmanlardan üç tanesinde sigmoid aktivasyon bulunurken bir tanesinde ise tanjant hiperbolik aktivasyon fonksiyonu vardır. Sigmoid aktivasyon fonksiyonu 0 ile 1 arasında değer üretirken tanjant hiperbolik fonksiyonu -1 ile 1 arasında değer üretir [38] .

Eş. 3.26'da girdi vektörü ve bir önceki gizli katmandan gelen bilgiye bakarak hangi bilginin unutulup hangi bilginin hatırlanacağına karar verir. Sigmoid fonksiyonu ile 0 ile 1 arasında çıktılar üretilir. 0 tamamen unut anlamına gelirken 1 tamamen hatırla anlamına gelir.

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (3.26)$$

Eş. 3.27 ile yeni potansiyel uzun dönemli hafıza oluşturulur. Eş. 3.28 ile oluşturulan bu potansiyel hafızanın ne kadarının hatırlanacağı belirlenir.

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (3.27)$$

$$\tilde{C}_t = \tanh\left(W_c * [h_{t-1}, x_t] + b_c\right) \quad (3.28)$$

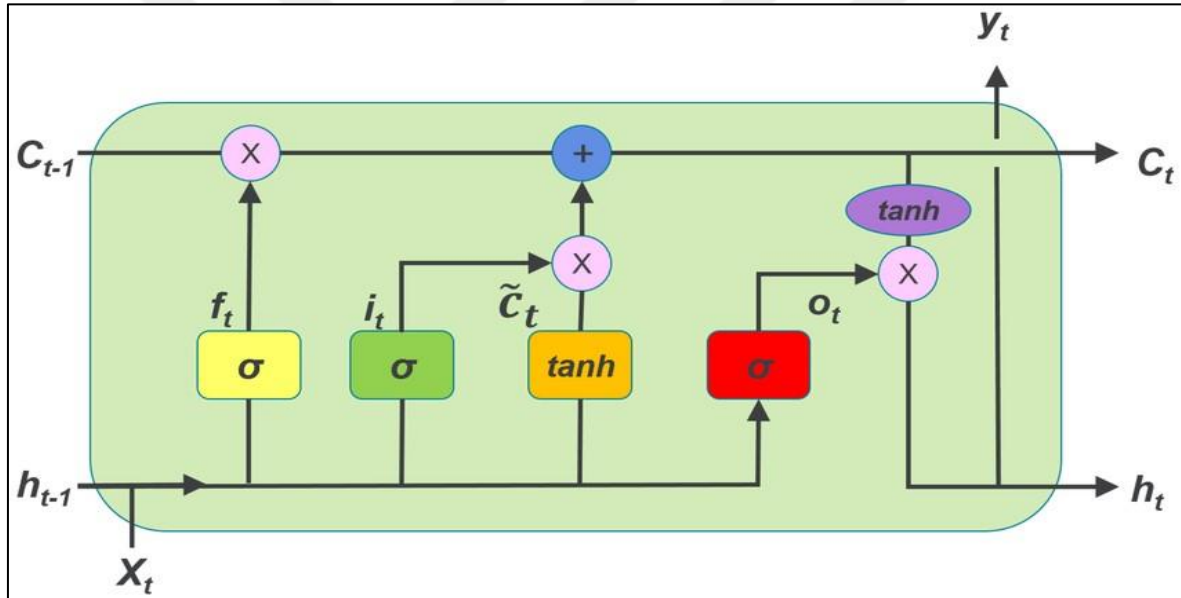
Eş. 3.29 ile uzun dönemli hafıza güncellenir.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (3.29)$$

Eş. 3.30 ile LSTM hücresinin çıktı vereceği değerlere karar verilir. Ardında Eş. 3.31 ile -1 ile 1 arasında değerler üretilir.

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \quad (3.30)$$

$$h_t = o_t * \tanh(C_t) \quad (3.31)$$



Şekil 3.6. UKSB hücresinin yapısı [39]

3.11. Değerlendirme Ölçütleri

Oluşturulan makine öğrenmesi modellerini değerlendirmek ve birbirleriyle kıyaslamak için R^2 , OMYH, SOMYH ve OMÖH olmak üzere dört farklı gösterge tanıtılmıştır [28]. Bu göstergeler Çizelge 3.1’de gösterilmiştir. Burada N test verisinin uzunluğunu, M eğitim verisinin uzunluğunu, S mevsimsel döngü uzunluğunu y_i gerçek veriyi, $\hat{y}_i$ ise tahmin edilen değeri göstermektedir.

Tüm bu hata ölçümleri her zaman serisi için tanımlandığından, model performansını ölçmek için bunların ortalama ve medyan değerleri test veri seti üzerinden hesaplanmıştır. Bu nedenle, her modeli değerlendirmek için toplamda altı hata ölçümü kullanılmıştır. Bunlar: Ortalama SOMYH, medyan SOMYH, ortalama OMÖH ve medyan OMÖH, R^2 ve OMYH.

Çizelge 3.1. Model değerlendirme ölçütleri

Ölçütler	Kısaltma	Formül
R-Kare	R^2	$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (\bar{y}_i - \hat{y}_i)^2}$
Ortalama Mutlak Yüzde Hata	OMYH	$OMYH = \frac{1}{N} \sum_{i=1}^N \left \frac{y_i - \hat{y}_i}{y_i} \right $
Ortalama Mutlak Ölçekli Hata	OMÖH	$OMÖH = \frac{\sum_{i=1}^N \hat{y}_i - y_i }{\frac{N}{M-S} \sum_{i=S+1}^M y_i - y_{i-S} }$
Simetrik Ortalama Mutlak Yüzde Hata	SOMYH	$SOMYH = \frac{1}{N} \sum_{i=1}^N \left(\frac{200(\hat{y}_i - y_i)}{ y_i + \hat{y}_i } \right)$

3.12. İstatistiksel Test

Makine öğrenmesi algoritmalarının ürettiği tahminler arasındaki farklılığın değerlendirilmesi için test verileri üzerinde parametrik veya parametrik olmayan istatistiksel testler kullanılabilir. Test kümesinin normal dağılıma uygunluğu bilinmediğinden, zaman serisi verileri için parametrik olmayan Friedman sıralama toplam testi sıklıkla kullanılmaktadır. Bu çalışmada Friedman testi her bir makine öğrenimi algoritmasına karşılık gelen SOMYH hatalarına dayanarak farklı makine öğrenimi tekniklerinin istatistiksel önemini değerlendirmek için kullanılır [40-43]. Eş. 3.32 ile test istatistiğinin değeri hesaplanır [44]. Çalışmada α anlamlılık düzeyi 0,05 alınmıştır.

Çizelge 3.2. Fridmen testi için veri tablosu

BLOK	Algoritma(1)	Algoritma(2)		Algoritma(k)
1	X_{11}	X_{12}		X_{1k}
2	X_{21}	X_{22}		X_{2k}
.	.	.		
.	.	.		
h	X_{h1}	X_{h2}		X_{hk}
Toplam	R_1	R_2		R_k

$$T = \frac{12}{hk(k+1)} \sum_{j=1}^k R_j^2 - 3h(k+1) \quad (3.32)$$

Sıfır hipotezine (H_0) göre makine öğrenimi algoritmaları arasında fark yoktur. Sıfır hipotezinin aksi durumlar H_1 hipotezi olarak düşünülür. Friedman testi, bir α anlamlılık düzeyi seçimi göz önüne alındığında, test tarafından hesaplanan χ^2 test istatistiğinde karşılık gelen değer aşağıdaki gibi yorumlanabilir:

$\chi^2 < T$ ise H_0 hipotezi reddedilir. Yani en az bir model diğerlerinden dikkate alınabilir bir farklılık göstermektedir.

$\chi^2 > T$ ise H_0 hipotezi reddedilemez. Yani modeller arasında dikkate alınabilir bir farklılık yoktur.



4. UYGULAMA

Bu tez çalışmasında bir doğal gaz dağıtım şirketinin Türkiye'nin beş farklı şehrindeki merkezleri için günlük doğalgaz tüketim tahmini gerçekleştirilmiştir. Çalışmanın bu bölümünde ise geliştirilen doğal gaz talep tahmin modeli detaylı olarak anlatılmaktadır. Tez çalışmasının uygulama kısmı beş bölümden oluşmaktadır:

- BDÖİ bazında günlük tüketim verileri bir araya getirilmesi,
- Bağımsız değişken oluşturma ve özellik mühendisliği,
- Veri setinin eğitim ve test olarak ayrılması,
- Makine öğrenmesi modellerinin veriye uygulanması,
- Günlük doğalgaz talep tahmininin gerçekleştirilmesi ve modellerin değerlendirmesi.

Önerilen tahmin süreci Şekil 4.1'de gösterilmiştir.

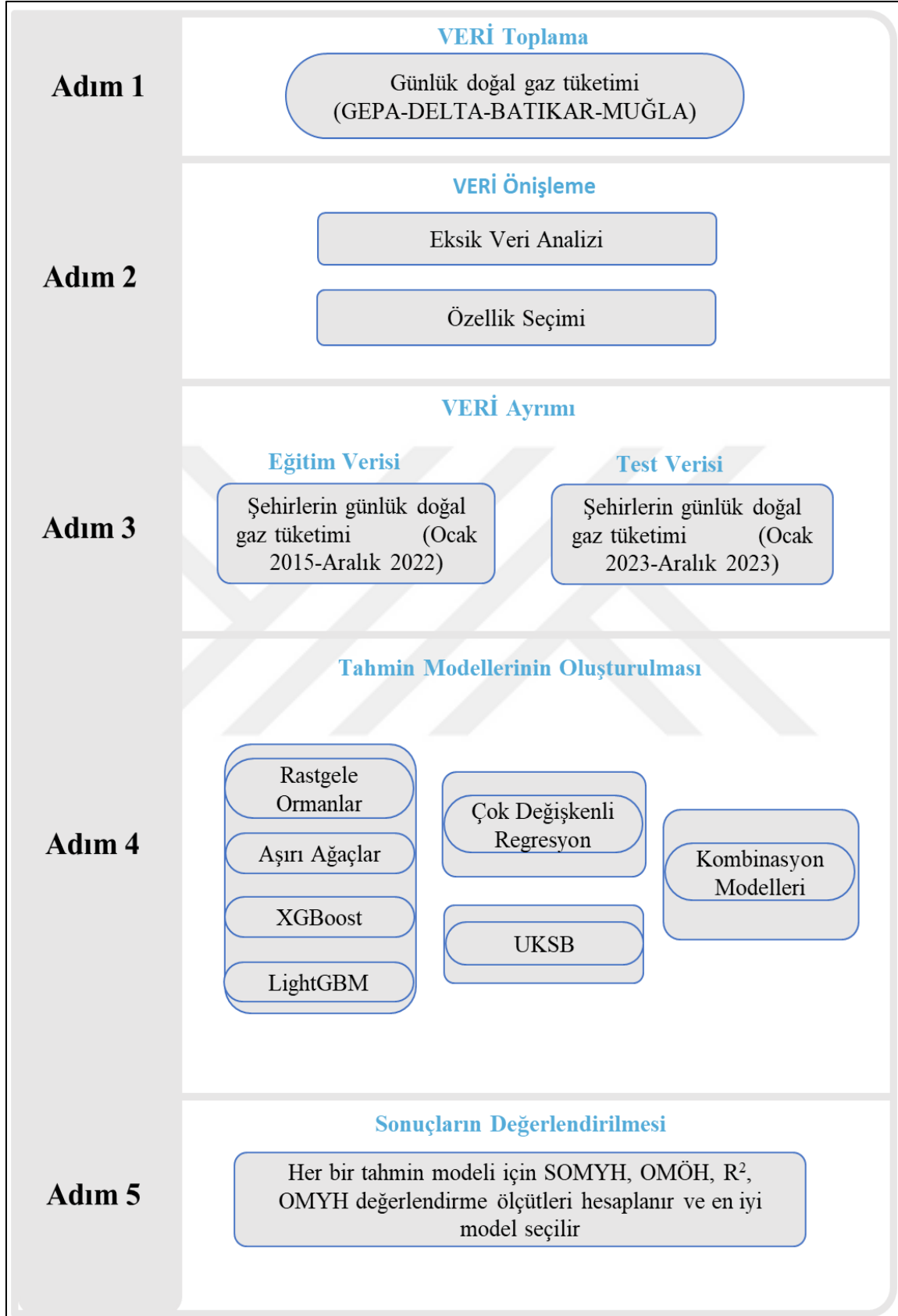
Adım 1: Doğalgaz dağıtım şirketinden BDÖİ bazlı elde edilen tüketim verilerinin toplanarak bir araya getirilmesi ve ilgili şehirlerin günlük doğalgaz tüketimlerinin hesaplanması.

