

**T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE ANORMAL İÇME
SUYU TÜKETİMLERİNİN TESPİT EDİLMESİ ve TAHMİN
MODELLERİNİN GELİŞTİRİLMESİ**

YÜKSEK LİSANS TEZİ

İsmail GÜNEY

Bilişim Sistemleri Mühendisliği Anabilim Dalı

TEMMUZ 2023

**T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE ANORMAL İÇME
SUYU TÜKETİMLERİNİN TESPİT EDİLMESİ ve TAHMİN
MODELLERİNİN GELİŞTİRİLMESİ**

YÜKSEK LİSANS TEZİ

İsmail GÜNEY

Bilişim Sistemleri Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. İhsan Hakan SELVİ

TEMMUZ 2023

İsmail GÜNEY tarafından hazırlanan “Makine Öğrenmesi Yöntemleriyle Anormal İçme Suyu Tüketimlerinin Tespit Edilmesi Ve Tahmin Modellerinin Geliştirilmesi” adlı tez çalışması 05.07.2023 tarihinde aşağıdaki jüri tarafından oy birliği ile Sakarya Üniversitesi Fen Bilimleri Enstitüsü Bilişim Sistemleri Mühendisliği Anabilim Dalı Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı : **Prof. Dr. Numan ÇELEBİ**
Sakarya Üniversitesi

Jüri Üyesi : **Doç. Dr. İhsan Hakan SELVİ** (Danışman)
Sakarya Üniversitesi

Jüri Üyesi : **Prof. Dr. Orhan ER**
Bakırçay Üniversitesi



ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Sakarya Üniversitesi Fen Bilimleri Enstitüsü Lisansüstü Eğitim-Öğretim Yönetmeliğine ve Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesine uygun olarak hazırlamış olduğum “MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE ANORMAL İÇME SUYU TÜKETİMLERİNİN TESPİT EDİLMESİ ve TAHMİN MODELLERİNİN GELİŞTİRİLMESİ” başlıklı tezin bana ait, özgün bir çalışma olduğunu; çalışmamın tüm aşamalarında yukarıda belirtilen yönetmelik ve yönergeye uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, bu tezi başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve 20.04.2016 tarihli Resmi Gazete’de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince Sakarya Üniversitesi’nin abonesi olduğu intihal yazılım programı kullanılarak Enstitü tarafından belirlenmiş ölçütlere uygun rapor alındığını, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun ortaya çıkması halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

(...../...../2023).

İsmail GÜNEY





*Çalışmam süresince destek ve katkılarını esirgemeyen, umut ve ümitlerini bir
nebze olsun eksiltmeyen değerli aileme...*



TEŞEKKÜR

Yüksek lisans eğitimim boyunca değerli bilgi ve deneyimlerinden yararlandığım, her konuda bilgi ve desteğini almaktan çekinmediğim, araştırmanın planlanmasından yazılmasına kadar tüm aşamalarında yardımlarını esirgemeyen, teşvik eden, aynı titizlikte beni yönlendiren değerli danışman hocam Doç. Dr. İhsan HAKAN SELVİ'ye teşekkürlerimi sunarım.

Abone verileri konusunda kurum desteği ve katkısını esirgemeyen Kayseri Su Kanalizasyon İdaresi'ne ve Genel Müdür Doç. Dr. Özgür ÖZDEMİR'e ayrıca su ile ilgili birikimlerinden istifade ettiğim her konuda bilgi ve desteğini esirgemeyen kıymetli hocam Prof. Dr. Mahmut FIRAT'a teşekkür ederim.

İsmail GÜNEY



İÇİNDEKİLER

Sayfa

TEŞEKKÜR	ix
İÇİNDEKİLER	xi
KISALTMALAR	xiii
SİMGELER	xix
TABLO LİSTESİ	xxi
ŞEKİL LİSTESİ	xxiii
ÖZET	xxv
SUMMARY	xxvii
1. GİRİŞ.....	1
1.1. Tezin Kapsamı.....	9
1.2. Tezin Amacı	9
2. KAYNAK ARAŞTIRMASI	11
2.1. Su Tüketimi ve Suyun Durumu.....	11
2.2. Anomali Analizleri	29
3. MATERYAL VE YÖNTEM	35
3.1. Kullanılan Yöntemin Şeması	35
3.2. Çalışma Alanı ve Veriseti	37
3.3. Veri Önışlemleri	42
3.3.1. Veri hazırlama	42
3.3.2. Veri dönüştürme.....	44
3.3.3. Veri birleştirme	48
3.3.4. Gürültü tanımlama	48
3.3.5. Kayıp değerler.....	48
3.3.6. Öznitelik seçimi	51
3.3.6.1. Entropi.....	52
3.3.6.2. Bilgi kazancı.....	53
3.3.6.3. Kazanç oranı.....	54
3.3.6.4. Simetrik belirsizlik katsayısı	54
3.3.6.5. Pearson korelasyon katsayısı.....	54
3.3.6.6. F-skör	55
3.3.6.7. RF sınıflandırma	55
3.3.7. Veri etiketleme	57
3.4. Anomali analizleri	57
3.4.1. Z-Score anomali analizi	59
3.4.2. TOL anomali analizi	59
3.4.3. IF anomali analizi.....	60
3.4.4. COPOD anomali analizi.....	60
3.4.5. MAD anomali analizi.....	62
3.4.6. LOF anomali analizi.....	63

3.4.7. EE anomali analizi	65
3.5. Performans Ölçümü.....	66
3.5.1. Doğruluk.....	67
3.5.2. Duyarlılık	67
3.5.3. Kesinlik	68
3.5.4. Hassasiyet.....	68
3.5.5. Özgüllük	68
3.5.6. MAE	68
3.5.7. MSE.....	69
3.6. ML tüketim tahmin modelleri	69
4. MODELİN UYGULANMASI	71
4.1. Modelin Uygulanması ve Analizi.....	71
5. ARAŞTIRMA BULGULARI	85
5.1. Anomali Analiz Sonuçları	85
5.2. ML Analiz Sonuçları	93
6. SONUÇ VE ÖNERİLER.....	95
KAYNAKLAR.....	99
EKLER.....	107
ÖZGEÇMİŞ.....	117

KISALTMALAR

%AA	: Ortalama Yaklaşıklık Oranı
AARE	: Ortalama Mutlak Bağlı Hata
AAT	: Atıksu Arıtma Tesisi
ABC	: Yapay Arı Kolonisi
ABYS	: Abone Bilgi Yönetim Sistemi
ACC	: Doğruluk
ADASU	: Adapazarı Su ve Kanalizasyon İdaresi
ANFIS	: Uyarlamalı-Ağ Tabanlı Bulanık Çıkarım Sistemi
ANN	: Yapay Sinir Ağı
ANOVA	: Varyans Analizi
ARCH	: Otoregresif Koşullu Değişen Varyans
ARIMA	: Otoregresif Bütünleşik Hareketli Ortalama
ASAT	: Antalya Su ve Atıksu İdaresi
ASKİ	: Ankara Su ve Kanalizasyon İdaresi
AUC	: ROC Eğrisi Altındaki Alan
B-DNN	: İkili Sınıflandırılmalı DNN
BFGS	: Broyden-Fletcher-Goldfarb-Shannon
BİST	: Borsa İstanbul
BM	: Birleşmiş Milletler
BOİ	: Biyolojik Oksijen İhtiyacı
BOTAŞ	: Boru Hatları İle Petrol Taşıma A.Ş.
BPNN	: Geri Yayılım Sinir Ağı
CAAD	: Güvene Duyarlı Anomali Tespiti
CART	: Sınıflandırma ve Regresyon Ağacı
CBS	: Coğrafi Bilgi Sistemi
CE	: Verimlilik Katsayısı
CGPB-NN	: Eşlenik Gradyan Powell-Beale NN

CHI	: Ki-Kare
CNN	: Evrişimsel Sinir Ağları
COPOD	: Copula-Tabanlı Aykırı Değer Tespiti
CPA	: Çapraz Tahmin Doğruluğu
CV	: Varyasyon Katsayısı
DAN2	: Dinamik ANN
DDoS	: Dağıtık DoS
DLM	: Dinamik Liner Model
DMA	: Ölçülebilir Alt Bölge
DNN	: Derin Sinir Ağı
DoS	: Hizmet Reddi
DP-NN	: Diferansiyel Polinom NN
DSİ	: Devlet Su İşleri Genel Müdürlüğü
DT	: Karar Ağaçları
DWT	: Ayrık WT
EBYS	: Elektronik Belge Yönetim Sistemi
EE	: Eliptik Zarf
Eİ	: Elektriksel İletkenlik
ELM	: Aşırı Öğrenme Makineleri
EPA	: Çevre Koruma Ajansı
EPDK	: Enerji Piyasası Düzenleme Kurumu
EV	: Açıklanan Varyans
FB-NN	: Geri Beslemeli NN
FFBP-NN	: İleri Beslemeli Geri Yayılım NN
FF-NN	: İleri Beslemeli NN
FL	: Bulanık Mantık
GA	: Genetik Algoritma
GARCH	: Genelleştirilmiş ARCH
GB-LR	: Geri Beslemeli LR
GB	: Gradient Boosting
GEE	: Google Earth Motoru
GEP	: Gen İfade Programlama

GGSO	: Gelir Getirmeyen Su Oranı
GMM	: Gauss Karışım Modeli
GR	: Kazanç Oranı
GR-NN	: Genelleştirilmiş Regresyon Sinir Ağı
GROM	: Altın Oran Optimizasyon
GRU	: Kapılı Yineleyen Birim
GSMH	: Gayrisafı Millî Hasıla
GSYH	: Gayrı Safı Yurtiçi Hâsıla
IF	: İzolasyon Ormanı
IG	: Bilgi Kazancı
IOA	: Anlaşma İndeksi
IOT	: Nesnelerin İnterneti
IQR	: Çeyrekler Arası Aralık
IWA	: Uluslararası Su Birliği
IWQI	: Sulama Suyu Kalite İndeksi
İLBANK	: İller Bankası A.Ş.
İSKİ	: İstanbul Su ve Kanalizasyon İdaresi
İSU	: Kocaeli Su ve Kanalizasyon İdaresi
KASKİ	: Kayseri Su ve Kanalizasyon İdaresi
KDE	: Çekirdek Yoğunluk Tahmini
KNN	: K-En Yakın Komşular
KNORA-E	: K-Nearest Oracles Eliminate
KNORA-U	: K-Nearest Oracles Union
KOİ	: Kimyasal Oksijen İhtiyacı
KOSKİ	: Konya Su ve Kanalizasyon İdaresi
LASSO	: En Az Mutlak Büzülme Seçici Operatörü
LCA	: Yerel Sınıf Doğruluğu
LDA	: Lineer Diskriminant Analizi
LJR	: Lojistik Regresyon
LM-NN	: Levenberg-Marquardt NN
LOF	: Yerel Aykırı Değer Faktörü
LR	: Doğrusal Regresyon

LS-SVM	: En Küçük Kareler SVM
LSTM	: Çift Yönlü Uzun Kısa Süreli Bellek
L-SVM	: Doğrusal Çekirdek Fonksiyonlu SVM
MAD	: Medyan Mutlak Sapma
MAE	: Ortalama Mutlak Hata
MAE-MAD	: Ortalama Mutlak Sapma
MAPE	: Ortalama Mutlak Yüzde Hata
MARE	: Ortalama Mutlak Bağlı Hata
MARS	: Çok Değişkenli Uyarlanabilir Regresyon Eğrileri
MASE	: Ortalama Mutlak Ölçekli Hata
MBM	: Meteoroloji Bölge Müdürlüğü
MCE	: Modifiye Verimlilik Katsayısı
MCMC	: Monte Carlo Markov Zinciri
MdAPE	: Medyan Mutlak Yüzde Hata
MD-NN	: Modüler NN
M-DNN	: Çoklu Sınıflandırılmalı DNN
META-DES	: Dinamik Topluluk Seçimi İçin Meta-Öğrenme
MGM	: Meteoroloji Genel Müdürlüğü
MIOA	: Modifiye Anlaşma İndeksi
missForest	: Parametrik Olmayan RF Tabanlı Atama
MLP	: Çok Katmanlı Algılayıcı
MLP-NN	: MLP Sinir Ağı
MLR	: Çoklu Doğrusal Regresyon
MÖ	: Makine Öğrenmesi
MRAE	: Ortalama Mutlak Bağlı Hata
MSE	: Ortalama Kare Hata
NASA	: Ulusal Havacılık ve Uzay Dairesi
NB	: Gaussian Naive Bayes
NLA	: Doğrusal Olmayan Yerel Yaklaşım
NMSE	: Normalleştirilmiş Ortalama Kare Hata
NN	: Sinir Ağı
NSE	: Nash–Sutcliffe Model Verimlilik Katsayısı

ODDS	: Aykırı Değer Algılama Verisetleri
OECD	: Ekonomik Kalkınma ve İş Birliği Örgütü
OLA	: Genel-Yerel Doğruluk
One-class SVM	: Tek-sınıf SVM
PCA	: Temel Bileşen Analizi
PICP	: Tahmin Aralığı Kapsama Yüzdesi
PK-SVM	: Poli Çekirdek Fonksiyonlu SVM
PPR	: Projeksiyon İzleme Regresyonu
Probe	: Bilgi Tarama
PSO	: Parçacık Sürüsü Optimizasyonu
R2L	: Yönetici Hesabı ile Yerel Oturum Açma
R4MS4E	: Dördüncü Kök Ortalama Dördüncü Kuvvet Hata
Rank	: Modifiye Sınıf Sıralaması
RBF-NN	: Radyal Temel Fonksiyonlu NN
RBF-SVR	: Radyal Çekirdek Fonksiyonlu SVR
RBK-SVM	: Radyal Çekirdek Fonksiyonlu SVM
Recall	: Duyarlılık
RF	: Rastgele Orman
RFE	: Tekrarlayan Özellik Eleme
rMB	: Bağlı Ortalama Sapma
RMSE	: Kök Ortalama Kare Hata
RNN	: Yinelemeli NN
ROC	: Alıcı İşletim Karakteristiği
RP-NN	: Esnek Geri Yayılım NN
rRMS	: Bağlı Ortalama Karekök
RTU	: Uzak Terminal Ünitesi
SARIMA	: Dönemsel ARIMA
SCADA	: Yönetmel Kontrol ve Veri Toplama
SCEM-UA	: Karıştırılmış Karmaşık Evrim Metropolis Algoritması
SCORM	: Paylaşılabilir İçerik Nesnesi Referans Modeli
SEDAŞ	: Sakarya Elektrik Dağıtım A.Ş.
SKH	: Sürdürülebilir Kalkınma Hedefleri

SMOTE	: Sentetik Azınlık Aşırı Örnekleme
SNN-CAD	: Benzerlik Tabanlı En Yakın Komşu Uyumlu Anomali Dedektörü
SPT-CAD	: Tek Noktalı Yörünge Uyumlu Anomali Dedektörü
S-RNN	: Basit RNN
SU	: Simetrik Belirsizlik Katsayısı
SVM	: Destek Vektör Makineleri
SVR	: Destek Vektör Regresyon
ŞUSKİ	: Şanlıurfa Su ve Kanalizasyon İdaresi
TEİAŞ	: Türkiye Elektrik İletim A.Ş.
TESKİ	: Tekirdağ Su ve Kanalizasyon İdaresi
TOL	: Tukey Aykırı Değer Etiketleme
TS	: Zaman Serisi
TÜFE	: Tüketici Fiyat Endeksi
TÜİK	: Türkiye İstatistik Kurumu
U2R	: Kullanıcı Hesabını Yönetici Hesabına Yükseltme
VMD	: Varyasyon Modu Ayrıştırma
WT	: Dalgacık Dönüşümü
WT-NN	: Dalgacık Dönüşümlü NN
YZ	: Yapay Zekâ
Z-Skor	: Z -Skor

SİMGELER

c	: Tutarlılık Katsayısı
d	: Gözlem Noktaları Arası Mesafe
$\widehat{F}_d$	: Ampirik Kopula Gözlemleri
E	: Verimlilik
$H(X)$	: Girdi Entropi
$H(Y)$	: Hedef Entropi
Imp	: Öznitelik Önem Puanı
K	: İçmesuyu Tüketimi Yapan Kullanıcı
n	: Gözlem Sayısı
p	: Boyut (Öznitelik) Sayısı
R	: Korelasyon Katsayısı
R^2	: Determinasyon Katsayısı
S	: Entropi
t	: Zaman
X	: Bağımsız Değişken Gözlem Değeri
y	: Bağımlı Değişken Gerçek Değeri
$\hat{y}$	: Bağımlı Değişken Tahmin Değeri
μ	: Gözlemlerin Ortalaması
σ	: Standart Sapma
$\parallel$	: Lojik İşlemi



TABLO LİSTESİ

Sayfa

Tablo 1.1. Tarihte su ile ilgili yapılan çatışma nedeni ve sayıları.	1
Tablo 1.2. Kentsel paydaşların “Sürdürülebilir Kentsel Su” hizmeti sunmaları için dört aşamalı eylem planı.	2
Tablo 1.3. AI uygulamalarının temel alanları ve Birleşmiş Milletler Sürdürülebilir Kalkınma Hedefleri.	4
Tablo 1.4. Suyun hane içinde nihai kullanım yerleri ve oranları.	6
Tablo 1.5. Su tasarrufu ile ilgili öneriler.	6
Tablo 2.1. Aykırı değerlerin literatüre dayalı işleme teknikleri.	30
Tablo 3.1. KASKİ okuma dönem sayılarına göre abone sayısı tablosu.	38
Tablo 3.2. Birleştirilmiş öznitelik bilgileri tablosu.	43
Tablo 3.3. Tüketim yapılan günlerin ilgili aylara dağılımı tablosu.	44
Tablo 3.4. Verisetindeki endeks hareketleri ile sayaç bilgisi dönüşüm tablosu.	45
Tablo 3.5. Verisetindeki sözleşme numarası ile kullanıcı bilgisi dönüşüm tablosu.	46
Tablo 3.6. Verisetindeki kademe tutar bilgi alanları ve dönüşüm tablosu.	46
Tablo 3.7. Birleştirilmiş verisetinde boş hücrelerin kontrol tablosu.	49
Tablo 3.8. Aykırılık puanlarına göre sınıflandırma etiketleri tablosu.	57
Tablo 3.9. İçmesuyu tüketim anomali tespit modellerine ait hata matrisi tanımlama tablosu.	67
Tablo 4.1. Öznitelik seçiminde kullanılan temel istatistikî değerler tablosu.	71
Tablo 4.2. Veriseti öznitelik bilgileri seçim yöntemlerinin karşılaştırılması.	73
Tablo 4.3. Veriseti öznitelik seçim puanlanması.	74
Tablo 4.4. Alt verisetler ve seçimi yapılan öznitelikler tablosu.	75
Tablo 4.5. Çalışma kapsamında kullanılan anomali analizi algoritmaları.	77
Tablo 4.6. Anomali analizi sonucunda tespit edilen gözlemlerin toplam aykırılık değerleri ve sınıfları.	81
Tablo 4.7. İçmesuyu tüketim gözlem noktaları mesafe değerleri.	82
Tablo 4.8. Geliştirilen modeller ve performans metrik değerleri.	83
Tablo 4.9. Anomali analizi sonucunda tespit edilen gözlemlerin toplam aykırılık değerleri ve sınıfları.	84
Tablo 5.1. Öznitelik seçim yöntemleriyle elde edilerek oluşturulan alt verisetleri ve bütün özniteliklerin olduğu veriseti sayıları tablosu.	86
Tablo 5.2. Öznitelik seçim yöntemleriyle aynı özniteliklerin seçilme sayıları ve oranları tablosu.	86
Tablo 5.3. Alt verisetlerinde öznitelik seçim yöntemlerinin seçim adetleri ve oranları tablosu.	87
Tablo 5.4. Sayaç, kullanıcı, kademe ve ödeme durumları özet tablosu.	90
Tablo 5.5. Model kapsamında geliştirilen anomali tespit model sayıları tablosu.	90
Tablo 5.6. Her bir aboneye ait doğruluk performans metriği maksimum olan en iyi model tablosu.	91

Tablo 5.7. İçme suyu tüketim sınıf tahmin modelleri performans tablosu.....	93
Tablo A.1. Literatürde geçen aykırı değer tanımları.	109



ŞEKİL LİSTESİ

Sayfa

Şekil 1.1. Konut suyu koruma modeli.	8
Şekil 3.1. Çalışma kapsamında önerilen model şeması.	36
Şekil 3.2. Uygulama alanı.	37
Şekil 3.3. KASKİ geçmiş yıllara ait mesken abonesi sayaç okuma sayıları.	38
Şekil 3.4. Abonelere ait analizi yapılan verilerin tutulduğu ilişkisel veritabanı modeli.	39
Şekil 3.5. Abone ve tahakkuk tablosuna göre sayacı en az bir defa okunmuş abone sayısı grafiği.	40
Şekil 3.6. Mesken ve diğer abone türlerine göre tahakkuk sayıları grafiği.	41
Şekil 3.7. Tahakkuk gören abonelere ait tarife türü sayıları grafiği.	41
Şekil 3.8. Çalışma kapsamında kullanılan veriler.	42
Şekil 3.9. Bağımsız değişkenlerin bağımlı değişkene mesafelerinin minkowski ile ölçüm gösterimi.	47
Şekil 3.10. Verisetinin Arttırılmış Dickey-Fuller yöntemi ile analiz grafiği.	50
Şekil 3.11. Veriseti dönem ortalaması ve dönem standart sapma grafiği.	50
Şekil 3.12. Öznitelik seçme yöntemleri şeması.	51
Şekil 3.13. Öznitelik seçimi yöntemleri kullanılarak oluşturulan verisetleri şeması.	52
Şekil 3.14. Entropi ve bilgi kazancı arasındaki ilişki.	53
Şekil 3.15. Topluluk öğrenmesi yöntemi ile aykırı tüketim tespit şeması.	58
Şekil 3.16. Anomali analizi model şeması.	58
Şekil 3.17. Kutu Grafiği yönteminin hesaplanması ve görselleştirilmesi.	59
Şekil 3.18. İki farklı gözlem noktası arasındaki mesafe gösterimi.	63
Şekil 3.19. Komşu sayısı 3 olacak şekilde gözlem noktaları k-mesafesi ve erişebilirlik mesafesi gösterimi.	64
Şekil 3.20. Eliptik zarf yöntemi ile aykırı gözlem noktaları tespiti.	65
Şekil 3.21. Topluluk öğrenmesi yöntemi ile aykırı tüketim tespit şeması.	69
Şekil 3.22. ML tüketim sınıfı tahmin modeli şeması.	70
Şekil 4.1. Kümülatif varyans oranına göre öznitelik sayıları grafiği.	72
Şekil 4.2. Veriseti öznitelik bilgilerinin korelasyon matrisi.	75
Şekil 4.3. Veriseti öznitelik bilgilerinin RF yöntemi ile elde edilen önem puanları grafiği.	76
Şekil 4.4. Z-Score anomali analizi ile tespit edilen anormal tüketim grafiği.	77
Şekil 4.5. TOL-Tukey Outlier Labeling algoritması ile tespit edilen anormal tüketim grafiği.	78
Şekil 4.6. IF-Isolating Forest Outlier Detection algoritması ile tespit edilen anormal tüketim grafiği.	78
Şekil 4.7. COPOD-Copula Based Outlier Detection algoritması ile tespit edilen anormal tüketim grafiği.	79

Şekil 4.8. MAD-The Median Absolute Deviation algoritması ile tespit edilen anormal tüketim grafiği.	79
Şekil 4.9. LOF-Local Outlier Factor algoritması ile tespit edilen anormal tüketim grafiği.	80
Şekil 4.10. EE-Eliptic Envelope Anomaly Dedection algoritması ile tespit edilen anormal tüketim grafiği.	80
Şekil 4.11. Etiketlenmiş gözlem noktalarına ait normal ve anormal tüketim durumları grafiği.	82
Şekil 5.1. Sayaç okuma dönem periyot grafiği.	85
Şekil 5.2. Abonelerin gerçek tüketim değerleriyle tespit edilen normal ve anormal tüketim sınıfları grafiği.	88
Şekil 5.3. Abonelerin minkowski tüketim değerleriyle tespit edilen normal ve anormal tüketim sınıfları grafiği.	89
Şekil 5.4. Tüketim süresince abonelerin kullandıkları sayaç durum grafiği.	89
Şekil 5.5. Tüketim süresince sözleşme yapan kullanıcı durum grafiği.	90
Şekil 5.6. Maksimum doğruluk sağlayan anomali analiz yöntem oranları grafiği.	91
Şekil 5.7. Doğruluk performansı %90 ve üzeri olan tahmin modellerinde kullanılan veriseti oranları grafiği.	92
Şekil B.1. Tarihte su ile ilgili yapılmış çatışmaların dünya haritasında gösterimi. .	115

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE ANORMAL İÇME SUYU TÜKETİMLERİNİN TESPİT EDİLMESİ ve TAHMİN MODELLERİNİN GELİŞTİRİLMESİ

ÖZET

Bu çalışmada, içme suyu gibi önemli bir ihtiyacın hane halkı tarafından tüketiminde belirli bir düzen olabileceği gibi, farklı etkenlere bağlı olarak düzensiz tüketimin de olabileceği öngörülmektedir. Artan nüfus, sınırlı içme suyu kaynakları, gelişen alt yapı ve teknoloji, içme ve kullanma suyuna olan talebi artırmıştır. Bu talebi karşılamak için alternatif su kaynakları arayışları olsa da mevcut suların israf edilmemesi ve daha verimli kullanılmasıyla taleplerin karşılanabileceği öngörülmektedir.

Yapay zekanın (AI) bir alt dalı olan makine öğrenmesi (ML) yöntemleriyle geçmiş dönemlerdeki içme suyu tüketim verileri analiz edilmiş, olağan ve olağan dışı tüketim davranış modelleri çıkarılmıştır. İçme suyu tüketimlerinde meydana gelebilecek anormal tüketimlerin tespit edilmesi ve abonelerin bu konuda bilgilendirilmesi ile hane içindeki tüketimlerin normal tüketim aralığı içinde kalmasının sağlanacağı öngörülmektedir. Günümüz bilişim dünyasında toplanan, kaydedilen ve işlenen veri miktarının önemli ölçüde artmasına rağmen, tam olarak analizlerinin zaman ve imkân açısından zor olduğu bilinmektedir.

Bu çalışmada Kayseri ili genelinde 2006 – 2022 (ilk 6 ay) tarihleri arasında sayaç endeks okuması 160 dönemden fazla olan 8.224 adet mesken abonesine ait abone, sayaç, tüketim ve tarife verileri dikkate alınmıştır. Veriler konumsal olarak abone temelinde birleştirilmiş, 41 öznitelikli bir veri kümesi oluşturulmuştur.

Hazırlanan veri kümesi; veri entegrasyonu, veri dönüştürme, gürültü tanımlama ve temizleme, tespit edilen kayıp değerleri tamamlama veya temizleme gibi veri ön işlemleri kullanılarak 24 özellikli bir verisete dönüştürülmüştür.

Temel istatistiksel değerler ve öznitelik seçimi yöntemleri kullanılarak içme suyu tüketimini etkileyen özniteliklerin her birinin puanları hesaplanmıştır. Özniteliklerin içme suyu tüketimini etkileyip etkilemediği, etkiliyorsa ne kadar etkilediği tespit edilmiştir. Öznitelik seçimlerinde amacı karşılamaya yeterli öznitelik adetinin tespiti açıklanan varyans kriteri (Variance Explained Criteria) oranı %90 alınarak belirlenmiştir. Böylece en az 6 en fazla 11 özneneliğin yeterli olacağı ve her öznitelik seçim yönteminde bu sayıda özneneliğin kalmasını sağlayacak eşik değerler belirlenmiştir.

Aykırı değer tüketim tespit modellerinde kullanılacak en uygun alt veriseti ve tüketimi etkileyen özniteliklerin seçimi için bilgi kazancı (IG), kazanç oranı (GR), simetrik belirsizlik katsayısı (SU), pearson korelasyon katsayısı (r), f-skoru ve rastgele orman (RF) öznitelik seçim yöntemleri kullanılarak belirlenen eşik değerler üzerinden 6 farklı alt veriseti elde edilmiştir. Ayrıca en az 3 öznitelik seçim yöntemiyle seçilen özniteliklerden oluşan 7. alt veriseti elde edilmiştir.

Her bir abone için sayaç sayısı, kullanıcı sayısı, tüketim ölçüm günleri, tüketim tutarları, fatura tutarları ve ödeme durumu verileri farklı olduğundan seçilen öznitelikler ve oluşan alt verisetleri de aboneye özel olarak elde edilmiştir. Her bir abone için veri önışlemleri yöntemleriyle en az 16 en fazla 24 öznitelikli alt verisetler elde edilmiştir.

Bütün verisetler tukey aykırı değer etiketleme (TOL), izolasyon ormanı (IF), z-skör, copula-tabanlı aykırı değer tespiti (COPOD), medyan mutlak sapma (MAD), yerel aykırı değer faktörü (LOF) ve eliptik zarf (EE) olmak üzere 7 farklı ML anomali analizi yöntemi kullanılarak anormal ve normal içme suyu tüketimleri tespit edilmiştir.

Anomali analizleri sonucunda her bir yöntem ile tespit edilen anormal tüketimler puanlanmış, 7 farklı anomali yönteminin puanları toplamı ile her bir anomali tüketim davranışının aykırılık puanı hesaplanmıştır. Aykırılık puanlarına göre bütün tüketim değerleri 4 farklı tüketim sınıfı (Normal, Dikkat, Riskli, Aşırı) ile etiketlenmiş, veriseti gözetimli hale getirilmiştir. Elde edilen gözetimli veriseti kullanılarak karar ağaçları (DT), gaussian naive bayes (NB), k-en yakın komşular (KNN), lojistik regresyon (LJR), çok katmanlı algılayıcı sinir ağı (MLP-NN), RF ve gradient boosting (GB) olmak üzere 7 farklı ML yöntemi ile de tüketim sınıf tahmin modelleri geliştirilmiştir.

Ayrıca geliştirilen modeller Doğruluk, Duyarlılık, Kesinlik, Hassasiyet, Özgüllük, MAE ve MSE performans metrikleri ile karşılaştırılmış, her bir abone için geliştirilen anormal tüketim tespit modellerinden en yüksek Doğruluk oranındaki model en iyi model olarak belirlenmiştir.

Anormal tüketim tespit modellerinden Doğruluk performansı aynı olanlar Duyarlılık ve Kesinlik oranlarına göre karşılaştırılmış, performansı yüksek olan tercih edilmiştir. Toplamda 921.088 adet model geliştirilmiş, bunlardan 16.077 adeti en iyi model olarak maksimum doğruluk performansı göstermiştir. Modellerin Doğruluk performans oranları en düşük %43, en yüksek %100 ve ortalama %85 olarak tespit edilmiş olup, EE yöntemiyle elde edilen modellerin çoğunlukla en iyi modeller olduğu, Z Skor ve LOF yöntemleri ile elde edilen modellerin çok düşük oranlarda maksimum Doğruluk sağladığı anlaşılmıştır.

Çalışmada elde edilen tüketim sınıfı tahmin modellerinde ise Doğruluk performans metriği %90 ve üzeri olanlar en iyi model olarak tercih edilmiştir. Böylece örneklem olarak alınan 8.224 abonenin 7.150 tanesine (%86,94) ait alt verisetlerinden (V1, V2, V3, V4, V5, V6, V7), diğer 1.074 adet abonenin (%13,06) ise temel verisetinden (V0) en iyi tüketim sınıfı tahmin modellerinin elde edildiği görülmüştür. Her bir aboneye ait geliştirilen maksimum Doğruluk performanslı model en fazla oranda F-Skor öznitelik seçim yöntemi ile elde edilen V5 alt veriseti kullanılarak elde edilmiştir. Diğer alt verisetleri ile de birbirine yakın oranlarda başarılı sonuçlar elde edildiği anlaşılmıştır.

Çalışma sonucunda anormal içme suyu tüketimlerinin ML yöntemleri ile tespit edilebileceği ve tüketim sınıflarının ML yöntemleriyle tahmin edilebileceği ispatlanmıştır. Ayrıca bireysel tüketim davranışlarını etkilemeye yönelik geri dönütler ile içme suyunun israf edilmeden daha verimli kullanılabileceği gösterilmiş, su yöneticilerinin yatırım planlamalarına ve yönetim yaklaşımlarına farklı bir bakış açısı sunulmuştur. Böylece çok yüksek taleplerin üstesinden gelinmesine, kayıp kaçak kullanım potansiyeli olan abonelerin tespit ve takip edilmesine, su tasarrufuna yönelik toplumsal bilincin arttırılmasına katkı sağlayabilecek bir model geliştirilmiştir.

DETECTING ABNORMAL DRINKING WATER CONSUMPTIONS AND DEVELOPMENT OF FORECAST MODELS BY MACHINE LEARNING METHODS

SUMMARY

In this study, it is predicted that there may be a certain order in the consumption of an important need such as drinking water by the household, as well as irregular consumption depending on different factors. Increasing population, limited drinking water resources, developing infrastructure and technology have increased the demand for drinking and utility water. There is a search for alternative water sources to meet this demand, but with this study, it is foreseen that these demands can be met by not wasting existing water and using it more efficiently.

By using machine learning (ML) methods, which is a sub-branch of artificial intelligence (AI), drinking water consumption data in the past periods were analyzed, and ordinary and unusual consumption behavior models were extracted. It is envisaged that by detecting abnormal consumptions that may occur in drinking water consumption and informing the subscribers about this issue, it will be ensured that the consumption in the household remains within the normal consumption range. Although the amount of data collected, recorded and processed in today's IT world has increased significantly, it is known that the exact analysis is difficult in terms of time and cost.

In this study, subscriber, meter, consumption, bill and payment data of 8,224 residential subscribers, whose water meter index reading is more than 160 periods throughout the province of Kayseri, between 2006 and 2022 (first 6 months) were taken into account. The data are combined on a spatial subscriber basis and a dataset which have 41-features is obtained. Prepared data set; It has been transformed into a 24-featured dataset using data preprocessing such as data integration, data transformation, noise identification and cleaning, and completion or cleaning of detected missing values. The scores of each of the features affecting drinking water consumption were calculated using basic statistical values and feature selection methods. It has been determined whether the features affect drinking water consumption, and if so, how much.

In feature selections, the determination of the number of features sufficient to meet the purpose was determined by taking the ratio of the variance explained criterion (Variance Explained Criteria) as 90%. Thus, threshold values were determined that minimum 6 and maximum 11 features would be sufficient and that this number of features would remain in each feature selection method.

Information gain (IG), gain ratio (GR), symmetrical uncertainty coefficient (SU), pearson correlation coefficient (r), f-score and random forest (RF) feature selection methods were used for the selection of features affecting consumption. Thus, 6

different sub-datasets were obtained over the determined threshold values. In addition, the 7th sub-dataset at least 3 times selected features which consisting of by the feature selection method was obtained.

The number of meters exchanged for each subscriber, the number of contracted users, the days of consumption measurement, consumption amounts, invoice amounts, payment status data are likely to be different, and the number of features is also considered to be different. Therefore, sub-datasets consisting of the selected features were recorded in the study specifically for each subscriber.