Adım 2: Veri setinden tüketim verileri ve zamansal özellikleri kullanılarak çeşitli bağımsız değişkenler oluşturulması. Oluşturulan bağımsız değişkenlerden tahmin performansına katkısı en yüksek olan değişkenlerin LASSO regresyonu ile seçilmesi.

Adım 3: Veri setinin eğitim (Ocak 2015-Aralık 2022) ve test (Ocak 2023-Aralık 2023) olarak ikiye ayrılması.

Adım 4: İstatistiksel, topluluk ve derin öğrenme stratejilerinin veri setine uygulanması.

Adım 5: Kurulan modellerin tahmin performanslarının belirlenmesi. Performans metrikleri açısından bütün modellerin karşılaştırılması ve en uygun model ile bir sonraki gün için gaz tüketim miktarının tahmin edilmesi.



Şekil 4.1. Tahmin sürecine ait akış diyagramı

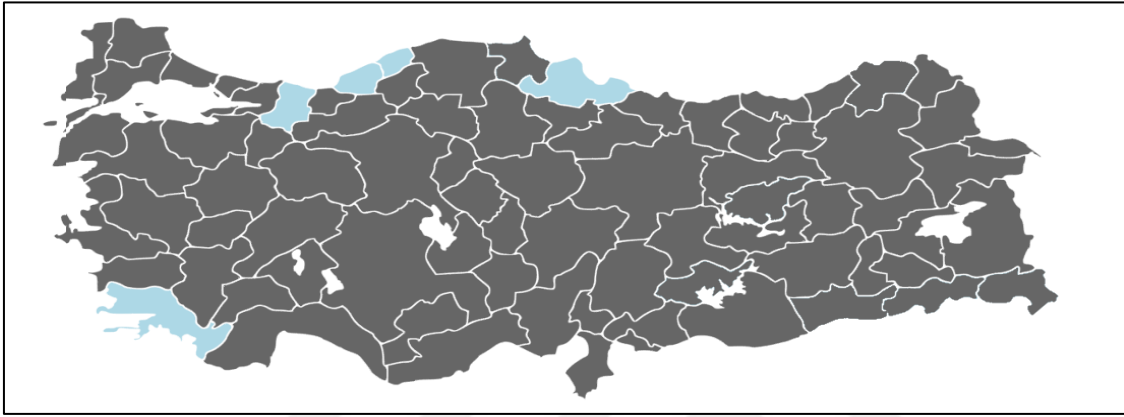
4.1. Çalışma Alanı ve Verinin Tanımlanması

Akmercan Doğalgaz Dağıtım A.Ş. Türkiye genelinde toplamda 12 farklı şehirde doğalgaz dağıtım faaliyetlerini gerçekleştirmektedir. Bu çalışmada tüketim miktarı ve şehirlerin doğalgaz dağıtım lisanslarının alınma tarihleri (tüketim veri adedi) göz önünde bulundurularak Çizelgede 4.1’de gösterilen 5 şehir için günlük doğalgaz tüketim tahmin işlemi gerçekleştirilmiştir. Şekil 4.2’de doğalgaz tüketimi tahmin edilen şehirler görsel olarak sunulmuştur. Bu beş şehir için, 2015- 2023 yılları arasında günlük doğalgaz tüketim verileri dağıtım şirketinden temin edilmiştir. Seçilen beş şehir yaklaşık olarak 23 139 km² bir alanı kaplamaktadır. Bu şehirler Türkiye nüfusunun %2,44’ünü barındırmakta ve ülkenin yüz ölçümünün yaklaşık %2,9’unu oluşturmaktadır. Çalışmanın odaklandığı şehirlere ilişkin bazı tanımlayıcı bilgiler Çizelge 4.1’de sunulmuştur.

Çizelge 4.1. Çalışmaya konu edinilen şehirlere ait bazı tanımlayıcı bilgiler

Şehir	İlçe	Nüfus	Yüz Ölçümü (km ²)	Abone Sayısı	Lisans Alınma Yılı	İklim
Sakarya	Gevye	51 676	780	25 153	2011	Ilıman İklim
	Pamukova	31 205	432			
Samsun	Alaçam	25 148	632	78 094	2011	Ilıman İklim
	Bafra	142 210	175			
	Havza	38 824	788			
	Lâdik	16 907	596			
	Vezirköprü	90 388	1 713			
	19 Mayıs	27 910	363			
	Yakakent	9 145	204			
Bartın	Tamamı	207 238	2 330	143 666	2012	Ilıman İklim
Zonguldak	Merkez	117 360	272			
	Çaycuma	91 968	490			
	Devrek	57 562	934			
	Gökçebey	21 344	189			
	Kilimli	33 639	90			
	Kozlu	50 908	177			
Muğla	Tamamı	1 066 736	12 974	54 877	2013	Akdeniz İklimi

Şehirlerin günlük doğalgaz tüketim miktarı, ilgili şehirde bulunan BDÖİ'lerden günlük olarak geçen doğalgazın toplamı olarak hesaplanmıştır. Çizelge 4.2'de beş farklı şehrin doğalgaz tüketim verilerine ilişkin özet istatistiksel değerler gösterilmiştir. Buna göre şehirlerin doğalgaz tüketim verileri incelendiğinde, en büyük tüketim Batıkar bölgesinde meydana geldiği gözlemlenmektedir. Batıkar bölgesini sırasıyla Gepa, Delta ve Muğla bölgeleri izlemektedir.



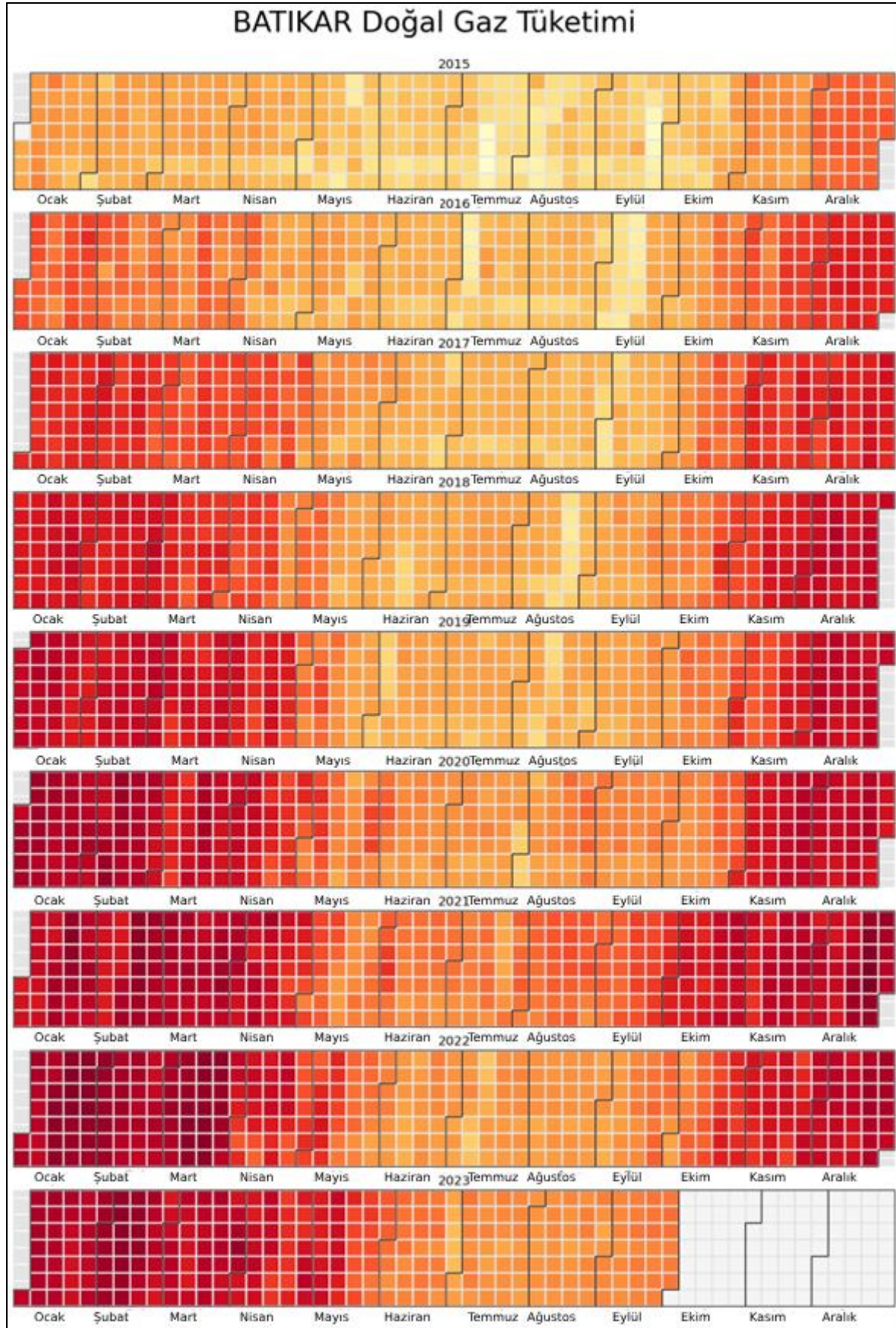
Şekil 4.2. Tahmin işlemi gerçekleştirilecek dört bölge

Gepa bölgesinde çok sayıda sanayi tesisinin bulunması ve yılın her döneminde üretimin gerçekleşmesi nedeniyle tüketim değişkenliği en küçük olan bölge olarak görülmektedir. Benzer şekilde Delta ve Muğla bölgelerinde bulunan elektrik üretim amaçlı çevrim tesisleri de tüketim değişkenliğinin görece düşük olmasının sebebidir. Muğla bölgesindeki gaz dağıtım faaliyetleri diğer bölgelere kıyasla çok sonra başladığı için tüketim açısından değerlendirildiğinde ölçek olarak diğer tüm bölgelerden daha küçüktür. Şekil 4.3 ve Ek 1'de bölgelerin tüketim ısı grafikleri gösterilmiştir. Grafikler incelendiğinde; Gepa bölgesi dışındaki diğer bölgelerde tüketimler Mayıs- Ekim aylarında azalırken, kış döneminde ısınma amaçlı doğalgaz kullanımı nedeniyle bu bölgelerdeki tüketim ciddi şekilde artış göstermektedir. Ancak Gepa bölgesindeki yoğun sanayi tüketimi nedeniyle tüketim açısından dönemler arasında ciddi farklılıklar bulunmamaktadır. Doğal gaz tüketim verilerinin incelenmesi sonucunda, Muğla bölgesi dışındaki bölgelerdeki tüketim örüntüsünün durağan bir seyir izlediği belirlenmiştir. Bu durum, %95 güven düzeyi ile yapılan Dickey-Fuller (DF) testi sonuçlarıyla da doğrulanmıştır. Yapılan test sonuçlarına göre, Muğla bölgesi dışındaki bölgelerin doğal gaz tüketim serisinin birim kök içermediği

ve dolayısıyla durağan olduğu tespit edilmiştir (p-değeri <0,05). Bu bulgular, söz konusu bölgelerdeki doğal gaz tüketimindeki istikrarlı bir eğilimin varlığını ortaya koymaktadır.