With all datasets, Tukey outlier labeling (TOL), isolation forest (IF), z-score, copula-based outlier detection (COPOD), median absolute deviation (MAD), local outlier factor (LOF), and elliptic envelope (EE) ML anomaly analysis methods have been used. Abnormal and normal drinking water consumptions were determined by using these 7 different ML anomaly analysis methods. As a result of the anomaly analysis, abnormal consumptions detected by each method were scored, and the abnormality score of each anomaly consumption behavior was calculated with the sum of the scores of 7 different anomaly methods. Each observation point detected by each algorithm is weighted with 1 point. Since 7 algorithms are used, a maximum of 7 points will be weighted for each observation point.

According to the outlier scores, all consumption values were labeled with 4 different consumption classes (Normal, Caution, Risky, Extreme), and the dataset was made supervised. With the obtained supervised data sets, consumption class estimation models were developed using decision trees (DT), gaussian naive bayes (NB), k-nearest neighbors (KNN), logistic regression (LJR), multilayer perceptron neural network (MLP-NN), RF and gradient augmentation (GB) methods.

Drinking water consumption prediction models are compared with ACC, R2 performance metrics, and the best prediction model is selected for each drinking water subscriber. Since the consumption data consists of TS data, the first years are education data and the last years are control data, and they are used by dividing them by 90% and 10%.

The error matrices of all anomaly detection models were calculated and evaluated as a basis for the performance measurement between the predictive values and the actual values of the models detecting abnormal drinking water consumption classes. In addition, the developed models were compared with Accuracy, Sensitivity, Sensitivity, Sensitivity, Specificity, MAE and MSE performance metrics.

Among the abnormal consumption detection models developed for each subscriber, the model with the highest accuracy was determined as the best model. Among the abnormal consumption detection models, those with the same Accuracy performance were compared according to their Sensitivity and Precision ratios, and the one with higher performance was preferred.

A total of 921,088 models were developed, of which 16,077 showed maximum accuracy performance as the best model. The Accuracy performance ratios of the models were determined as the lowest 43%, the highest 100% and the average 85%, and it was understood that the models obtained by the EE method were mostly the best models, and the models obtained with the Z Score and LOF methods provided maximum Accuracy at very low rates.

In the consumption class prediction models obtained in the study, the accuracy performance metric of 90% and above was preferred as the best model. Thus, it was seen that the best consumption class prediction models were obtained from the sub-datasets (V1, V2, V3, V4, V5, V6, V7) of 7,150 (86.94%) of 8,224 subscribers taken as a sample. It was observed that the best consumption class prediction models were obtained from the basic dataset (V0) of the remaining 1,074 subscribers (13.06%).

The maximum Accuracy performance model developed for each subscriber was obtained by using the V5 sub-dataset obtained by the F-Score feature selection method at the highest rate. It has been understood that successful results have been obtained with other sub-datasets at rates close to each other.

As a result of the study, it has been proven that abnormal drinking water consumptions can be determined by ML methods and consumption classes can be estimated by ML methods. In addition, with the feedbacks aimed at influencing individual consumption behaviors, it has been shown that drinking water can be used more efficiently without wasting, and a different perspective has been presented to the investment planning and management approaches of water managers. Thus, a model has been developed that can contribute to overcoming very high demands, detecting and tracking subscribers with potential for loss and illegal use, and raising social awareness for water saving.

The main purpose of the study is to sensitize the subscribers and to reduce the abnormal consumptions in their consumption behaviors, in other words, to reduce them to normal levels. In this case, it is predicted that 139,427 m³ of water will be used with 20% savings in the consumptions in the 'Extreme' class, 154,058 m³ with 15% savings in the consumptions in the 'Risky' class, and 169,402 m³ of water with 10% savings in the consumptions in the 'Caution' class. Thus, a more efficient use of 462,887 m³ of 2.58% water in the system will be achieved.

The study helped to meet the rapidly increasing urban population and water needs. Contributed to the delivery of equitable water services in a world facing increasing water scarcity and environmental degradation. It has been shown what factors can affect drinking and domestic water consumption. It has been shown that there is a regular behavior pattern and seasonality in water consumption. It has been determined that there are periods that disrupt the consumption order. It has been revealed that abnormal consumptions that disrupt the consumption pattern may be caused by meter reading error, lost and illegal consumption, meter measurement error and wasteful consumption. It has been shown that drinking water consumption behavior can be modeled with ML algorithms, which is a sub-branch of AI, and can be disciplined with feedback.

The subject of the study is related to the consumption of drinking and utility water in households and a model has been developed that can be used in all service areas where mass consumption such as electricity, natural gas, internet, GSM is required and where a distribution network, subscriber management system and infrastructure is needed. In addition, these systems include many requirements such as consumption measurements and follow-up, quality controls, fault detection, maintenance and repair follow-up, inspection, cost-reducing measures. From this point of view, it is foreseen that the study can contribute to these areas as well.



1. GİRİŞ

Su, canlıların beslenme, temizlenme, ulaşım ve taşıma gibi temel nedenlerle sürekli ihtiyaç duyduğu bir araç olmakla birlikte özelde insanoğlunun da üzerinde her zaman hesap yapageldiği önemli bir değer olmuştur. Milattan önceki çağlardan günümüze kadar insanlık suyun temini, yönetilmesi, paylaşılması, taşınması ve kullanılması hususlarına önem vermiş yer yer bu önem Tablo 1.1’de özetlenen ve Şekil B.1’de de harita üzerinde gösterilen çatışmaların yaşanmasına neden olmuştur. Çatışmalar suyun bir silah olarak veya savaşların tetikleyici nedeni olarak görülmesiyle veya su kaynaklı kazaların oluşmasıyla gerçekleşmiş böylece insanlığın tarih içindeki mücadelesinde önemli izler bırakan su, halen tarihe, coğrafyaya şekil vermeye devam edegelmiştir (Institute, 2023a).

Tablo 1.1. Tarihte su ile ilgili yapılan çatışma nedeni ve sayıları.

Dönem Aralığı (Yıllar)	Çatışma Sayısı	Çatışmada Su Kullanım Türü	Ülke Sayısı	Coğrafi Bölge Sayısı
Milattan Önce	26	Silah, Kaza	10	6
0 - 1000	2	Silah, Tetikleyici,	1	2
1001 - 1900	25	Silah, Kaza, Tetikleyici	13	9
1901 - 2000	209	Silah, Kaza, Tetikleyici	101	14
2001 - 2022	1.036	Silah, Kaza, Tetikleyici	109	16
Toplam:	1.298	Silah, Kaza, Tetikleyici	178	16

Çatışmaların büyük çoğunluğunun son 20 yılda meydana geldiği, ayrıca 20. yüzyılda nüfus artışının dört kat, su talebinin dokuz kat artması dikkate alındığında su güvenliği konusunun günümüz dünyasında hissedilecek kadar hızla büyüyen sosyal, politik ve ekonomik sorunları tetiklediği, giderek yayılan bir çevresel krize dönüştüğü artık kabul edilen bir gerçektir. Yapılan araştırmalarda 2030 yılına kadar küresel tatlı su talebinin mevcut arzı %40'ın üzerinde aşacağı ve etkilenecek insan sayısının 4 milyara yaklaşacağı, artan rekabet koşulları ve beraberinde su üzerinde oluşturduğu stresin,

dünya genelinde gıda, enerji, üretim ve insan güvenliği açısından önemli bir risk ve etki oluşturacağı öngörülmektedir (Cini ve ark, 2014).

Kentsel yaşamın giderek yaygınlaştığı günümüzde en önemli tüketim kaynaklarından biri olan su, potansiyel olarak yüzeyde, yeraltında, havada farklı formlarda bulunmakla birlikte formlararası değişim ve geçiş döngüsüne sahip olarak hayatın içinde yer almaktadır. Ülkemizde ve dünyada 20 yılı aşkın bir süre zarfında nesnelerin interneti (IOT) tabanlı cihazların, kablosuz haberleşmenin ve YZ'nın gelişimiyle akıllı şehir konseptleri hızla yaygınlaşmış ve giderek alanını genişletmiş, hatta oldukça kompleks sorunları da çözebilecek etkinliğe kavuşmuştur.

Akıllı şehirler kavramı beraberinde akıllı su şehirleri kavramına zemin oluşturmuş böylece Uluslararası Su Birliği (IWA) tarafından Tablo 1.2'de gösterilen 4 seviyeli Akıllı Su Şehri İlkeleri Uygulama Planı oluşturulmuştur. Tez konusu ve sonuçlarının bu plan içerisinde tüketilen su miktarının azaltılması, mevcut suyun en verimli şekilde kullanılması, tüketim davranış analizlerinin sunduğu geri dönütler ile abonelerin bilinçlendirilmesi, tüketimin bu sayede disipline edilmesi ve nihai olarak da tespit edilen aykırı tüketimlerin dikkate alındığı kayıp ve kaçak denetimleri ile su yönetiminin daha bilimsel yapılabilmesi hususlarında katkı sağlayacağı düşünülmektedir (International Water Association, 2017).

Tablo 1.2. Kentsel paydaşların “Sürdürülebilir Kentsel Su” hizmeti sunmaları için dört aşamalı eylem planı (International Water Association, 2017).

Aşama 1	Aşama 2
- Yenileyici, Canlandırıcı Su Hizmetleri	- Suya Duyarlı Şehir Dizaynı
• Su kaynakları ekosistemini ve su kaynaklarını canlandır • Kullanılan su ve enerji miktarını azalt • Tekrar kullan, kurtar ve yeniden kullanım döngüsüne sok, geri dönüştür • Diğer hizmetlerle entegre olan sistemik bir yaklaşım kullan • Sistemin modüler özelliklerini arttır ve çoklu seçenekler oluştur	• Canlandırıcı su hizmetlerini etkinleştir, • Kentsel mekanlardaki taşkın risklerini azalt, • Mevcut suyu en verimli şekilde kullanma anlayışını güçlendir, • Çevresel etkilerini en aza indirmek için kentlerdeki bazı uygulamaları değiştir veya uyumlu duruma getir

Tablo 1.2. (Devamı) Kentsel paydaşların “Sürdürülebilir Kentsel Su” hizmeti sunmaları için dört aşamalı eylem planı (International Water Association, 2017).

Aşama 3	Aşama 4
- Havza Bağlantılı Şehirler	- Akıllı Su Toplumu
• Su kaynaklarını koruma ve suyu önleme planı yap • Su kaynaklarının kalitesini koru • Suya bağlı olağandışı riskler için hazırlık yap	• Bilinçlendirilmiş vatandaşlar • Suyun ikincil faydalarının farkında olan profesyoneller • Disiplinler ötesi planlama ekipleri • Akıllı su yönetimi kararları alabilecek yetenekte politika yapıcılar • Konuyla ilgili ve güven yaratan liderler

Su dağıtımının Yönetmelik Kontrol ve Veri Toplama (SCADA) sistemleri ile yönetilmeye başlanması, verilerin toplanması ve kayıt altına alınması ile ciddi bir veri havuzu oluşmuş, beraberinde verilerle anlamlı sonuçlar elde edilebilecek analizlerin önü açılmış daha kompleks çözümler sunan AI modelleri geliştirilmiştir. Mevcut durumda su sektöründe AI ile geliştirilen uygulamalar su alt yapısının fiziki durumunun tespiti ve bakımı, su talebi ve tüketiminin tahmini, su rezervuarlarının ve barajların sağlık ve çevresel etkilerinin izlenmesi, su kalitesinin izlenmesi ve su ile ilgili felaketlerin öngörülmesi ve izlenmesi olarak 5 başlıkta raporlanmıştır. Öyle ki AI destekli yeniliklerin 2030 yılına kadar küresel ekonomiye 200 milyar dolar katkı sağlayacağı tahmin edilmektedir (Yıldız ve Özgüler, 2020).

Birleşmiş Milletler (BM) bünyesinde bağlı bütün üye ülkelerin 2030 yılına kadar uygulayabilecekleri Sürdürülebilir Kalkınma Hedefleri (BM SKH) belirlenmiş, bu hedeflerden özellikle su ile ilgili olanlar için geliştirilen AI çalışmaları ana konu başlıkları Tablo 1.3’te ilişkisel olarak gösterilmiştir. Hedeflenen en önemli gaye 2030 yılına kadar herkesin güvenli ve erişilebilir içme suyuna kavuşmasını sağlamak için, altyapıya yatırım yapmak, sıhhi tesisleri inşa etmek ve her düzeyde hijyeni teşvik etmek zorunluluğu olmuştur. Ayrıca hedefler arasında su kıtlığını hafifletmek için ormanlar, dağlar, sulak alanlar ve nehirler gibi suyla bağlantılı eko-sistemleri korumak ve eski haline getirmek zorunluluğu yanında gelişmekte olan ülkelerde su verimliliğini

teşvik etmek ve arıtma teknolojilerini desteklemek için uluslararası iş birliğine de vurgu yapılmıştır (Türkiye Cumhuriyeti Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı, 2020).

Tablo 1.3. AI uygulamalarının temel alanları ve Birleşmiş Milletler Sürdürülebilir Kalkınma Hedefleri (Türkiye Cumhuriyeti Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı, 2020).

Su ile İlgili Yapılmış Yapay Zekâ Çalışmaları Ana Konu Başlıkları ve Birleşmiş Milletler Üye Ülkeler 2030 Yılına Kadar Sürdürülebilir Kalkınma Hedefleri	Su altyapısının önleyici bakımı	Su talebinin ve tüketiminin tahmini	Barajların ve rezervuarların izlenmesi	Su kalitesinin izlenmesi	Suya bağlı felaketlerin tahmini ve izlenmesi
Hedef 3.3: 2030'a kadar AIDS, tüberküloz, sıtma ve ihmal edilen tropikal hastalıkların salgınlarına son verilmesi ve hepatit, su yoluyla bulaşan hastalıklar ve diğer bulaşıcı hastalıklarla mücadele edilmesi,					
Hedef 3.9: 2030'a kadar tehlikeli kimyasallardan ve hava, su ve toprak kirliliği ve kontaminasyonundan kaynaklanan ölüm ve hastalıkların önemli ölçüde azaltılması,					
Hedef 6.1: 2030'a kadar küresel ölçekte herkes için güvenli, adil ve uygun fiyatlı içme suyuna erişimin sağlanması,					
Hedef 6.3: 2030'a kadar zararlı kimyasalların ve maddelerin boşaltımının bitirilmesi ve salınımının minimize edilmesi, arıtılmayan atık su oranının yarıya düşürülmesi, küresel ölçekte geri dönüşümün ve güvenli geri kazanımın kayda değer miktarda çoğaltılması ile kirliliğin azaltılması ve su kalitesinin artırılması,					
Hedef 6.4: 2030'a kadar, tüm sektörlerde su kullanım verimliliğinin önemli ölçüde artırılması ve su kısıtının dikkate alınması için tatlı suyun sürdürülebilir çıkarılması ve tedarik edilmesinin sağlanması ve su kıtlığı çeken insanların sayısının önemli ölçüde azaltılması,					
Hedef 6.5: 2030'a kadar, uygun olduğu durumlarda sınır ötesi iş birliği dâhil olmak üzere her düzeyde entegre su kaynakları yönetiminin uygulanması,					
Hedef 6.6: 2020'ye kadar dağlar, ormanlar, sulak alanlar, nehirler, akiferler ve göller de dahil olmak üzere suyla ilgili ekosistemlerin korunması ve yenilenmesi					

Tablo 1.3. (Devamı) AI uygulamalarının temel alanları ve Birleşmiş Milletler Sürdürülebilir Kalkınma Hedefleri (Türkiye Cumhuriyeti Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı, 2020).

Su ile İlgili Yapılmış Yapay Zekâ Çalışmaları Ana Konu Başlıkları ve Birleşmiş Milletler Üye Ülkeler 2030 Yılına Kadar Sürdürülebilir Kalkınma Hedefleri	Su alt yapısının önleyici bakımı	Su talebinin ve tüketiminin tahmini	Barajların ve rezervuarların izlenmesi	Su kalitesinin izlenmesi	Suya bağlı felaketlerin tahmini ve izlenmesi
Hedef 11.5: 2030'a kadar, ölüm sayısının ve etkilenen insan sayısının su ile ilgili felaketler de dahil olmak üzere felaketlerin, oluşan doğrudan ekonomik kayıpların küresel gayri safi yurtiçi hâsıla (GSYH) içindeki oranının önemli ölçüde azaltılması, yoksulların ve savunmasız insanların korunmasına daha çok önem verilmesi,					
Hedef 15.1: 2020'ye kadar kara ve iç tatlı su ekosistemlerinin ve hizmetlerinin, özellikle ormanlar, sulak alanlar, dağlar ve kurak alanların uluslararası sözleşmeler kapsamında korunması, restorasyonunun ve sürdürülebilir kullanımının sağlanması, aynı zamanda uzun süreli su temini için havza ve su kaynaklarının korunması ve acil durum hazırlığı yapılması.					

Gelişen teknoloji ile birlikte gelecekte su ihtiyacı tahminine yönelik senaryolara göre 2040 yılına gelindiğinde yerel su kaynaklarının yeniden kullanılacağı, geri dönüşümün zorunluluk halini alacağı, yağmur suyu yönetiminin yaygınlaşacağı, yeterince kullanılmayan su kaynaklarına erişimlerin sağlanacağı, doğrudan suyu arıtma ve arıtım yollarının yeşil alanlara veya parklara dönüştürüleceği, kentsel su hizmetlerinin görünür hale getirileceği ve teknolojinin haneler üzerinde büyük bir etkisi olacağı, bir mobil uygulama aracılığıyla gün boyunca su kullanımlarının gerçek zamanlı takip edilmesinin sağlanacağı, hangi amaçlarla ne kadar su kullanıldığının ve hangi saatlerde kullanıldığının, şebekedeki su kayıp ve kaçaklarının tespitinin, su kullanımlarının faturaları nasıl etkileyeceğinin görülebileceği bir kentsel su modeli bizleri beklemektedir (Arup and Sydney Water, 2015).

Hane içi tüketimlerin Tablo 1.4'te gösterildiği oranlarda olduğu, hane büyüklüğü ile orantılı olduğu ancak büyüklüğe oranla tüketimin daha az arttığı ve bir tasarruf yapılacaksa nerelerde yapılabileceği öngörülebilecektir (Mayer P. W. ve ark, 1999).

Tablo 1.4. Suyun hane içinde nihai kullanım yerleri ve oranları (Mayer P. W. ve ark, 1999).

Son Kullanım	İç Mekân Su Kullanım Yüzdeleri
Tuvaletler	%26,7
Çamaşır Makineleri	%21,7
Duşlar	%16,8
Musluklar	%15,7
Sızıntı	%13,7
Diğer Lokal	%2,2
Banyolar	%1,7
Bulaşık Makineleri	%1,4
Toplam *	%100,0

*ondalık tek haneye indirildiğinden %99,9

Ayrıca bilimsel çalışmalarda hane halkı dışındaki tüketim oranları her ne kadar kültüre ve coğrafyaya göre değişiklikler gösterse de nüfus ve yüz ölçüm ölçekleri dikkate alınarak yaklaşık olarak hesaplanabilmektedir (Mayer P. ve ark, 2000).

Bu tezin amaçlarından biri olan su tasarrufuyla ilgili Tablo 1.5'te gösterildiği gibi 5 ana başlık altında bir dizi önerilerde bulunulmuş, su yöneticileri ve nihai tüketicilerin bu önerilerden birçoğunu yaptıkları, ilgili planlamaların giderek yaygınlaştığı sonraki dönemlerde daha fazlasının hayatın içinde yer alacağı görülmektedir (Campbell ve ark, 2004).

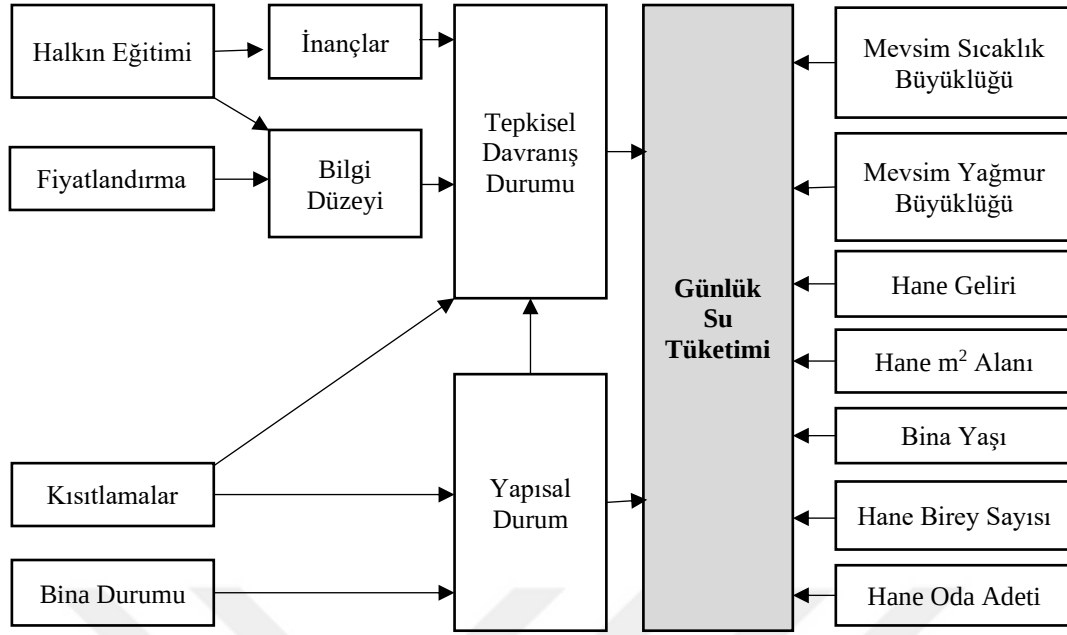
Tablo 1.5. Su tasarrufu ile ilgili öneriler (Campbell ve ark, 2004).

Yönetmelikler	Kamuya Yönelik Bilgilendirmeler
• Az su kullanan armatürleri belirten tesisat kodları • Düşük su kullanımı gerektiren peyzaj düzenlemeleri • Su israfı yasağı • Belirli dış mekân amaçları için ham veya içilemez su gereksinimleri • Yeniden satışta güçlendirme gereksinimleri • Yeni nesil tuvaletler için ücretsiz dağıtım veya sübvansiyonlar	• Genel tanıtım: reklam panoları, postalar, su faturası ekleri • Herkese açık forumlar • Gazete, radyo, TV, internet kullanımı • İlköğretim ve ortaöğretimlerde eğitim programları • Verimli sulama sistemlerini teşvik etmek için çalıştaylar • Su bazlı çevre düzenlemesini teşvik eden atölyeler (xeriscape) • Günlük hava durumu raporlarına peyzaj sulama ihtiyaçlarının eklenmesi

Tablo 1.5. (Devamı) Su tasarrufu ile ilgili öneriler (Campbell ve ark, 2004).

Mühendislik	Fiyatla İlgili Stratejiler
• Yeni nesil tuvaletler için ücretsiz dağıtım veya sübvansiyonlar • Düşük akışlı duş başlıkları ve musluk havalandırıcılarının ücretsiz dağıtımı • Az su kullanan giysiler veya bulaşık makineleri için sübvansiyonlar • Peyzaj ve yağmurlama sistemi dönüşümü için sübvansiyonlar • Yağmur suyunun toplanması ve peyzajda kullanımı için sübvansiyonlar • Çevre düzenlemesinde gri suyun yeniden kullanımı için sübvansiyonlar • Geri kazanılmış su dağıtım sisteminin genişletilmesi • Sızıntı tespit programları	• Emtia ücretleri olan sayaçlar • Su dağıtımlarının evrensel ölçümü • Suyun marjinal fiyatının faturalarda belirgin şekilde gösterilmesi • Su fiyatındaki artış • En yoğun talep dönemleri için daha yüksek su fiyatları • Su için artan kademe oranları • Su bütçeleri • Ekonomik açıdan çekici hale getirmek için geri kazanılan suyun sübvansiyonu
Dolaylı Mühendislik	
• Kişisel etkileşim içeren herhangi bir mühendislik çözümü	• Yüksek su kullanımı olan müşterilere yönelik su kullanımı denetimleri

Hanelerin günlük su tüketimlerini etkileyen Şekil 1.1’de gösterildiği gibi birçok faktör belirlenmiş ve günlük su tüketim değerlendirmesi yapılırken bunlar göz önünde bulundurulmuş, bunların haricinde de birçok etmenin var olduğu dikkate alınarak bu etmenler arasındaki ilişkilerin tüketimi tetiklediği gerçeğinden hareketle bir konut suyu koruma modeli uyarlanmıştır (Billings ve Jones, 2008).



Şekil 1.1. Konut suyu koruma modeli (Billings ve Jones, 2008).

Tez ile birlikte gerek sular idarelerinin gerekse abonelerin kendilerine sorabilecekleri hızla artan kent nüfusunun su ihtiyacı nasıl karşılanabilir, giderek artan bir şekilde hissedilen su kıtlığı ve çevresel bozulma ile karşı karşıya kalınan bir dünyada adil su hizmetleri sağlanabilir mi, içme ve kullanma su tüketimini etkileyen faktörler nelerdir, su tüketiminde bir düzen bir davranış kalıbı var mıdır, varsa bu bir mevsimsellik taşıyor mudur, tüketim düzenini bozan dönemler var mıdır, tüketim düzenini bozan aykırı tüketimler ile sayaç okuma hatası, kayıp kaçak tüketim, sayaç ölçüm arızası, tüketimde israf vb. tespitler yapmak mümkün müdür, tüketim davranışı AI alt dalı olan ML algoritmaları ile modellenilebilir mi, tüketim davranışı disipline edilebilir mi, tüketim üzerinden su tasarrufu ile kaynak arayışına ciddi bir katkı sağlanabilir mi sorularına cevaplar sunulmaya çalışılmıştır.

Çalışmanın ikinci bölümünde kaynak araştırması başlığı altında içme ve kullanma suyunun durumu, toplu hizmetlerde abone davranışları, ML ve anomali analizi konularında yapılmış çalışmalar araştırılmış ve elde edilen sonuçlar paylaşılmıştır.

Üçüncü bölümde materyal ve yöntem başlığı altında analiz çalışması adım adım anlatılmıştır. Bu bölümde öncelikle analizde kullanılan ve analizi oluşturan bütün argümanlar kısaca özetlenmiştir.

Dördüncü bölümde modelin uygulanması başlığı altında çalışmada önerilen materyal ve yöntemlerin nerede ve nasıl kullanıldığı gösterilmiştir.

Beşinci araştırma bulguları başlıklı bölümde ise analiz sonucunda elde edilen sonuçlar ve araştırmaya konu olan soruların cevapları anlatılmıştır.

Son olarak tartışma ve sonuç başlığı altında elde edilen cevaplar ile ilgili değerlendirmeler ve öneriler sıralanmıştır.

1.1. Tezin Kapsamı

Bu tez kapsamında Kayseri ili içme ve kullanma su abonelerinin hane içi su tüketimleri dikkate alınmış, dünyada ve ülkemizde her geçen gün üzerinde titizlikle durulan su tasarrufuyla ilgili konularda toplumsal bilincin arttırılmasına ve abone bazlı tüketimlerden hareketle sayaç ölçüm hassasiyetlerine, su kayıp kaçak denetimlerine yönelik yapılan ML çalışmalarına katkıda bulunacak bir model geliştirilmiş, bununla birlikte sadece bireysel davranışları etkileyen faktörler üzerinde durulmuş tüketim davranışını etkilemeye yönelik geri dönütler ile bir tasarrufun sağlanması, beraberinde su yöneticilerinin yatırım planlamalarına ve yönetim yaklaşımlarına farklı bir bakış açısı sunacak analizlerin yapılması ve bulguların paylaşılması hedeflenmiştir.

1.2. Tezin Amacı

İçmesuyu tüketim davranışlarını tespit etmek, alternatif su kaynağı arayışının aksine varolan kaynakları daha verimli kullanmaya katkı sağlamak, tüketimi etkileyen faktörlere göre aykırı tüketim davranış sınırları belirlemek ve bunları takip etmek, abonelerin israf etmeden su kullanımına ve dolaylı olarak da ekonomisine katkıda bulunmak, kurumun çok yüksek taleplerde yaşayacağı zorlukların üstesinden gelmesine katkıda bulunmak, tüketim sonucunda ortaya çıkan atıksu yükünü minimize etmek ve oluşan arıtma maliyetlerini azaltmak, kayıp kaçak kullanım potansiyeli olan aboneleri tespit etmek ve takip edilmesini sağlamak.



2. KAYNAK ARAŞTIRMASI

Çalışmada kaynak araştırmaları ve özetleri iki başlık altında toplanmıştır. Öncelikli olarak çalışmanın konusu içme ve kullanma suyunun hane içindeki tüketimiyle ilgili olduğu için özellikle su ile ilgili yapılan çalışmalar incelenmiş, konusu itibariyle birbirinden farklı çalışmalar dikkate alınmaya çalışılmış, su tüketimi ve suyun durumu başlığı altında su ile ilgili kapsamın, ihtiyaçların, beklentilerin neler olduğu hususlarına yönelik paylaşımlar yapılmıştır.

Son olarak, çalışma aynı zamanda bir anomali analizi çalışması olduğundan hareketle anomali analizlerinin farklı alanlardaki kullanımı, nasıl kullanıldığı, ne gibi faydalar sağladığı hususları ile ilgili bilgilendirmeler anomali analizleri başlığı altında incelenmiş ve özet olarak aktarılmıştır.

2.1. Su Tüketimi ve Suyun Durumu

Gökçen ve Yeşil (2021) çalışmalarında ultrasonik sensörlü su sayacı tarafından farklı akış hızlardaki ölçümler ile akış hattının dolu olduğu ve akışın sıfır olduğu fiziksel bozucu etkilerden kaynaklı aynı anda diğer sayaç ölçümlerinin de karşılaştırılmasıyla tespit edilen yanlış-pozitif zaman serisi ölçümler toplanmış ve oluşan veriseti LJR, Destek Vektör Makineleri (SVM), Radyal Çekirdek Fonksiyonlu SVM (RBK-SVM), Lineer Diskriminant Analizi (LDA) algoritmaları kullanılarak yanlış pozitif ultrasonik sensör ölçümlerini tespit eden modeller oluşturulmuştur. Modeller Doğruluk (Accuracy), Hassasiyet (Sensitivity), Özgüllük (Specificity), Keskinlik (Precision) performans metrikleriyle kıyaslanmış RBK-SVM %99,6 doğruluk oranı ile en iyi sonucu vermiş, diğer model performansları da yüksek olduğundan karmaşıklık açısından oldukça basit olan lojistik regresyon algoritması gerçek sistem çalışması için seçilmiştir.

Kadinski ve Ostfeld (2021) çalışmalarında çeşitli alt bölgelere ayrılan su şebekesinde kirliliği gerçek zamanlı tespit etmek, şebekedeki tehlikeli alanları belirlemek ve tüketicileri uyarmak için sensörler yerleştirmenin çok etkili olduğu gösterilmiş, bu sensörler ile şebekede tespit edilen anormal kirlilikler sonucunda tepkisel su tüketim

davranışlarından anlaşılan ayrıca tüketicilerin su dağıtım kurumlarına ve tıbbi tesislere verdikleri tıbbi, coğrafi, sosyal geçmiş, konum ve kirlenmiş suyun yutulmasından kaynaklanan olası semptom bilgileri kullanılmış, bulanık bir mantık sistemine dayalı, suyun kalitesine göre karar vermenin temel alındığı hidrolik simülasyona bağlı bir model önerilmiştir.

Lund ve arkadaşları (2021) çalışmalarında Danimarka Elsinore şehrindeki kanalizasyon atıksu akış ölçümü için sensörler yerleştirilmiş, sensörlerden elde edilen ölçümler ile yaklaşık 2.000 tüketiciden Ekim 2018 ve Mart 2019 tarihleri arasında 3 farklı zaman diliminde 1 haftalık periyotla elde edilen, saatlik ölçümler yapan akıllı sayaçlara ait tüketim verileri, su dağıtım çıkış verileri, Atıksu Arıtma Tesisi (AAT) giriş verileri ve hanelerin yıllık su tüketim verileri kullanılarak simüle edilen bir kanalizasyon sisteminde elde edilen atıksu miktarları karşılaştırılmıştır. Çalışma sonunda şebekedeki sızma, sızıntı ve sensör hataları gibi sistem ve veri anormalliklerini tespit eden bir model önerilmiştir.

Dogo ve arkadaşları (2021) çalışmalarında Alman kamu su hizmetleri şirketi Thüringer Fernwasserversorgung bünyesinde 03.08.2016 ve 13.02.2017 arasında 98 gün boyunca dakikada bir elde edilmiş, 2 sınıflı (normal, anormal) 10 öznitelikli veriseti kullanılmıştır. Çalışmada su kalitesi anomali tespiti bir sınıflandırma problemi olarak değerlendirilmiş, problemin çözümü için topluluk öğrenme modelleri ile birleştirilen Modifiye Sınıf Sıralaması (Rank), Yerel Sınıf Doğruluğu (LCA), Genel-Yerel Doğruluk (OLA), K-Nearest Oracles Eliminate (KNORA-E), K-Nearest Oracles Union (KNORA-U) ve Dinamik Topluluk Seçimi için Meta-Öğrenme (META-DES) olmak üzere altı dinamik seçim tekniği kullanılarak alt veri kümeleri elde edilmiş sonrasında bu dengesiz veri kümeleri üç Sentetik Azınlık Aşırı Örnekleme (SMOTE) tekniği tabanlı yeniden örnekleme algoritmaları kullanılarak dengeli hale getirilmiş ve parametrik olmayan RF tabanlı atama (missForest) yöntemi kullanılarak eksik veriler işlenmiştir. Dengeli hale getirilmiş yeni verisetleri ile tekrar oluşturulan modeller Doğruluk, F1-Skor, G-Ortalama performans metrikleriyle karşılaştırılmıştır. Çalışma sonucunda, dinamik seçim teknikleri kullanılarak elde edilen dengeli veriseti sınıflandırıcılarının topluluk modellerinin performansını artırabileceği gösterilmiştir.

Kızılöz ve Şişman (2021) çalışmalarında Kocaeli'nin en fazla su kaybı yaşanan altı ilçesine ait 144 adet veriden oluşan 2018 ve 2019 yılları arasındaki aylık hesaplanan bir veriseti kullanılmıştır. Veriseti %65 (93 adet) eğitim, %20 (29 adet) doğrulama ve kalan %15 (22 adet) test seti olarak rastgele şekilde bölünmüş, tüketilen su miktarı, şebeke uzunluğu, servis bağlantı uzunluğu, toplam şebeke uzunluğu parametrelerinden 2 tanesi girdi şebeke basıncı ve şebeke yaşı parametrelerinden 1 tanesi de çıktı olarak alınmış ve ayrı ayrı Yapay Sinir Ağı (ANN) ve Kriging yöntemi ile tahmin modelleri geliştirilmiştir. Çalışma sonunda tüketilen su miktarı ve servis bağlantı uzunluğu faktörlerinin girdi şebeke yaşı faktörünün çıktı olduğu Kriging yöntemi ile elde edilen modelin determinasyon (belirlilik) katsayısı (R^2) 0,89 ve kök ortalama kare hata (RMSE) 0,022 performans metrikleriyle ANN modellerine göre daha iyi Gelir Getirmeyen Su Oranı (GGSO) tahmini yapabildiği gösterilmiştir.