Çizelge 4.2. Veri setlerinin istatistiksel özellikleri (Sm³)

İstatistiksel Özellik	GEPA	DELTA	BATIKAR	MUĞLA
Adet	3 287	3 287	3 287	2 922
Ortalama	232 242	188 725	376 345	127 221
Standart Sapma	56 866	132 151	257 587	144 070
Minimum Değer	47 589	9 132	35 379	0
%25	196 488	85 347	168 366	27 631
%50	221 743	134 797	272 173	61 055
%75	268 506	291 783	563 061	201 912
Maksimum Değer	447 161	605 007	1 240 532	914 794
Toplam	763 380 100	620 339 600	1 237 048 644	371 741 000
Çarpıklık	0.48	0.84	0.88	1.62
Basıklık	0.26	-0.27	-0.26	2.55
Durağanlık Testi (p değeri)	0.003	0.031	0.040	0.122



Şekil 4.3. Batıkar bölgesinin tüketim ısı grafiği

4.2. Eksik Değer Analizi

Şehirlerin doğalgaz tüketim verileri incelendiğinde Gega, Delta ve Batıkar bölgelerinde herhangi bir eksik değer bulunmamaktadır. Ancak Muğla bölgesinin 2015 yılı gaz tüketim verisi çok sayıda eksik değer içerdiği için analiz dışında bırakılmıştır.

4.3. Bağımsız Değişken Oluşturma ve Seçme

Muğla bölgesindeki eksik doğalgaz tüketim verileri giderildikten sonra, makine öğrenmesi modellerin kullanacağı bağımlı ve bağımsız değişkenlerin belirlenmesi gerekmektedir. Bu çalışmada tüketim verisi kullanılarak 33 adet bağımsız değişken oluşturulmuştur. Çalışmada kullanılabilecek özelliklerin tam listesi aşağıdaki gibidir:

Her bölge için 1, 2, 3, 7 ve 365 günlük gecikmeli doğalgaz tüketim verisi: (5 bağımsız değişken)

- Her bölge için ortalama, standart sapma, maksimum ve minimum metrikleri ile 7, 30, 365 günlük hesaplanan pencere değişkenleri: (12 bağımsız değişken)
- Her bölge için içinde bulunulan zamanı gösteren değişkenler. (Ay, Yıl, Mevsim, Yılın Haftası, Haftanın Günü, Ayın Günü, Yılın Günü, Hafta Sonu): (15 bağımsız değişken)
- Resmî tatil günlerini gösteren değişken: (1 bağımsız değişken)

Bu çalışmada zamanı ifade eden değişkenler tüm veri seti boyunca düzenli olarak tekrarlandığı için bu verilerin sayısallaştırılmasında döngüsel kodlama tekniği uygulanmıştır. Her bir değişkenin sinüs ve kosinüs değeri hesaplandığı için toplamda 15 değişken oluşturulmuştur.

Doğrusal regresyon türlerinden biri olan LASSO regresyonu bağımsız değişken seçiminde kullanılmıştır [35, 45]. Bu çalışmada λ 'nin optimal değerini seçmek için 5 katlı çapraz doğrulama seti kullanılmıştır. Her çapraz doğrulama seti için, $[10^{-3} - 2 \cdot 10^1]$ aralığındaki λ değerleri üzerinde 0,1 adım boyutunda bir ızgara araması gerçekleştirilmiştir. OKH'yi en aza indiren λ değeri, en uygun değer olarak seçilmiştir. Çizelge 4.3'te her bir dağıtım bölgesi için LASSO regresyonunun seçtiği bağımsız değişkenler gösterilmiştir. Özellik seçim süreci

sonrasında Gepa, Delta, Batıkar ve Muğla bölgelerinde sırası ile 4, 9, 1, 4 adet bağımsız değişkenin katsayıları LASSO regresyonu tarafından sıfır olarak belirlenmiştir. Bu özellikler çalışma kapsamı dışında tutulmuştur.

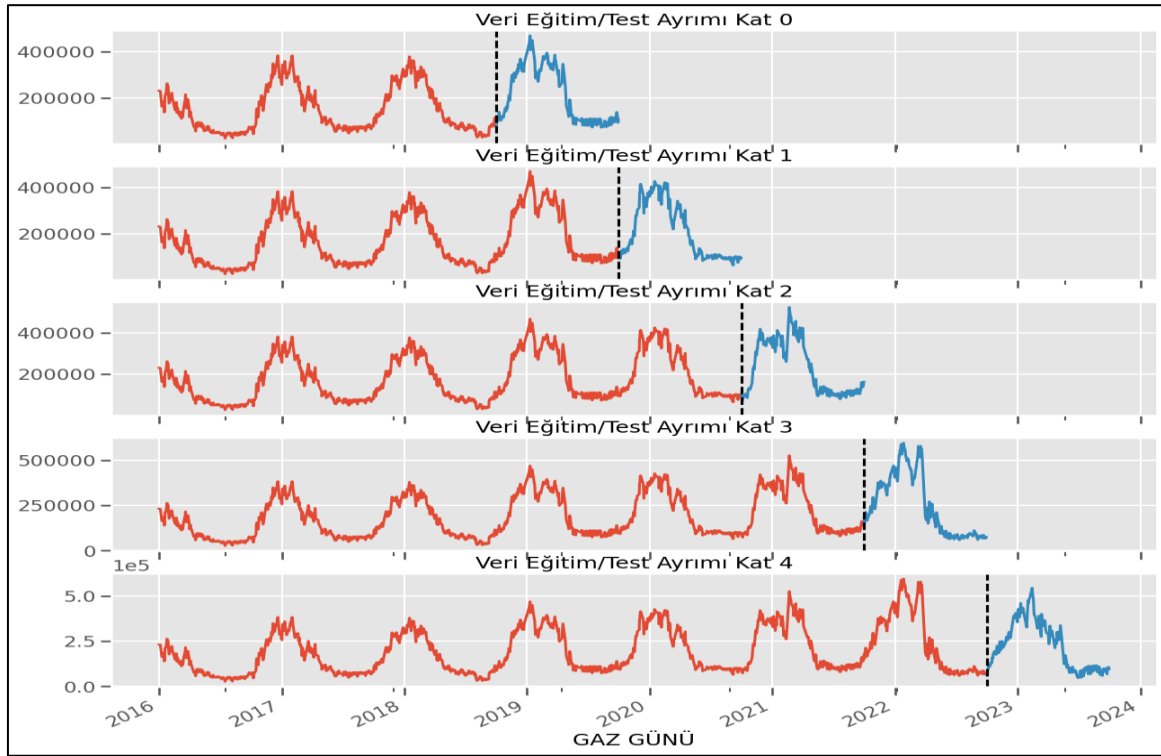
Çizelge 4.3. LASSO regresyonu tarafından seçilen bağımsız değişkenler

BÖLGE	SEÇİLEN BAĞIMSIZ DEĞİŞKENLER
Gepa	Gecikme1, Gecikme2, Gecikme3, Gecikme7, Gecikme365, Pencere7_ort, Pencere7_std, Pencere7_max, Pencere7_min, Pencere30_ort, Pencere30_std, Pencere30_max, Pencere30_min, Pencere365_max, Pencere365_min, Sin_Mevsim, Cos_Mevsim, Sin_Yıl, Cos_Yıl, Sin_Hafta, Cos_Hafta, Sin Haftanın_Günü, Cos Haftanın_Günü, Sin Ayın_Günü, Cos Ayın_Günü, Hafta_Sonu, Resmi_Tatil, Sin_Ay, Cos_Ay
Delta	Gecikme1, Gecikme 2, Gecikme 3, Gecikme 365, Pencere7_std, Pencere7_max, Pencere7_min, Pencere30_ort, Pencere30_max, Pencere30_min, Pencere365_ort, Pencere365_std, Pencere365_min, Sin_Mevsim, Cos_Mevsim, Cos_Yıl, Sin_Hafta, Cos_Hafta, Sin Haftanın_Günü, Cos Haftanın_Günü, Sin Ayın_Günü, Cos_Ayın_Günü, Hafta_Sonu, Resmi_Tatil
Batıkar	Gecikme1, Gecikme 2, Gecikme3, Gecikme7, Gecikme365, Pencere7_ort, Pencere7_std, Pencere7_max, Pencere7_min, Pencere30_ort, Pencere30_std, Pencere30_max, Pencere30_min, Pencere365_ort, Pencere365_std, Pencere365_max, Pencere365_min, Sin_Mevsim, Cos_Mevsim, Sin_Yıl, Cos_Yıl, Sin_Hafta, Sin Haftanın_Günü, Cos Haftanın_Günü, Sin Ayın_Günü, Cos Ayın_Günü, Hafta_Sonu, Resmi_Tatil, Sin_Ay, Cos_Ay, Sin_Yılın_Günü, Cos_Yılın_Günü
Muğla	Gecikme1, Gecikme2, Gecikme3, Gecikme7, Gecikme365, Pencere7_ort, Pencere7_std, Pencere7_max, Pencere30_ort, Pencere30_std, Pencere30_max, Pencere30_min, Pencere365_ort, Pencere365_std, Pencere365_min, Pencere365_max, Sin_Mevsim, Cos_Mevsim, Cos_Yıl, Sin_Hafta, Cos_Hafta, Sin Haftanın_Günü, Cos Haftanın_Günü, Hafta_Sonu, Resmi_Tatil, Cos_Ay, Sin_Ayın_Günü, Cos_Ayın_Günü, Cos_Yılın_Günü

4.4. Modellerin Uygulanması

Veri setleri oluşturulduktan sonra, Türkiye'nin beş farklı şehrine ait günlük doğal gaz tüketimini tahmin etmek amacıyla üç farklı makine öğrenmesi stratejisi ve toplamda altı farklı makine öğrenmesi algoritması performansı karşılaştırılmıştır. Bu amaçla ilk olarak veri, eğitim ve test verileri olmak üzere ikiye ayrılır. 01.01.2015-31.12.2022 dönemine ait veriler eğitim verileri olarak kullanılmıştır. Önerilen modellerin performanslarını değerlendirmek için kullanılan 01.01.2023-31.12.2023 dönemine ait veriler ise test verisi olarak kullanılmıştır. Çalışmada kullanılan makine öğrenmesi algoritmalarının en iyi hiper parametre kombinasyonunu rastgele arama yöntemi kullanılarak belirlenmiştir. Zaman serisi verileriyle çalıştığımız için çapraz doğrulama süreci sırasında verilerin sıralı yapısının korunmasını sağlamak çok önemlidir [35]. Bu amaçla modeller eğitilirken zaman serisi çapraz doğrulama şeması kullanılmıştır. Bu çalışmada her katlama için eğitim setinin uzunluğu değişken olacak şekilde ayarlanmıştır. Her test seti ise eğitim setinin hemen

ardından gelen 365 günden oluşur. Bu değerler, tüm veri kümesinin eşit uzunlukta 5 kata bölüneceği şekilde seçilir. Şekil 4.4'te eğitim verisi için çapraz doğrulama şemasını göstermektedir. Çalışma Python programlama dili kullanılarak gerçekleştirilmiştir. Çalışmada numpy, pandas, pytorch, sklearn, sktime, scipy ve matplotlib kütüphaneleri kullanılmıştır.



Şekil 4.4. Model eğitimi için çapraz doğrulama şeması

Farklı modellerin performansını değerlendirmek için ortalama SOMYH, medyan SOMYH, ortalama OMÖH ve medyan OMÖH, R^2 ve OMYH olmak üzere altı farklı metrik kullanılmıştır. Bu ölçütlerin dağıtım bölgeleri kırılımında eğitim ve test verileri için aldığı değerler Çizelge 4.5 ve Çizelge 4.6'te sunulmuştur. Çizelge 4.6'ya göre doğalgaz talebi Gepa, Delta ve Batıkar bölgelerinde daha doğru bir şekilde tahmin edilirken, Muğla bölgesinde değerlendirme ölçütleri büyük ölçüde sapmaktadır. Bu izlenim, Şekil 4.9'da gösterilen tahmin performansı metrikleri temelinde doğrulanmıştır. Örneğin, üç bölgede için SOMYH değerleri %8-%12 arasında iken Muğla bölgesinde bu değer %24-%28 arasında değişmektedir.

Eğitim verileri göz önüne alındığında Muğla bölgesi için LightGBM en iyi performansı gösterir iken diğer üç bölge için ARA modeli tüm değerlendirme metrikleri açısından en iyi performansı gösteren model olmuştur. LSTM ise diğer modellere kıyasla tüm bölgeler için düşük performans elde etmiştir. Test verileri göz önüne alındığında çoklu doğrusal regresyon Gepa bölgesi için en iyi performansı göstermiştir. RO modeli Muğla bölgesinde en yüksek performansı gösterir iken, Delta bölgesi için LightGBM, Batıkar bölgesi için ise ARA en yüksek doğruluğu elde etmiştir. Çizelge 4.4'te dağıtım bölgelerinin tüketim tahmini yapılırken kullanılacak modeller gösterilmiştir. Kullanılacak bu modeller, test verisi üzerinde gösterdikleri performansa dayanarak belirlenmiştir.