Çıtakoğlu ve Özeren (2021) çalışmalarında Sakarya Havzasının 1995 – 2014 (20 yıl) yılları arasında Şubat, Nisan, Haziran, Ağustos ve Kasım aylarında ölçülmüş su kalite parametrelerinden oluşan 1.388 adetlik bir veriseti kullanılmış, sıcaklık (T), elektriksel iletkenlik (Eİ), pH, biyolojik oksijen ihtiyacı (BOİ) ve aylar parametreleri giriş, çözünmüş oksijen parametresi ise çıkış verisi olarak değerlendirilmiştir. Ayrıca veriseti %80 (1.107 veri) eğitim, %20 (281 veri) test verisi olarak ikiye bölünmüş, girdi parametrelerinin farklı kombinasyonları ile çok katmanlı ANN kullanılarak çözünmüş oksijen miktarını tahmin eden modeller geliştirilmiş ve Ortalama Mutlak Hata (MAE), Ortalama Mutlak Bağlı Hata (MRAE) ve Nash–Sutcliffe Model Verimlilik Katsayısı (NSE) performans metrikleriyle karşılaştırılmıştır. Çalışma sonucunda performans kriterlerinin tamamında en iyi sonuçların BOİ–T–pH–Eİ modelinden alındığı görülmüş, klasik performans değerlendirme kriterlerinden farklı olarak NSE değeri 0,719 ile 0,7–0,9 aralığında kalarak yüksek tahmin başarısı elde edildiği gösterilmiştir.

Amini ve arkadaşları (2021) çalışmalarında Meşhed Şehri Su ve Kanalizasyon Kurumuna ait 24 farklı kuyudan su debisi, hidrolik ve elektriksel parametreler toplanmış, elde edilen veriseti %70 eğitim, kalan %30 test seti olarak bölünmüş, su akış hızını tahmin etmek için destek vektör regresyon (SVR), aşırı öğrenme makinesi (ELM), sınıflandırma ve regresyon ağacı (CART), RF, ANN, genelleştirilmiş

regresyon sinir ağı (GR-NN), doğrusal regresyon (LR) ve KNN olmak üzere 8 farklı model oluşturulmuş, ortalama kare hata (MSE), RMSE, MAE ve çapraz tahmin doğruluğu (CPA) performans metrikleri ile karşılaştırılmıştır. Çalışma sonucunda, LR, RF ve CART modellerinin diğer ML yöntemlerinden daha doğru tahmin sonuçları ürettikleri gösterilmiştir.

Abdel-fattah ve arkadaşları (2021) çalışmalarında Mısır'daki Bahr El-Baqr drenajı için sulama suyu kalite indeksinin (IWQI) belirlenmesi, izlenmesi ve sulama için uygunluk tahmini amaçlanmış, Mart 2016 ile Şubat 2019 tarihleri arasında aylık pH, Kalsiyum, Magnezyum, Sodyum, Potasyum, Bikarbonat, Klor ve Sülfat oranlarından oluşan parametreler ile aylık IWQI hesaplanmıştır. Oluşan veriseti ile ANN ve otoregresif bütünleşik hareketli ortalama (ARIMA) algoritmaları kullanılarak sulama suyu kalite tahmini yapan modeller geliştirilmiştir. Çalışma sonucunda modeller R^2 performans metriğiyle karşılaştırılmış, gizli katmanı 10 nöronlu olan ANN modelinin 0,98 üstü performans ile en iyi tahmin değeri üreten model olduğu gösterilmiştir.

Lattawit ve arkadaşları (2021) çalışmalarında Tayland genelinde 2017-2020 yılları arasında 10 dakika periyotla 4 farklı ölçüm noktasından elde edilen su seviyesi TS verileri kullanılmış, medyan ve MAD algoritmaları ile aykırı değer tespiti, doğrusal enterpolasyon, spline yöntemi ve çift yönlü uzun kısa süreli bellek (LSTM) yöntemi ile de eksik veri tamamlama işlemleri gerçekleştirilmiştir. Çalışma sonucunda eşik ve pencere boyutları değiştirilerek geliştirilen modeller F1-skor ve RMSE performans metrikleriyle karşılaştırılmış modellerin gerçek zamanlı aykırı değer tespitinde istenen performansları sağladığı gösterilmiştir.

Doorn (Doorn, 2020) çalışmasında araştırma yayınları veri tabanı olan ISI Web of Knowledge kullanılarak su alanındaki AI uygulamalarının yapıldığı ingilizce dergi makaleleri incelenmiş, 130 adet %22 modelleme, 78 adet %13 tahmin ve öngörü, 183 adet %30 erken uyarı veya olay tespiti odaklı karar destek ve operasyonel yönetim ve 211 adet %35 genellikle su planlama ve yönetim kararlarını sistematik olarak belirlemeyi sağlamaya dayalı optimizasyon olmak üzere toplamda 601 adet çalışma dört farklı kategoride değerlendirilmiştir. Çalışma sonucunda öncelikle çeşitli boyutlardaki verisetlerinin AI tabanlı teknikler kullanarak optimize edilmesi, ikinci olarak veriye dayalı modellerdeki eksikliklerin teoriye dayalı veri bilimi yardımıyla

giderilmesi, son olarak AI teknikleri ile model geliştirilmesinin ve uygulanmasının yalnızca veri bilimcilerine bırakılmaması su uzmanları ile birlikte yapılmasının gerekliliği olmak üzere üç önemli husus vurgulanmıştır.

Yousefi (Yousefi, 2020) çalışmasında Kanada Kelowna şehrinin su tüketim veriseti kullanılmış, doğrusal olmayan yerel yaklaşım (NLA), MLP-NN, gen ifade programlama (GEP) ve çoklu doğrusal regresyon (MLR) algoritmaları ile kısa vadeli su tüketim tahmin modelleri üretilmiştir. Çalışma sonunda RMSE, MAE performans metrikleri ile modeller karşılaştırılmış, NLA ile elde edilen modelin tüm modeller arasında %98 doğrulukla en iyi performansa sahip olduğu gösterilmiştir.

Arslan (Arslan, 2020) çalışmasında Eskişehir kent merkezi Abone Bilgi Yönetim Sistemi (ABYS) ve Coğrafi Bilgi Sistemi (CBS) verileri incelenmiş kurumların olmadığı sadece hanelere ait tüketimlerin olduğu giriş ve çıkışlarında içme suyu şebekesi SCADA sistemi üzerinden basınç kırıcı vanalar ile kontrol edilebilen ve 15 dakikada bir ölçüm yapılan 4 ölçülebilir alt bölge (DMA) belirlenmiştir. Ayrıca çalışmada minimum gece debisinin konut su kullanımının en düşük, basıncın en yüksek olduğu 02.00 ve 04.00 saatleri arasında gözlemleneceği yaklaşımı benimsenmiş bu zaman aralığında elde edilen günlük en düşük gece debi bilgilerinden bölgelerin aylık fiziksel su kaybı oranları hesaplanmıştır. Çalışma sonucunda, aylık fiziksel su kaybı oranı Gündoğdu'da 2011'de %11,5, 2019'da %23,59, Sazova'da 2011'de %3,9, 2019'da %28,5, Şirintepe'de 2019'da %20,0, Ertuğrulgazi'de 2019'da %46,35 olarak tespit edilmiştir. Fiziksel su kaybı oranının yüksek olduğu Ertuğrulgazi'de su dağıtım sistemi basıncı birer haftalık aralıklarla debiye duyarlı ve kritik noktaya duyarlı olarak yönetilmiş, sabit basınçla yönetildiği bir haftayla kıyaslanarak, debiye duyarlı yönetimde %3,84'lük, kritik noktaya duyarlı yönetimde %9,36'lık su tasarrufu sağlandığı tespit edilmiştir.

Bulut (Bulut, 2020) çalışmasında Erzincan ili Ocak 2014-Ocak 2020 yılları arasındaki içme suyu verileri kullanılarak elde edilen su dengesi tablosu incelenmiş, Mart 2018'de şebeke yenileme ve SCADA sisteminin aktif edilmesiyle birlikte sistem öncesi ve sonrası içme suyu verileri yıllık olarak karşılaştırılmıştır. Çalışma sonunda SCADA öncesi toplam fiziksel ve idari kayıp su oranı %64,46, fiziksel su kayıp oranı ise %27,89 iken sonrasında fiziksel ve idari kayıp su oranı %36,53, fiziksel su kayıp

oranı ise %13,64 olarak tespit edilmiş, SCADA sisteminin su kayıplarını azalttığı gösterilmiştir.

Akpınar ve Özçelik (2020) çalışmalarında, Avrupa'da en çok tercih edilen sabit gerçek zamanlı protokollerden biri olan EtherCAT tabanlı bir su seviye kontrol sistemi test yatağı laboratuvar ortamında geliştirilmiş, 38 öznitelikli bir veriseti oluşturulmuştur. Veriseti Temel Bileşen Analizi (PCA) algoritması ile öznitelik seçimi yapılmış anormal su seviyesi tespiti için SVM, Simple Logistic, RF olmak üzere 3 denetimli ve Canopy, Expectation Maximization, Farthest First, Filtered Clusterer, Hierarchical Clusterer, Make Density Based Clusterer, Simple K Means olmak üzere 7 denetimsiz öğrenme yöntemi ile tahmin modelleri geliştirilmiş, modeller Doğruluk, Keskinlik, Duyarlılık performans metrikleriyle karşılaştırılmıştır. Çalışma sonucunda denetimli modellerin özellikle SVM ve RF'ın en kabul edilebilir sonuçlar verdiği, denetimsiz modellerden Hierarchical Clusterer, tüm izler için %86 en iyi Doğruluk oranına sahip, Farthest First %96 ile en yüksek Keskinlik oranına sahip, ayrıca Filtered Clusterer %92 ile en iyi Duyarlılık oranına sahip olduğu ve her iki yöntemin de EtherCAT tabanlı sistemlerde olayları algılamak için uygun olduğu gösterilmiştir.

Körlü (Körlü, 2019) çalışmasında Hatay Reyhanlı bölgesinde bulunan Devlet Su İşleri Genel Müdürlüğü (DSİ) rasat kuyusuna ait 2000 ile 2015 yılları arasındaki yeraltı su seviyesi aylık ölçüm değerleri ile bölgenin aylık ortalama yağış ve sıcaklık değerlerinden oluşan veriseti kullanılmış SVM, RBK-SVM, Poli Çekirdek Fonksiyonlu SVM (PK-SVM), Uyarlamalı-Ağ Tabanlı Bulanık Çıkarım Sistemi (ANFIS), MLR, ANN algoritmaları ile yeraltı suyu seviyesi tahmin modelleri geliştirilmiş modeller MSE, MAE ve R performans metrikleri kullanılarak karşılaştırılmıştır. Çalışma sonucunda RBK-SVM ve PK-SVM'nin yeraltı suyu seviyesini daha iyi tahmin ettiği gösterilmiştir.

Jalal ve arkadaşları (2019) çalışmalarında Tunus'un Ghedir El Golla su arıtma istasyonuna ait su dağıtım sisteminde kablosuz sensör ağları kullanılarak toplanan 2018 yılına ait kimyasal, fiziksel ve mikrobiyolojik kategorilerde olmak üzere 38 adet su parametresinden ve 103 kayıttan oluşan veriseti kullanılmıştır. Su dağıtım sistemine herhangi bir müdahalenin önlenmesi ve kirliliğin bir an önce tespit edilmesi için SVM, DT ve KNN ML algoritmaları kullanılarak içme suyu kalitesini sınıflandıran ve

tanımlayabilen AI modelleri geliştirilmiştir. Modeller doğruluk, kesinlik ve duyarlılık performans metrikleriyle karşılaştırılmış Doğrusal Çekirdek Fonksiyonlu SVM (L-SVM) modelinin çalışmada önerilen uygulama için yeterli olduğu gösterilmiştir.

Zhang Y. ve arkadaşları (2019) çalışmalarında Çin genelinde 45 adet içme suyu arıtma tesisinden aylık alınan su kalitesi ve işletme ile ilgili parametre bilgilerinden oluşan veriseti kullanılmış, işletmenin genel performansını tahmin etmeyi amaçlayan ANN ve Genetik Algoritma (GA) birleştirilmiş HANN adlı hibrit bir istatistiki model geliştirilmiştir. Çalışma sonucunda daha fazla eğitim verisi kullanılarak HANN model R^2 performansının 0,71'den 0,93'e çıktığı bu nedenle gelişme potansiyeline sahip olduğu gösterilmiştir.

Zhu ve arkadaşları (2019) çalışmalarında Tuna Nehri'nin en büyük kollarından biri olan Hırvatistan'daki Drava Nehri üzerindeki 4 istasyonun uzun gözlem verilerine sahip iki tanesine ait istasyon adı, kotu, havza alanı, kalibrasyon periyodu, validasyon periyodu, meteoroloji istasyonu adı, meteoroloji istasyonu kotu, uzaklık, parametrelerinden oluşan veriseti kullanılmış nehir suyu sıcaklık tahmini için MLP-NN, ANFIS ve MLR modelleri geliştirilmiştir. Modeller R, Willmott anlaşma indeksi, RMSE, MAE, performans metrikleriyle karşılaştırılmış, çalışma sonucunda nehir suyu sıcaklığı tahmininde dalgacık dönüşümü (WT), MLP-NN ve ANFIS modellerinin birleştirilmesi ile elde edilen hibrit modellerin aşırı iklim koşulları için nehir suyu sıcaklığı tahmininde normal modellerden daha iyi performans potansiyeli taşıdığı gösterilmiştir.

Liu Z. ve arkadaşları (2019) çalışmalarında bir anaerobik atık su arıtma sisteminde atık kimyasal oksijen ihtiyacının (KOİ) tahmini için en küçük kareler SVM (LS-SVM) kullanılarak bir tahmin modeli geliştirilmiş, R, MAPE ve RMSE performans metrikleriyle model performansı değerlendirilmiştir. Çalışma sonucunda dinamik durumlu LS-SVM modeliyle tahmin ve gerçek veriler arasındaki R 0,9752 olarak hesaplanmış çok iyi tahmin performansına sahip olduğu gösterilmiştir.

Jin ve arkadaşları (2019) çalışmalarında Çin'in Ashi Nehri'nde gerçek zamanlı amonyak azotu (NH_3-N), bulanıklık ve iletkenlik olmak üzere 3 su kalitesi parametresinin 8 Haziran 2014 (00.00) ile 26 Ekim 2014 (20.00) tarihleri arasında 4

saatlik aralıklarla ölçümü yapılarak elde edilen 846 adet TS gözlem verisi kullanılmış ve elde edilen veriseti %70 eğitim ve doğrulama, %30 test olmak üzere ikiye bölünmüştür. Erken uyarı için yüzey suyu kalite parametre değişimlerini tahmin eden GA ile BP-NN entegre edilerek GA-BP-NN ve normal BP-NN modelleri geliştirilmiş modeller tahmin aralığı kapsama yüzdesi (PICP) performans metriği kullanılarak karşılaştırılmıştır. Çalışma sonucunda önerilen model ile $\text{NH}_3\text{-N}$ %95,13, bulanıklık %100 ve iletkenlik %100 PICP performans değerleri elde edilmiş dolayısıyla modelinin mevcut yetersiz veriler kullanılarak belirli konumlar için su kalitesi değişimini analiz etmede etkili bir teknik olabileceği gösterilmiştir.

Gerger ve Aslan (2019) çalışmalarında Temmuz 2015 ile Haziran 2016 dönemleri arasında Şanlıurfa Su ve Kanalizasyon İdaresi (ŞUSKİ) tarafından elde edilen veriler kullanılmış, IWA tarafından geliştirilen WAS v5.0 programı ile standart su dengesi formu oluşturulmuş, yaklaşık %49,4 oranında kayıp kaçak olduğu tespit edilmiştir. Çalışma sonucunda fiziki kayıpların azaltılması için içme suyu şebekesinin 108 adet DMA planlanması, 85 DMA'ya basınç kırıcı vana takılarak hidrolik model oluşturulması, SCADA sistemi ile basınç kontrolü yapılması, sahada aktif kaçak kontrollerin yapılması, ölçüm hassasiyeti azalan sayaçların değiştirilmesi ve park, bahçe, cami gibi gelir getirmeyen su tüketimlerinin kayıt altına alınması önerilmiştir.

Zhang Q. ve arkadaşları (2019) çalışmalarında Kanada, Barrie'deki Barrie AAT'nin 1 Kasım 2015 ile 31 Ekim 2016 tarihleri arasında on beş dakika aralıklarla ölçülmüş, 1 yıllık 35.132 adet akış hızı verisi kullanılmış, veriseti %66 eğitim ve %34 test için olmak üzere iki bölüme ayrılmış ARIMA ve MLP-NN algoritmaları ile atık su akışını tahmin etmek için TS modeller geliştirilmiş, modeller RMSE, MAPE ve NSE performans metrikleriyle karşılaştırılmış, çalışma sonucunda ARIMA modelinin MLP-NN modelinden daha iyi bir tahmin performansına sahip olduğu ancak birbirine çok yakın olduğu gösterilmiştir.

Önen ve arkadaşları (2018) çalışmalarında, Diyarbakır kenti için 2005–2014 yılları arasında ilgili kurum ve kuruluşlardan temin edilen şebekeye verilen su miktarı, şebekede oluşan su kayıpları, tahakkuk, sıcaklık, nem, nüfus, yağış, gelişmişlik durumunu gösteren GSYH ve içme suyu abone sayısı gibi parametrelerden oluşan bir veriseti kullanılmış, gelecekte içme suyu ihtiyaç tahmini için bir yapay zekâ tekniği

olan GEP yöntemiyle verisetindeki girdi parametreleri rastgele çıkarılarak 19 farklı model geliştirilmiştir. Çalışma sonucunda R^2 ve RMSE performans metrikleriyle modeller karşılaştırılmış, tahakkuk, nem, yağış, sıcaklık, su kayıpları, nüfus, gelişmişlik ve içme suyu abone sayısı parametreleri ile oluşturulan modelin R^2 0,982 ve RMSE 3,89 performans değerleriyle en iyi talep tahmin modeli olduğu gösterilmiştir.

Seo ve arkadaşları (2018) çalışmalarında Güney Kore'nin altı şehrine (Anseong-si, Hwaseong-si, Pyeongtaek-si, Osan-si, Suwon-si ve Yongin-si) içme suyu sağlayan Suji su arıtma tesisini işleten K-water'dan temin edilen 2008-2017 yılları arasındaki günlük su talep verileri kullanılmış, veriseti %70 eğitim %30 test için bölünmüştür. Veriseti ile varyasyon modu ayrıştırma (VMD) ile ELM ve ANN kullanılarak su talebi tahmin modelleri geliştirilmiş modeller MAE, RMSE, dördüncü kök ortalama dördüncü kuvvet hata (R4MS4E), ortalama mutlak bağıl hata (MARE), medyan mutlak yüzde hata (MdAPE), modifiye verimlilik katsayısı (MCE), modifiye anlaşma indeksi (MIOA), performans metrikleriyle karşılaştırılmıştır. Çalışma sonucunda, ELM modelinin ANN modelinden VMD-ANN ve VMD-ELM modellerinin, ANN ve ELM modellerinden daha iyi performansa sahip oldukları ve VMD-ELM modelinin tüm modeller arasında en iyi performansı sağladığı gösterilmiştir.

Allawi ve arkadaşları (2018) çalışmalarında rezervuar ile ilgili daha önce yapılmış 30 çalışma seçilmiş, detaylı incelenmiş, ANN ve bulanık mantık (FL) yöntemlerinin rezervuar su giriş hızı ve buharlaşma oranı tahmininde yüksek düzeyde doğruluğa sahip sağlam modeller sağlayabildikleri gösterilmiştir.

Yüksel ve arkadaşları (2018) çalışmalarında SCADA, ABYS ve CBS bir arada kullanılmış, Diyarbakır Büyükşehir Belediyesi merkez ilçelerinin (Sur, Bağlar, Yenişehir ve Kayapınar) içme suyu şebeke ve abone boru metrajları, nüfus, abone sayıları ve abone türleri (meskenler, ticarethaneler, resmî kurumlar ve umuma açık yerler) gibi parametreler dikkate alınmış Şehitler-1 ve Şehitler-2 olarak belirlenen pilot bölgelerde şebekedeki su kayıpları tespit edilmiştir. Çalışma sonucunda Diyarbakır ili içme suyu şebekesindeki kayıpların bir hidrolik çalışma yöntemi ile belirlenmesine ve önlenmesine yönelik daha somut ve etkili çözüm önerileri sunulmuştur.

Şen (2018) çalışmasında Kocaeli Su ve Kanalizasyon İdaresi (ISU) Genel Müdürlüğündeki 222 istasyonlu içme suyu SCADA sisteminin 02.04.2014 tarihi ile 10.04.2014 tarihleri arasında Derince Çamlık Terfi istasyonuna ait üst kademe su deposunun su seviyesi (A_n), üst kademe su deposunun şebeke çıkışı debisi (B_n), sıcaklık (C_n), terfi istasyonundaki motorların çalışma ve durma süreleri (D_n), terfi istasyonundaki motorların çalışma ve durma sıraları (E_n) parametreleri giriş, terfi motor durumları (K_n) parametresi çıkış verisi olarak belirlenmiş, farklı giriş parametre kombinasyonları ($A_n-B_n-C_n-D_n-E_n$, $A_n-B_n-D_n-E_n$, $A_n-B_n-D_n$ veya A_n-D_n) ile oluşturulan ANN modellerinin %94,44 oranında başarılı sinyal ürettiği tespit edilmiştir. Çalışma sonucunda terfi motorlarını kontrol eden uzak terminal ünitesine (RTU) geliştirilen model lojik olarak kodlanmış gerçek çalışma koşullarında ve haberleşme kesintilerinde dahi SCADA sistemlerinde sezgisel yaklaşım ile çıkış sinyali üretilebileceği gösterilmiştir.

Başa ve Kurt (2017) çalışmalarında Tekirdağ Su ve Kanalizasyon İdaresi (TESKİ) tarafından gerçekleştirilen SCADA Uygulaması, CBS Uygulamaları, Bilgi İşlem Alt yapı Uygulamaları, ABYS, Elektronik Belge Yönetim Sistemi (EBYS) uygulamaları incelenmiştir. Çalışma sonucunda akıllı su yönetimi uygulamalarının su gelirlerini arttırdığı, işletme maliyetlerini düşürdüğü, su kalitesini yükselttiği, vatandaşlara yüksek standartlarda hizmet sağladığı, abone ve su idarelerinin memnuniyetini arttırdığı gösterilmiştir.

Ghalehkhondabi ve arkadaşları (2017) çalışmalarında 2005'ten 2015'e kadar su talep tahmin modelleri ile ilgili yayımlanmış makaleler gözden geçirilmiş, modellerde kullanılan verisetlerinin genellikle bir şehir veya bir ülkeye ait veriler olduğu, talep tahminini etkileyen değişkenlerin tüketim, iklimsel ve sosyoekonomik faktörler olmak üzere üç ana kategoride değerlendirildiği, denetimsiz öğrenme, derin öğrenme, metasezgisel yöntemler ve topluluk öğrenme yöntemleri ile modeller geliştirildiği, modellerin genelde kısa vadeli tahminlerde kullanıldığı, orta ve uzun vadeli tahminlerle sosyoekonomik etkenlerin dikkate alınmadığı tespit edilmiştir. Çalışma sonucunda modeller ile ilgili hibrit yapıların geliştirildiği, gelişmelerin etkisiyle zamanla iyileştirmeler olacağı daha doğru tahmin üreten modellere ihtiyaç duyulduğu belirtilmiştir.

March ve arkadaşları (2017) çalışmalarında akıllı su sayaçlarının kullanımı, uzaktan anlık ölçümlerin yapılabilmesi ve su tüketim taleplerinin dönemsel süreçlerde değişimi ile ilgili güneydoğu İspanya şehri Alicante şehrinin su dağıtım ve yönetim şirketinde uzaktan su sayacı okunmasından sorumlu yönetici ve çalışanlarla Mayıs ve Haziran 2016, Ocak ve Mart 2017 tarihlerinde dört görüşme gerçekleştirilmiş. Çalışma sonucunda akıllı su sayaçlarının anlık tüketim kalıplarını tespit etmede faydalı olabileceği bu verilerden hareketle fiyatlandırma ve bilgilendirme yapılarak tüketicilerde daha sorumlu davranışlar geliştirilebileceği ve fiyatlandırma politikaları belirlenebileceği tespitleri yapılmıştır.

Akdağ (2016) çalışmasında Diyarbakır kent merkezi 2003-2013 yıllarına ait su talebini etkileyen ortalama hava sıcaklığı, ortalama yağış miktarı, ortalama nisbi nem, kent merkezi nüfusu, tüketici fiyat endeksi (TÜFE), su fiyatı, abone başına düşen su faturası tutarı faktörlerine ve gerçekleşen su talebine ait aylık değerler kullanılmış, su talep tahmini için ANN, SVM ve Box-Jenkins yöntemleri ile tahmin modelleri geliştirilmiş modeller verimlilik (E), MSE, RMSE, MAPE, R, R^2 performans metrikleriyle karşılaştırılmıştır. Çalışmanın sonucunda tüm performans ölçütlerinde SVM ile ARIMA modelleri birbirine yakın olmakla birlikte ARIMA modelinin daha iyi olduğu, ANN modelinin de %3,61 hata payıyla en iyi performansı sağladığı dolayısıyla gerçeğe daha yakın tahmin ürettiği gösterilmiştir.

Aydoğdu ve Fırat (2015) çalışmalarında Malatya su dağıtım şebekesinde 2006-2012 yılları arasında gözlemlenen arıza kayıtları dikkate alınmış arıza oranına etkisi olan faktörlerin belirlenmesi için bulanık kümeleme yöntemi ile 9 alt bölge tanımlanmış her bir bölgeye ait veriseti ve kümeleme yapılmamış veriseti kullanılmış, LS-SVM, İleri Beslemeli Sinir Ağı (FF-NN) ve GR-NN algoritmaları ile arıza oranı tahmin modelleri geliştirilmiş, modeller R, E ve RMSE performans metrikleriyle karşılaştırılmıştır. Çalışma sonucunda her bir veri seti için LS-SVM modelinin FF-NN ve GR-NN modellerine göre daha yüksek performansa sahip olduğu gösterilmiştir.

Yang ve arkadaşları (2014) çalışmalarında Zhi-Tan AAT'nin Ocak 2006 ile Aralık 2010 arasındaki giriş suyunun bulanıklık verileri ve Taipei Su Kaynağı Alanında meydana gelen 1 ve 2 günlük yağış kayıtları varyans analizi (ANOVA) ile incelenmiş %3 üzerinde anlamlılığa sahip parametreler girdi değişkenleri olarak seçilmiş, ANN

algoritması kullanılarak bir bulanıklık tahmin modeli ve bir erken uyarı sistemi geliştirilmiştir. Çalışmanın sonucunda önerilen modelin %94,44 ile %98,61 oranları arasında doğru tatmin ürettiği gösterilmiştir.

Almalawi ve arkadaşları (2014) çalışmalarında su dağıtım sistemine ait bir SCADA sisteminde su akışı, basınç, talep miktarı, seviye, vana durumu ve ayarı, pompa durumu ve hızı verilerinden oluşan üç farklı veriseti kullanılmış, verisetlerinden biri internet ortamında halka açık paylaşılan gerçek bir veriseti, diğer ikisi de simülasyon sistemi tarafından üretilmiştir. Verisetlerindeki her bir gözlem değerinin KNN algoritması ile bir kümeleme yapılarak aykırı anormal gözlemlerin tespitini sağlayan bir denetimsiz model geliştirilmiştir. Çalışma sonucunda modeller tespit oranı, yanlış pozitif, keskinlik, F-skor performans metrikleriyle karşılaştırılmış, gerçek verisetiyle geliştirilen modelin hesaplama süresi açısından uzun sürdüğü ancak tutarlı ve tutarsız durumları %92,86 oranında doğrulukta tespit edebildiği gösterilmiştir.

Akdağ (2014) çalışmasında Diyarbakır ili kent merkezi içme suyu tüketimine dair 2009-2013 yılları arasındaki su faturası, su fiyatı, enflasyon oranı, hane halkı geliri, hane halkı büyüklüğü, yağış miktarı, hava sıcaklığı, nem oranı, rüzgâr hızı ve nüfus verileri kullanılmış, korelasyon matrisi yöntemi ile talebi etkileyen faktörler belirlenmiş, elde edilen veriseti kullanılarak ANN, Winters Üstel Düzeltme ve ARIMA algoritmaları ile talep tahmin modelleri geliştirilmiştir. Çalışma sonucunda modeller E, MSE, RMSE, MAPE, R ve R^2 performans metrikleriyle karşılaştırılmış, önerilen ANN modeline ait E %92,4, MSE 57.358.296.063, RMSE 239.496 ve MAPE değeri ise %3,6 olarak elde edilmiş, dolayısıyla çok küçük oranda hatalı sonuçlar ürettiği, tahmin tutarlılığının oldukça yüksek olduğu gösterilmiştir.

Bennett ve arkadaşları (2013) çalışmalarında South East Queensland Konut Son Kullanım Çalışması (SEQREUS) tarafından sağlanan 205 haneye ait ayrıştırılmış su son kullanım talebi, hane geliri, eğitim seviyesi, yetişkin sayısı, çocuk sayısı, genç sayısı, tuvalet, duş, çamaşır makinesi, musluk, bulaşık makinesi gibi demografik, sosyo-ekonomik ve su cihazı kullanım bilgilerini içeren kapsamlı bir veriseti kullanılmış, geleneksel ANN algoritması ile tuvalet, çamaşır yıkama, duş, bulaşık makinesi, musluk ve toplam iç talep tahmini ile ilgili üç farklı tahmin modeli geliştirilmiştir. Modeller ARE, AAE, RMSE ve R^2 performans metrikleriyle

karşılaştırılmış bulaşık makinesi ve banyo talebi tahmin modellerinin düşük, diğer tahmin modellerinin de orta düzey güvenilir olduğu gösterilmiştir.

Adamowski J. ve arkadaşları (2012) çalışmalarında Kanada Montreal şehrine ait Mayıs 2001-Ağustos 2009 tarihleri arasında günlük maksimum sıcaklık, toplam yağış ve günlük su tüketimi verileri kullanılmış, MLR, MNLR, ARIMA, ANN ve dalgacık dönüşümlü NN (WT-NN) algoritmaları ile günlük su talep tahmin modelleri geliştirilmiştir. Çalışma sonucunda modeller R^2 , E, RMSE, RRMSE performans metrikleriyle karşılaştırılmış WT-NN modellerinin MLR, MNLR, ARIMA ve ANN modellerinden daha doğru günlük su talep tahmini sağladığı gösterilmiştir.

Ajbar ve Ali (2012) çalışmalarında Suudi Arabistan'daki Mekke şehrinin 2003-2006 yılları arasındaki yıllık hane geliri, yıllık hane yoğunluğu, yıllık şehir nüfusu, aylık beklenen ziyaretçi sayısı, aylık maksimum sıcaklık, yıl boyunca ve ay takvimi aylarına göre Hac ve Umre ibadetleri için gelen misafirlerin ve ülke halkının şehir nüfusuna olan yansımaları ve resmî kurumların uyguladığı kotalar ve planlamalara ilişkin veriler kullanılmış, ANN algoritması ile aylık ve yıllık su talebi tahmin modelleri geliştirilmiştir. Çalışma sonucunda hem aylık hem de yıllık su talebi tahmin modelleri 2009-2010 su tüketim verileriyle doğrulanmış, modellerin girdi parametrelerinin etkisini analiz etmede karar vericilerin elinde faydalı bir araç olabileceği ancak sistemdeki sızıntılar, değişen politikalar, sosyal alışkanlıklar nedeniyle anormallikler ve düzensizlikler olabileceği gösterilmiştir.

Yıldız (2011) çalışmasında Ankara Su ve Kanalizasyon İdaresi (ASKİ) 2007 yılı başından 2010 yılı ilk çeyreğine kadarki saatlik SCADA verileri ile Meteoroloji İşleri Genel Müdürlüğü (MGM) hava sıcaklığı ve bağıl nem gibi iklimsel verileri kullanılmış, ANN algoritması ile günlük su harcama eğrisi tahmin modeli geliştirilmiş, ASAT Konyaaltı bölgesindeki mevsimsellik özelliği ihtiva eden ve belirlenmiş tüketim kategorileri olan SCADA verileriyle 16 gün için test edilmiştir. Çalışma sonucunda model R^2 performans metriği ile değerlendirilmiş, Ankara veriseti ile ortalama %75, Antalya veriseti ile %85 doğrulukta tahmin üretebildiği ayrıca model ile kısa dönem planlama ihtiyaçlarının tahmininin yapılabilmesi ve bu tahmin değerlerinin basınç yönetimi için optimum vana ayarı çalışmalarında kullanılabileceği gösterilmiştir.

Adamowski J. ve Karapataki (2010) çalışmalarında Kıbrıs'ın Lefkoşa şehrinde iki farklı bölgenin (Athalassa ve Halk Bahçesi) 2002-2007 yılları arasında toplamda 6 yıla ait haftalık en yüksek su talebi, maksimum sıcaklık, toplam yağıştan oluşan meteorolojik verileri kullanılmış, veriseti %70 eğitim %30 test için bölünmüş, MLR, Levenberg-Marquardt NN (LM-NN), esnek geri yayılım NN (RP-NN) ve eşlenik gradyan Powell-Beale NN (CGPB-NN) algoritmalarıyla her birinden 20 farklı haftalık maksimum su talep tahmin modeli geliştirilmiş, modeller R^2 , RMSE, ortalama mutlak bağıl hata (AARE), maksimum mutlak bağıl hata (maxARE) performans metrikleriyle karşılaştırılmıştır. Çalışma sonucunda LM-NN modelinin, MLR, CGPB-NN ve RP-NN modellerinden daha doğru tahmin ürettiği ayrıca haftalık en yüksek su talebinin belirlenmesinde yağış miktarından ziyade yağışın oluşumunun daha önemli bir etken olduğu gösterilmiştir.

Herrera ve arkadaşları (2010) çalışmalarında İspanya'da yaklaşık 5.000 nüfuslu bir DMA'den Ocak 2005'ten Nisan 2005'e kadar saatlik olarak ölçülmüş 8 haftalık 1.344 adet su tüketim verisi ile günlük olarak ölçülmüş sıcaklık, rüzgâr hızı, yağmur miktarı ve atmosfer basıncı verileri kullanılmış, SVR, çok değişkenli uyarlanabilir regresyon eğrileri (MARS), projeksiyon izleme regresyonu (PPR) ve RF algoritmaları ile su tüketim tahmin modelleri geliştirilmiştir. Çalışma sonucunda modeller E, MAE ve RMSE performans metrikleriyle karşılaştırılmış, en doğru tahmin üreten modelin SVR olduğu sonrasında sırasıyla MARS, PPR ve RF modelleri olduğu gösterilmiştir.

Karakuş ve arkadaşları (2010) çalışmalarında Sivas Belediyesi tarafından iki mahalle (Esentepe ve Yunus Emre) ultrasonik sayaç ve basınç ölçerler takılarak bir DMA olarak belirlenmiş, 24 saat boyunca kayıt altına alınmış basınç ve debi verileri ile aynı bölgedeki konut, cami, okul, park-bahçe abonelerinin aylık ve günlük su sarfıyat bilgileri kullanılmıştır. Çalışma öncesinde içme suyu şebekesindeki kayıp-kaçak oranı %40,43 olarak hesaplanmış, sahada gece saatlerinde ana borularda kaçak arama cihazı ile dinleme yapılmış, 7 adet şebeke arızası tespit edilmiştir. Çalışma sonucunda arızalar giderilmiş, kayıp-kaçak oranı %27,22'ye düşürülmüş ve yapılan çalışmaların şehir geneline yayılması halinde oranın daha da aşağılara düşürülebileceği gösterilmiştir.