Çizelge 4.4. Seçilen modeller ve tahmin hataları

BÖLGE	Tercih Edilen Model	Tahmin Hatası (Ortalama SOMYH %)	Tahmin Hatası (Ortalama OMÖH)
Gepa	Çok Değişkenli Regresyon	%7,39	0,5304
Delta	LightGBM	%10,54	0,3925
Batıkar	Aşırı Rastgeleleşmiş Ağaçlar	%8,55	0,3957
Muğla	Rastgele Ormanlar	%24,28	0,7210

Şekil 4.5'te test verileri için modellerin performansları görsel olarak kıyaslanmıştır.

Çizelge 4.5. Eğitim verisi için modellerin tahmin hatası

Şehir	Gösterge	Aşırı Rastgeleleşmiş Ağaçlar	LightGBM	Çok Değişkenli Regresyon	Rastgele Ormanlar	XGBoost	UKSB
GEPA	R ²	0,8807	0,8771	0,8593	0,8649	0,8678	0,8568
	OYMH	%6,14	%6,17	%6,56	%6,40	%6,41	%6,62
	Ortalama OMÖH	0,3662	0,3671	0,3922	0,3823	0,3819	0,3910
	Medyan OMÖH	0,3518	0,3470	0,3696	0,3508	0,3609	0,3620
	Ortalama SOMYH	%5,85	%5,89	%6,28	%6,13	%6,11	%6,26
	Medyan SOMYH	%4,05	%3,95	%4,27	%4,01	%4,15	%4,11
DELTA	R ²	0,9815	0,9766	0,9754	0,9771	0,9761	0,9718
	OYMH	%8,12	%8,97	%9,05	%8,91	%8,99	%10,30
	Ortalama OMÖH	0,2430	0,2673	0,2762	0,2677	0,2709	0,3042

Çizelge 4.6. (devam) Eğitim verisi için modellerin tahmin hatası

	Medyan OMÖH	0,2458	0,2604	0,2829	0,2710	0,2611	0,3262
	Ortalama SOMYH	%7,84	%8,57	%8,97	%8,64	%8,62	%10,25
	Medyan SOMYH	%5,85	%6,11	%6,57	%6,31	%6,29	%7,84
BATIKAR	R ²	0,9817	0,9773	0,9751	0,9768	0,9777	0,9623
	OYMH	%7,53	%8,25%	%8,86	%8,49	%8,18	%10,55
	Ortalama OMÖH	0,2371	0,2571	0,2765	0,2651	0,2570	0,3320
	Medyan OMÖH	0,2398	0,2518	0,2898	0,2606	0,2495	0,3237
	Ortalama SOMYH	%7,33	%7,95	%8,79	%8,24	%7,92	%10,05
	Medyan SOMYH	%5,43	%5,49	%6,67	%5,88	%5,53	%7,35
MUĞLA	R ²	0,9619	0,9805	0,9527	0,9586	0,9635	0,9453
	OYMH	%23,03	%18,90	%24,37	%23,22	%30,82	%30,20
	Ortalama OMÖH	0,2945	0,2143	0,3126	0,2987	0,2937	0,3450
	Medyan OMÖH	0,2670	0,2085	0,2620	0,2694	0,2498	0,2991
	Ortalama SOMYH	%19,21	%16,05	%22,61	%19,68	%22,12	%23,20
	Medyan SOMYH	%12,72	%9,38	%12,60	%12,70	%12,55	%14,42

Çizelge 4.7. Test verisi için modellerin tahmin hatası

Şehir	Gösterge	Aşırı Rastgeleleşmiş Ağaçlar	LightGBM	Çok Değişkenli Regresyon	Rastgele Ormanlar	XGBoost	UKSB
GEPA	R ²	0,8249	0,8005	0,8846	0,8388	0,8057	0,8766
	OYMH	%9,60	%9,73	%7,74	%8,80	%9,71	%7,99
	Ortalama OMÖH	0,7057	0,7476	0,5304	0,6648	0,7410	0,5574
	Medyan OMÖH	0,7000	0,7096	0,4822	0,6315	0,7359	0,5315
	Ortalama SOMYH	%9,28	%9,58	%7,39	%8,65	%9,52	%7,67
	Medyan SOMYH	%7,35	%7,36	%4,58	%6,30	%7,20	%4,98
DELTA	R ²	0,9548	0,9576	0,9583	0,9519	0,9554	0,9533
	OYMH	%11,38	%10,86	%11,26	%12,00	%11,11	%11,70
	Ortalama OMÖH	0,4109	0,3925	0,4005	0,4261	0,4075	0,4215
	Medyan OMÖH	0,4454	0,4041	0,4376	0,4600	0,4463	0,4611
	Ortalama SOMYH	%10,98	%10,54	%10,93	%11,64	%10,78%	%12,12
	Medyan SOMYH	%8,62	%8,49	%8,51	%9,04	%8,61%	%8,16
BATIKAR	R ²	0,9581	0,9602	0,9527	0,9577	0,9591	0,9468

Çizelge 4.8. (devam) Test verisi için modellerin tahmin hatası

	OYMH	%8,68	%8,88	%11,85	%8,80	%8,91	%9,59
	Ortalama OMÖH	0,3957	0,3942	0,4610	0,3972	0,3993	0,4451
	Medyan OMÖH	0,3761	0,4052	0,5305	0,3800	0,4082	0,4300
	Ortalama SOMYH	%8,55	%8,80	%11,01	%8,67	%8,82	%9,42
	Medyan SOMYH	%6,94	%7,03	%9,92	%7,02	%7,18	%7,26
MUĞLA	R ²	0,8985	0,8934	0,8907	0,8932	0,8943	0,8972
	OYMH	%30,97	%34,79	%38,14	%27,64	%30,25	%30,20
	Ortalama OMÖH	0,7411	0,7736	0,8096	0,7210	0,7474	0,7519
	Medyan OMÖH	0,8071	0,8994	1,0381	0,7549	0,8187	0,8247
	Ortalama SOMYH	%25,85	%27,70	%28,66	%24,28	%25,40	%25,68
	Medyan SOMYH	%20,18	%20,21	%21,25	%17,05	%20,57	%19,89

Yukarıda belirtilen performans değerlendirme kriterlerine ek olarak, parametrik olmayan Friedman testi, topluluk modelleri ve diğer modeller arasındaki tahmin doğruluğundaki farkların istatistiksel önemini değerlendirmek için yeni bir kriter olarak kullanılmıştır. Ardından nemenyi testi, tüm modeller arasındaki farkların istatistiksel önemini incelemek için kullanılmıştır. İstatistiksel testleri gerçekleştirmek için her bir modelden elde edilen SOMYH değerleri kullanılmıştır. Çizelge 4.7’de yapılan istatistik testlerine ait p değerleri gösterilmiştir. 0,05’ten küçük p değerleri, modeller arasında anlamlı bir farkın olduğunu belirtir ve bu, %95 güven düzeyinde önemlidir. Çizelge 4.7’ye göre, tüm dağıtım bölgeleri için Friedman testinin sonucu 0,05’ten küçüktür. Bu H₀ hipotezinin reddedilmesi anlamına gelir; yani her bir bölge için en az bir model diğerlerinden anlamlı şekilde farklılık göstermektedir. İkinci aşamada, her bir bölge için oluşturulan tahmin modellerinin ikili karşılaştırmaları nemenyi testi ile gerçekleştirilmiştir [16, 46]. İstatistiksel olarak anlamlı bulunan model çiftlerine ait p değerleri koyu renkle işaretlenmiştir.

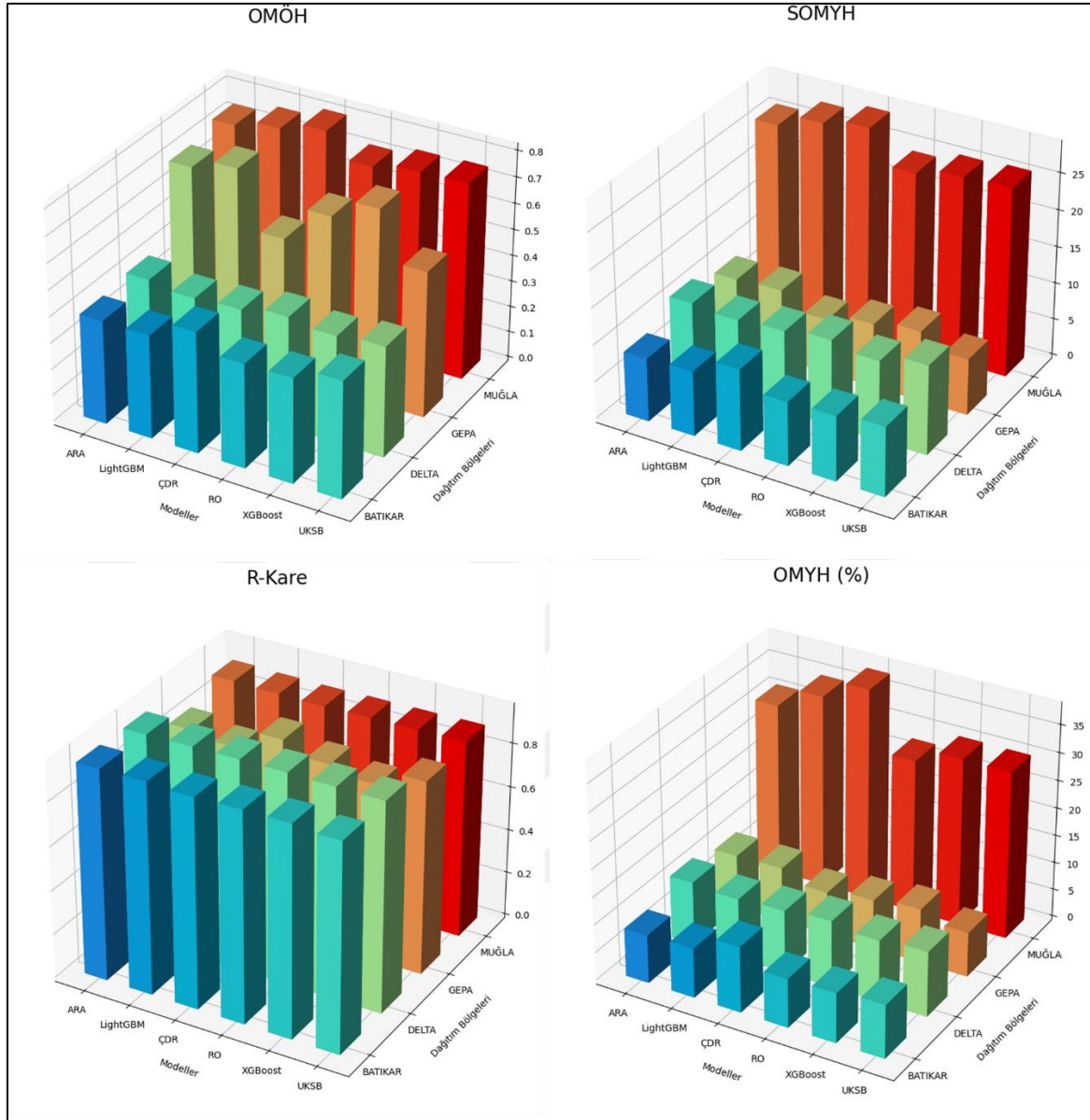
Çizelge 4.9. İstatistiksel test sonuçları (p değeri)

	Batıkar	Delta	Gepa	Muğla
Friedman Testi	0,0000	0,0001	0,0000	0,0001
XGBoost - RO	0,9000	0,0053	0,0010	0,8000
XGBoost - ÇDR	0,0010	0,9000	0,0010	0,0170
XGBoost - LightGBM	0,9000	0,8234	0,9000	0,3553

Çizelge 4.10. (devam) İstatistiksel test sonuçları (p değeri)