Yavuz (2009) çalışmasında Sakarya Büyükşehir Belediyesi Adapazarı Su ve Kanalizasyon İdaresi (ADASU) abonelerinden rastgele seçilmiş 20.000 abonenin abone sicil no, abone türü, sayaç durumu verileri ile 01.01.2007–20.08.2009 tarihleri arasındaki sarfiyat, tahakkuk, tahsilat ve varsa kaçak su tüketimi tespit edilenlerin ceza tutarı, ceza yılı, ceza ayı, ceza günü, ceza türü, tahakkuk dönemi (gün, ay, yıl), tahsilat tipi, tahsilat türü, gecikme tutarı ve ödeme tarihi verileri kullanılmış, veriseti %30 eğitim %70 test için bölünmüştür. Çalışma sonucunda veri madenciliği tekniklerinden ANN, LJR ve DT algoritmalarıyla kaçak su tüketimi yapan abone tahmin modelleri geliştirilmiş, modeller birbirine yakın değerlerde yaklaşık %85 doğrulukta tahmin üretmiş ve en iyi modelin ANN modeli olmasının yanısıra DT modelinin tercih edilebilir uygunlukta olduğu gösterilmiştir.

Cinal (2009) çalışmasında Sakarya ili Adapazarı ilçe merkezinde su şebekesinin genelini temsil edebilecek karakteristiğe sahip 2.440 adet konut, 654 adet ticari olmak üzere toplamda 3.094 abonenin bulunduğu ve tek bir noktadan su girişi sağlanan bir DMA pilot bölge olarak oluşturulmuş, su girişinin olduğu yere debimetre ve basınç düzenleyici vana yerleştirilmiş, 22-28 Ocak 2009 tarihlerinde basınç düzenleyici vana tam açık iken, 29.01.2009-04.02.2009 tarihleri arasında vana açıklığı sabit 48m basınçta iken ve 12-18 Şubat 2009 tarihlerinde ise vana açıklığı gece 41m, gündüz 48m basınçta iken şebeke test edilmiş ve basınç, debi ve sarfiyat verileri kaydedilmiştir. Çalışma sonucunda önerilen basınç kontrolü yöntemiyle hizmette herhangi bir aksamaya mahal vermeden %38 kayıp su oranının %21 seviyelerine kadar azaltıldığı gösterilmiştir.

Adamowski J. F. (2008) çalışmasında Kanada, Ottawa şehrinde dış mekân su kullanımının yüksek olduğu bir alan için her yılın Mayıs-Ağustos ayları süresince 10 yıllık en yüksek günlük su talebi verileri ve günlük meteorolojik değişkenler (maksimum sıcaklık ve yağış) verileri kullanılmış, günlük yaz suyu talebi tahmini için MLR, ARIMA ve ANN algoritmaları ile modeller geliştirilmiştir. Çalışma sonucunda modeller AARE, MaxARE ve R^2 performans metrikleriyle karşılaştırılmış en iyi tahminin giriş verileri bir önceki günün en yüksek su talebi, mevcut ve önceki günün maksimum sıcaklığı ve beş gün öncesine ait yağışların oluşup oluşmadığı

kullanıldığında ANN modeliyle elde edildiği yağış miktarından ziyade yağış oluşumunun tüketimde etkili olduğu gösterilmiştir.

Ghiassi ve arkadaşları (2008) çalışmalarında Amerika'daki San Jose Su Şirketi'nden temin edilen haftalık değerler için altı aylık döneme (26 hafta) ve aylık değerler için iki yıllık döneme (24 ay) ait günlük pompalanan su miktarı ve saatlik su talep miktarı verileri kullanılmış, dinamik ANN (DAN2), ARIMA ve ileri beslemeli geri yayılım NN (FFBP-NN) algoritmaları ile aylık, haftalık, günlük ve saatlik su talebi modelleri geliştirilmiştir. Çalışma sonucunda modeller MAPE performans metriğiyle karşılaştırılmış DAN2 aylık, haftalık ve günlük modeli %99 ve saatlik modeli %97 oranlarında doğru tahmin ürettiği ARIMA ve FFBP-NN'dan daha iyi performansa sahip olduğu gösterilmiştir.

Caiado (2008) çalışmasında İspanya'nın 1 Ocak 2001 ile 30 Haziran 2006 tarihleri arasında günlük su tüketim verileri kullanılmış, 1 Ocak 2001-31 Mayıs 2006 tarihleri arasındaki 1.976 gözlem eğitim 1-30 Haziran 2006 tarihleri arasındaki 30 gözlem test verisi olmak üzere veriseti ikiye bölünmüş, günlük ve haftalık talep tahmini için Holt-Winters, ARIMA ve genelleştirilmiş otoregresif koşullu değişen varyans (GARCH) algoritmaları ile tek değişkenli TS modelleri geliştirilmiştir. Çalışma sonucunda modeller MSE performans metriği ile karşılaştırılmış, ARIMA ve GARCH modellerinin hem günlük hem de haftalık tahmin periyotlarında aynı performansa sahip olduğu, Holt-Winters modelinin uzun vadeli tahminlerde ARIMA ve GARCH modellerinden daha iyi olduğu, bir ila dört adım (gün veya hafta) ilerisi için ARIMA ve GARCH modellerinin Holt-Winters modellerinden daha iyi performansa sahip olduğu gösterilmiştir.

Cutore ve arkadaşları (2008) çalışmalarında İtalya Katanya şehri su dağıtım şebekesinden beslenen ve şehrin karakteristiğini taşıdığı düşünülen pilot bir alanda 01 Ocak 2003–31 Aralık 2004 tarihleri arasında uzaktan ölçüm yapılarak kayıt altına alınan tüketim verileri kullanılmış, günlük su talep tahmini için Monte Carlo Markov zinciri (MCMC) yöntemi ile geliştirilmiş, Karıştırılmış Karmaşık Evrim Metropolis algoritması (SCM-UA) ile optimize edilen bir ANN, Bayes yöntemi ile optimize edilen bir ANN, Regresyon ve ANFIS algoritmaları ile modeller geliştirilmiştir. Çalışma sonucunda modeller RMSE ve E performans metrikleriyle karşılaştırılmış

önerilen modelin diğer 3 modele yakın oranlarda günlük su tüketim tahmini yapabildiği gösterilmiştir.

Düğenci (2007) çalışmasında Kaliforniya Irvine Üniversitesi ML kütüphanesinde yer alan 9 giriş, 6 çıkış parametresinden oluşan atıksu veritabanı (UCI, 2006) ile ADASU Karaman AAT'nin 6 giriş, 4 çıkış parametresinden oluşan veriler kullanılmış, verisetlerindeki giriş parametreleri ile ANN, GA ve yapay arı kolonisi (ABC) algoritmaları ile eğitilmiş AAT çıkış parametreleri tahmin modelleri geliştirilmiştir. Çalışma sonucunda modellerin Askıda Katı Madde, BOİ, KOİ, pH gibi çıkış kirlilik parametre değerlerini yaklaşık %85 doğrulukla tahmin ettiği gösterilmiştir.

Alvisi ve arkadaşları (2007) çalışmalarında Ferrara Üniversitesinin İçme Suyu Dağıtım Yönetimi araştırma projesi kapsamında İtalya'nın Emilia Romagna şehrinde Castelfranco Emilia Belediyesinden kısa vadeli debi, basınç ve su talebi ölçümleri kullanılmış, ANN algoritması ile saatlik ve günlük su talep tahmin modeli geliştirilmiştir. Çalışma sonucunda modeller açıklanan varyans (EV), RMSE ve MAE performans metrikleriyle karşılaştırılmış gelecekteki su taleplerinin saatlik olarak doğru ve sağlam bir tahmininin sağlanabildiği gösterilmiştir.

Altunkaynak ve arkadaşları (2005) çalışmalarında İstanbul'un 1995-2004 dönemini kapsayan 9 yılın aylık su tüketim verileri kullanılmış, ilk 7 yıla ait veriler eğitim, kalan 18 aya ait olanlar ise test verisi olmak üzere veriseti ikiye bölünmüş, önceki 3 tüketim kaydı girdi olarak belirlenmiş, The Takagi Sugeno bulanık TS yöntemi ile bir tüketim tahmin modeli geliştirilmiştir. Çalışma sonucunda model ARIMA modelleriyle karşılaştırılmış, önerilen yöntemin %90 üzerinde doğru tahmin üretebildiği The Takagi Sugeno bulanık mantığının ARIMA modellerine üstünlüğü gösterilmiştir.

Athanasiadis ve arkadaşları (2005) çalışmalarında gerçek su tüketicilerinin dinamik davranışlarını su fiyatı, simülatör, su tedarikçisi, tüketici topluluğu, meteorolojik parametreleri kullanarak simüle eden DAWN modeli kullanılmış, belirli su fiyatlandırma politikalarının uygulanması ve sonrası toplumsal su tüketim davranışları simüle edilerek beş farklı senaryo üzerinden 1 aylık su tüketim tahmin modelleri geliştirilmiştir. Çalışma sonucunda önerilen model Yunanistan'daki Selanik Belediye Su Temini ve Kanalizasyon Kurumu verileriyle test edilerek doğrulanmış, modellerin

su karar vericilerine su fiyatlandırma ve su yönetimi politikaları hususunda yardımcı olabileceği gösterilmiştir.

Bougadis ve arkadaşları (2005) çalışmalarında Kanada Ottawa şehrinin 1993-2002 yılları arasında sadece yaz ayları (1 Mayıs-3 Eylül) dönemlerine ait haftalık en yüksek su talebi, yağış ve maksimum hava sıcaklığı verileri kullanılmış, verisetinin 1993-2001 yılları arasında 162 haftalık kısmı eğitim, 2002 yılı 18 haftalık kısmı da test için ayrılmış, haftalık en yüksek su talep tahmini için parametrik değişiklikler ile üç farklı ANN, üç farklı MLR ve yedi farklı ARIMA modeli geliştirilmiştir. Çalışma sonucunda modeller AARE, MaxARE ve R^2 performans metrikleriyle karşılaştırılmış, yağış miktarının yağış oluşumundan daha önemli olduğu, en iyi tahmin sonuçlarının, mevcut haftanın yağış ve sıcaklık verisiyle birlikte önceki haftalık talep verisinin kullanımından elde edildiği, ANN modelinin MLR ve ARIMA modellerinden önemli ölçüde daha iyi performansa sahip olduğu gösterilmiştir.

Jain ve Ormsbee (2002) çalışmalarında ABD Kentucky Lexington şehrinin 1982-92 yılları arasında günlük periyotlarda su talebi, maksimum hava sıcaklığı ve toplam yağış verileri kullanılmış, ARIMA ve ANN algoritmaları ile günlük su talep tahmin modelleri geliştirilmiştir. Çalışma sonucunda modeller AARE, MaxARE ve R^2 performans metrikleriyle karşılaştırılmış ANN tekniklerinin, kısa vadeli su talebini tahmin etmek için geleneksel yöntemlerden daha iyi bir performansa sahip olduğu gösterilmiştir.

Jain ve arkadaşları (2001) çalışmalarında Hindistan Teknoloji Enstitüsü, Kanpur kampüsündeki 1989-1998 yılları arasında 98 haftalık su talebi, toplam haftalık yağış ve haftalık ortalama maksimum hava sıcaklığı verileri kullanılmış, verisetinin ilk 78 haftası eğitim, kalan 20 haftası test için olmak üzere ikiye bölünmüş, kısa vadeli talep tahmini için altı farklı ANN, beş regresyon ve iki TS modeli olmak üzere 13 farklı model geliştirilmiştir. Çalışma sonucunda modeller AARE, MaxARE ve R^2 performans metrikleriyle karşılaştırılmış ANN tekniği kullanılarak geliştirilen modellerin, sürekli olarak regresyon ve TS analizi modellerinden daha iyi performansa sahip olduğu gösterilmiştir.

Froukh (2001) çalışmasında İngiltere Wiltshire şehri Swindon bölgesine ait su temini, hane sayısı, hane geliri, yaşam standardı, çeşitli anket ve çalışmalardan elde edilen son kullanım tüketimi ve ACRON sınıflandırması verileri kullanılmış, evsel su talebi tahmini için bir karar destek sistemi modeli geliştirilmiştir. Çalışma sonucunda önerilen sistem ile kullanıcı belirli bir talep alanı seçimi yaparak mevcut verilerin kural tabanlı çıkarım ve soru-cevap şeklinde seçimiyle evsel su talebi tahmini yapabildiği gösterilmiştir.

Billings ve arkadaşları (1999) çalışmalarında sıcaklık, yağış, gelir durumu ve su fiyatı verileri kullanılmış, durum uzayı ve eşit ağırlıklı doğrusal çoklu regresyon yöntemleri ile su talep tahmin modelleri geliştirilmiştir. Çalışma sonucunda su fiyatının sadece birkaç ay denebilecek kısa bir vadede tüketime etkisinin olduğu birkaç ay sonra fiyatı ne olursa olsun her hane temel ihtiyaçları için minimum su tüketimini sürdürdüğü ayrıca su tasarrufu ile ilgili etkili toplumsal eğitimler ile kullanımın değiştirilebileceği gösterilmiştir.

Clarke ve arkadaşları (1997) çalışmalarında İngiltere Leeds, Batı Yorkshire kentinde 4.000 kişilik bir coğrafi bölge örneklem alan olarak seçilmiş yaş, hanedeki birey sayısı, meslek vb. veriler kullanılarak bir mikro simülasyon modeli oluşturulmuştur. Çalışma sonucunda elde edilen simülasyon modeli kullanılarak 200.000 nüfus için su talep tahmini yapılmıştır.

2.2. Anomali Analizleri

Literatürde geçen aykırı değer tanımlamaları Tablo A.1’de detaylı verilmiş olup özetle gözlem noktaları arasında alışılmışın dışında düşük veya yüksek bir değeri, gözlem hatalarını, kayıt hatalarını, veri hazırlama hatalarını, hesaplama hatalarını, kodlama hatalarını veya veri işleme hataları gibi olası değer aralığının dışında kalan bir değeri, bir hata sonucu oluşmamış ama daha özel bir çok anlamlar barındıran beklenmedik ilginç bir değeri, bir tahmin modeli oluşturmaya engel olacak kadar farklı bir değeri, parametrelerle oluşturulmuş bir tahmin modelini ciddi oranda etkileyen veya parametre tahminlerini etkileyen bir değeri, birden fazla veriseti birleştirilirken etkiyi ve ilişkisel yapıyı değiştiren düşük veya yüksek bir değeri, farklı çalışmaları birleştirirken farklı büyüklüklerdeki örneklemelerin etkisiyle gözlemlenen çok farklı bir

değeri, farklı çalışmaları birleştirirken sonuçları ciddi değiştiren bir değeri, zaman serisi (TS) bir verisetinde sürekli bir gözlem şeklini değiştirecek ve/veya aynı zamanda sonraki değerleri de etkileyecek bir değeri, gözlemlerde etkili bir sıçrama ile kalıcı veya geçici bir seviye kaymasına neden olan bir değeri temsil edebilir (Aguinis ve ark, 2013).

Herhangi bir aykırı değer tespiti ile aykırı değerın nasıl bir anlam ifade ettiğı, ne kadar gözardı edilebileceğı, nasıl değerlendirileceğı konusu her olay için farklılıklar taşıdığı gibi müdahale işlemleri konusu da araştırmaya veya verinin kendisine bağılı olarak değişiklikler gösterebilmektedir. Aykırı değerin tespiti ve müdahale şekli, konunun uzmanları tarafından ayrı ayrı tanımlanması ve modele kodlanması durumunda otomatik olarak da gerçekleştirilebilir. Aykırı değeri tespitinden sonra bir aykırı değerin işlenmesiyle ilgili literatürde geçen teknikler Tablo 2.1’de verilmiştir (Aguinis ve ark, 2013).

Tablo 2.1. Aykırı değeri literatüre dayalı işleme teknikleri (Aguinis ve ark, 2013).

	Aykırı Değeri İşleme Teknikleri	Açıklama
1	Doğru değeri ile değiştirilmesi	Bir veri noktasını uygun değeriine düzeltme.
2	Aykırı değeri kaldırılması	Veri noktasının analizden çıkarılması.
3	Aykırı değeri ayrıntılı olarak incelenmesi	Vakayı benzersiz bir ilgi fenomeni olarak incelemek için takip çalışması yapmak.
4	Saklanması	Bir aykırı değeri varlığını kabul etmek, ancak analizden önce aykırı değeri hiçbir şey yapmamak.
5	Aykırı değeri içeren, içermeyen bulguların raporlanması	Aykırı değeri olan ve olmayan önemli sonuçların ve sonuçlardaki herhangi bir farklılık varsa bunun da bir açıklamasının yapılarak raporlanması
6	Winsorization	Uç değeri belirlı bir veri yüzdesine dönüştürülmesi. Örneğın 5. yüzdelik dilimin altındaki tüm verileri 5. yüzdelik dilime ve 95. yüzdelik dilimin üzerindeki tüm veriler 95. yüzdelik dilime ayarlanır.
7	Kesme yapılması	Gözlenen değeri inandır bir aralık içinde ayarlanması ve veri kümesinden diğeri değeri çıkarılması.

Tablo 2.1. (Devamı) Aykırı değerlerin literatüre dayalı işleme teknikleri (Aguinis ve ark, 2013).

Aykırı Değer İşleme Teknikleri	Açıklama
8 Dönüşüm yapılması	Her bir değere deterministik bir matematiksel fonksiyon (ör. log fonksiyonu, ln fonksiyonu) uygulamak, sadece analizde uzaktaki veri noktasını ve veri noktaları arasındaki bağıl sıralamayı korumakla kalmaz, ama aynı zamanda yapıdaki veri noktalarının hata varyansını ve çarpıklığını da azaltır.
9 Değişiklik yapılması	Aykırı bir değeri başka, daha az aşırı bir değere manuel olarak değiştirme.
10 En küçük mutlak sapma	Sıradan en küçük kareler (OLS) yöntemine benzer şekilde, bir veri kümesine yakından yaklaşan bir fonksiyon üreterek artıkları sınırlayan regresyon katsayılarının değerlerini seçer.
11 En küçük kırılmış kareler	Her bir durum için kalan kareyi en yüksekten en düşüğe sıralar ve en yüksek değeri budar veya kaldırır.
12 M tahmini	Karesi alınmış artıkları, artıkların başka bir işleviyle değiştirilerek etkili aykırı değerlerin etkisini azaltan bir sağlam teknikler sınıfı. Özellikle bu yöntem, etkili TS inovasyon aykırı değerleri belirlendiğinde kullanılır.
13 Bayes istatistikleri	Önceki bilgileri ve eldeki gözlemlenen verileri tartarak parametre tahminlerini türetir. Bu tür ön bilgilerin kullanımı, "küçültmeye" yardımcı olur veya uzaktaki veri noktalarını verilerin merkezine veya geometrik olarak ağırlık merkezine yaklaştırır.
14 İki aşamalı sağlam prosedür	Her bir veri noktasına ağırlık atamak için Mahalanobis mesafesini kullanır, tahminlerde aşırı olan durumların ağırlığı azaltılır. İki aşamalı bir süreçle tamamlanır.
15 Yinelemeli olarak yeniden ağırlıklı en küçük kareleri kullanan doğrudan sağlam yöntem	İki aşamalı sağlam prosedürlere benzer, bu yöntem her bir veri noktasına ağırlık atamak için Mahalanobis mesafesini kullanır. Bununla birlikte, ağırlıkların atanması, yinelemeli olarak yeniden ağırlıklandırılmış bir en küçük kareler algoritması kullanılarak tamamlanır.
16 Genel-leştirilmiş tahmin denklemleri (GEE) yöntemleri	Bu yöntemler, çok seviyeli modelin rastgele kısımdaki varyansları ve kovaryansları doğrudan artıklardan tahmin eder ve bireysel ve grup farklılıklarını modellemekten ziyade ortalama nüfus etkilerini tahmin eder. GEE tahminleri maksimum olabilirlik tahminlerinden daha az verimli olsa da GEE tahminleri çok seviyeli modelin rastgele kısmının yapısı hakkında daha zayıf varsayımlarda bulunur ve böylece etkili aykırı değerlerin etkisini sınırlar.
17 Önyükeme yöntemleri	Bu yöntemler, bir modelin parametrelerini ve bunların standart hatalarını teorik bir örnekleme dağılımına başvurmadan örnekten tahmin eder. Sonuç olarak araştırmacı, deneysel bir örnekleme dağılımından alınan istatistiklerin beklenen değeri ve değişkenliği tahminlerini hesaplayabilir.

Tablo 2.1. (Devamı) Aykırı değerlerin literatüre dayalı işleme teknikleri (Aguinis ve ark, 2013).

Aykırı Değer İşleme Teknikleri		Açıklama
18	Parametrik olmayan yöntemler	Bu yöntemler, gözlemlerin belirli bir şekilde dağıtıldığını varsaymaz. Bu esneklik, aykırı değerlerin varlığında sağlam sonuçlar elde edilmesine yardımcı olur. Örneğin, verilerden parametre tahminleri elde etmek için, bir tür parametrik olmayan yöntem, gözlemlerin ham değerlerinin aksine, bazıları o kadar aşırı olabilir ki, karşılık gelen gözlemlerin etkili aykırı değerler olmasına neden olabilecek gözlemlerin sırasını kullanır. Sonuç olarak, sıraya dayalı parametrik olmayan bir analiz, aykırı değerlerin etkisini azaltır (yani, ağırlıkları düşürür).
19	Ağırlıksız meta-analiz	Daha büyük örneklem boyutlarına sahip birincil düzeydeki çalışmalara daha fazla ağırlık vermeyen meta-analitik istatistiklerin (örneğin, ortalama, standart sapma) elde edilmesi.
20	Genelleştirilmiş M-tahmin	Karesi alınmış artıkları, artıkların başka bir işleviyle değiştirilerek aykırı değerlerin etkisini azaltan bir yöntem. Özellikle, bir TS analizinde, bu yöntem, etkili TS toplam aykırı değerleri belirlendiğinde kullanılır.

Abro (2021) çalışmasında internet üzerinden temin edilebilen Aykırı Değer Algılama Verisetleri (ODDS) Glass, Letter Recognition Shuttle, Forest Cover, Vertebral, Wisconsin-Breast Cancer, Wine ve Vowels olmak üzere 8 tanesi kullanılmış, aykırı değer tespiti için SVM, KNN, ANN ve Rotasyon ormanı sınıflandırma algoritmalarını birleştiren yığınlama tabanlı bir topluluk öğrenme modeli geliştirilmiştir. Çalışma sonucunda modeller doğruluk, AUC, F-Skor, kesinlik ve duyarlılık performans metrikleriyle karşılaştırılmış, topluluk öğrenme yöntemlerinin ML sınıflandırıcılarının başarılarını arttırdığı ve yığınlama yönteminin diğer topluluk öğrenme yöntemlerine göre daha başarılı olduğu gösterilmiştir.

Topuz (2014) çalışmasında ML algoritmalarıyla sınıflandırmaya, en yakın k komşuluğa, kümelemeye, bilgi kazancına, istatistiksel değerlere dayalı anomali analizi yaklaşımları anlatılmıştır. MÖ, veri madenciliği, istatistik, veri analizi, yapay zekâ, bioinformatik ve matematik güçlü şekilde birbiriyle ilişkili alanlardır. İstenmeyen e-postaların tanınması (Spam detection), dolandırıcılığın ortaya çıkarılması (fraud detection), basamak tanıma (digit recognition), konuşma tanıma ve işleme (natural language processing), yüz tanıma (face detection), ürün tavsiye etme (product recommendation), medikal teşhis (medical diagnosis), stok ticareti (stock trading), müşteri bölümlendirme (customer segmentation), şekil tanıma (shape detection) ML

algoritmalarının meşhur uygulamalarıdır. Aslında MÖ'nin tanımı çözülen problemin yapısına ve ortamına bağlıdır.

Cousineau ve Chartier (2010) çalışmalarında olası aykırı değerlerin tespit edilmesini amaçlayan çeşitli teknikler gözden geçirilmiştir. Bu teknikler, tek değişkenli verilerle ilgili olanlar ve çok değişkenli verilerle ilgili olanlar olmak üzere iki sınıfa ayrılmış, popülasyon dağılımının normal olduğu, popülasyonun normal olmadığı ancak bilindiği veya popülasyonun bilinmediği durumlar, bir bağımlı değişken veya birden fazla bağımlı değişken olan durumlar ayrı ayrı değerlendirilmiş her durum için duruma uygun bir model geliştirilmesi gerektiği ve kesin tekniklerin bulunamadığı gösterilmiştir.

Chandola ve arkadaşları (2009) çalışmalarında sınıflandırma, kümeleme, en yakın komşu, istatistiksel, bilgi teorisi ve spektral bazlı tekniklerin her biriyle temel bir anormallik tespit modeli geliştirilmiş ve ardından o kategorideki mevcut farklı tekniklerin nasıl temel tekniğin varyantları olduğu, avantaj ve dezavantajları gösterilmiştir.

Hodge ve Austin (2004) çalışmalarında aykırı değer tespitinde doğru dağıtım modeli, doğru öznitelik türleri, ölçeklenebilirlik, hız, yeni örnekler için artımlı yetenek ve veri kümesine uygun bir algoritma seçimi gerekliliği vurgulanmış, aykırı değer saptama yöntemlerinin istatistik (yakınlık tabanlı, parametrik, parametrik olmayan ve yarı parametrik), sınır ağları (denetimli ve denetimsiz) ve ML olmak üzere üç hesaplama alanından türetildiği, yaklaşımların temelde mesafe tabanlı, küme tabanlı, yoğunluk tabanlı, derinlik tabanlı, model tabanlı ve grafik tabanlı algoritmalar olduğu gösterilmiştir.

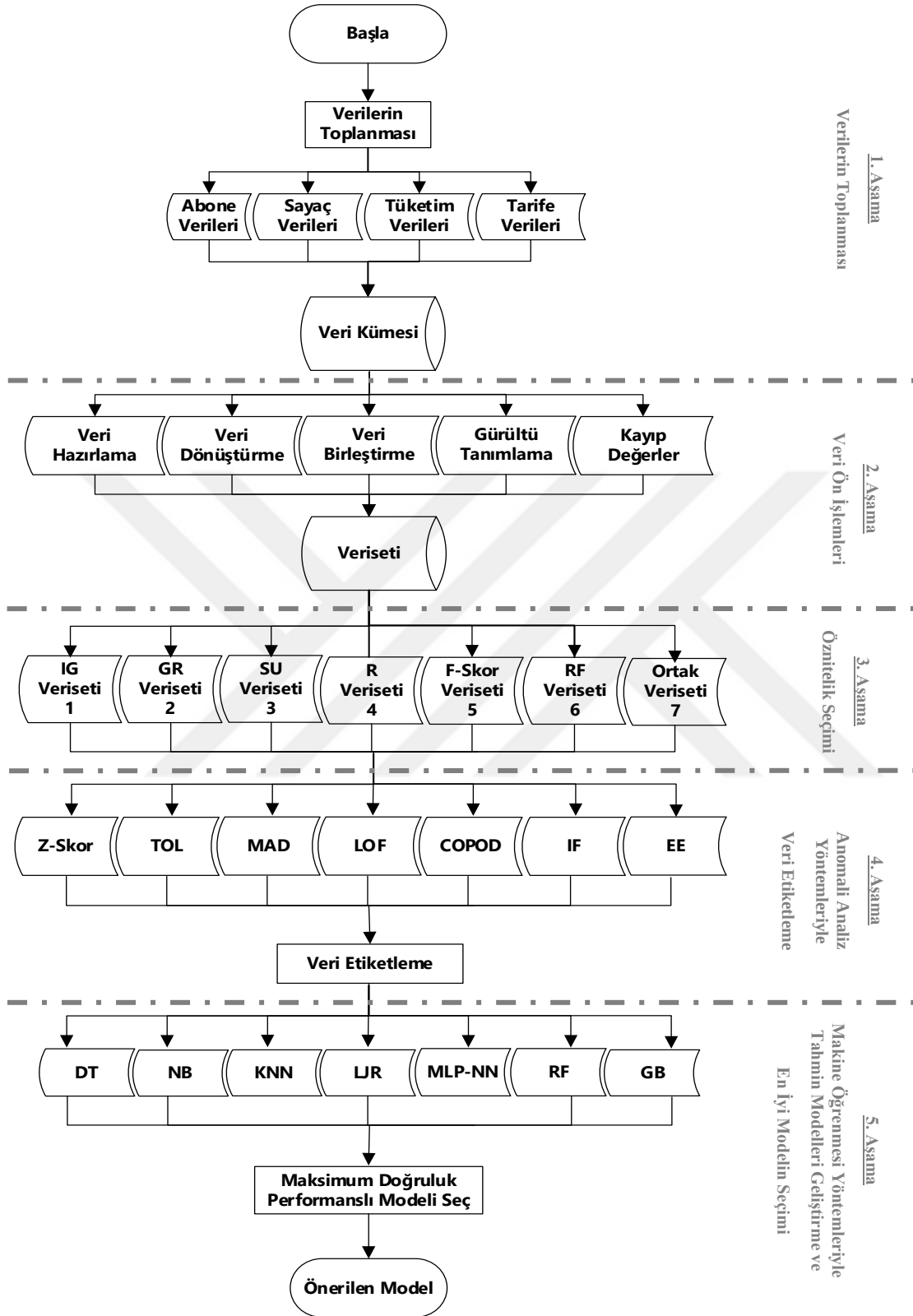
3. MATERYAL VE YÖNTEM

Çalışmada Kayseri Su ve Kanalizasyon İdaresi Genel Müdürlüğü (KASKİ) tarafından içme ve kullanma suyu hizmeti verilen, her ay düzenli olarak sayaç okuması yapılan ve geçmiş yıllara ait tüketimleri kayıt altına alınan konut abonelerine ait ABYS’de yer alan abone, sayaç, sayaç okuma, tahakkuk ve tahsilat verileri kullanılmıştır.

Çalışmada öncelikle toplanan ham veriler üzerinde veri ön işlemleri yapılarak analiz yapılacak veriseti ve alt verisetleri elde edilmiş, sonrasında anomali analiz algoritmaları ile aykırı tüketim davranışları tespit edilerek gözetimsiz veriseti dört sınıflı bir etiketleme yapılarak gözetimli hale getirilmiştir. Elde edilen yeni veriseti ile gözetimli ML algoritmaları kullanılarak TS tahmin modelleri geliştirilebileceği gösterilmiş, sahada sayaç okuma esnasında tüketimin hangi sınıfta olduğuyula ilgili aboneye makbuz üzerinden bilgi verilebileceği aynı zamanda kurum veritabanında abonelerin hangi sınıflarda tüketim yaptıklarının takibinin ve denetiminin yapılabileceği değerlendirilmiştir.

3.1. Kullanılan Yöntemin Şeması

Çalışmada önerilen modelin uygulama şeması Şekil 3.1’de gösterildiği gibi beş aşamada gerçekleştirilmiş olup ilk aşamada ihtiyaç duyulacağı öngörülen ham veriler temin edilerek bir veri kümesi elde edilmiş akabinde veri kümesi veri ön işlemlerinden geçirilerek analiz için uygun bir zaman serisi veriseti elde edilmiştir. Üçüncü aşamada öznitelik seçim yöntemleriyle alt verisetler türetilmiş ardından gözetimsiz ML ve istatistiki teknikler kullanılarak anormal tüketim tespitleri ve tüketim sınıf etiketlemesi yapılmıştır. Son aşamada ise etiketlenmiş gözetimli hale getirilmiş veriseti ile ML teknikleri kullanılarak tüketim sınıfı tahmin modelleri geliştirilmiş ve model performansları karşılaştırılarak en iyi model seçimi yapılmıştır.



Şekil 3.1. Çalışma kapsamında önerilen model şeması.

Çalışma ile veriseti anormal gözlem noktalarının etiketlenmesinin uzman operatör görüşü, abonenin kendisi veya model tarafından otomatik olarak yapılmasının mümkün olacağı esnek, kişiselleştirilebilir senaryolarla çalışan hibrit bir yapı oluşturulmuştur. Böylece sahada sayaç okuma esnasında abonelere hangi tüketim sınıfında tüketim yaptığı bilgisinin verilmesi ve sular idarelerine de potansiyel sorunlar barındırması muhtemel kapsamlı bir abone listesinin sunulması sağlanmıştır. Bu çalışma bir başlangıç kabul edilirse kurum tarafından yapılacak denetimler ve tespitler ile model daha da geliştirilebilecektir.

3.2. Çalışma Alanı ve Veriseti

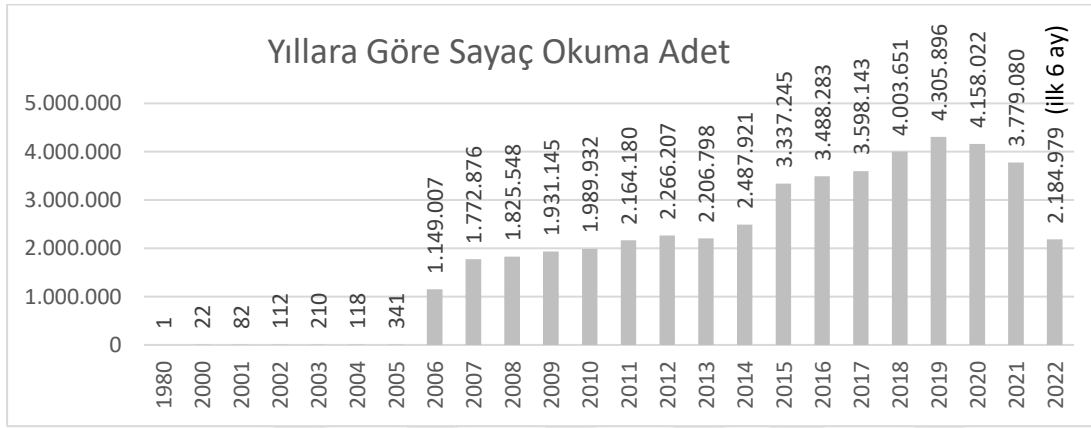
Kayseri ili Şekil 3.2’de gösterildiği gibi iç anadolu bölgesinin güney doğusunda 2022 yılı itibariyle 16 tanesi ilçe, bir tanesi de büyükşehir olmak üzere 17 belediye, 1.434,357 nüfus, 16.917,00 km² yüz ölçüme sahiptir.



Şekil 3.2. Uygulama alanı.

Kayseri 1988 yılından itibaren büyükşehir olmuş olup, KASKİ ise 1989’da büyükşehir bünyesinde kurulmuştur. KASKİ ile birlikte su ve kanalizasyon hizmetleri tek elden yönetilmeye başlanmış, öncesinde her belediye kendi mücavir alanında bu hizmetleri kendi imkanları ve yöntemleri ile verdiği ve yönettiği için Şekil 3.3’teki grafikten de

anlaşılacağı üzere bilgisayar ortamında geçmiş yıllara (1980-2005) ait kayıtların çok düşük miktarlarda olması veya hiç olmaması nedeniyle değerlendirilememiştir. Ayrıca bilgisayar ortamında sayaç okumaya yönelik sağlıklı denebilecek veriler 2006 ve sonrasında tutulmaya başlanmış, ilerleyen yıllarda artarak devam etmiş, 2014 sonrasında ise büyükşehir ve sular idareleri ile ilgili yapılan yasal düzenlemeler ile bütün bir ili kapsayan hizmet ağı ve beraberinde elde edilen veriler ile bütünleşik bir yapıya dönüşmüştür.