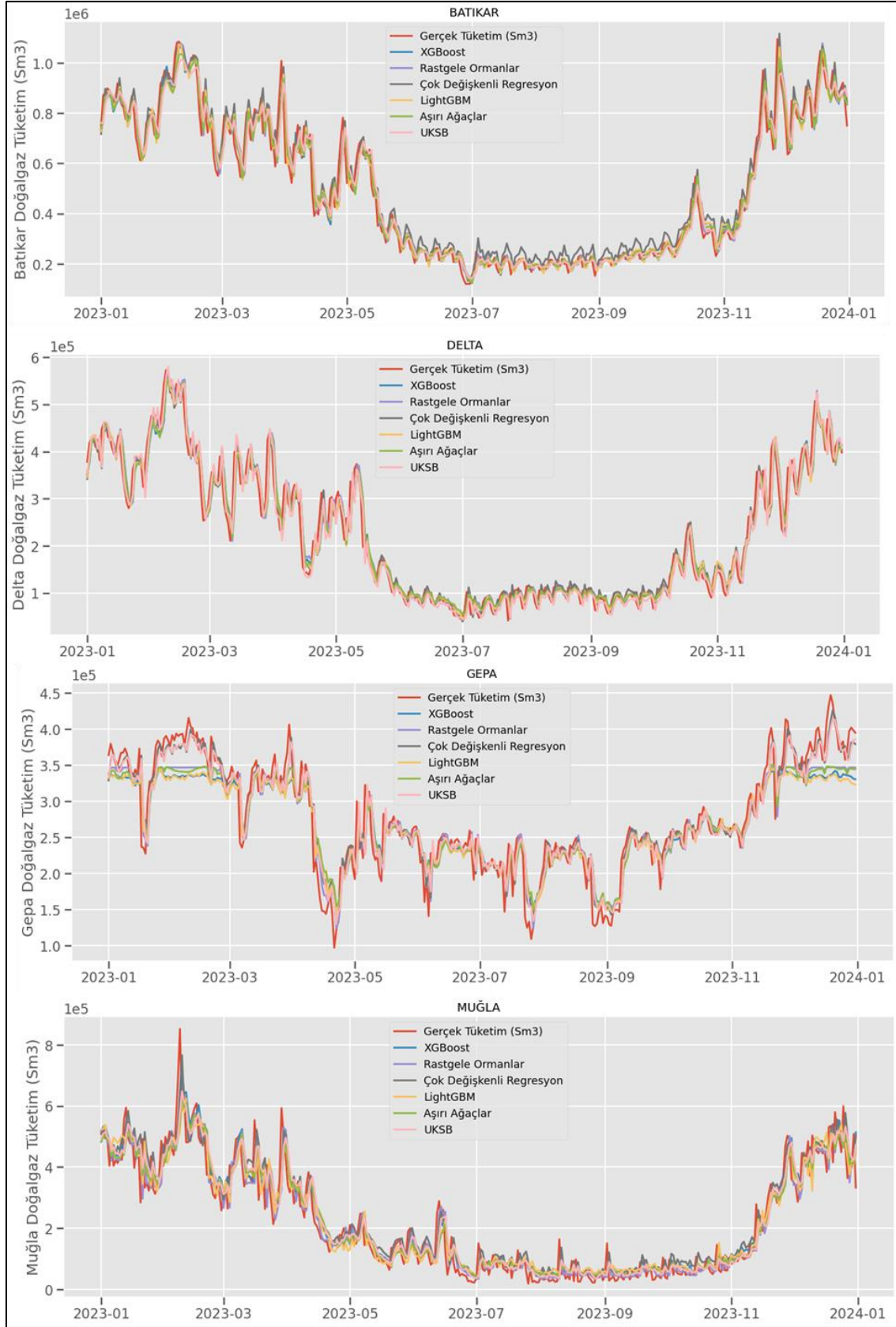
XGBoost - ARA	0,6749	0,6291	0,1878	0,9000
XGBoost - UKSB	0,1514	0,5949	0,0010	0,9000
RO - ÇDR	0,0010	0,0219	0,0037	0,0010
RO - LightGBM	0,9000	0,0010	0,0010	0,0170
RO - ARA	0,8557	0,3319	0,2303	0,9000
RO - UKSB	0,0696	0,3673	0,5149	0,7206
ÇDR - LightGBM	0,0010	0,5834	0,0010	0,8005
ÇDR - ARA	0,0010	0,8691	0,0010	0,0050
ÇDR - UKSB	0,6063	0,8348	0,3919	0,0264
LightGBM - ARA	0,9000	0,0734	0,3207	0,1800
LightGBM - UKSB	0,0181	0,0625	0,0010	0,4419
ARA - UKSB	0,0016	0,9000	0,0012	0,9000

Batıkar bölgesi için, ARA modeli ARA-RO ve ARA-LightGBM arasında yapılan Nemeyni testi sonucunda %95 güven düzeyi ile istatistiksel olarak anlamlı bir fark bulunamamıştır. Delta bölgesi için ise, LightGBM ile UKSB, ARA, ÇDR ve XGBoost arasında istatistiksel olarak anlamlı bir fark tespit edilmemiştir. Gepa bölgesi için seçilen ÇDR ile UKSB arasında ise anlamlı bir fark belirlenmemiştir. Son olarak, Muğla bölgesi incelendiğinde, RO algoritması ile XGBoost, ARA ve UKSB arasında istatistiksel olarak anlamlı bir fark saptanmamıştır.



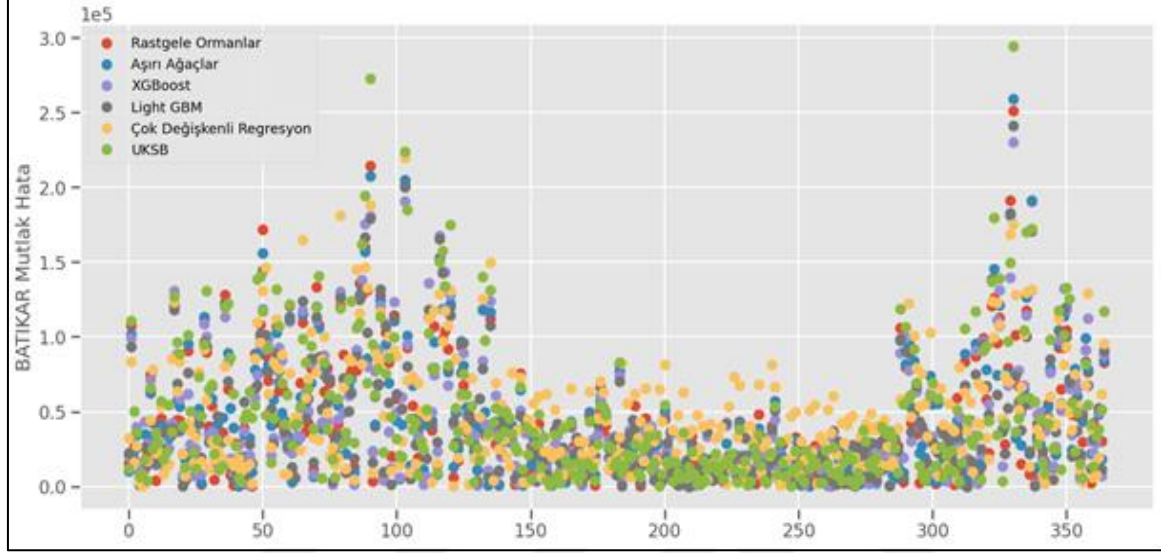
Şekil 4.5. Bölgelere göre modellerin ürettiği performans ölçütleri

Modellerin performansını daha iyi gözlemlemek için Şekil 4.6’da 1 Ocak 2023’ten 31 Aralık 2023’e kadar farklı coğrafi konumlarda bulunan 4 bölgenin fiili doğal gaz tüketimi ile tahmin edilen gaz tüketim değerlerinin karşılaştırmalı grafiği gösterilmiştir. Şekilden de görülebileceği gibi, şehirlerin farklı ekonomik yapıları ve iklim koşulları nedeniyle, doğal gaz talebinin tepe ve dip zaman noktaları ile gaz tüketim ölçekleri farklılık gösterebilmektedir. Gega ve Muğla bölgeleri için yapılan tahminler, gerçek tüketim verilerine diğer bölgelere kıyasla daha az uyum sağlamaktadır. Bu durum, R^2 katsayılarının diğer bölgelere göre daha düşük olmasına yol açmıştır.



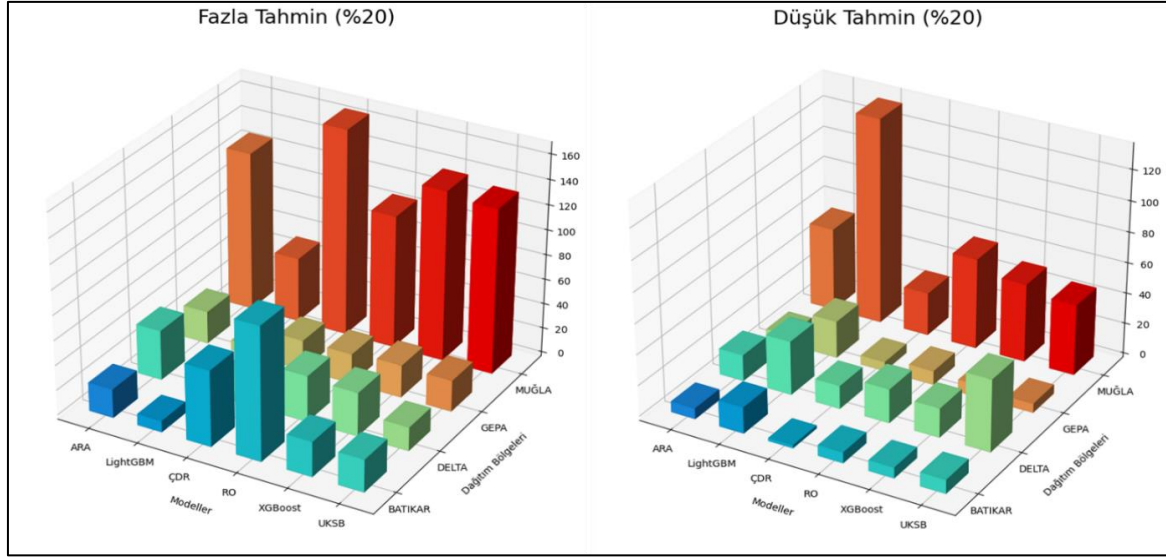
Şekil 4.6. Dört farklı şehrin gerçek ve tahmin edilen tüketim değerleri

Şekil 4.7’de Batıkar bölgesi için tahmin modellerinden elde edilen MAE değerlerinin dağılımı gösterilmiştir. Şekil incelendiğinde Batıkar bölgesi için ARA modelinin görece daha düşük ve istikrarlı mutlak hata değerleri ürettiği görülebilir. Diğer bölgelere ait mutlak hata değerlerinin saçılım grafiği Ek 2’de sunulmuştur.



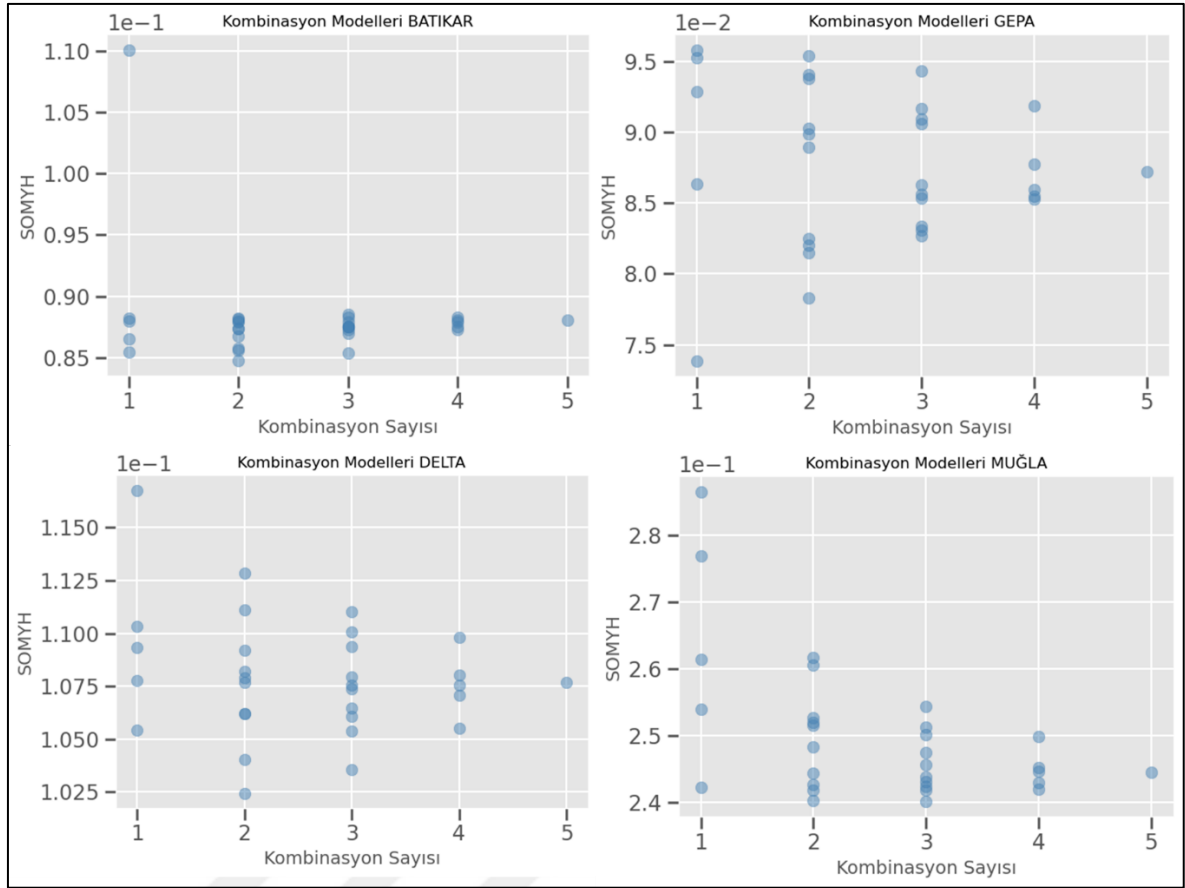
Şekil 4.7. Tahmin modellerinin MAE değerleri

Geleneksel tahmin doğruluğunun değerlendirilmesinin yanı sıra, aşırı tahminlerin incelemesi ve risk analizi yapılırken bu durumda göz önünde bulundurulması önemlidir. Şekil 4.8 incelendiğinde en fazla aşırı tahmin Muğla bölgesinde gerçekleşmiştir. Bu durum Muğla’nın tahmin doğruluğu göz önünde bulundurulduğunda beklenen bir sonuçtur. Tüm bölgeler için yapılan incelemelerde, ÇDR düşük tahmin üretmede en başarılı çalışan algoritma olarak belirlenmiştir. LightGBM algoritması ise aşırı tahmin konusunda tüm bölgelerde en başarılı algoritma olarak belirlenmiştir. Batıkar ve Gepa bölgeleri en az düşük tahmin gerçekleştirilen bölge olarak görülmektedir.



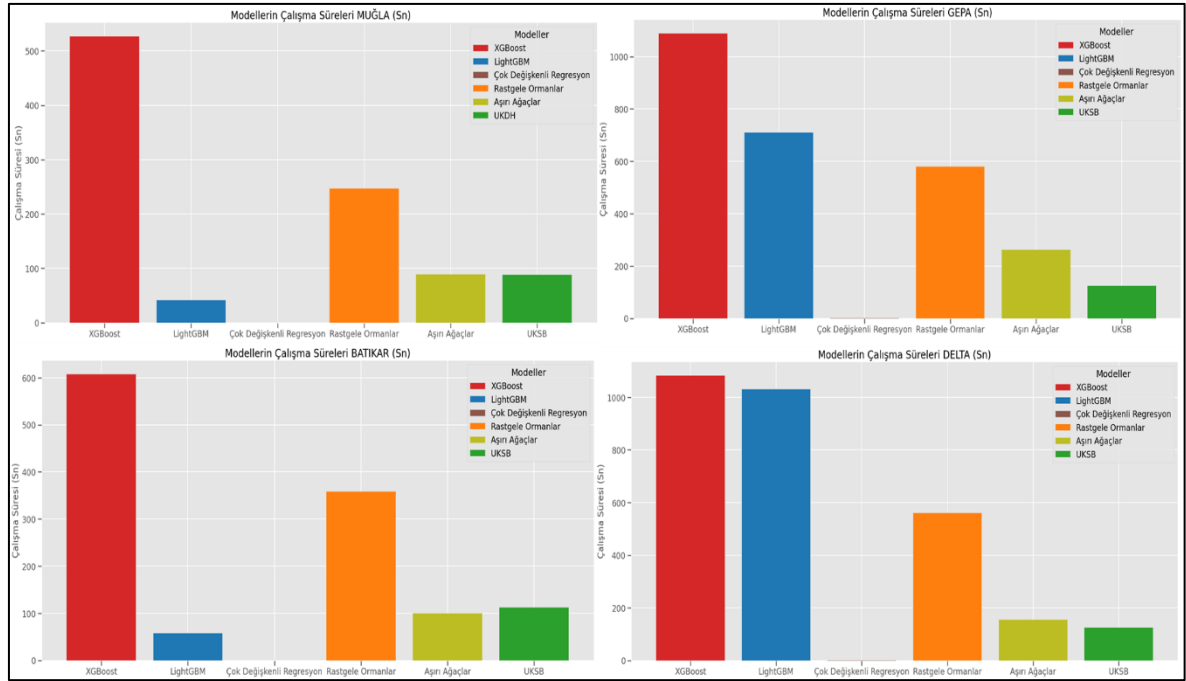
Şekil 4.8. Yirmi dört model (6 yöntem $\times$ 4 bölge) için ekstrem veri noktalarının sayısı

Araştırmanın son aşamasında, bireysel modellerden elde edilen tahminlerin tüm kombinasyonlarının ortalama, medyan, maksimum ve minimum değerleri üzerinden kapsamlı bir değerlendirme yapılmıştır. Her bir bölge için toplamda 4*31 farklı değerlendirme gerçekleştirilmiş olup, en etkili performansı sergileyen kombinasyonların belirlenmesi amaçlanmıştır. Şekil 4.9 tüm bu kombinasyonlar için hesaplanan SOMYH değerlerini göstermektedir. Yapılan değerlendirme sonucunda, Gepa bölgesinde, Çizelge 4.4'te sunulan modelin en iyi tahmin performansını sürdürdüğü gözlemlenmiştir. Batıkar bölgesinde, "ÇDR-ARA" kombinasyonunun minimum değerinin ürettiği tahmin sonuçları, %8,47 SOMYH değeri ile bireysel modellerden daha üstün bir performans sergilemiştir. Ayrıca, Delta ve Muğla bölgeleri için "ÇDR-LightGBM" ve "RO-XGBoost-ÇDR" kombinasyonlarının minimumları, sırasıyla %10,24 ve %24,02 SOMYH değerleri ile bireysel modellerden kısmen daha iyi performans göstermiştir. Bu bulgular, bölgesel tahminlerde optimal kombinasyonların belirlenmesi ve genel model performansının artırılması konusunda önemli bir katkı sağlamaktadır.



Şekil 4.9. Kombinasyon modellerin SOMYH değerleri

Makine öğrenmesi algoritmalarını eğitim süresi açısından kıyaslamak, algoritmaların karmaşıklığı, veri setinin büyüklüğü ve kullanılan donanım gibi faktörlere bağlı olarak değişkenlik gösterir. Genel olarak, çoklu doğrusal regresyon gibi düşük karmaşıklığa sahip algoritmalar, eğitim süreleri açısından daha hızlı sonuç verir. Bununla birlikte, derin ve topluluk öğrenme algoritmaları, özellikle büyük ve karmaşık veri setleri üzerinde çalıştığında daha uzun eğitim sürelerine ihtiyaç duyabilir. Şekil 4.10'da görülen dağıtım bölgesi kırımında her bir algoritmanın eğitim süreleri detaylı olarak incelenmiştir. Şekil incelendiğinde tüm bölgeler için en kısa eğitim süresine sahip olan algoritmaların sırasıyla MLR ve UKSB olduğu belirlenmiştir. Aksine, en uzun eğitim süresine sahip algoritma ise XGBoost olarak gözlemlenmiştir, bu durum tüm bölgeler için geçerlidir. Algoritmaların eğitim süreleri toplamı üzerinden bir değerlendirme yapıldığında, Delta bölgesi 2953 sn'lik eğitim süresi ile en uzun süreye sahip bölge olarak belirlenmiştir. Diğer bölgeler sırasıyla Gepa (2765 sn), Batıkar (1234 sn) ve Muğla (991 sn) olarak sıralanmıştır. Bu bulgular, her bir algoritmanın performansını belirleme ve bölgesel eğitim süreleri arasındaki farklılıkları anlama konusunda önemli bir içgörü sağlamaktadır.



Şekil 4.10. Dağıtım bölgesi kırılımında modellerin eğitim süreleri



5. SONUÇ VE DEĞERLENDİRME

Gün öncesi doğalgaz talep tahmini talep planlaması, operasyonel maliyetlerin azaltılması ve enerji arz güvenliğinin sağlanması gibi birçok kritik kararın temelini oluşturduğundan dolayı tahminlerin yüksek düzeyde doğruluk ile elde edilebilmesi büyük önem taşımaktadır.

Bu araştırma kapsamında, Türkiye'nin dört farklı dağıtım bölgesi (beş şehir) için 2015-2023 yılları arasında günlük doğalgaz tüketim verileri, lisans sahibi dağıtım şirketlerinden temin edilmiştir. Muğla bölgesi 2015 yılı tüketim verisinde çok sayıda eksik veri bulunduğu için ilgili yıl çalışma kapsamı dışında tutulmuştur. Ardından tüketim verisi ve kullanılarak gecikmeli, pencere ve zamansal özellikleri dikkate alınarak toplamda 33 farklı bağımsız değişken oluşturulmuştur. Daha basit bir özellik altkütmesi elde etmek için LASSO regresyonu bir özellik seçim tekniği olarak uygulanmıştır. Çalışmanın ikinci aşamasında Türkiye'nin 4 farklı bölgesinin günlük doğalgaz tüketimini öngörmek için ÇDR, topluluk öğrenme stratejileri ve UKSB olmak üzere üç farklı makine öğrenmesi tekniğinden yararlanılmıştır. Test verisi için Gega bölgesi %7,39 SOMYH değeri ile en yüksek tahmin doğruluğuna sahip bölge olurken onu sırasıyla Batıkar (SOMYH: %8,55), Delta (SOMYH: %10,54) ve Muğla (SOMYH: %24,28) bölgeleri takip etmiştir. Muğla bölgesinde gaz dağıtım faaliyetleri diğer bölgelere kıyasla çok daha sonra başladığı için doğalgaz tüketim paterni oldukça karmaşıktır. Muğla diğer bölgeler ile kıyaslandığında buradaki SOMHY değerinin çok daha yüksek olmasının sebebi, bu karmaşık doğal gaz tüketim örüntüsünden kaynaklanmaktadır.

Friedman sıra toplam testinin sonuçlarına göre, tüm bölgeler için, modellerin performansları arasında istatistiksel olarak anlamlı bir fark bulunmuştur. Ardından değerlendirme metrikleri göz önüne alınarak seçilen modeller ile diğer modeller nemeyi testi ile ikili olarak karşılaştırılmıştır. Yapılan ikili karşılaştırmalar sonucunda seçilen modeller ile bazı farklı alternatifleri arasında istatistiksel olarak anlamlı bir fark bulunmamıştır. Bu durum değerlendirme metriklerine göre seçilen modeller yerine farklı modellerin de seçilebileceğini göstermektedir.

Çalışmanın son aşamasında tahmin algoritmalarının tüm kombinasyonları ortalama, medyan, maksimum ve minimum değerleri ile birleştirilmiştir. Ardından en iyi makine

öğrenmesi topluluğu ve birleştirme metriği belirlenmeye çalışılmıştır. Gepa bölgesi dışındaki bölgelerde tahmin performansında kısmi bir artış gerçekleşmiştir.