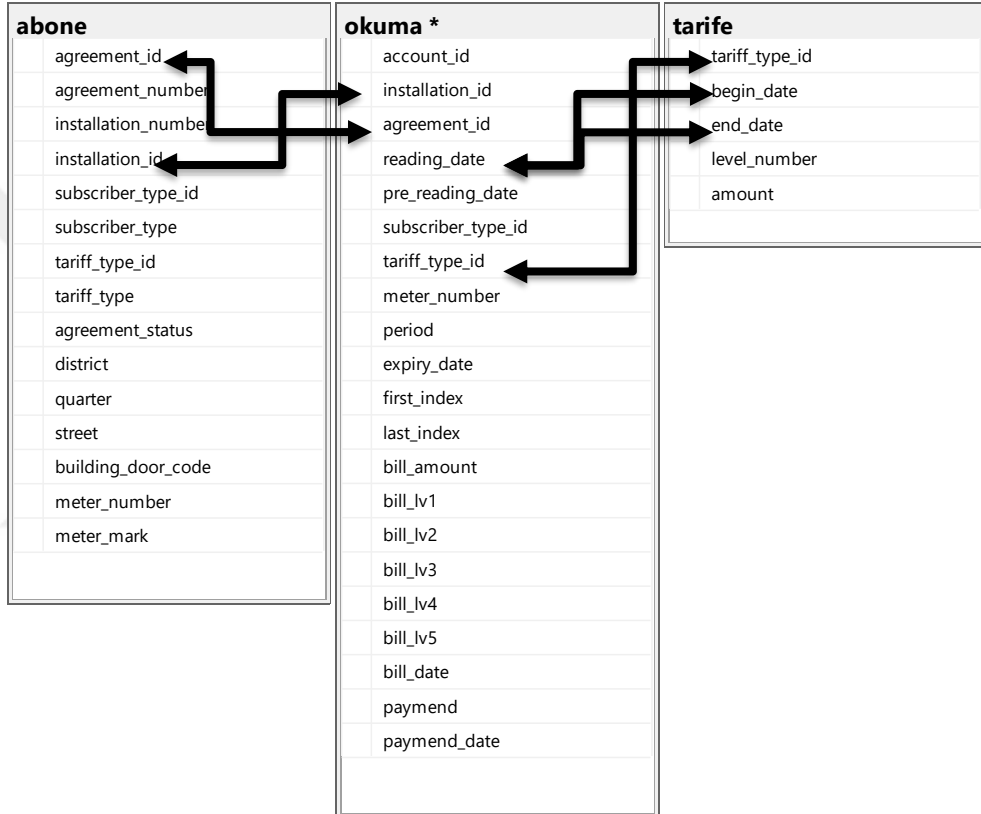
Şekil 3.3. KASKİ geçmiş yıllara ait mesken abonesi sayaç okuma sayıları.

Çalışma kapsamında veriseti olarak Kayseri il genelinde okuması yapılan toplam 668.823 aboneden yine okuması yapılan 610.821 adet konut abonesine ait 1980 – 2022 (ilk 6 ay dahil) tarihsel dönem aralığı dikkate alınmıştır. Çalışma kapsamında Tablo 3.1’de gösterildiği gibi çok yüksek sayıda abone analizinin yapılmasının çalışmaya fazladan bir katkı sunmayacağı ve çok fazla zaman ve işleme neden olacağı için okuma dönem adet sayısı 160 ve üzeri olan 8.224 adet konut abonesinin verileri incelenmiştir.

Tablo 3.1. KASKİ okuma dönem sayılarına göre abone sayısı tablosu.

Dönem Adet	Abone Sayısı	Mesken Abone Sayısı
120'den az	454.574	409.292
120 ve üzeri	214.249	201.529
140 ve üzeri	140.491	133.914
160 ve üzeri	8.286	8.224

KASKİ’den temin edilen abone verileri Şekil 3.4.’te gösterildiği gibi 3 farklı tabloda tutulmuş olup özetle adres numarataj, kullanıcı sözleşme, sayaç, sayaç okuma, tarife, tahakkuk ve tahsilat verilerinden oluşmaktadır. 21 öznitelikli okuma tablosu ile 15 öznitelikli abone tablosu abone numarası (installation_id) ve kullanıcı numarası (agreement_id) üzerinden, 5 öznitelikli tarife tablosu ile de tarife türü (tariff_type_id), tarife başlangıç tarihi (begin_date) ve bitiş tarihi (end_date) üzerinden ilişkilendirme yapılarak tablolar arası eşleştirme sağlanmıştır.



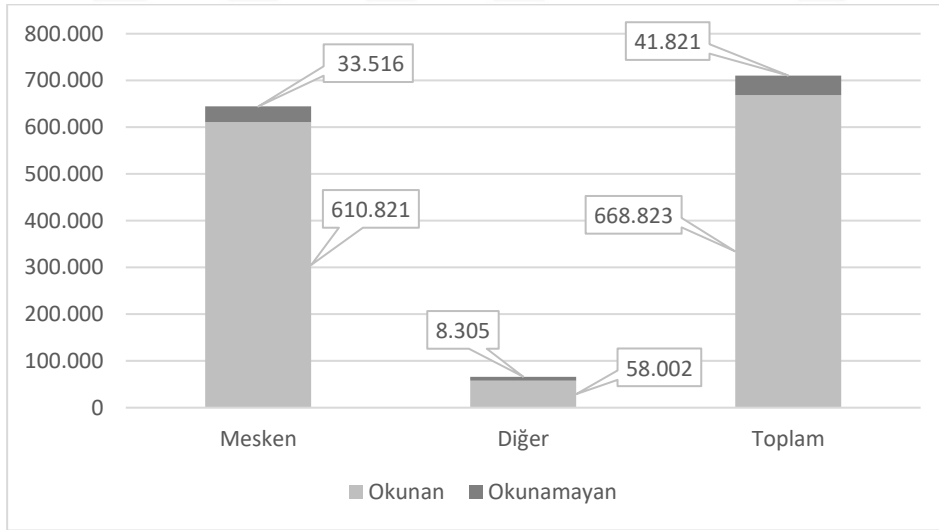
Şekil 3.4. Abonelere ait analizi yapılan verilerin tutulduğu ilişkisel veritabanı modeli.

Literatürdeki çalışmaların çoğu abonelerin tamamı üzerinden kümülatif talepleri dikkate almış, elde edilen modeller değerlendirilirken bütün abonelerin ortak davranışı varsayımıyla hareket edilmiş ve dolayısıyla il geneli hesaplanan hane başına düşen ortalama TÜFE ve nüfus artış/azalış oranları da dikkate alınmıştır. Bu çalışmada ise her bir aboneye özel aykırı tüketim tespit modeli oluşturulduğundan ekonomik göstergeler ve nüfusa ilişkin istatistiki verilerin her haneye özel, dikkate alınması gerekmektedir. Ancak aboneye özel istatistiki değerlerin temini ve sürekli güncel

tutulması birçok kurumda tutulan verilerin (maaş, harcama, hanede yaşayanların doğum, ölüm, evlenme, evcil hayvan, bitki vb.) online entegrasyonu ile veya su yönetimi idareleri tarafından ABYS’de haneye özel veri alanları tanımlanıp bilgilerin sahadan personel marifetiyle toplanması ile mümkün olacaktır. Söz konusu bu durumda abonenin rızası, sular idarelerinin ihtiyaç öncelikleri ve bazı yasal düzenlemeler dikkate alındığında mümkün olmadığı ancak uzun vadeli teknik ve teknolojik gelişmeler ile birlikte (e-Devlet vb.) elde edilebileceği öngörülmüş, haneye özel ekonomik göstergeler ve nüfusa ilişkin istatistiki veriler girdi değerlerine dahil edilmemiştir.

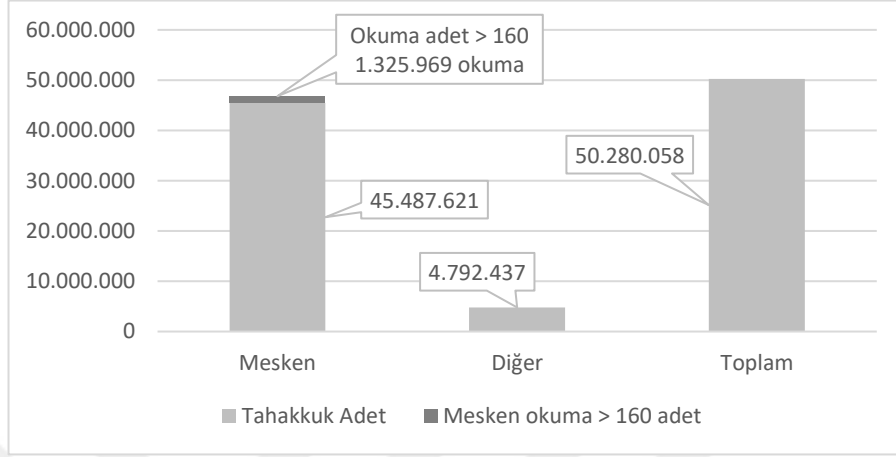
Ayrıca sayaç okuması yapılan bölgelerin iklimsel verileri okumacı veya sabit meteoroloji istasyonları marifetiyle tespit edilse dahi hane içi iklim şartlarının da haneden haneye farklılıklar taşıyacağı gerçeğinden hareketle iklim verileri girdi değeri olarak dahil edilmemiştir.

Abone ve okuma tabloları incelenmiş bazı abonelerin sayaç okumalarının yapılmadığı anlaşılmış, bu nedenle çalışmada Şekil 3.5’te gösterildiği gibi abone sayıları ve diğer hususlarda tahakkuk (sayaç okuma) görmüş aboneler dikkate alınmıştır.



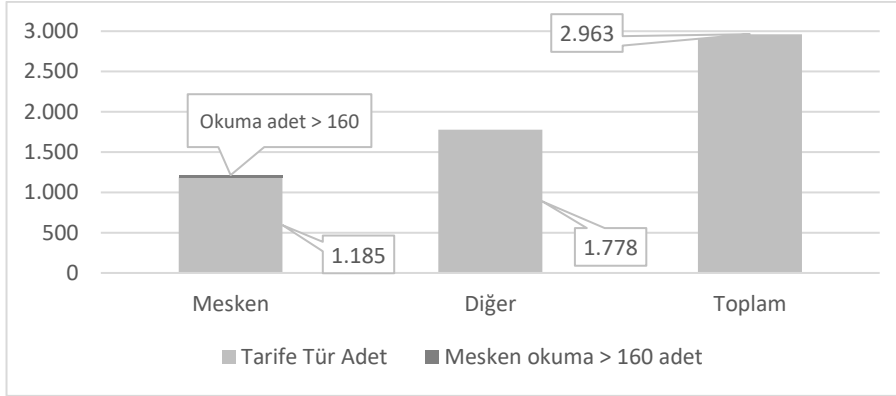
Şekil 3.5. Abone ve tahakkuk tablosuna göre sayacı en az bir defa okunmuş abone sayısı grafiği.

Analiz için temin edilen 2022 yılı 6. Ay sonu itibariyle toplam 50.280.058 adet okuma verisinden her biri en az 160 dönem sayaç okuma gören mesken abonelere ait Şekil 3.6’da da gösterildiği gibi 1.325.969 adeti analiz için değerlendirilmiştir.



Şekil 3.6. Mesken ve diğer abone türlerine göre tahakkuk sayıları grafiği.

Çalışma kapsamında temin edilen bir diğer tablo olan tarife tablosundan Şekil 3.7’de gösterildiği gibi örneklem kapsamında seçilen “Mesken” ve “Tutanaklı Mesken” olmak üzere 2 farklı mesken abone türü altında 29 farklı tarife türüne ait dönemsel tarife bilgileri değerlendirilmiştir.



Şekil 3.7. Tahakkuk gören abonelere ait tarife türü sayıları grafiği.

Çalışma ile birlikte her ne kadar hanelere özel veriler kullanılamamış olsa da çalışma sonunda elde edilen modele KASKİ veya aboneler tarafından tüketimi etkilediği düşünülen özel alanların da sonradan eklenebileceği bir modüler yapı sunulmuştur.

3.3. Veri Önışlemleri

3.3.1. Veri hazırlama

Kayseri ili içme suyu abonelerine ait Şekil 3.8’de temsili gösterilen adres numarataj, sözleşme yapılan kullanıcı, sayaç, sayaç okuma, tarife, tahakkuk ve tahsilat verileri çalışma kapsamında KASKİ’den temin edilmiştir.



Şekil 3.8. Çalışma kapsamında kullanılan veriler.

Çalışmada öncelikle temin edilen okuma, abone ve tarife tabloları her bir mesken abone numarası dikkate alınarak Tablo 3.2’de gösterildiği şekilde birleştirilmiştir. Okuma tablosu ana tablo olarak alınmış, akabinde abone tablosu abone numarası ve sözleşme numarası dikkate alınarak, tarife tablosu da tarife numarası ve tarife başlama ve bitiş tarihleri dikkate alınarak birleştirilmiştir. Elde edilen birleşim tablosunda abone numarası, abone tür numarası, tarife tür numarası, sayaç numarası bilgileri mükerrer olduğu için okuma tablosu esas alınarak diğer tablolardaki mükerrer öznitelikler çıkarılmış böylece 33 öznitelikli bir veriseti elde edilmiştir.

Tablo 3.2. Birleştirilmiş öznitelik bilgileri tablosu.

No	Öznitelik	Açıklama	Tablo
1	reading_date	Sayaç okuma tarihi	Okuma
2	account_id	Sicil numarası	Okuma
3	installation_id, installation_number	Abone numarası	Okuma, Abone
4	agreement_id, agreement_number	Sözleşme numarası	Okuma, Abone
5	pre_reading_date	Önceki okuma tarihi	Okuma
6	subscriber_type_id	Abone tür numarası	Okuma, Abone
7	tariff_type_id	Tarife tür numarası	Okuma, Abone, Tarife
8	meter_number	Sayaç numarası	Okuma, Tarife
9	period	Okuma dönemi	Okuma
10	expiry_date	Vade tarihi	Okuma
11	first_index	İlk index	Okuma
12	last_index	Son index	Okuma
13	fatura	Fatura tutarı	Okuma
14	bill_lv1	1. kademe tutar	Okuma
15	bill_lv2	2. kademe tutar	Okuma
16	bill_lv3	3. kademe tutar	Okuma
17	bill_lv4	4. kademe tutar	Okuma
18	bill_lv5	5. kademe tutar	Okuma
19	bill_date	Fatura tarihi	Okuma
20	odeme	Ödeme tutar	Okuma
21	odeme_date	Ödeme tarihi	Okuma
22	a_subscriber_type	Abone türü	Abone
23	a_tariff_type	Tarife türü	Abone
24	a_agreement_status	Abone durumu	Abone
25	a_district	İlçe	Abone
26	a_quarter	Mahalle	Abone
27	a_street	Cadde-Sokak	Abone
28	a_building_door_code	Dış kapı numarası	Abone
29	a_meter_mark	Sayaç markası	Abone
30	t_begin_date	Tarife başlama tarihi	Tarife
31	t_end_date	Tarife bitim tarihi	Tarife
32	t_level_number	Kademe numarası	Tarife
33	t_amount	Tarife birim fiyatı	Tarife

3.3.2. Veri dönüştürme

Sayaç okuma dönemlerinde sayaçların konut içinde olması, hane halkının evde bulunmaması, sayacın arızalı olması, sayaç camının buğulu olması, yeni okumacının adresi bulamaması veya okumacı ihmali gibi nedenlerle endeks okumasının yapılamadığı bilinmektedir. Bu durum dikkate alınarak biriken endekslerin bir sonraki okuma döneminde kayıt altına alındığı, dolayısıyla bazı dönemsel tüketimlerin olduğundan fazla görülebileceği, bu durumun analiz sonuçlarını yanıltabileceği öngörülmüş ve öncelikle denklem 3.1’de gösterildiği gibi son endeks ile ilk endeks öznitelik farkından tüketim miktarı özniteliği oluşturulmuştur.

$$su_toplamlam_m3 = last_index - first_index \quad (3.1)$$

Akabinde sayaç okuma tarihi ile önceki okuma tarihi farkından da denklem 3.2’de gösterildiği gibi tüketim süresi hesaplanmış yeni bir öznitelik olarak eklenmiştir.

$$gecen_gun = reading_date - pre_reading_date \quad (3.2)$$

Hesaplanan tüketim gün sayısının yılın hangi ayından kaç günü kapsadığı bilgisine ulaşmak için yılın her ayı için Tablo 3.3’te gösterildiği gibi 12 yeni öznitelik oluşturulmuş, hangi ayda kaç günlük bir tüketim olduğu hesaplanarak ilgili aya işlenmiştir.

Tablo 3.3. Tüketim yapılan günlerin ilgili aylara dağılımı tablosu.

Okuma Tarihi	Önceki Okuma Tarihi	Gecen Gün	1	2	3	4	5	6	7	8	9	10	11	12
22.02.2006	17.12.2005	67	30	23	0	0	0	0	0	0	0	0	0	14
18.04.2006	22.02.2006	55	0	6	31	18	0	0	0	0	0	0	0	0
20.06.2006	18.04.2006	63	0	0	0	12	31	20	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
29.04.2022	31.03.2022	29	0	0	0	29	0	0	0	0	0	0	0	0
12.05.2022	29.04.2022	13	0	0	0	1	12	0	0	0	0	0	0	0
6.06.2022	12.05.2022	25	0	0	0	0	19	6	0	0	0	0	0	0

Ayrıca her bir sayacın ölçüm hassasiyeti farklı olabileceği gibi sayaç ömrünün 10 yıl olması ve sayaçların zamanla ölçüm hassasiyetlerini kaybetmesi durumları dikkate alındığında sayaç değişimlerinin tüketim değerlerini etkileyeceği düşünülmüştür. Bu nedenle sayaç ilk endeks bilgisinden hareketle, yani endeks 0 olduğunda yeni bir sayaç takıldığı anlamı çıkarılmış dolayısıyla abone sayacının yenilendiği tespiti yapılmıştır. Bu yöntemle her bir abone için bütün okuma dönemleri incelenerek tüketimlerinin kaçınıcı sayaç ile yapıldığı bilgisini gösteren yeni bir öznitelik (“sayac_durum”) Tablo 3.4’te gösterildiği gibi geliştirilmiş anormal tüketim tespiti durumunda da dikkate alınacak bir unsur olarak değerlendirilmiştir.

Tablo 3.4. Verisetindeki endeks hareketleri ile sayaç bilgisi dönüşüm tablosu.

Okuma Tarihi	İlk İndeks	Son İndeks	Sayac_durum	S1	S2	S3
22.02.2006	262	268	S1	1	0	0
18.04.2006	268	284	S1	1	0	0
20.06.2006	284	307	S1	1	0	0
---	---	---	---	---	---	---
14.11.2007	0	3	S2	0	1	0
28.12.2007	3	13	S2	0	1	0
6.02.2008	13	20	S2	0	1	0
---	---	---	---	---	---	---
29.04.2022	0	10	S3	0	0	1
12.05.2022	10	17	S3	0	0	1
6.06.2022	17	29	S3	0	0	1

Sözleşme numarası bilgisinden hareketle ilgili okuma dönemi sözleşme numarası ile önceki okuma dönemi sözleşme numarası karşılaştırılmış, başka bir sözleşme numarası söz konusu ise kullanıcı değişimi olduğu anlamı çıkarılmıştır. İçmesuyu tüketiminin başka bir kullanıcı tarafından yapıldığı bilgisi yeni bir öznitelik (“kullanici_durum”) oluşturularak kaçınıcı kullanıcıya ait bir tüketim olduğu bilgisi elde edilmiştir. Ayrıca her kullanıcı tüketim alışkanlığının ve hane içi nüfusunun farklı olabileceğinden hareketle her bir abone için kullanıcı durum bilgisi de Tablo 3.5’te gösterildiği gibi ayrı ayrı özniteliklere dönüştürülmüştür. Böylece kullanıcıların tüketime etkilerinin analizi sağlanmış, anormal tüketim tespitinde dikkate alınacak bir unsur olarak değerlendirilmiştir.

Tablo 3.5. Verisetindeki sözleşme numarası ile kullanıcı bilgisi dönüşüm tablosu.

reading_date	agreement_id	kullanici_durum	K1	K2	K3
22.02.2006	12001815	K1	1	0	0
18.04.2006	12001815	K1	1	0	0
20.06.2006	12001815	K1	1	0	0
---	---	---	---	---	---
17.05.2018	357910	K2	0	1	0
12.06.2018	357910	K2	0	1	0
10.07.2018	357910	K2	0	1	0
---	---	---	---	---	---
13.11.2019	402841	K3	0	0	1
11.12.2019	402841	K3	0	0	1
7.01.2020	402841	K3	0	0	1
---	---	---	---	---	---
29.04.2022	402841	K3	0	0	1
12.05.2022	402841	K3	0	0	1
6.06.2022	402841	K3	0	0	1

Verisetinde her bir okuma dönemine ait 5 kademeli bir faturalandırma bilgisi (bill_lv1, bill_lv2, bill_lv3, bill_lv4, bill_lv5) mevcut olduğundan, bu bilgiler kullanılarak kademe durumunu gösteren bir öznitelik (“kademe”) Tablo 3.6’da gösterildiği gibi geliştirilmiş, kaçınıcı kademeye kadar bir tüketim yapıldığı bilgisi bu alana girilmiştir. Fatura toplam tutar bilgisi öznitelik olarak mevcut olduğundan kademe tutar öznitelikleri verisetinden çıkarılmıştır.

Tablo 3.6. Verisetindeki kademe tutar bilgi alanları ve dönüşüm tablosu.

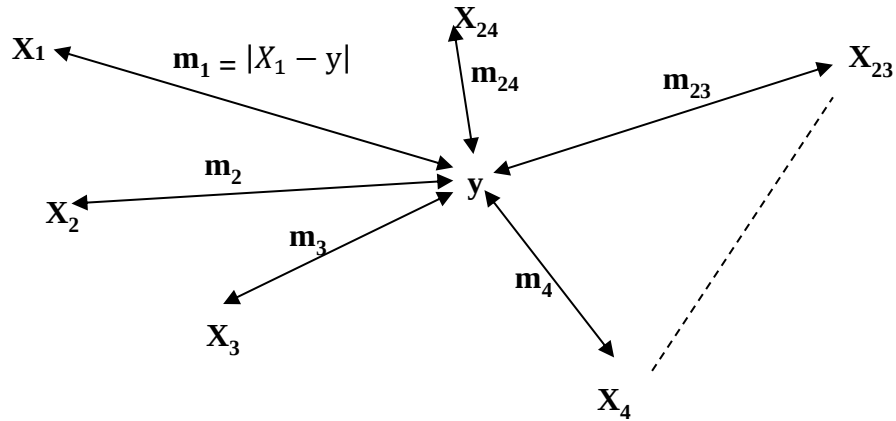
Okuma Tarihi	bill_lv1	bill_lv2	bill_lv3	bill_lv4	bill_lv5	kademe
22.02.2006	6,67	0,00	0,00	0,00	0,00	1
18.04.2006	17,78	0,00	0,00	0,00	0,00	1
20.06.2006	25,56	0,00	0,00	0,00	0,00	1
---	---	---	---	---	---	---
31.03.2022	13,82	16,38	0,00	0,00	0,00	2
29.04.2022	65,52	0,00	0,00	0,00	0,00	1
12.05.2022	39,04	0,00	0,00	0,00	0,00	1
6.06.2022	72,36	0,00	0,00	0,00	0,00	1

Ayrıca verisetini oluşturan özniteliklerin birden fazla olması, değerlerinin gün, ay, yıl, ₺, m³ gibi farklı birimlerden oluşması ve grafik olarak gerçek değerlerinde göstermenin zor olması nedeniyle, grafik gösterim ve performans hesaplamalarının daha sağlıklı yapılması için her bir öznitelige ayrı ayrı olmak üzere bütün gözlem noktaları dikkate alınarak denklem 3.3'te gösterildiği gibi normalize işlemi yapılmış, yine her bir özniteliğin gözlem noktalarına ait değerler 0 ile 1 aralığına indirgenmiştir.

$$X_n = \frac{(X_g - X_{min})}{(X_{max} - X_{min})} \quad (3.3)$$

Burada X_n normalize edilmiş gözlem değeri, X_g gözlemin gerçek değeri ve X_{max} , X_{min} 'de sırasıyla ilgili gözlemin en büyük ve en küçük değerleridir.

Ayrıca çalışmada gerçek değerler ile analiz yapılmasının yanında mesafe ölçüm yöntemi kullanılarak üretilen değerler ile de analiz yapılmıştır. Mesafe ölçüm yöntemlerinden girdi öznitelik sayıları ikiden fazla olduğu için tek boyutlu mesafe ölçüm yöntemi Manhattan, iki boyutlu mesafe ölçüm yöntemi Öklid kullanılmamış, Şekil 3.9'da da gösterildiği gibi minkowski yöntemi kullanılarak gözlem noktaları mesafe değerleri hesaplanmıştır.



Şekil 3.9. Bağımsız değişkenlerin bağımlı değişkene mesafelerinin minkowski ile ölçüm gösterimi.

Bağımsız değişkenler olan X_i girdi değerlerinin, bağımlı değişken olan y değerine mesafeleri m_i olarak alınmış olup denklem 3.4'te gösterildiği şekilde her bir gözlem

noktasına yani sayaç okuma noktası özniteliklerine ait değerler tek bir mesafe değerine dönüştürülmüştür.

$$Minkowski = d(X, y) = \left[\sum_{i=1}^n (|X_i - y|^p) \right]^{\frac{1}{p}} \quad (3.4)$$

Burada p girdi öznitelik yani boyut sayısı, y çıktı değeri ve X_i 'de her bir girdi öznitelik değeridir. Ayrıca negatif korelasyona sahip özniteliklerin X değerleri -X olarak alınmış, her bir tüketim gözlem noktası mesafesi toplam mesafe olarak hesaplanmıştır.

3.3.3. Veri birleştirme

Access veritabanı kullanılarak csv formatında temin edilen veriler tek bir tabloda birleştirilmiştir. Ayrıca abonelere, sayaçlara ve tarifelere ait değerler tablolar arası ilişkisel öznitelikler dikkate alınarak eşleştirilmiş bütün özniteliklerin yer aldığı TS bir veriseti elde edilmiştir.

3.3.4. Gürültü tanımlama

Çalışma, hanelerin tüketimlerine odaklandığı için öncelikle içme suyu abonelerinden sadece mesken abonelerine ait veriler seçilmiştir. Ayrıca her dönem sayaç okuması yapılamadığı abone sayaç okuma dönem adetlerinin farklılıklar gösterdiği tespit edilmiş, mesken abonelerinden de en az 160 okuma dönemi okuma yapılan aboneler dikkate alınmamıştır.

Verisetinde her bir aboneye ait özniteliklerden birinin veya birkaçının bütün gözlem noktalarındaki değerlerinin aynı olması durumunda ilgili özniteliğin analizi etkilemeyeceği ve analiz öncesinde bu özniteliklerin kontrol edilip ayıklanmasının işlemleri hızlandıracağı değerlendirilmiş, ancak bu tercih öznitelik seçim yöntemlerine bırakılmıştır.

3.3.5. Kayıp değerler

Abone sayaç değişimlerinin tespiti için okuma dönemlerinde sayaç numara bilgilerinin değişimleri öncelikle dikkate alınmıştır. Geçmiş sistem kayıtlarında sayaç değişimlerine ait iş emirlerinin ve sayaç numara bilgilerinin sağlıklı olmaması

nedeniyle, sayaç deęişim bilgileri, okuma endeks hareketlerinde endeks deęerlerinin tekrar sıfır olması bilgisi üzerinden tespit edilerek elde edilmiştir.

İlgili okuma dönemi birim fiyat bilgisi tarife tablosundan okuma tarihi dikkate alınarak geçerli tarifeye ait tutarın tespit edilmesiyle elde edilmektedir. Ancak geçmiş dönemlere ait tarife bilgisine ulaşamadığı durumlar söz konusu olmuş, eksik veriler atama (imputation) yöntemi kullanılarak ilgili tarihten bir önceki dönemde başlayan birim fiyat bilgisi veya toplam dönem fatura tutarı toplam dönem tüketim miktarına bölünerek elde edilen tutar bilgisi dikkate alınarak eksik veriler tamamlanmıştır.

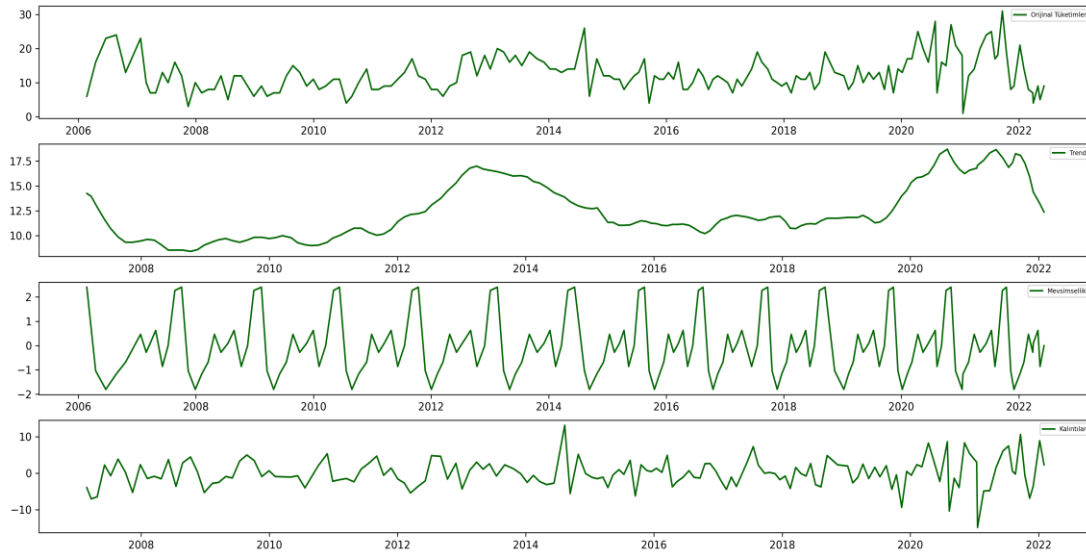
Ayrıca birleştirilmiş veriseti içinde her bir öznitelik altında boş olan deęerlerin olup olmadığıyla ilgili olarak Tablo 3.7’de gösterildiğı gibi bir kontrol yapılmış, %10’dan fazla eksik verisi olan özniteliklere ait eksik veriler tamamlanmaya çalışılmış, tamamlanması mümkün olmadığı durumlarda ise analizden çıkarılmıştır.

Tablo 3.7. Birleştirilmiş verisetinde boş hücrelerin kontrol tablosu.

Öznitelikler	Kayıt Sayısı	Dolu Deęerler	Eksik Deęerler	Eksik %
installation_id	167	167	0	0,00%
agreement_id	167	167	0	0,00%
pre_reading_date	167	167	0	0,00%
subscriber_type_id	167	167	0	0,00%
tariff_type_id	167	167	0	0,00%
meter_number	167	5	162	97,01%
---	---	---	---	---
a_tariff_type	167	167	0	0,00%
a_quarter	167	167	0	0,00%
a_street	167	167	0	0,00%
a_meter_number	167	167	0	0,00%
t_begin_date	167	167	0	0,00%
t_end_date	167	167	0	0,00%
t_level_number	167	167	0	0,00%
t_amount	167	167	0	0,00%

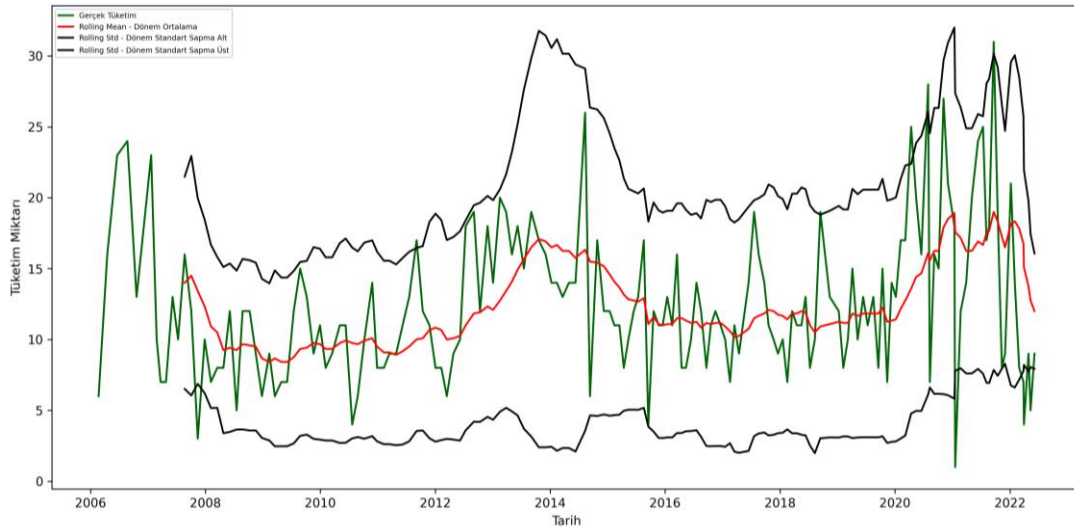
Ayrıca her bir aboneye ait veriseti, Artırılmış Dickey-Fuller (ADF) testi ile analiz edilmiş, p deęeri %99 olarak tespit edilmiştir. Anlamlılık seviyesi %5'ten daha büyük olduğu için içme suyu tüketim davranışlarının mevsimsellik karakteri taşıdığı, bazı

yıllar artan bazı yıllar azalan bir trendin olduğu, sıfır hipotezinin reddedilemeyeceği ve TS'nin durağan olmadığı Şekil 3.10'da gösterildiği gibi tespit edilmiştir.



Şekil 3.10. Verisetinin Arttırılmış Dickey-Fuller yöntemi ile analiz grafiği.

İçmesuyu tüketim davranışları mevsimsellik karakteri taşıdığı için ilgili dönem tüketim alt ve üst sınırları tespit edilirken ortalama değer (mean value) yöntemi yerine Şekil 3.11'de gösterilen dönemsel ortalama (rolling mean) değerler dikkate alınmıştır.

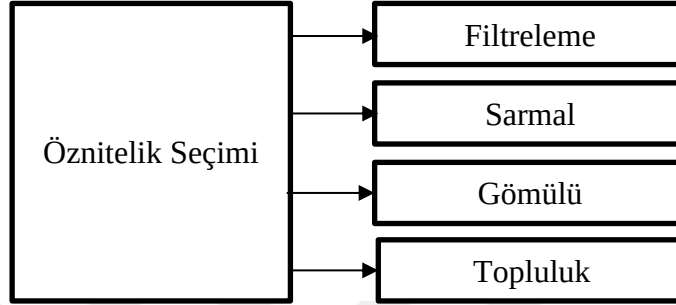


Şekil 3.11. Veriseti dönem ortalaması ve dönem standart sapma grafiği.

Önerilen yöntem okuma yapılan dönem ve tüketimin yapıldığı aylara ait gün sayılarından hareketle önceki yıllara ait aynı ayların oranlarına göre hesaplanması sonucu elde edilen en iyi örnek metod (simplest method) yöntemi kullanılmıştır.

3.3.6. Öznitelik seçimi

Öznitelik seçimi Şekil 3.12’de gösterildiği gibi filtreleme, sarmal, gömülü ve topluluk öznitelik seçim yöntemleri olmak üzere 4 başlıkta incelenebilir (Cai ve ark, 2018).



Şekil 3.12. Öznitelik seçme yöntemleri şeması.

Filtreleme öznitelik seçim yaklaşımı ile veriseti istatistiksel olarak analiz edilir. Temel olarak verisetindeki bütün bilgiyi içeren benzersiz özniteliklerin öncelikle puanlanmasına sonrasında sıralanmasına dayalı olarak değerlendirilmesini ve seçilmesini sağlayan bir yöntemdir (Cai ve ark, 2018; Zhang R. ve ark, 2019).

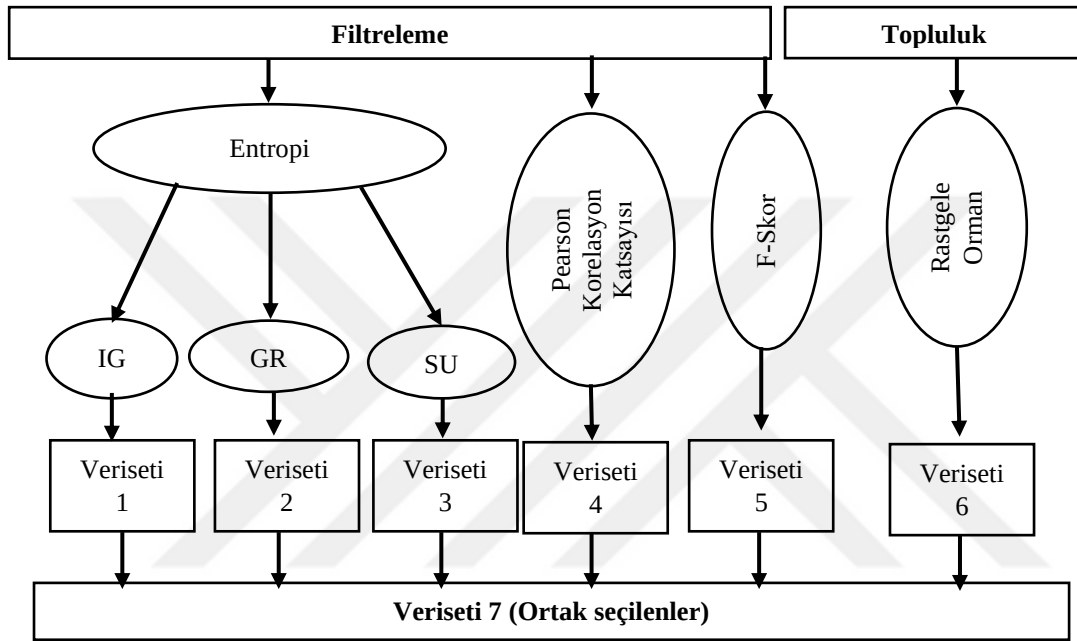
Sarmal öznitelik seçim yaklaşımında, veriseti denetimli ise sınıflandırıcı algoritmalar kullanılarak, denimsiz ise kümeleme algoritmaları ile öznitelikler üzerinde arama işlemi yapılarak alt kümeler oluşturulmakta ve belli sınıflar sadece değerlendirmeye alınmaktadır. Çok büyük ölçekli verisetlerinde bu yaklaşımın kullanılması durumunda her bir alt kümenin değerlendirilmesi için yüksek oranlarda hesaplama işlemleri gerektirmekte olduğu için zaman ve kapasite açısından bir değerlendirme yapılarak tercih edilmesi gerekmektedir (Cai ve ark, 2018; Zhang R. ve ark, 2019).

Gömülü öznitelik seçim yönteminde model geliştirme sürecinde harcanan zamanın azaltılması amaçlanmakta olup hem sınıflandırma hem de seçim işlemini eşzamanlı olarak model eğitimi aşamasında gerçekleştiren DT, SVM gibi algoritmalar kullanılmaktadır (Cai ve ark, 2018; Zhang R. ve ark, 2019).

Topluluk öznitelik seçim yönteminde farklı öznitelik alt kümeleri elde edilerek oluşturulan verisetleri tarafından eğitilen temel sınıflandırma algoritmaları

kullanılarak yapılmış öznitelik seçim topluluklarının rastgele alt uzay yöntemi (RSM) ve RF yöntemi gibi çoğunluk oylamasıyla daha doğru sınıflandırma ve seçim yapılmasına dayanmaktadır (Cai ve ark, 2018).

Bu çalışmada öznitelik seçimi yöntemlerinden Şekil 3.13'te gösterildiği gibi IG, GR, SU, R, F-Skor ve RF sınıflandırma yöntemleri kullanılmıştır. Bu yöntemlerden IG, GR ve SU yöntemleri entropi kullanılarak hesaplanmaktadır.



Şekil 3.13. Öznitelik seçimi yöntemleri kullanılarak oluşturulan verisetleri şeması.

3.3.6.1. Entropi

Entropi (S), bir sistemin belirsizliğinin veya düzensizliğinin bir ölçüsüdür. Gözlem değerleri ne kadar düzenli ise S o kadar düşük, ne kadar düzensiz ise S o kadar yüksek olacaktır. Başka bir ifadeyle gözlem değerinin olasılığı ne kadar yüksekse S o kadar düşük veya tam tersi olacaktır. Ayrıca düzensizlik arttıkça bilgi miktarı da artacağından S aynı zamanda bilgi miktarı olarak da yorumlanabilir. Bir Y öznitelğine ait bütün gözlem değerlerinin entropisi denklem 3.5'te gösterildiği gibi hesaplanır. Burada $p(y)$, veri kümesindeki y veri sınıfının tüm sınıf içerisindeki bulunma olasılığını ifade eder.

$$\text{Entropi} = S = H(Y) = - \sum_{y \in \gamma} p(y) \log_2 p(y) \quad (3.5)$$

Bir Y özniteliğini tanımlarken bir X özniteliği de kullanılıyorsa Y ile X arasında denklem 3.6'daki gibi bir ilişki söz konusu olacaktır.

$$H(Y/X) = - \sum_{x \in X} p(x) \sum_{y \in \gamma} p(y/x) \log_2 p(y/x) \quad (3.6)$$

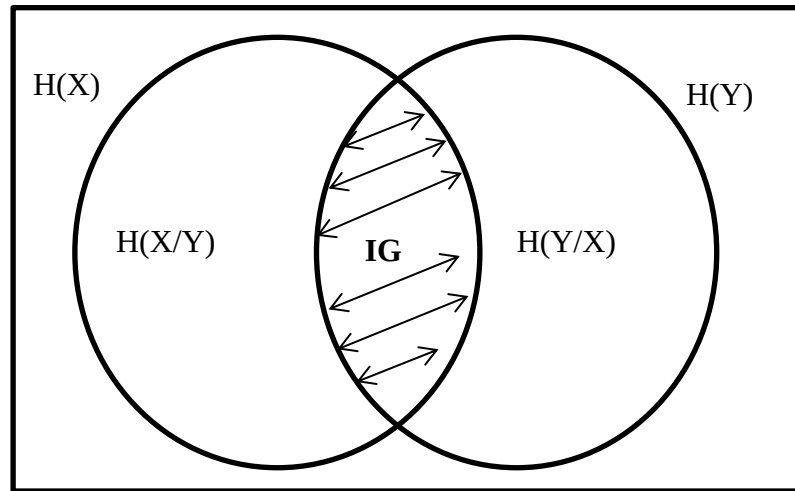
Burada $H(Y/X)$ değeri X biliniyorsa Y'deki belirsizliğin azalma miktarı veya $H(X)$ X'in $H(Y)$ y'nin bilgi miktarı ise $H(Y/X)$ sadece Y'ye ait bilgi miktarı olarak da yorumlanabilir.

3.3.6.2. Bilgi kazancı

X özelliğine bağlı olarak Y özelliğinin entropi değerindeki azalmayı gösterir. Bilgi kazancı (IG) denklem 3.7'de gösterildiği şekilde hesaplanır.

$$\text{Bilgi Kazancı} = \text{IG} = H(Y) - H(Y/X) \quad (3.7)$$

Ayrıca X'in ve Y'nin ortak bilgi miktarı olarak da Şekil 3.14'te gösterildiği gibi ifade edilebilir (Hall, 1999).



Şekil 3.14. Entropi ve bilgi kazancı arasındaki ilişki (Kadılar, 2019).

3.3.6.3. Kazanç oranı

Kazanç oranı (GR), IG yönteminin açığı olan çeşitliliğe karşı X öznitelikliğin entropisi vasıtasıyla normalize edilmesiyle elde edilir. GR değeri denklem 3.8’de gösterildiği şekilde hesaplanır (Hall, 1999; Budak, 2018; Novakovic, 2009).

$$\text{Kazanç Oranı} = \text{GR} = \frac{\text{Bilgi Kazancı}}{H(X)} \quad (3.8)$$

3.3.6.4. Simetrik belirsizlik katsayısı

GR’deki gibi IG yönteminin zayıf yanını X ve Y özniteliklerinin entropilerinin toplamı vasıtasıyla normalize edilmesi sonucu elde edilir. Simetrik belirsizlik katsayısı (SU) değeri denklem 3.9’da gösterildiği şekilde hesaplanır (Hall, 1999; Budak, 2018; Novakovic, 2009).

$$\text{Simetrik Belirsizlik Katsayısı} = \text{SU} = 2 * \frac{\text{Bilgi Kazancı}}{H(Y) + H(X)} \quad (3.9)$$

3.3.6.5. Pearson korelasyon katsayısı

Pearson korelasyon katsayısı yöntemi, iki sayısal ve sürekli olan değişkenler arasındaki doğrusal bağımlılık ilişkisini istatistiksel olarak ölçmek amacı ile kullanılan en yaygın yöntemlerden birisidir. Kovaryans yöntemine dayanmaktadır. Bu teknik ile veri kümesindeki her bir öznitelik için denklem 3.10’da gösterildiği gibi öznitelik gözlem değerlerinin ortalaması bulunur, elde edilen ortalama ($\bar{x}$) değeri kullanılarak denklem 3.11’de gösterildiği gibi standart sapma (σ) değeri bulunur ve denklem 3.12’de gösterildiği gibi korelasyon katsayısı bulunursa, öznitelikler arasında bulunan doğrusal bağımlılık ilişkisi tespit edilmiş olur.

$$\text{Ortalama} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.10)$$

$$\text{Standart Sapma} = \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.11)$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y} \quad (3.12)$$

Burada r korelasyon katsayısı, x ve y , n uzunluğunda iki vektör, $\bar{x}$ x 'in ve $\bar{y}$ y 'nin ortalaması, σ_x x 'in σ_y y 'nin standart sapmasıdır. Ayrıca r , -1 ile +1 arasında bir değer almaktadır. denklem 3.13'te gösterildiği gibi r 'nin değeri 0 ise iki değişken arasında bir ilişki olmadığını, negatif ise birbirleri arasında zıt yönde, pozitif ise birbirleri arasında aynı yönde bir ilişki olduğunu göstermektedir (Özmen ve ark, 2012).

$$f(r) = \begin{cases} \text{negatif korelasyon, zıt ilişki, } r < 0 \\ \text{pozitif korelasyon, ilişkili, } r > 0 \\ \text{ilişki yok, } r = 0 \end{cases} \quad (3.13)$$

3.3.6.6. F-skor

Fisher-skor filtresi denklem 3.14'te gösterildiği şekilde her bir özneliğin maksimum ayırt ediciliğine dair bir skor hesaplar. Hesaplanan skoru fazla olan özneliklerin daha önemli oldukları kabul edilir ve uzman kararıyla belirlenen sayıda öznelik seçimi yapılır.

$$F_i = \frac{(\bar{x}_i - \bar{y})^2}{\sigma_{x_i}^2 + \sigma_y^2} \quad (3.14)$$

Burada F_i i 'ninci niteliğin f-skor değerini, $\bar{x}_i$ i 'ninci özneliğin ve $\bar{y}$ y 'nin ortalamasını, $\sigma_{x_i}^2$ i 'ninci özneliğin ve σ_y^2 y 'nin varyans değerlerini ifade eder (Morán-Fernández ve ark, 2016).

3.3.6.7. RF sınıflandırma

RF, rastgele seçilen DT'ye dayanan topluluk öğrenicisidir, topluluktaki her ağaç orijinal eğitim verisinin değiştirilmesiyle orijinal verinin rastgele bir örneğinden

oluşturulmaktadır. RF algoritması, oluşturulan her DT'nın bir sınıf için oy kullandığı ve ormandaki tüm ağaçlar arasından en çok oyu alan sınıflandırmayı seçmeye dayalı bir yöntemdir. Yöntemin en önemli avantajlarından biri, her özneliğin etkisini öğrenmeyi sağlayan bir önem puanını ölçebilmesidir. Ancak özneliklerin sayısının çok fazla olması durumunda RF kullanımı öznelik önem puanlarının hesaplanmasını ve sınıflandırma için en uygun özneliklerin seçimini zorlaştırmaktadır. DT'yi oluşturmak ve her ağaçtaki son sınıfı belirlemek için denklem 3.15'te gösterildiği şekilde hesaplanan bir gini indeksi kullanılmaktadır (Uddin ve Uddin, 2015).

$$Gini(v) = \sum_{i=1}^l f_i (1 - f_i) \quad (3.15)$$

Burada f_i , v düğümündeki i sınıfı kayıtlarının kesridir. Yine v ağaç düğümünü bölmek için X_i özneliğinin Gini bilgi kazancı denklem 3.16'daki şekilde tanımlanmıştır.

$$gain(X_i, v) = Gini(X_i, v) - (W_L(Gini(X_i, v^L)) + W_R(Gini(X_i, v^R))) \quad (3.16)$$

Burada $Gini(X_i, v)$, v düğümündeki kirliliktir. W_L ve W_R sırasıyla v düğümünün sol ve sağ alt düğümleridir. W_L ve W_R sol ve sağ alt düğüme atanan örneklerin kesridir. Son olarak, kirlilikteki azalmayı maksimize eden öznelik, seçilen öznelik olarak değerlendirilmektedir. X_i özneliği için önem puanı denklem 3.17'deki gibi $gain(X_i, v)$ 'den hesaplanmaktadır.

$$Imp_i = \frac{1}{n_{tree}} \sum_{k \in S_{X_i}} gain(X_i, v) \quad (3.17)$$

Burada n_{tree} , RF'daki veya topluluk boyutundaki ağaçların sayısıdır ve $k \in S_{X_i}$, bölünmüş düğümler kümesidir. Bu çalışmada $n_{tree} = 50$ alınmıştır. Bütün özneliklerin önem puanlarının normalleştirilmesi denklem 3.18'deki şekilde tanımlanır:

$$Imp_{max} = \frac{Imp_i}{Imp_{max}} \quad (3.18)$$

Imp_i RF'lerden X_i 'nin önem puanını temsil eder ve Imp_{max} maksimum önemi temsil eder ve normalleştirilmiş önem puanı $0 \leq Imp_{max} \leq 1$ arasında yer alır (Uddin ve Uddin, 2015).

3.3.7. Veri etiketleme

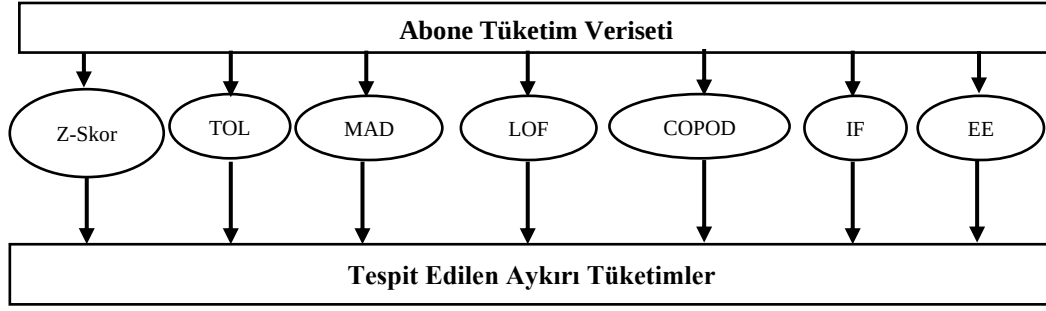
Anomali analizi gözetimsiz veriseti dikkate alınarak yapılmış aykırı, anormal veya harici tüketim değerleri tespit edilerek kurum uzman görüşü ile veya çalışma kapsamında önerilen model ile yapılmıştır. Abone tüketimleri için “Normal”, “Dikkat”, “Riskli”, “Aşırı” olmak üzere 4 farklı tüketim sınıfı ve 4 farklı renk belirlenmiş, tüketim gözlem değerlerinin etiketlenmesi bu sınıflara ve renklere göre yapılmıştır. Çalışmada kullanılan anormal tüketim tespit yöntemlerinin sonuçlarına göre puanlama işlemi yapılmış, sınıflandırmalar bu puanlara göre Tablo 3.8’de gösterildiği şekilde yapılmıştır.

Tablo 3.8. Aykırılık puanlarına göre sınıflandırma etiketleri tablosu.

Aykırılık Puanı	Tüketim Sınıfı
0-1	Normal
2-3	Dikkat
4-5	Riskli
6-7	Aşırı

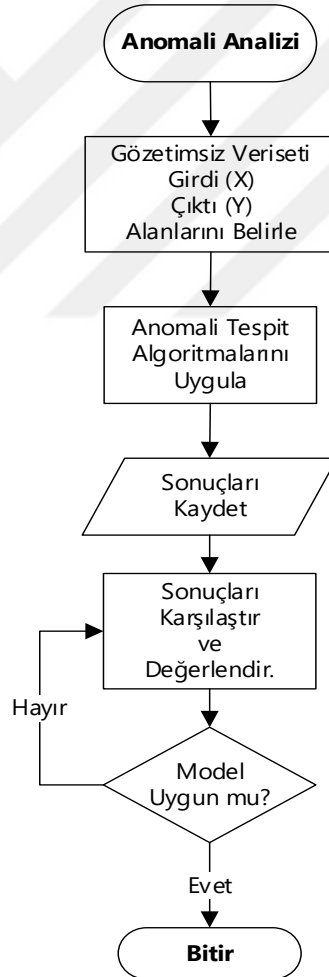
3.4. Anomali analizleri

Çalışmada temin edilen gözetimsiz verilerden oluşan TS verisetindeki aşırı yüksek veya aşırı düşük anormal içme suyu tüketim tespitleri, literatürde de detaylı bir şekilde açıklanan yöntemlerden Şekil 3.15’te gösterildiği gibi 7 tanesi seçilerek yapılmış, örneklem kapsamındaki abonelerin içme suyu tüketimlerini sınıflandırmak ve etiketlemek için yeterli olmuştur. Daha spesifik sınıflandırma ve etiketleme için diğer ML sınıflandırma tekniklerinden de istifade edilebilir.



Şekil 3.15. Topluluk öğrenmesi yöntemi ile aykırı tüketim tespit şeması.

Aykırı tüketim tespiti yapan her bir yöntemin çalışma şeması Şekil 3.16’da gösterilmiş olup, devamında bütün yöntemlerden elde edilen sonuçlar aykırılık puanlarını elde etmek için değerlendirilmiştir.



Şekil 3.16. Anomali analizi model şeması.

3.4.1. Z-Score anomali analizi

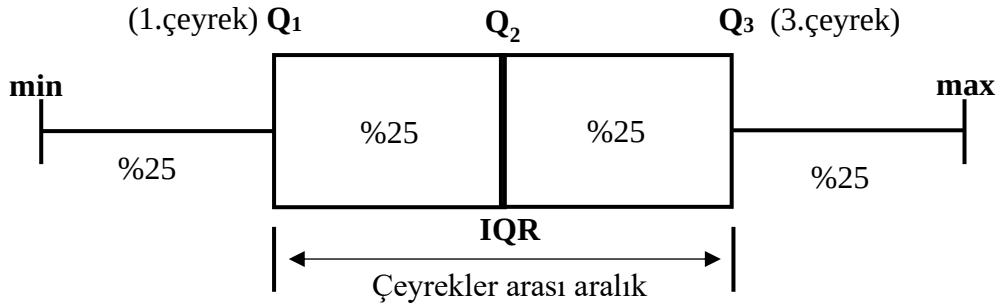
Aykırı gözlem belirlemede en temel yöntemlerden biridir. Gözlem değerlerinden oluşan bir dağılımın standartlaştırılması veya diğer bir tabirle normalizasyonu, dağılım ortalaması 0 ve standart sapması 1 olacak şekilde yeniden hesaplanarak doğrusal bir dönüşüm yapmak demektir. Dönüşümden önceki ve sonraki noktalar arasındaki bağıl mesafeler aynı kalır ve bu standartlaştırılmış puanlara z puanları denir. her bir gözlem için Z puanı denklem 3.19’da gösterildiği gibi hesaplanır (Denis, 2021).

$$z_i = \frac{(x_i - \bar{x})}{\sigma} \quad (3.19)$$

Z puanı, karar değeri ile karşılaştırılarak gözlemin normal dağılım aralığında yer alıp almadığı belirlenir. Karar değeri ise farklı kaynaklara göre değişmektedir. Bu çalışmada karar değeri 2 olarak alınmıştır (Tabachnick ve Fidell, 2013).

3.4.2. TOL anomali analizi

Kutu grafikleri (Box plot) veri dağılımı, merkezi eğilim, çarpıklık ve basıklık hakkında araştırmacılara bilgiler sunan ve beş temel hesaplamanın kullanıldığı bir yöntemdir. Kutu grafiğinin oluşturulmasına ilişkin hesaplamalar Şekil 3.17’de verilmiştir.



Şekil 3.17. Kutu Grafiği yönteminin hesaplanması ve görselleştirilmesi.

Kutu grafiği oluşturulurken öncelikle veri kümesi içindeki y_i gözlem değerleri küçükten büyüğe doğru sıralanır. Veri kümesinin ortanca değeri (Q₂) belirlenir. Medyanın solunda ve sağında kalan alt veri grupları için de ortanca değerleri bulunur. Bulunan bu ortanca değerleri Q₁ (birinci çeyrek) ve Q₃ (üçüncü çeyrek) olarak isimlendirilir. En büyük ve en küçük değerler kullanılarak denklem 3.20’de

gösterildiği gibi çeyrekler arası aralık (IQR) belirlenir. Bu değerlerin ötesinde yer alan değerlere sahip gözlemler denklem 3.21’de gösterildiği gibi aykırı gözlem olarak belirlenir (Tukey, 1977).

$$IQR = Q_3 - Q_1 \quad (3.20)$$

$$\text{Aykırı Değer} = \begin{cases} y_i < \min = Q_1 - 1.5(IQR) \\ y_i > \max = Q_3 + 1.5(IQR) \end{cases} \quad (3.21)$$

3.4.3. IF anomali analizi

İzolasyon ormanı (IF), RF algoritmasına dayalı denetimsiz bir anomali tespit yöntemidir. Algoritma, rastgele bir özellik seçerek ve ardından seçilen özelliğin maksimum ve minimum değerleri arasında rastgele bir bölünmüş değer seçerek gözlemleri alt örnek noktalarına böler. Gereken bölme sayısı, kök düğümünden sonlandırıcı düğüme kadar olan yol uzunluğuna eşittir. Birden çok izolasyon ağacı oluşturulduktan sonra, algoritma anormal olma potansiyelini ölçmek için veri noktalarını yol uzunluğunu referans alarak puanlar. Öncelikle denklem 3.22’de gösterildiği şekilde ikili arama ağacında başarısız aramanın ortalama yol uzunluğu $c(n)$ hesaplanır ve denklem 3.23’te gösterilen puan fonksiyonu kullanılarak her bir gözlemin puanı hesaplanır.

$$c(n) = 2H(n - 1) - (2(n - 1)/n) \quad (3.22)$$

$$s(x_t, n) = 2^{-\frac{E(h(x_t))}{c(n)}} \quad (3.23)$$

Burada x_t gözlem değeri, n gözlem sayısı, $h(x_t)$ kökten x_t veri noktasına geçilen kenarların sayısı, $E(h(x_t))$ $h(x_t)$ ’nin ortalamasıdır ve μ_t olarak da tanımlanabilir. Anormallikler izolasyon ağacının köküne daha yakın olduğundan puanları düşük olacaktır (Liu F. T. ve ark, 2008).

3.4.4. COPOD anomali analizi

COPOD, yüksek boyutlu verisetlerinde sonuçlarında yorumlanabilir özellikleri barındıran bir aykırı değer algılama algoritmasıdır. Kopula, çok değişkenli dağılımın

bağımlılık yapısından aykırı değerleri ayırmayı sağlayan denklem 3.24'te gösterildiği gibi U_t aykırı değerleri olan gözlemlerden oluşan bir vektör, d değişken sayısı, C_u değer aralığı $[0,1]$ olan bir kümülatif dağılım fonksiyonudur (CDF).

$$C_u(U) = P(U_1 \leq u_1, \dots, U_d \leq u_d) \quad (3.24)$$

COPOD, üç aşamalı bir algoritmadır. İlk olarak denklem 3.25'in tersinden yararlanılarak $\hat{F}_d$ ampirik kopula gözlemleri elde edilir.

$$\hat{U}_1, \dots, \hat{U}_d = \hat{F}_1(x), \dots, \hat{F}_d(x) \quad (3.25)$$

İkinci aşamada denklem 3.26, denklem 3.27, denklem 3.28'de gösterildiği gibi n tane gözlemin her biri için üç farklı ampirik gözlem değeri hesaplanır.

$$\text{Sol kuyruk ampirik kopula} = \hat{F}_d(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \leq x) \quad (3.26)$$

$$\text{Sağ kuyruk ampirik kopula} = \hat{\bar{F}}_d(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(-X_i \leq -x) \quad (3.27)$$

$$\text{Düzeltilmiş çarpıklık ampirik kopula} = b_i = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_i)^3}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_i)^2}} \quad (3.28)$$

Burada $\mathbb{I}$ (lojik) işlemi parantez içi şart sağlanıyorsa 1, sağlanmıyorsa 0 değeri verilir. Son olarak ampirik kopulalar tarafından denklem 3.29, denklem 3.30, denklem 3.31'de gösterildiği gibi üretilen olasılığın negatif logunun maksimumunu aykırı değer olarak alıyoruz.

$$p_L = -\sum_1^d \log(\hat{F}_d(x_i)) \quad (3.29)$$

$$p_R = -\sum_1^d \log(\hat{\bar{F}}_d(x_i)) \quad (3.30)$$

$$p_s = \begin{cases} -\sum_1^d \log(\hat{F}_d(x_i)), & b_d < 0 \\ -\sum_1^d \log(\hat{\bar{F}}_d(x_i)), & b_d \geq 0 \end{cases} \quad (3.31)$$

Her bir gözlem için hesaplanan üç farklı ampirik gözlem değeri kullanılarak denklem 3.32’de gösterildiği gibi aykırılık puanları belirlenir. Burada hesaplanan aykırılık puanı, aykırı değer olma olasılığını değil, ilgili gözlem noktasının veri kümesindeki diğer noktalarla karşılaştırıldığında ne kadar olası olduğunun göreceli ölçüsünü gösterir. Aykırılık puanı ne kadar büyükse, gözlemin bir aykırı değer olma olasılığı o kadar yüksektir (Li ve ark, 2020).

$$O(x_i) = \max\{p_l, p_r, p_s\} \quad (3.32)$$

Aykırı değer olma olasılık puanı belirlenen bir karar değeri ile karşılaştırılır ve bu eşik değerinden büyük olanlar aykırı değer olarak belirlenir. Bu çalışmada karar değeri ~2,15 alınmıştır.

3.4.5. MAD anomali analizi

Hampel yöntemi olarak ta bilinen Frank Rudolf Hampel tarafından literatüre kazandırılan ortanca mutlak sapma değeri (MAD), verisetindeki gözlemlerin ortanca değerden ne kadar uzakta olduğunun ölçüsüdür ve denklem 3.33’te gösterildiği gibi her bir x_i gözleminin ortanca ($\check{x}$) ile farkının mutlak değeri olarak hesaplanır. Elde edilen MAD değerleri kullanılarak denklem 3.34’te gösterildiği gibi alt sınır ve üst sınır karar değerleri ortanca ($\overline{MAD}$) ve tutarlılık katsayısı (c) kullanılarak hesaplanır. Hesaplanan alt sınır karar değerinden küçük ve üst sınır karar değerinden büyük olan değerler aykırı gözlem olarak belirlenir (Durivage, 2014).

$$MAD = |x_i - \check{x}| \quad (3.33)$$

$$Karar\ Değerleri \begin{cases} Alt\ sınır = \check{x} - (c * \overline{MAD}) \\ Üst\ sınır = \check{x} + (c * \overline{MAD}) \end{cases} \quad (3.34)$$

Farklı kaynaklara göre c değeri değişmektedir. Bu çalışmada c, 4 olarak alınmıştır.

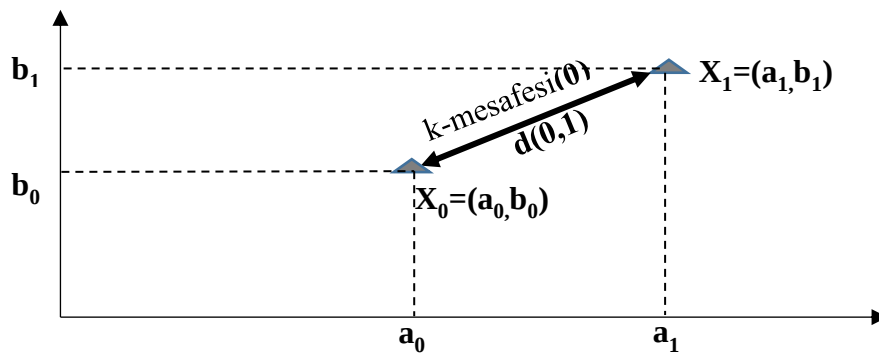
3.4.6. LOF anomali analizi

Yerel aykırı değer faktörü bir gözlem noktasının komşu gözlem noktalarına göre yerel yoğunluğundaki sapmayı hesaplar. Yerel yoğunluğu komşularından çok daha düşük olan nesneler aykırı değer olarak kabul edilir. Öncelikle uzmanlar tarafından k adet komşu sayısı belirlenir bu çalışmada k üç olarak alınmıştır. Her bir X_i gözlem noktasının komşularına olan mesafesi denklem 3.4, denklem 3.36, ve denklem 3.36'da gösterilen Minkowski, Öklid ve Manhattan benzeri mesafe ölçüm yöntemlerinden herhangi biriyle hesaplanır.

$$\text{Öklid} = d(i, j) = \sqrt{\sum_{k=1}^p (x_{i,j} - x_{j,k})^2} \quad (3.35)$$

$$\text{Manhattan} = d(i, j) = \sum_{k=1}^p (|x_{i,j} - x_{j,k}|) \quad (3.36)$$

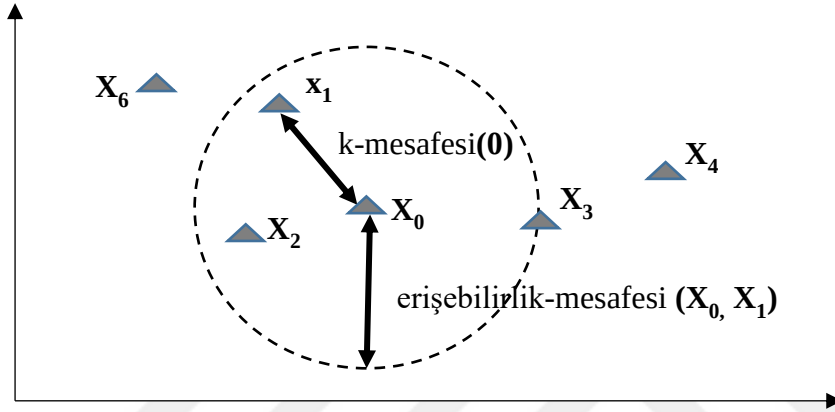
Burada i, j gözlem noktalarını, p ise her bir gözleme ait parametre sayısını temsil etmektedir. Hesaplanan bütün komşu mesafeler küçükten büyüğe sıralanır ve k'ncı sıradaki mesafe, gözlem noktasının k-mesafesi olarak alınır. Şekil 3.18'de gösterildiği gibi komşu mesafeler her bir gözlem noktası için ayrı ayrı yapılır ve n tane gözlem noktası için toplamda $\frac{n(n-1)}{2}$ adet k-mesafesinin hesaplanması gerekmektedir.



Şekil 3.18. İki farklı gözlem noktası arasındaki mesafe gösterimi.

İkinci aşama olarak erişilebilirlik mesafesi denklem 3.37’deki gibi normal mesafe ve k-mesafelerinden en büyük olanı alınarak hesaplanır. Şekil 3.19’da çember içinde kalan noktalar ile X_0 noktası arasındaki erişilebilirlik mesafesi X_0 ’ın k-mesafesi olarak alınır. Çember dışında kalan noktalar ile olan erişilebilirlik mesafesi ise normal mesafe değerleri olarak kalacaktır.

$$erişebilirlik_mesafesi(X_0, X_1) = \max\{k_mesafesi(X_1), mesafe(X_0, X_1)\} \quad (3.37)$$



Şekil 3.19. Komşu sayısı 3 olacak şekilde gözlem noktaları k-mesafesi ve erişilebilirlik mesafesi gösterimi.

Üçüncü aşama olarak yerel erişilebilirlik yoğunluğu (Local reachability density – Lrd) denklem 3.38’de gösterildiği gibi hesaplanır.

$$Lrd(X_0) = \frac{1}{\frac{\sum_{i=1}^k erişebilirlik_mesafesi(X_0, X_i)}{k}} \quad (3.38)$$

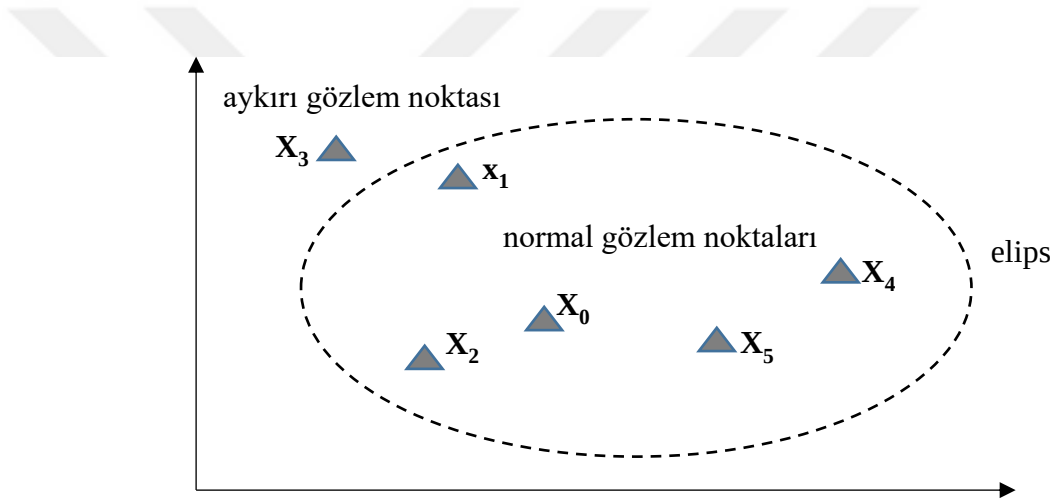
Bir örneğin yerel erişilebilirlik yoğunluğunun ve k-en yakın komşularının oranlarının ortalamasıdır. Bir x gözlem noktasının LOF puanı denklem 3.39’da gösterildiği gibi hesaplanır.

$$LOF(X_0) = \frac{\sum_{i=1}^k \frac{Lrd(X_i)}{Lrd(X_0)}}{k} \quad (3.39)$$

Son olarak, her veri noktasına bir LOF puanı hesaplanır. LOF puanları 1'den küçük ise daha yoğun bir bölgeyi işaret eder. 1'den oldukça büyük skorlar ise aykırı değerleri gösterir. Fakat veri setinin yapısı ve problemin şekline göre farklı kararlar alınabilir. X_0 veri noktasının bir aykırı değer olup olmadığını belirlemek için belirlenen bir karar değeri ile karşılaştırılır ve bu eşik değerinden büyük olanlar aykırı değer olarak belirlenir. Bu çalışmada karar değeri 4 olarak alınmıştır (Breunig ve ark., 2000).