Gelecekteki planlanması gereken çalışmalar olarak; Genetik algoritma, parçacık sürüsü optimizasyonu gibi doğadan ilham alan meta sezgisel tekniklerin kullanılması, bağımsız değişken seçimi veya hiper parametre ayarlaması yapıp, tahmin performansında ve hesaplama süresinde iyileşme elde edilmesi hedeflenmelidir. Buna ek olarak dalgacık dönüşümü, ampirik mod ayrıştırması gibi araçlar ile veri setindeki gürültü elimine edilerek de tahmin algoritmaları yeniden uygulanabilir.



KAYNAKLAR

1. İnternet: British Petroleum (BP). Statistical Review of World Energy 2023. URL: <https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html>, Son Erişim Tarihi: 09.05.2023.
2. İnternet: International Energy Agency 2023 (IEA). Natural Gas. URL: <https://www.iea.org/energy-system/fossil-fuels/natural-gas>. Son Erişim Tarihi: 09.05.2023.
3. Birol, F. (2020). *International Energy Agency Outlook Flagship Report*. IAE. Paris. 20-21.
4. İnternet: Birleşmiş Milletler (BM). Sürdürülebilir Kalkınma Amaçları. URL: <https://turkiye.un.org/tr/sdgs>, Son Erişim Tarihi: 23.04.2024.
5. Beyca, O.F., Ervural, B.C., Tatoglu, E., Ozuyar, P.G. Zaim, S. (2019). Using machine learning tools for forecasting natural gas consumption in the province of Istanbul. *Energy Economics*, 80, 937-949.
6. Enerji Piyasası Düzenleme Kurumu. (2022). *Doğalgaz Piyasası Sektör Raporu*. EPDK. Ankara. 18-19.
7. Enerji Piyasası Düzenleme Kurumu. (2022). *Doğalgaz Piyasası Sektör Raporu*. EPDK. Ankara. 70-71.
8. Sharma, V., Cali, Ü., Sardana, B., Kuzlu, M., Banga, D. Pipattanasomporn, M. (2021). Data-driven short-term natural gas demand forecasting with machine learning techniques, *Journal of Petroleum Science and Engineering*. 206.
9. Li, F., Zheng, H., Li, X. Yang, F. (2021). Day-ahead city natural gas load forecasting based on decomposition-fusion technique and diversified ensemble learning model. *Applied Energy*, 303.
10. Liu, J., Wang, S., Wei, N., Chen, X., Xie, H. Wang, J. (2021). Natural gas consumption forecasting: A discussion on forecasting history and future challenges. *Journal of Natural Gas Science and Engineering*, 90.
11. Özmen, A. (2021). Sparse regression modeling for short- and long-term natural gas demand prediction. *Annals of Operations Research*, 322(2), 921-946.
12. Taşpınar, F., Çelebi, N. Tutkun, N. (2013). Forecasting of daily natural gas consumption on regional basis in Turkey using various computational methods. *Energy and Buildings*, 56, 23-31.
13. Ozdemir, G., Aydemir, E., Olgun, M.O. Mulbay, Z. (2016). Forecasting of Turkey natural gas demand using a hybrid algorithm. *Energy Sources Part B: Economics Planning and Policy*, 11(4), 295-302.

14. Es, H.A. (2020). Monthly natural gas demand forecasting by adjusted seasonal grey forecasting model. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 43(1), 54-69.
15. Bilici, Z., Özdemir, D. Temurtaş, H. (2023). Comparative analysis of metaheuristic algorithms for natural gas demand forecasting based on meteorological indicators. *Journal of Engineering Research (Kuwait)*, 11(3), 259-265.
16. Ceylan, Z. (2023). Comparative analysis of deep learning and classical time series methods to forecast natural gas demand during COVID-19 pandemic. *Energy Sources, Part B: Economics, Planning and Policy*, 18(1).
17. Arık, O.A. (2019). Artificial bee colony algorithm to forecast natural gas consumption of Turkey. *SN Applied Sciences*, 1(10), 1138.
18. Sen, D., Günay, M.E. Tunç, K.M.M. (2019). Forecasting annual natural gas consumption using socio-economic indicators for making future policies. *Energy*, 173, 1106-1118.
19. Akpınar, M., Adak, M.F. Yumusak, N. (2017). Day-ahead natural gas demand forecasting using optimized ABC-based neural network with sliding window technique: The case study of regional basis in Turkey. *Energies*, 10(6).
20. Yucesan, M., Pekel, E., Celik, E., Gul, M. Serin, F. (2021). Forecasting daily natural gas consumption with regression, time series and machine learning based methods. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*.
21. Szoplik, J. (2015). Forecasting of natural gas consumption with artificial neural networks. *Energy*, 85, 208-220.
22. Rehman, A., Zhu, J. J., Segovia, J. Anderson, P.R. (2022). Assessment of deep learning and classical statistical methods on forecasting hourly natural gas demand at multiple sites in Spain. *Energy*, 244.
23. Hribar, R., Potočnik, P., Šilc, J. Papa, G. (2019). A comparison of models for forecasting the residential natural gas demand of an urban area. *Energy*, 167, 511-522.
24. Du, J., Zheng, J., Liang, Y., Lu, X., Klemeš, J.J., Varbanov, P.S., Shahzad, K., Rashid, M.I., Ali, A.M., Liao, Q. Wang, B. (2022). A hybrid deep learning framework for predicting daily natural gas consumption. *Energy*, 257.
25. Song, J., Zhang, L., Jiang, Q., Ma, Y., Zhang, X., Xue, G., Shen, X. Wu, X. (2022). Estimate the daily consumption of natural gas in district heating system based on a hybrid seasonal decomposition and temporal convolutional network model. *Applied Energy*, 309.
26. Laib, O., Khadir, M.T. Mihaylova, L. (2019). Toward efficient energy systems based on natural gas consumption prediction with LSTM Recurrent Neural Networks. *Energy*, 177, 530-542.

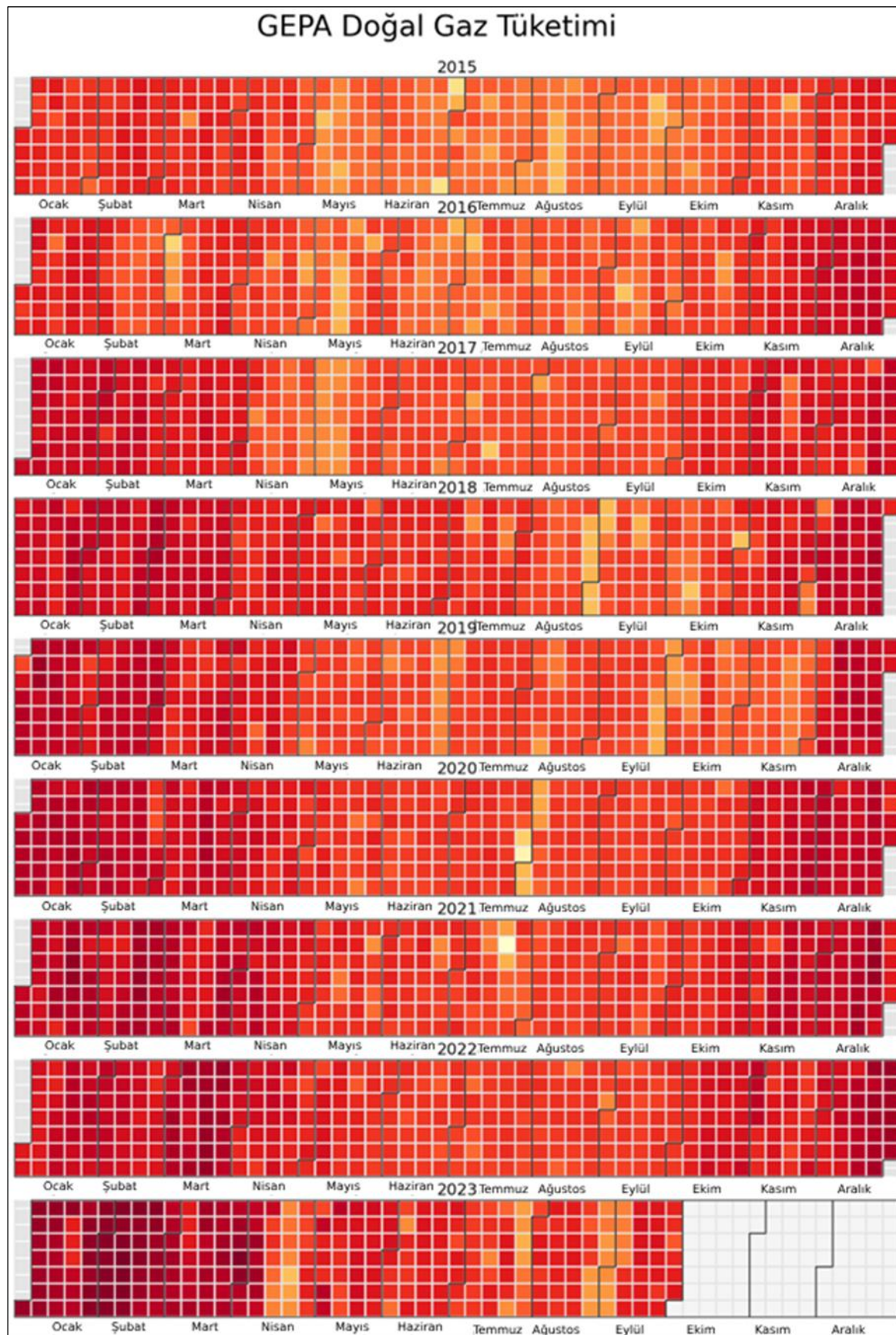
27. Wei, N., Yin, L., Li, C., Liu, J., Li, C., Huang, Y. Zeng, F. (2022). Data complexity of daily natural gas consumption: Measurement and impact on forecasting performance. *Energy*, 238.
28. Wei, N., Li, C., Peng, X., Li, Y. Zeng, F. (2019). Daily natural gas consumption forecasting via the application of a novel hybrid model. *Applied Energy*, 250, 358-368.
29. Panapakidis, I.P. Dagoumas, A.S. (2017). Day-ahead natural gas demand forecasting based on the combination of wavelet transform and ANFIS/genetic algorithm/neural network model. *Energy*, 118, 231-245.
30. Su, H., Zio, E., Zhang, J., Xu, M., Li, X. Zhang, Z. (2019). A hybrid hourly natural gas demand forecasting method based on the integration of wavelet transform and enhanced Deep-RNN model. *Energy*, 178, 585-597.
31. Kunapuli, G. (2022), Ensemble Methods for Machine Learning (6. Baskı). New York; Manning Publications, 8.
32. Altunkaynak, B. (2017). *Veri Madenciliği Yöntemleri ve R Uygulamaları* (1. Baskı). Ankara: Seçkin Yayıncılık, 123.
33. Drucker, H. (1997). Improving Regressors using Boosting Techniques. *Proceedings of the Fourteenth International Conference on Machine Learning*, 9, 107-115.
34. Newbold, P., Carlson, W.L. Thorne, B.M. (2017). *Statistics for Business and Economics*, Boston; Pearson Education
35. Gaurav Kapoor, N.W. (2023). Electricity price forecasting in New Zealand: A comparative analysis of statistical and machine learning models with feature selection. *Applied Energy*, 347.
36. Bentejac, C., Csörgö, A., Martinez-Munoz, G. (2020). A Comparative Analysis of XGBoost. *Artificial Intelligence Review*, 54.
37. Ke, G., Men, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems*, 3149-3157.
38. Kılınç, D. ve Başeğmez, N. (2019). *Uygulamalarla Veri Bilimi* (3. Baskı). İstanbul: Abaküs Yayınları, 144.
39. Ismail, A. A., Wood, T., Bravo, H. C. (2018). Improving Long-Horizon Forecasts with Expectation-Biased LSTM Networks. *Knowledge Discovery and Data Mining*, August
40. Bandara, K., Bergmeir, C. Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, 140.

41. Bandara, K., Hewamalage, H., Liu, Y.-H., Kang, Y. Bergmeir, C. (2021). Improving the accuracy of global forecasting models using time series data augmentation. *Pattern Recognition*, 120.
42. Godahewa, R., Webb, G.I., Schmidt, D. Bergmeir, C. (2023). SETAR-Tree: a novel and accurate tree algorithm for global time series forecasting. *Machine Learning*, 112(7), 2555-2591.
43. Hewamalage, H., Bergmeir, C. Bandara, K. (2021). Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1), 388-427.
44. Aytaç, M. (2017). *Matematiksel İstatistik* (6. Baskı). Bursa: Ezgi Kitapevi, 461-462.
45. He, Y., Qin, Y., Wang, S., Wang, X. Wang, C. (2019). Electricity consumption probability density forecasting method based on LASSO-Quantile Regression Neural Network. *Applied Energy*, 233, 565-575.
46. Erol, B., İnkaya, T. (2024). Long short-term memory network based deep transfer learning approach for sales forecasting. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 12, 191-202.

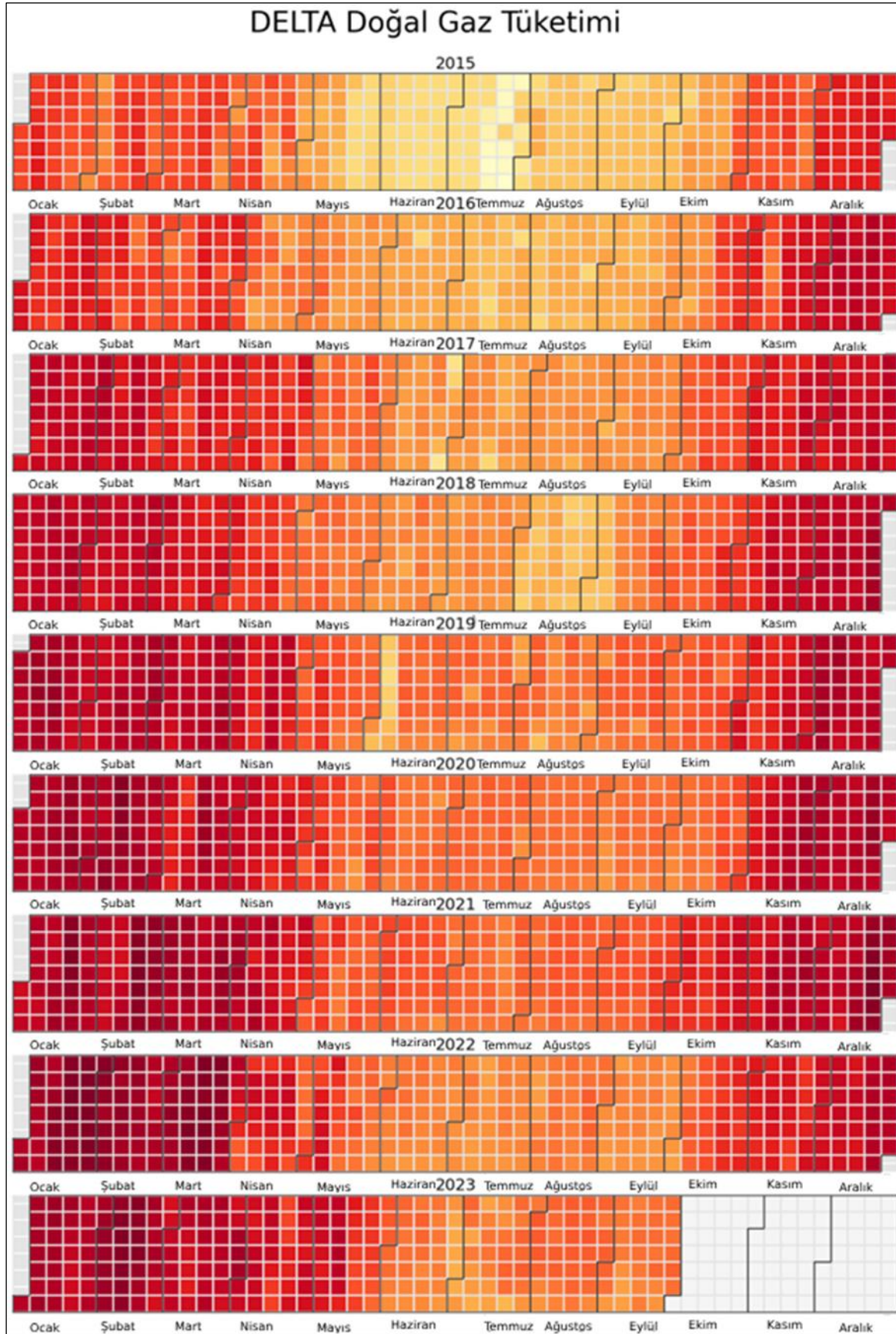


EKLER

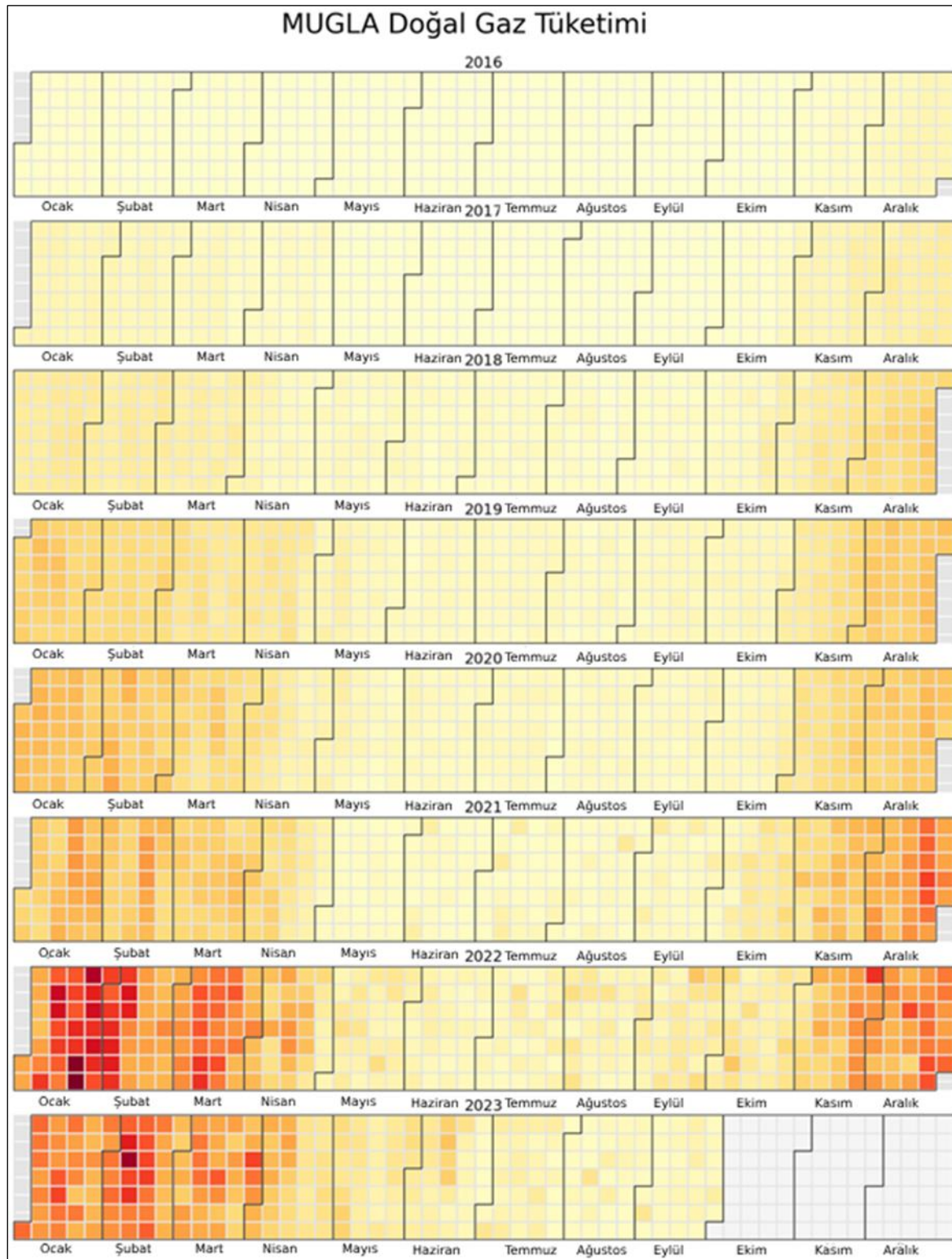
EK-1. Bölgelerin tüketim ısı grafikleri



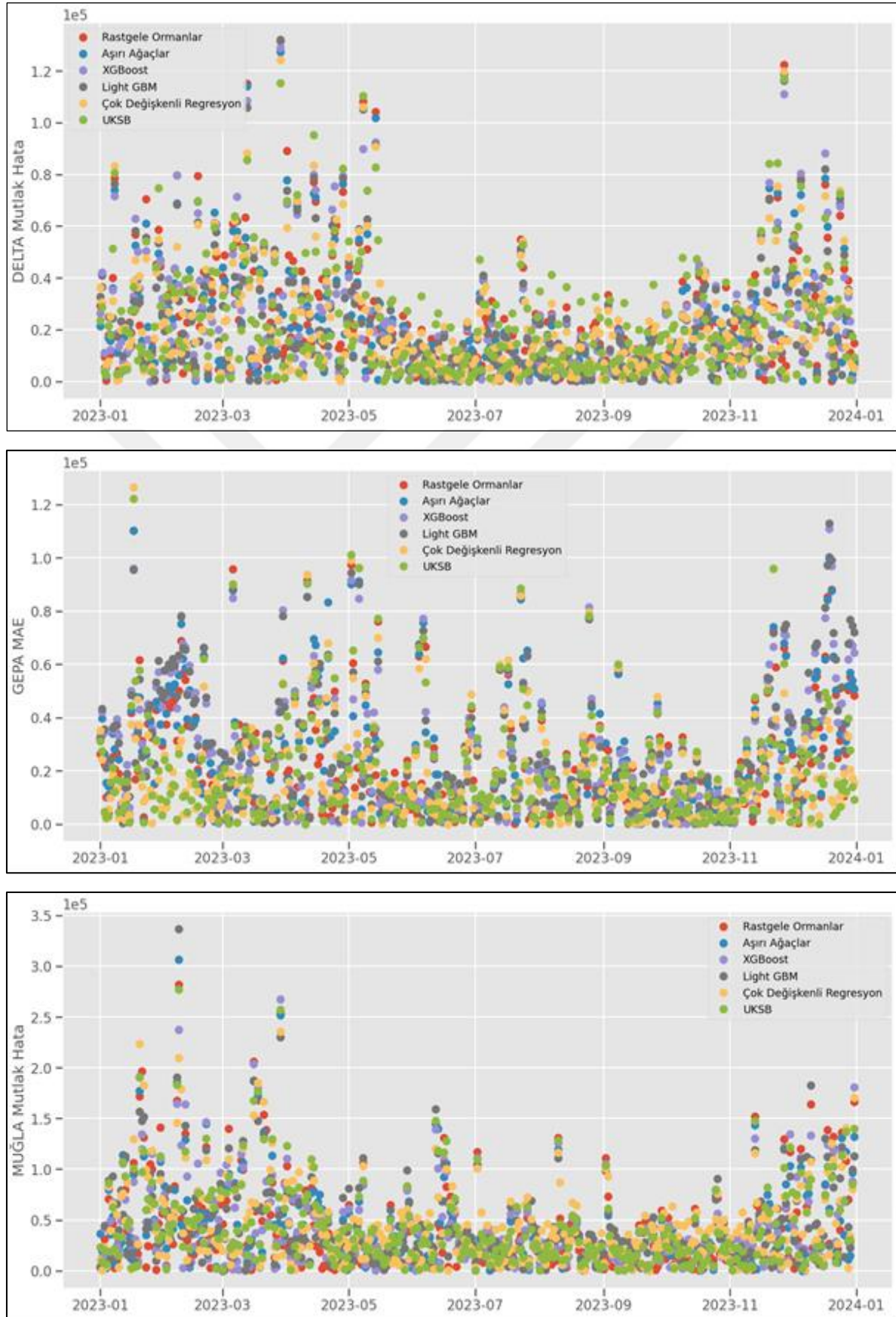
EK-1. (devam) Bölgelerin tüketim ısı grafikleri



EK-1. (devam) Bölgelerin tüketim ısı grafikleri



EK-2. Bölgelerin tüketim mutlak hata grafikleri





Gazili olmak ayrıcalıktır