3.4.7. EE anomali analizi

Eliptik zarf yöntemi, verileri, veri özellikleri arasındaki olası kovaryanslarla birlikte yüksek boyutlu bir gauss dağılımı olarak modellenir. Gözlem noktalarının çoğunluğunun Şekil 3.20'de gösterildiği gibi bir elipse sığması için çalışılır.



Şekil 3.20. Eliptik zarf yöntemi ile aykırı gözlem noktaları tespiti.

Elipsin çok dışında bulunan veri örnekleri daha sonra anormallikler veya aykırı değerler olarak kabul edilir. Eliptik zarf yöntemi, elipsin şeklini ve boyutunu tahmin etmek için hızlı minimum kovaryans belirleyicisini (FAST-MCD) kullanır (Rousseeuw, 1999).

Yöntemde amaç, n adet gözlem p adet boyut içerisinde denklemler 3.40 veya denklemler 3.41'de gösterildiği gibi belirlenen h adet gözlem rastgele seçilir ve n 'nin h kombinasyonu kadar alt gözlem kümeleri oluşturulur.

$$h = \frac{n + p + 1}{2} \quad (3.40)$$

$$\frac{n + p + 1}{2} \leq h \leq n \quad (3.41)$$

Her bir alt kümenin varyans-kovaryans matrisi ve determinanı hesaplanır ve kontrol edilir. Eğer grup determinanı sıfır ise sıfırdan büyük oluncaya kadar rastgele gözlemler alt gözlem gruplarına eklenir. Sonrasında gözlem noktalarının denklem 3.42’de gösterildiği şekilde Mahalanobis mesafeleri (d_{MH}) hesaplanır ve küçükten büyüğe doğru sıralanır.

$$d_{MH} = \sqrt{(x - \check{x})^T C^{-1} (x - \check{x})} \quad (3.42)$$

Burada $\check{x}$ gözlemlerin ortalaması, T matrisin devriği ve C denklem 3.43’te gösterildiği gibi hesaplanan kovaryans matrisidir. Mahalanobis mesafesi d_{MH} bir veri noktasının bir dağılımın ortalamasından kaç σ uzakta olduğunu verir.

$$C = \frac{1}{n} \sum_{i=1}^n (x_i - \check{x}) \quad (3.43)$$

Son olarak en küçük değerlere sahip alt kümenin ortalaması konum tahmincisi, varyans-kovaryans matrisi de dağılım ölçüsü olarak belirlenir ve verilerin çoğunu kapsayan bir elips oluşturulur. Elipsin dışında kalan gözlem noktaları aykırı değer olarak belirlenir (Rousseeuw, 1999).

3.5. Performans Ölçümü

Çalışmada bütün anomali tespit modellerinin Tablo 3.9’da gösterildiği gibi öncelikle TP, TN, FP, FN değerlerinden oluşan hata matrisleri hesaplanmıştır. Karışıklık matrisi de denen hata matrisi, çalışmada anormal içme suyu tüketim sınıflarını tespit eden modellerin tahmin değerleri ile gerçek değerler arasındaki performans ölçümünde kullanılan yöntemler için bir altlık olarak değerlendirilmiştir.

Ayrıca geliştirilen modeller Doğruluk, Duyarlılık, Kesinlik, Hassasiyet, Özgüllük, MAE ve MSE performans metrikleri ile karşılaştırılmış, her bir abone için geliştirilen anormal tüketim tespit modellerinden en yüksek doğruluk oranındaki model en iyi model olarak belirlenmiştir.

Anormal tüketim tespit modellerinden doğruluk oranı aynı olanlar var ise anomali duyarlılık ve anomali kesinlik oranlarına göre yüksek oranda olan tercih edilmiştir.

Tablo 3.9. İçmesuyu tüketim anomali tespit modellerine ait hata matrisi tanımlama tablosu.

		Gerçek Değer	
		Positive	Negative
		Normal	Aykırı
		-0-	-1-
Öngörülen Değer	True Normal -0-	TP (True positive) Normal tüketim olarak tespit edilmiş.	FN (False negative) Aykırı tüketim normal olarak tespit edilmiş.
	False Aykırı -1-	FP (False positive) Normal tüketim aykırı olarak tespit edilmiş.	TN (True Negative) Aykırı tüketim aykırı olarak tespit edilmiş.

3.5.1. Doğruluk

Gerçek ve öngörülen tüketim gözlem noktalarının normal ve anormal tüketim sınıfına ait tespit oranıdır.

$$Doğruluk = Accuracy = ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.44)$$

3.5.2. Duyarlılık

Normal tüketim olarak tahmin etmemiz gereken gözlemlerin ne kadarını normal tüketim olarak tahmin ettiğimizi gösteren bir metriktir.

$$Duyarlılık = Recall = \frac{TP}{TP + FN} \quad (3.45)$$

3.5.3. Kesinlik

Kesinlik, normal tüketim olarak tahmin edilen değerlerin gerçekten normal olan tüm gözlem değerlerine oranıdır.

$$Kesinlik = Precision = \frac{TP}{TP + FP} \quad (3.46)$$

3.5.4. Hassasiyet

Gerçek Pozitif Oran (TPR) olarak da ifade edilir ve normal tespit edilen normal tüketim gözlem değerlerinin tüm normal tespit edilen tüketim gözlem değerlerine oranıdır.

$$Hassasiyet = TPR = \frac{TP}{FN + TP} \quad (3.47)$$

Örneğin, Hassasiyet bize kalp hastalığı olan insanların yüzde kaçının gerçekte doğru olarak tanımlandığını söyler.

3.5.5. Özgüllük

Yanlış Pozitif Oran (FPR) olarak da ifade edilir ve yanlışlıkla aykırı tüketim olarak tespit edilmiş gözlem noktalarının tüm anormal tüketim noktalarına oranıdır.

$$Özgüllük = FPR = \frac{FP}{FP + TN} \quad (3.48)$$

Örneğin, Özgüllük bize kalp hastalığı olmayan insanların yüzde kaçının gerçekte doğru bir şekilde tanımlandığını söyler.

3.5.6. MAE

Analizi yapılan her bir aboneye ait verisetindeki bütün tüketim sınıfı gerçek değerleri ile öngörülen değer arasındaki her bir farkın mutlak değerinin ortalamasıdır.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3.49)$$

Burada y gerçek, $\hat{y}$ öngörülen değer, n örnek sayısıdır. Bu değer ne kadar küçükse modelin o kadar daha iyi doğru tespit yapabildiğini gösterir.

3.5.7. MSE

Analizi yapılan her bir aboneye ait verisetindeki bütün tüketim sınıfı gerçek değerleri ile öngörülen değer arasındaki fark karesinin ortalamasıdır.

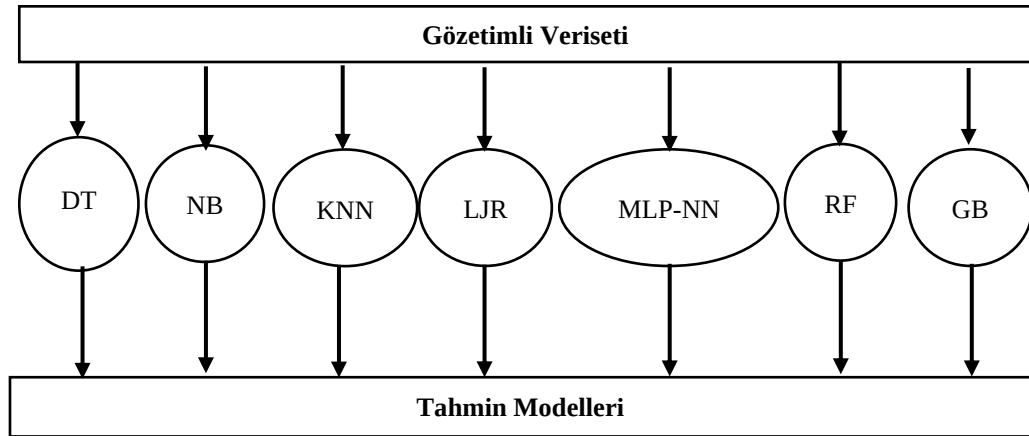
$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.50)$$

Burada y gerçek, $\hat{y}$ öngörülen değer, n örnek sayısıdır. Bu değer ne kadar küçükse modelin o kadar daha iyi doğru tespit yapabildiğini gösterir.

3.6. ML tüketim tahmin modelleri

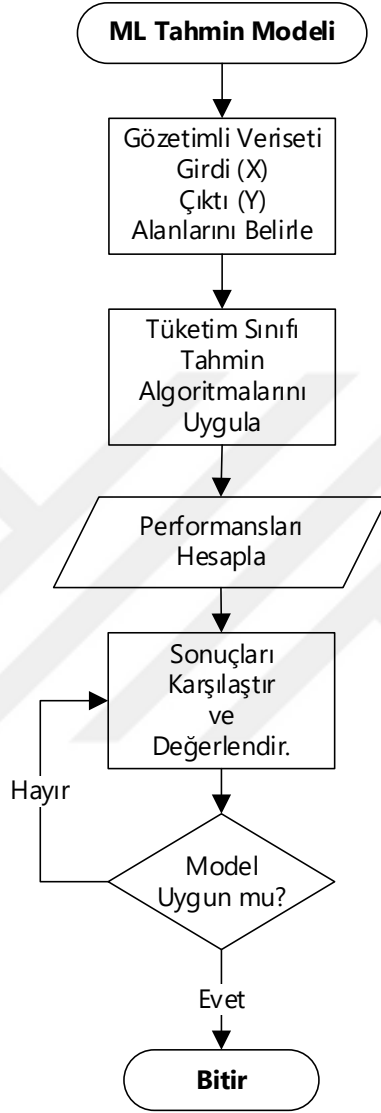
Çalışma sonunda tüketim sınıfı etiketlemesi ile gözetimli hale getirilmiş veriseti, Şekil 3.21’de gösterildiği gibi 7 farklı ML sınıflandırma algoritması kullanılarak içme suyu tüketim sınıfı tahmin modelleri geliştirilmiştir.

Tüketim verileri TS verilerden oluştuğu için ilk yıllar eğitim, son yıllar da kontrol verisi olmak üzere %90’a, %10 oranlarında bölünerek ve bölünmeden kullanılmıştır.



Şekil 3.21. Topluluk öğrenmesi yöntemi ile aykırı tüketim tespit şeması.

İçmesuyu tüketim sınıfı tahmin modellerinin çalışma şeması Şekil 3.22’de gösterilmiş olup, ACC ve determinasyon katsayısı (R^2) performans metrikleriyle karşılaştırılmış, her bir içme suyu aboneliği için en iyi tahmin modeli seçilmiştir.



Şekil 3.22. ML tüketim sınıfı tahmin modeli şeması.

4. MODELİN UYGULANMASI

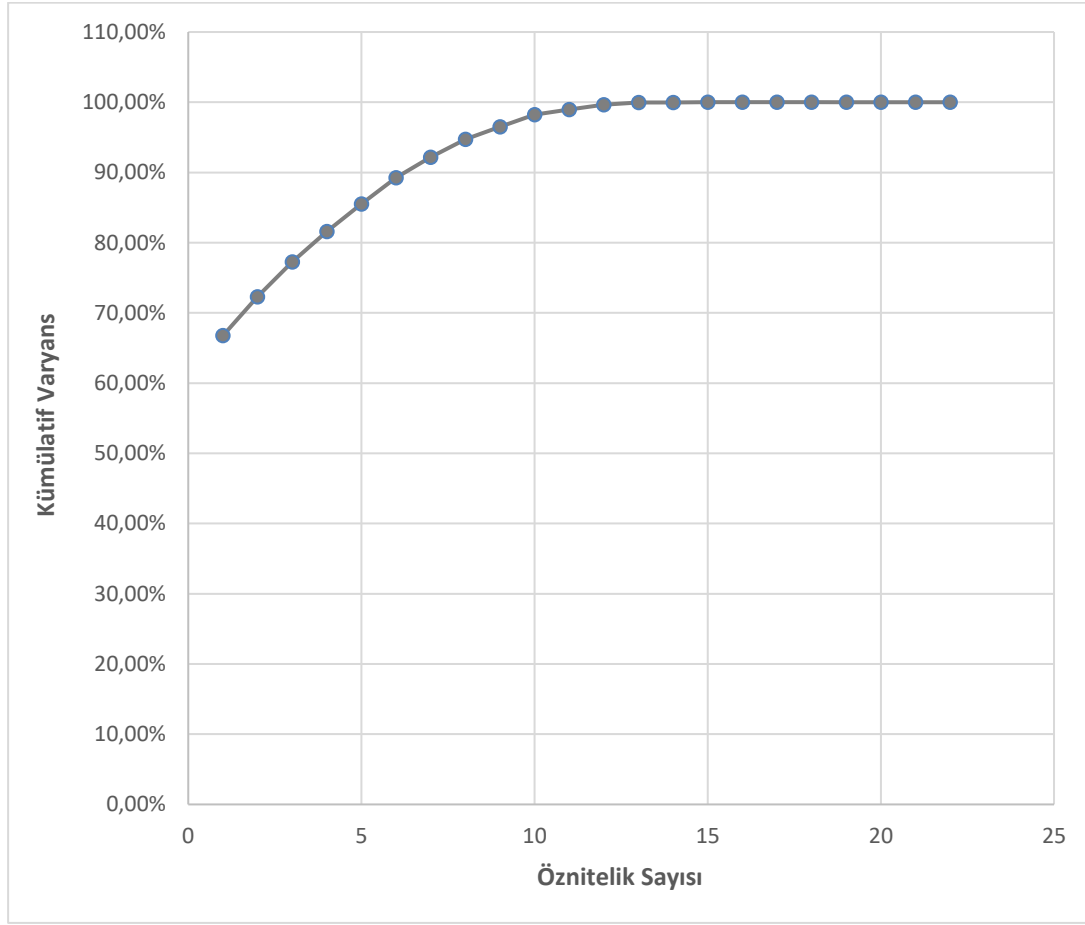
4.1. Modelin Uygulanması ve Analizi

Çalışmada içme suyu abonelerinin tüketim miktarı çıktı değişkeni olarak belirlenmiş hedef entropi $H(Y)$ tüketim değerleri kullanılarak hesaplanmıştır. Öncelikle tüketimi etkileyen öz niteliklerin seçiminde kullanılacak temel istatistikî değerler Tablo 4.1’de gösterildiği gibi hesaplanmıştır.

Tablo 4.1. Öz nitelik seçiminde kullanılan temel istatistikî değerler tablosu.

Öz nitelikler	Min	Max	Ortalama - μ -	Ortanca Medyan	Standart Sapma - σ -	Varyans	Hedef Entropi $H(Y)$	Girdi Entropi $H(X)$
gecen_gun	3	93	35,49	35	10,19	103,84	4,24	1,98
fatura	4,35	165,37	35,85	29,24	26,85	720,72	4,24	4,14
odeme	4,35	165,37	35,93	29,24	26,88	722,65	4,24	4,14
ocak	0	30	3,13	0	7,62	58,12	4,24	0,55
subat	0	27	2,90	0	6,51	42,43	4,24	0,64
mart	0	31	3,16	0	7,08	50,08	4,24	0,71
nisan	0	30	2,96	0	7,63	58,19	4,24	0,61
mayıs	0	31	3,06	0	7,99	63,77	4,24	0,52
haziran	0	26	2,91	0	6,42	41,23	4,24	0,78
temmuz	0	31	2,97	0	6,91	47,73	4,24	0,71
agustos	0	31	2,97	0	7,22	52,10	4,24	0,76
eylul	0	30	2,78	0	7,38	54,47	4,24	0,59
ekim	0	27	2,90	0	6,70	44,94	4,24	0,71
kasım	0	30	2,77	0	7,06	49,91	4,24	0,60
aralık	0	31	2,98	0	7,83	61,31	4,24	0,54
kademe	1	2	1,01	1	0,08	0,01	4,24	0,04
K1	0	1	0,69	1	0,46	0,21	4,24	0,17
K2	0	1	0,10	0	0,29	0,09	4,24	0,12
K3	0	1	0,21	0	0,41	0,17	4,24	0,25
S1	0	1	0,08	0	0,27	0,07	4,24	0,15
S2	0	1	0,59	1	0,49	0,24	4,24	0,17
S3	0	1	0,33	0	0,47	0,22	4,24	0,15

Öznitelik seçiminde eşik değeri belirlenirken yüksek korelasyona sahip veya puanı yüksek olan öznitelikler dikkate alınmış, amacı karşılayacak kadar öznitelik seçimi esas alınmıştır. Seçilecek öznitelik sayısı veriseti özdeğerlerinin (Eigen Value) temsilini hesaplayan açıklanan varyans kriteri (Variance Explained Criteria) oranı %90 alınmıştır. Şekil 4.1’deki grafikten de anlaşılacağı gibi en az 6, en fazla 11 adet faktörün yeterli olacağı, bu nedenle her bir öznitelik seçim yönteminde bu adetler kadar özniteliğin kalmasını sağlayan eşik değerleri belirlenmiştir.



Şekil 4.1. Kümülatif varyans oranına göre öznitelik sayıları grafiği.

Hesaplanan temel istatistiki değerler kullanılarak IG, GR, SU, R, F-Skor, RF sınıflandırması, CHI, ANOVA ve PCA olmak üzere veriseti öznitelik seçim yöntemleri ile özniteliklerin her birinin içme suyu tüketimini etkileme puanları Tablo 4.2’de gösterildiği gibi hesaplanmış, özniteliklerin içme suyu tüketimini etkileyip etkilemediği ve etkiliyorsa etkisinin ne kadar olduğu belirlenmiştir.

Tablo 4.2. Veriseti öznitelik bilgileri seçim yöntemlerinin eşik değeri ile karşılaştırılması.

Öznitelikler	IG	GR	SU	R	F	RF	CHI	ANOVA	PCA
gecen_gun	0,22	0,11	0,07	0,32	4,05	0,12	1,00	1,87	0,19
fatura	0,89	0,21	0,21	0,73	0,73	0,17	0,92	0,00	-0,02
odeme	0,89	0,21	0,21	0,73	0,73	0,17	0,92	0,00	-3,40
ocak	0,07	0,14	0,03	-0,09	1,03	0,04	1,00	0,92	0,00
subat	0,11	0,17	0,05	-0,17	1,32	0,04	1,00	0,89	0,00
mart	0,13	0,18	0,05	-0,07	1,13	0,04	1,00	0,80	0,00
nisan	0,10	0,17	0,04	-0,05	1,07	0,03	0,88	0,41	0,01
mayıs	0,09	0,18	0,04	0,06	0,98	0,03	0,98	0,83	0,02
haziran	0,11	0,14	0,04	0,17	1,35	0,04	0,84	0,10	0,02
temmuz	0,10	0,14	0,04	0,29	1,21	0,04	0,22	1,49	0,31
agustos	0,14	0,18	0,05	0,18	1,15	0,05	0,68	0,03	0,67
eylul	0,10	0,17	0,04	0,07	1,16	0,03	0,88	0,40	0,75
ekim	0,13	0,18	0,05	0,05	1,28	0,04	0,73	0,06	1,73
kasım	0,10	0,16	0,04	0,02	1,23	0,03	0,97	0,61	1,79
aralik	0,08	0,15	0,03	0,00	1,02	0,03	0,96	0,68	2,53
kademe	0,01	0,15	0,00	-0,13	4,90	0,00	0,00	0,00	2,92
K1	0,01	0,04	0,00	-0,22	5,13	0,01	1,00	0,10	3,74
K2	0,01	0,07	0,00	-0,04	5,69	0,01	0,56	0,44	3,92
K3	0,01	0,04	0,00	0,28	5,57	0,02	0,02	0,00	4,33
S1	0,01	0,05	0,00	0,08	5,71	0,01	0,03	0,01	4,98
S2	0,01	0,04	0,00	-0,23	5,21	0,02	0,97	0,12	5,54
S3	0,00	0,03	0,00	0,19	5,45	0,01	0,78	0,26	66,76
Eşik Değer	0,12	0,17	0,045	0,20	1,50	0,042	0,90	0,80	2,5

Aykırı tüketim tespit modellerinde kullanılacak en uygun alt verisetini oluşturan ve tüketimi etkileyen özniteliklerin seçimi için öznitelik seçim algoritmalarından 6 tanesi tercih edilmiş, belirlenen eşik değerleri üzerinden Tablo 4.3'te gösterildiği gibi 6 farklı alt veriseti elde edilmiştir. Ayrıca bütün yöntemlerde seçilmiş öznitelikler oylama yöntemi kullanılarak 3 ve daha fazla kez seçilmiş özniteliklerden oluşan 7. bir alt veriseti elde edilmiştir.

Tablo 4.3. Veriseti öznitelik seçim puanlanması.

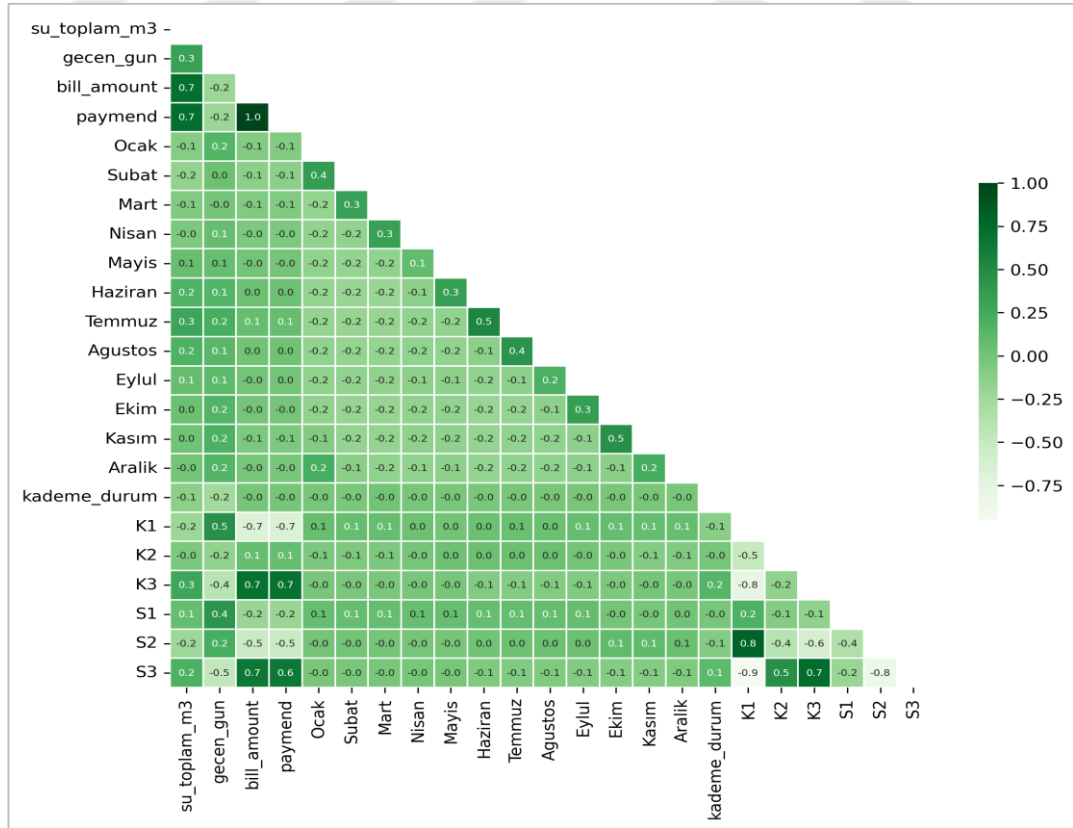
Öznitelikler	IG	GR	SU	R	F	RF	Puan
gecen_gun	1	0	1	1	1	1	5
fatura	1	1	1	1	0	1	5
odeme	1	1	1	1	0	1	5
ocak	0	0	0	0	0	0	0
subat	0	1	1	0	0	1	3
mart	1	1	1	0	0	1	4
nisan	0	0	0	0	0	0	0
mayıs	0	1	0	0	0	0	1
haziran	0	0	0	0	0	0	0
temmuz	0	0	0	1	0	0	1
agustos	1	1	1	0	0	1	4
eylul	0	0	0	0	0	0	0
ekim	1	1	1	0	0	0	3
kasım	0	0	0	0	0	0	0
aralik	0	0	0	0	0	0	0
kademe	0	0	0	0	1	0	1
K1	0	0	0	1	1	0	2
K2	0	0	0	0	1	0	1
K3	0	0	0	1	1	0	2
S1	0	0	0	0	1	0	1
S2	0	0	0	1	1	0	2
S3	0	0	0	0	1	0	1
Alt Verisetler	V1	V2	V3	V4	V5	V6	V7

Her bir yöntem ile seçimi yapılan özniteliklerden oluşan alt verisetleri Tablo 4.4’te gösterildiği gibi belirlenmiştir. Burada her abone için değişimi yapılan sayaç sayısı, sözleşme yapılan kullanıcı sayısı, tüketim ölçümü yapılan günler, tüketim miktarları, fatura tutarları, ödeme durumu verileri farklı olması muhtemel olacağından öznitelik sayılarının da farklı olması tabii olacaktır. Bundan dolayı alt verisetleri her aboneye özel olarak çalışmada kaydedilmiştir.

Tablo 4.4. Alt verisetler ve seçimi yapılan öznitelikler tablosu.

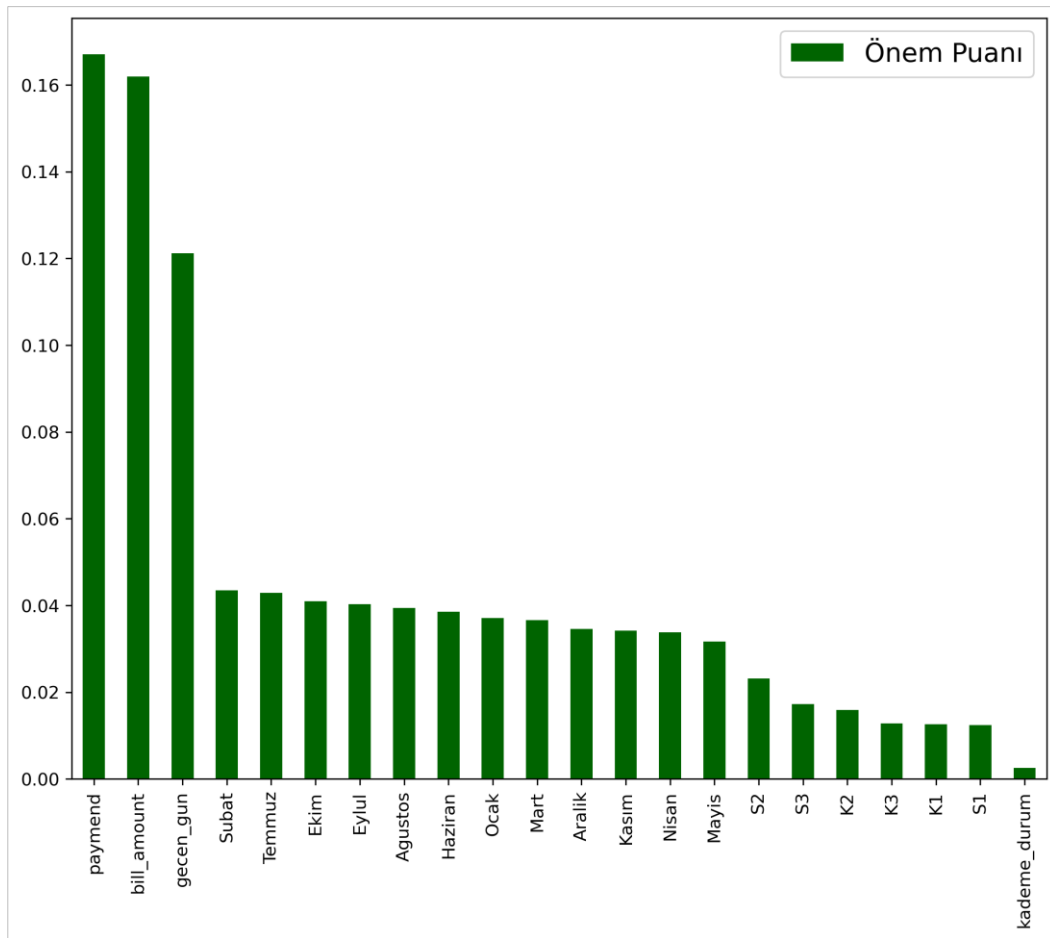
Alt Verisetler	Yöntem	Seçilen Öznitelikler	Öznitelik Adet
Veriseti 1	IG	gecen_gun, fatura, odeme, mart, agustos, ekim	6
Veriseti 2	GR	fatura, odeme, subat, mart, mayis, agustos, ekim	7
Veriseti 3	SU	gecen_gun, fatura, odeme, subat, mart, agustos, ekim	7
Veriseti 4	R	gecen_gun, fatura, odeme, temmuz, K1, K3, S2	7
Veriseti 5	F	gecen_gun, kademe, K1, K2, K3, S1, S2, S3	8
Veriseti 6	RF	gecen_gun, fatura, odeme, subat, mart, agustos	6
Veriseti 7	Ortak	gecen_gun, fatura, odeme, subat, mart, agustos, ekim	7

Çalışmada veri ön işlemleri sonucu elde edilen nihai verisetinde çıktı değişkeni olarak alınan dönemsel su tüketim miktarı ile girdi değişkenleri arasındaki korelasyon oranları hesaplanmış, Şekil 4.2’de gösterildiği gibi korelasyon matrisi çıkarılmıştır.

**Şekil 4.2.** Veriseti öznitelik bilgilerinin korelasyon matrisi.

Korelasyon matrisinden de anlaşılaçağı üzere eşik değeri $\pm 0,2$ alındığında 7 öznitelikli bir veriseti (Veriseti 4) elde edilmiştir.

Çalışmada tercih edilen öznitelik seçim algoritmalarından RF sınıflandırma yöntemi ile aylık su tüketimini etkileyen öznitelik bilgilerinin önem puanları hesaplanmış ve büyükten küçüğe doğru sıralanmıştır. Eşik değeri 0,04 alındığında Şekil 4.3'te gösterilen grafikten de anlaşılaçağı gibi “Ekim” özniteliğı ve daha büyük önem puanına sahip diğer 5 özniteliğın seçimi yapılmış, toplamda altı öznitelikli bir veriseti (Veriseti 6) elde edilmiştir.



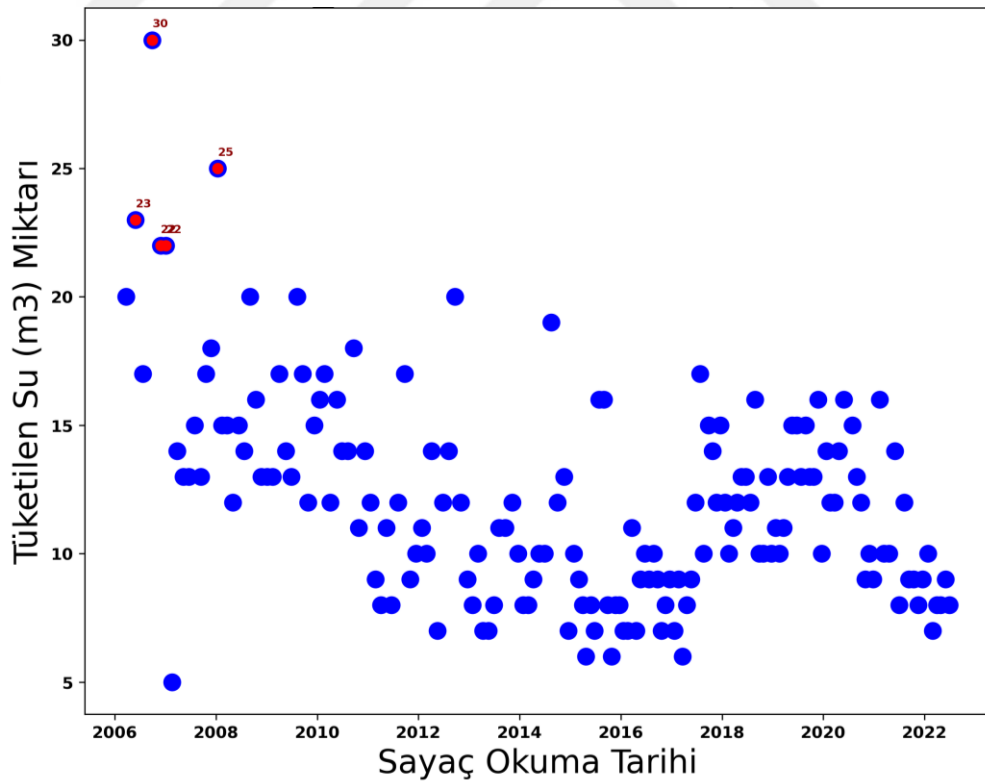
Şekil 4.3. Veriseti öznitelik bilgilerinin RF yöntemi ile elde edilen önem puanları grafiğı.

İçmesuyu tüketimlerindeki anormal durumları tespit etmek için Tablo 4.5'te gösterildiğı gibi 7 farklı anomali analiz algoritması kullanılmıştır.

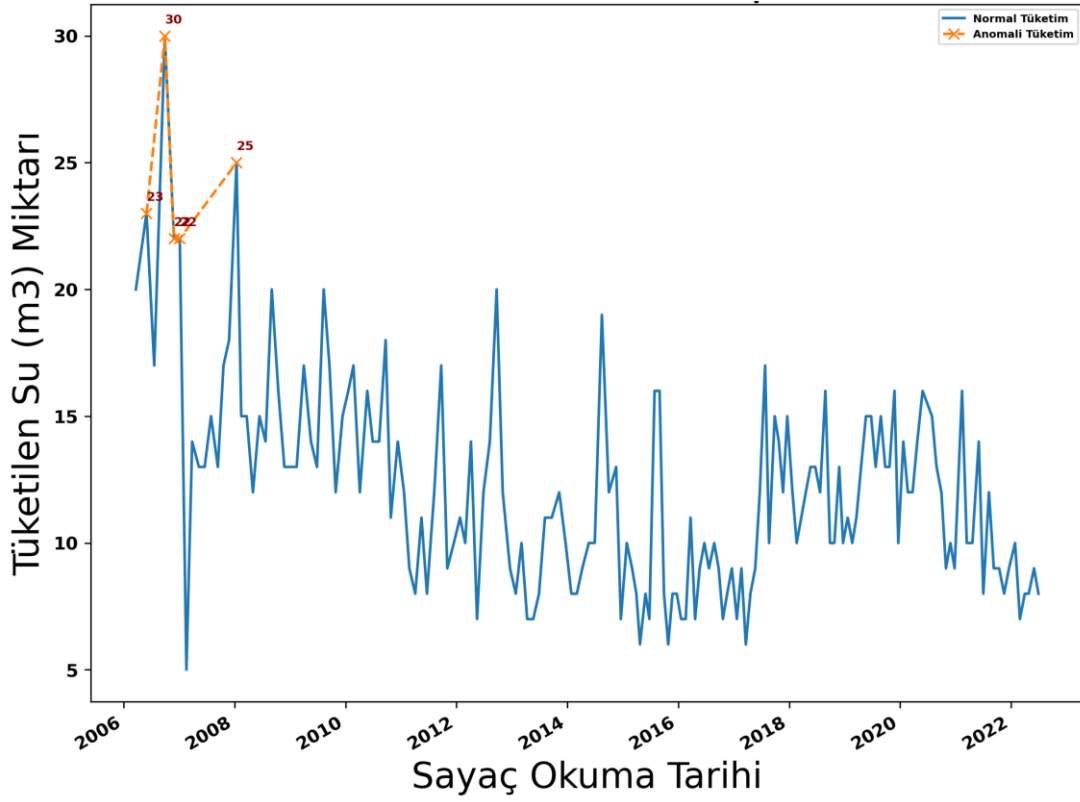
Tablo 4.5. Çalışma kapsamında kullanılan anomali analizi algoritmaları.

ML_Model	Grup	Karar Değeri
1 Tukey Outlier Labeling (TOL)	Anomali	Min, Max
2 Isolating Forest (IF)	Anomali	Puanı düşük, köke yakın olan
3 Z_Score	Anomali	> 2
4 Copula-Based Outlier Detection (COPOD)	Anomali	> 2,15
5 The Median Absolute Devision (MAD)	Anomali	Alt sınır, Üst sınır
6 Local Outlier Factor (LOF)	Anomali	> 4
7 Elliptic Envelope (EE)	Anomali	Elips dışında kalan

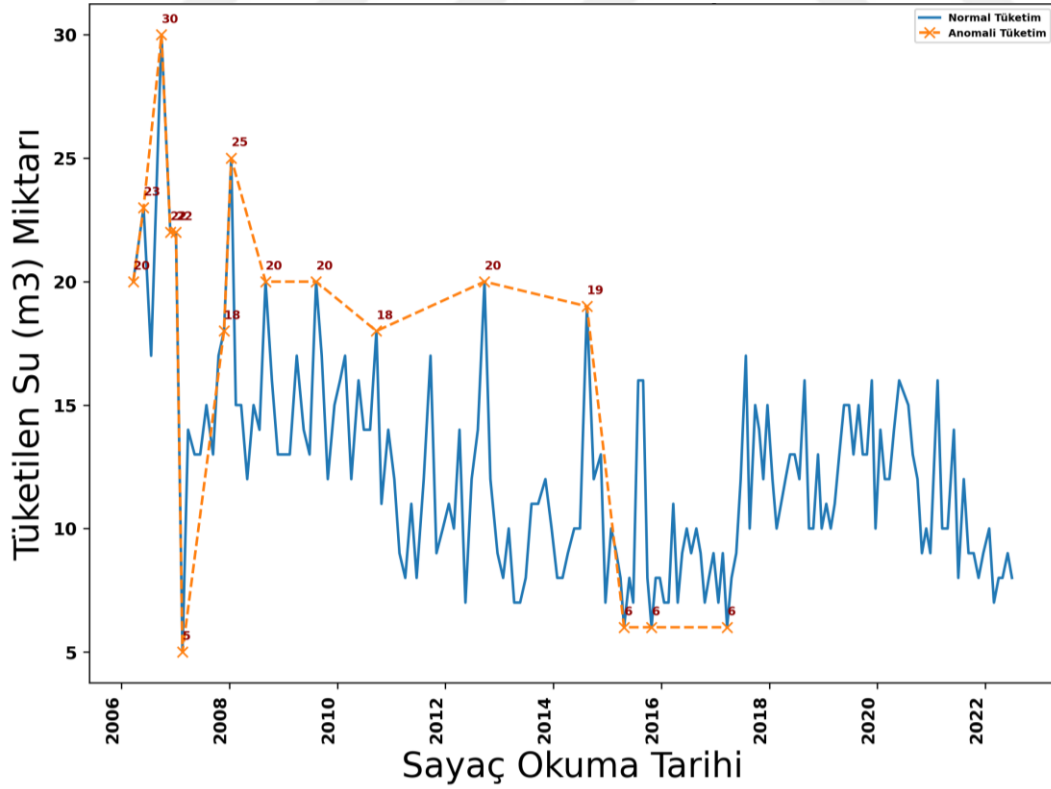
Kullanılan anomali tespit yöntemlerinin her birine özel olmak üzere eşik değerler belirlenmiştir. Eşik değerler, kurumdaki uzman görüşü, sahadan elde edilecek geri dönütler ile daha spesifik olarak ayarlanabilmektedir. Anomali tespit yöntemleriyle elde edilen anormal gözlem noktaları Şekil 4.4, Şekil 4.5, Şekil 4.6, Şekil 4.7, Şekil 4.8, Şekil 4.9 ve Şekil 4.10'daki grafiklerde gösterilmiştir.



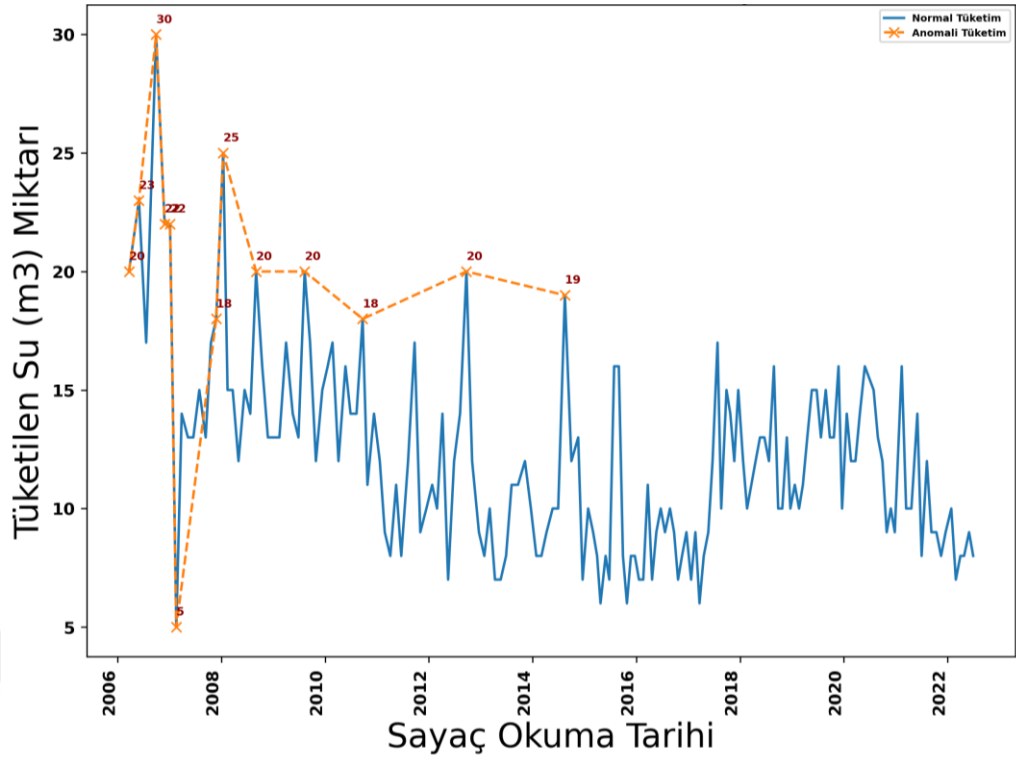
Şekil 4.4. Z-Score anomali analizi ile tespit edilen anormal tüketim grafiği.



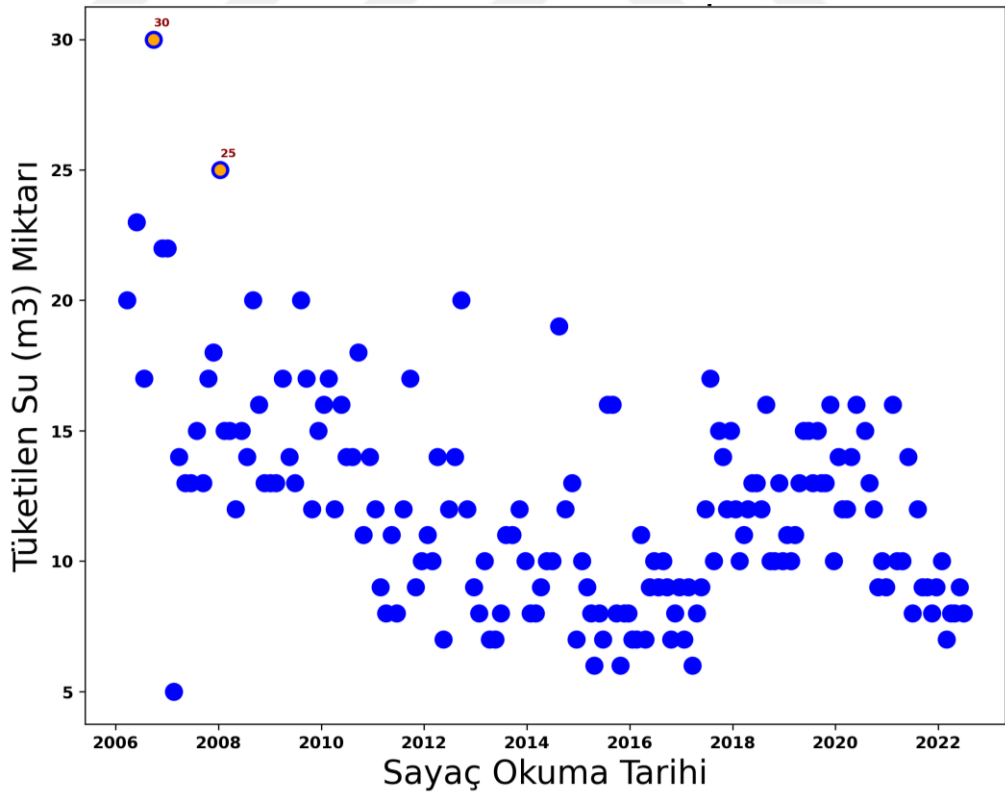
Şekil 4.5. TOL-Tukey Outlier Labeling algoritması ile tespit edilen anormal tüketim grafiği.



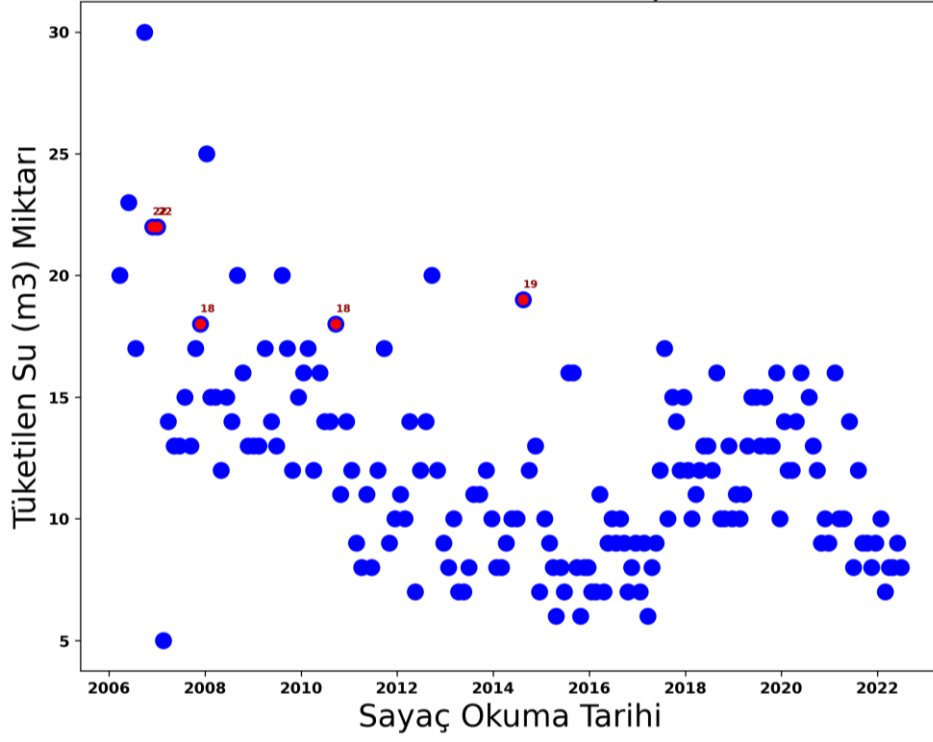
Şekil 4.6. IF-Isolating Forest Outlier Detection algoritması ile tespit edilen anormal tüketim grafiği.



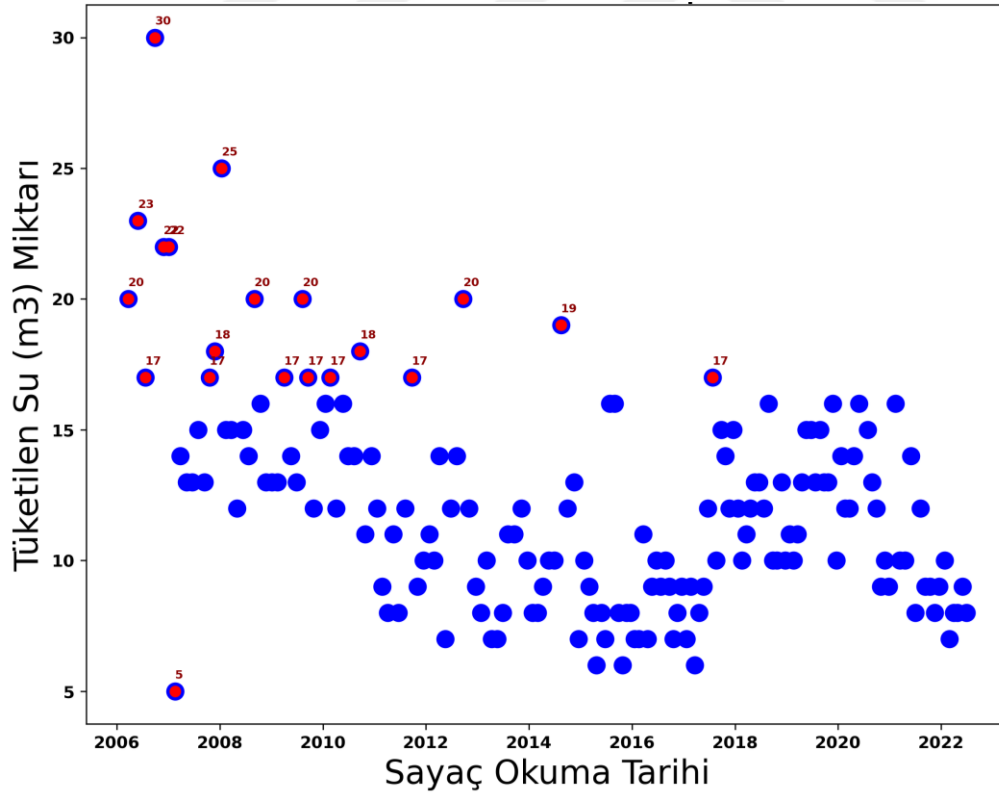
Şekil 4.7. COPOD-Copula Based Outlier Detection algoritması ile tespit edilen anormal tüketim grafiği.



Şekil 4.8. MAD-The Median Absolute Deviation algoritması ile tespit edilen anormal tüketim grafiği.



Şekil 4.9. LOF-Local Outlier Factor algoritması ile tespit edilen anormal tüketim grafiği.



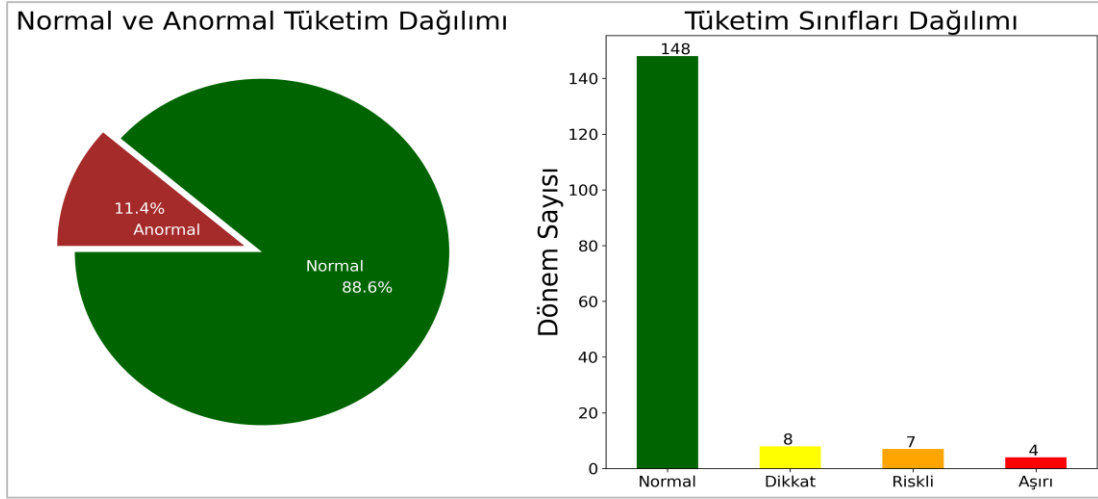
Şekil 4.10. EE-Elliptic Envelope Anomaly Dedection algoritması ile tespit edilen anormal tüketim grafiği.

Çalışmada kullanılan tüketim gözlem noktalarını “Normal”, “Dikkat”, “Riskli” ve “Aşırı” olmak üzere 4 farklı ifadeye göre sınıflandırmak ve gerekli olan tüketim sınırlarının belirlenmesi için Tablo 4.6’da gösterildiği gibi her algoritma için 1 puan verilerek 1 ile 7 puan arasında toplam aykırılık puanları hesaplanmıştır. Aykırılık puan değerleri 0 veya 1 ise “Normal”, 2 veya 3 ise “Dikkat”, 4 veya 5 ise “Riskli” ve 6 veya 7 ise “Aşırı” tüketim yapıldığı anlamında uyarı ifadeleri kullanılmış olup “Yeşil”, “Sarı”, “Turuncu” ve “Kırmızı” renk kartelaları ile kategorize edilmiş sayaç okuma bildirim makbuzlarında bu renkler ile kullanıcıların bilgilendirilmesi sağlanmış olacaktır.

Tablo 4.6. Anomali analizi sonucunda tespit edilen gözlemlerin toplam aykırılık değerleri ve sınıfları.

Okuma Tarihi	Z Score	MAD	TOL	COPOD	IF	LOF	EE	Aykırılık Puanı	Tüketim Sınıfı
22.02.2006	0	0	0	0	0	0	0	0	Normal
18.04.2006	0	0	0	0	0	0	0	0	Normal
20.06.2006	1	0	0	1	1	0	1	4	Riskli
...	...	...	...	...	...	...	...	...	...
22.08.2021	0	0	0	0	0	0	0	0	Normal
20.09.2021	1	1	1	1	1	0	1	6	Aşırı
15.10.2021	0	0	0	0	0	0	0	0	Normal
11.11.2021	0	0	0	0	0	0	0	0	Normal
1.12.2021	0	0	0	0	0	0	0	0	Normal
7.01.2022	0	0	0	1	1	0	1	3	Dikkat
1.02.2022	0	0	0	0	0	0	0	0	Normal
1.03.2022	0	0	0	0	0	0	0	0	Normal
28.03.2022	0	0	0	0	0	0	0	0	Normal
31.03.2022	0	0	0	0	1	0	0	1	Normal
29.04.2022	0	0	0	0	0	0	0	0	Normal
12.05.2022	0	0	0	0	0	1	0	1	Normal
6.06.2022	0	0	0	0	0	0	0	0	Normal

Son olarak etiketlenmiş yani gözetimli hale getirilmiş veriseti içeriğinde Şekil 4.11’de tüketim sınıf dağılımı gösterilen tüketim gözlem noktalarından oluşan verisetinde anormal veri sayısı çok düşük olmakla birlikte ML sınıflandırma algoritmaları ile tüketim sınıf tahmin modelleri geliştirilebilecek bir veriseti alt yapısı elde edilmiştir.



Şekil 4.11. Etiketlenmiş gözlem noktalarına ait normal ve anormal tüketim durumları grafiği.

Veri önışlemleri sonucunda elde edilen ve analiz uygulamasında kullanılan nihai 24 öznitelikli veriseti ve alt verisetlerine ait girdi ve çıktı gerçek değerleri, normalize edilerek bütün sayaç okuma gözlem noktalarının girdi değerleri ile çıktı değeri arası mesafeleri minkowski yöntemi ile hesaplanmış, Tablo 4.7’de gösterildiği gibi tek öznitelikli bir veriseti elde edilmiştir.

Tablo 4.7. İçmesuyu tüketim gözlem noktaları mesafe değerleri.

okuma_tarihi	su_toplam_m3	m_manhattan	m_oklid	m_minkowski
2006-03-04	28	10,71	2,758568469	0,887913293
2006-05-02	29	11,1	2,923627883	0,938718245
2006-07-11	30	11,13	2,96224577	0,972211258
2006-09-06	35	13,13	3,57056018	1,157521705
2006-11-16	32	11,57	3,145933884	1,041654657
...	...	...	...	...
2022-01-28	13	6,54	1,920416621	1,290003469
2022-03-03	11	5,79	1,835020436	1,230050385
2022-03-30	10	5,34	1,743559577	1,190106262
2022-04-27	12	6,25	1,864537476	1,260009459
2022-06-02	10	5,3	1,784376642	1,190202062

Çalışmanın model geliştirme uygulaması bu değerler ile tekrar analiz edilmiş, yeni anomali tespit modelleri ayrıca geliştirilmiştir. Hem gerçek değerler hem de normalize edilen minkowski mesafe değerleri kullanılarak elde edilen modeller doğruluk,

duyarlılık, kesinlik, hassasiyet, özgüllük, MAE ve MSE performans metrikleri ile Tablo 4.8’de gösterildiği gibi karşılaştırılmıştır.

Tablo 4.8. Geliştirilen modeller ve performans metrik değerleri.

Model	Okuma	Normal	Anormal	Aşırı	Dikkat	Riskli	TP	FP	FN	TN	F1	Doğruluk	Duyarlılık	Kesinlik	Özgüllük	R ²	MAE	MSE
c_Anomali	160	141	19	0	0	0	141	0	0	19	1,00	1,00	1,00	1,00	1,00	1,00	0,00	0,00
c_COPOD	160	144	16	0	0	0	141	3	0	16	0,91	0,98	0,84	1,00	1,00	0,82	0,02	0,02
c_EE	160	144	16	0	0	0	141	3	0	16	0,91	0,98	0,84	1,00	1,00	0,82	0,02	0,02
c_IF	160	144	16	0	0	0	133	11	8	8	0,46	0,88	0,42	0,50	0,94	-0,13	0,12	0,12
c_LOF	160	120	40	0	0	0	103	17	38	2	0,07	0,66	0,11	0,05	0,73	-2,28	0,34	0,34
c_MAD	160	140	20	0	0	0	139	1	2	18	0,92	0,98	0,95	0,90	0,99	0,82	0,02	0,02
c_TOL	160	145	15	0	0	0	141	4	0	15	0,88	0,98	0,79	1,00	1,00	0,76	0,03	0,03
c_V1_Anomali	160	142	18	0	0	0	124	18	17	1	0,05	0,78	0,05	0,06	0,88	-1,09	0,22	0,22
c_V1_COPOD	160	144	16	0	0	0	126	18	15	1	0,06	0,79	0,05	0,06	0,89	-0,97	0,21	0,21
c_V1_EE	160	144	16	0	0	0	126	18	15	1	0,06	0,79	0,05	0,06	0,89	-0,97	0,21	0,21
c_V1_IF	160	144	16	0	0	0	126	18	15	1	0,06	0,79	0,05	0,06	0,89	-0,97	0,21	0,21
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
g_V7_LOF	160	152	8	0	0	0	134	18	8	0	0,00	0,84	0,00	0,00	0,94	-0,63	0,16	0,16
g_V7_MAD	160	154	6	0	0	0	137	17	5	1	0,08	0,86	0,06	0,17	0,96	-0,38	0,14	0,14
g_V7_TOL	160	154	6	0	0	0	137	17	5	1	0,08	0,86	0,06	0,17	0,96	-0,38	0,14	0,14

Burada “g_” ile gerçek değerler ile, “c_” ile başlayan modeller minkowski mesafe değerleri ile elde edilen modellerdir. Her bir abone için gerçek değerler ile 56, minkowski mesafe değerleriyle ile 56 adet olmak üzere toplamda 112 adet model geliştirilmiştir.

Bütün modeller ile tespit edilen anomali tüketimler Tablo 4.9’de gösterildiği gibi karşılaştırılmış, performans metriklerinden en yüksek doğruluk oranına sahip model en iyi model olarak seçilmiştir. En iyi anormal tüketim tespit modeli birden fazla ise yani doğruluk değeri aynı ise kesinlik ve duyarlılık oranları dikkate alınmıştır.

Tablo 4.9. Anomali analizi sonucunda tespit edilen gözlemlerin toplam aykırılık değerleri ve sınıfları.

Okuma Tarihi	g_V1_Z_score	g_V1_MAD	g_V1_TOL	...	c_V7_COPOD	c_V7_IF	c_V7_LOF	c_V7_EE	c_V7_a_puani	c_V7_a_sınıf	c_V7_Anomali
2006-03-04	1	1	0	...	1	0	0	1	4	Riskli	1
2006-05-02	1	1	0	...	1	1	0	1	5	Riskli	1
2006-07-11	1	1	0	...	1	1	1	1	6	Aşırı	1
2006-09-06	1	1	1	...	1	1	0	1	6	Aşırı	1
2006-11-16	1	1	0	...	1	1	1	1	6	Aşırı	1
...	...	...	...	...	...	...	...	...	...	...	...
2022-01-28	0	0	0	...	0	0	0	0	0	Normal	0
2022-03-03	0	0	0	...	0	0	1	0	1	Dikkat	0
2022-03-30	0	0	0	...	0	0	0	0	0	Normal	0
2022-04-27	0	0	0	...	0	0	0	0	0	Normal	0
2022-06-02	0	0	0	...	0	0	0	0	0	Normal	0

Tüketim sınıf etiketlemesi ile gözetimli hale getirilmiş veriseti, 7 farklı ML algoritması kullanılarak tüketim sınıfı tahmin modelleri geliştirilmiştir. Bütün modeller ACC ve R^2 performans metrikleri ile karşılaştırılmış, en yüksek doğruluk oranına sahip model en iyi model olarak seçilmiştir. Tahmin modelleri aboneye özel olup sayaç okuma esnasında abone bilgilendirmelerinde kullanılabilecektir.

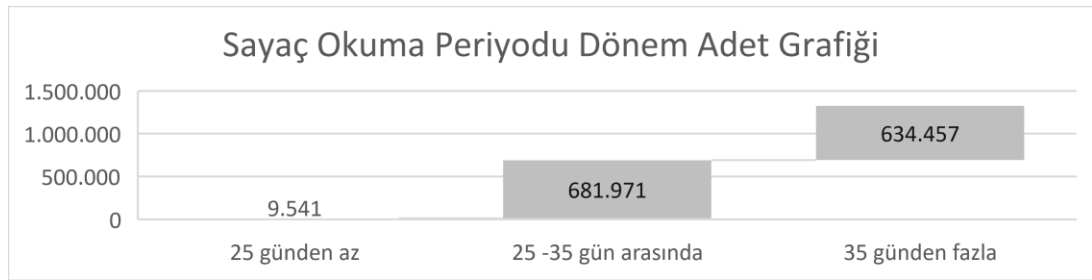
5. ARAŞTIRMA BULGULARI

5.1. Anomali Analiz Sonuçları

Veriseti ön işlemleri sonucunda 2 tablo oluşturulmuş, tabloların birinde her bir abonenin her bir okuma dönemine ait 24 öznitelikli veriseti, ikinci tabloda ise her bir aboneye ait alt verisetlerinin hangi özniteliklerden oluştuğu bilgisi yer almaktadır. Toplamda ilk tablo 1.325.969, ikinci tablo da 142.322 satır veriden oluşmaktadır.

Veriseti incelendiğinde özetle 2006–2022 (ilk 6 ay) yılları arasında kış aylarında içmesuyu tüketimlerinin düştüğü, yaz aylarında tüketim değerlerinin diğer aylara göre daha yüksek olduğu, çalışma kapsamında analizi yapılan abonelere ait toplamda 17.908.783 m³ tüketim yapıldığı, 1 okuma döneminde konut/mesken su tüketim miktarı en yüksek 818 m³, en düşük 1 m³ ve ortalama 13 m³ olduğu, yine 1 okuma döneminin en geç 819 gün, en erken aynı gün ve ortalama 36 gün olduğu tespit edilmiştir.

Abonelere ait tüketim verileri incelendiğinde Şekil 5.1’de de gösterildiği gibi 1.325.969 adet okumanın %0,72’sinin 25 günden az bir zamanda, %47,85’sinin 35 günden fazla bir zamanda ve %51,43’ünün ise 25 ile 35 gün aralığında yapıldığı tespit edilmiştir.



Şekil 5.1. Sayaç okuma dönem periyot grafiği.

Temel veriseti 24 öznitelikten, öznitelik seçim yöntemleri ile elde edilen alt verisetleri ise en az 7 en fazla 12 adet öznitelikten oluşmuştur. Burada kümülatif varyans %90 ve üzerinde olacak şekilde öznitelik adetleri belirlenmiş, bir sonraki öznitelik kümülatif varyansı %95’in üzerine çıkarıyorsa alt veriseti öznitelik sayısının bir fazlası olarak belirlenmiştir.

Her bir aboneye ait Tablo 5.1’de gösterildiği gibi gerçek değerlerden oluşan ve veri ön işlemleri ile elde edilen 24 öznitelikğin tamamının olduğu ve öznitelik seçim yöntemleri sonrasında elde edilen alt verisetleri olmak üzere toplamda 131.584 adet alt veriseti kullanılmıştır.

Tablo 5.1. Öznitelik seçim yöntemleriyle elde edilerek oluşturulan alt verisetleri ve bütün özniteliklerin olduğu veriseti sayıları tablosu.

	Abone Sayısı	Öznitelik Seçim Yöntem Sayısı	Alt Veriseti Sayısı	Tüm Özniteliklerin Olduğu Veriseti Sayısı
Gerçek değerler ile	8.224	7	57.568	8.254
Minkowski değerleri ile	8.224	7	57.568	8.254
Toplam Model			115.136	16.508

Alt verisetler 6 farklı öznitelik seçim algoritması ile belirlenmiş olup, öznitelik seçme algoritmalarının Tablo 5.2’de aynı öznitelikleri seçme sayıları ve oranları gösterilmiştir.

Tablo 5.2. Öznitelik seçim yöntemleriyle aynı özniteliklerin seçilme sayıları ve oranları.

Öznitelik Seçim Yöntemleri				Bilgi Kazancı	Kazanç Oranı	Simetrik Belirsizlik Katsayısı	Korelasyon	F Skor	Rastgele Orman	Ortak
				IG	GR	SU	R	F	RF	V7
				V1	V2	V3	V4	V5	V6	V7
Bilgi Kazancı	IG	V1	Adet		495	7.066	84	0	1.018	3.669
			Oran		%6	%86	%1	%0	%12	%45
Kazanç Oranı	GR	V2	Adet			587	114	0	340	345
			Oran			%7	%1	%0	%4	%4
Simetrik Belirsizlik Katsayısı	SU	V3	Adet				85	0	972	3.641
			Oran				%1	%0	%12	%44
Korelasyon	R	V4	Adet					11	51	44
			Oran					%0	%1	%1
F Skor	F	V5	Adet						0	0
			Oran						%0	%0
Rastgele Orman	RF	V6	Adet							707
			Oran							%9

Ayrıca öznitelik seçme yöntemlerinden IG ve SU yöntemlerinin ortak olarak belirlenen özniteliklere en yakın oranlarda öznitelik seçimi yaptıkları anlaşılmıştır. Sırasıyla RF ve GR yöntemlerinin çok az da olsa aynı öznitelikleri seçtikleri, korelasyon ve f skor yöntemleri kullanılarak yapılan öznitelik seçimlerinin diğer yöntemlere göre neredeyse hiç benzerlik göstermediği, birbirleriyle de çok az bir benzerlik taşıdıkları anlaşılmıştır.

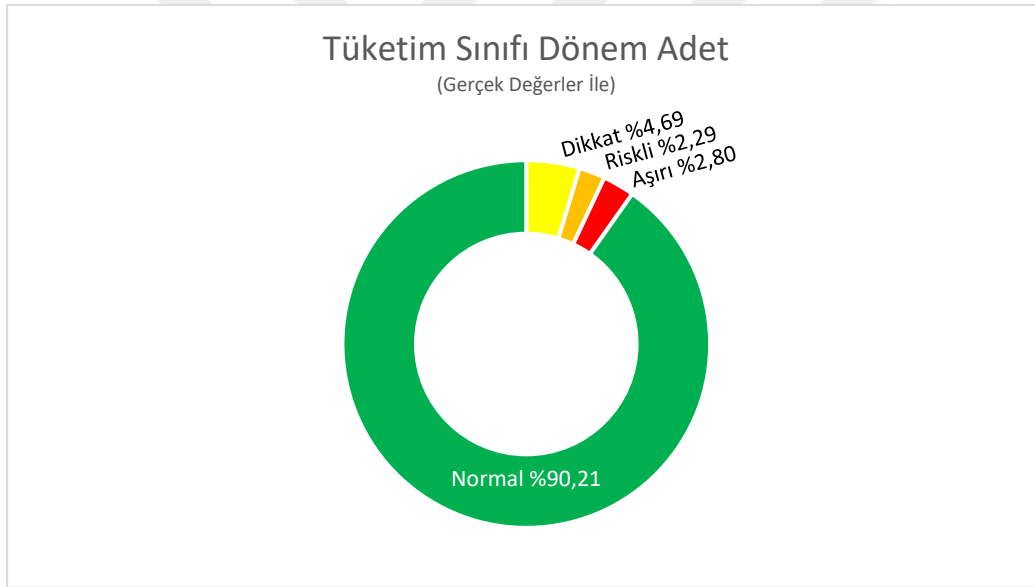
Ayrıca veri ön işlemleri yöntemleriyle her bir abone için en az 16 en fazla 24 öznitelikli temel verisetleri elde edilmiştir. Bu temel verisetleri kullanılarak öznitelik seçim yöntemleriyle en az 6 en fazla 11 öznitelikli alt verisetleri elde edilmiştir. Toplamda 142.322 adet öznitelik kullanılmış olup, Tablo 5.3'te de gösterildiği gibi seçilen özniteliklerin seçilme sayısı ve oranları belirlenmiştir.

Tablo 5.3. Alt verisetlerinde öznitelik seçim yöntemlerinin seçim adetleri ve oranları tablosu.

Öznitelik	Alt Verisetlerinde	Seçilme Oranı
Haziran	33.406	%23,47
Temmuz	32.408	%22,77
Ekim	30.908	%21,72
Şubat	30.665	%21,55
Eylül	30.451	%21,40
Mart	29.798	%20,94
Kasım	28.814	%20,25
Aralık	26.991	%18,96
Mayıs	26.172	%18,39
Ocak	24.735	%17,38
Ağustos	23.464	%16,49
Nisan	22.197	%15,60
S1	14.732	%10,35
Ödeme	13.248	%9,31
Kademe	10.852	%7,62
K1	10.150	%7,13
S2	9.856	%6,93
S3	1.879	%1,32
K2	1.428	%1,00
K3	148	%0,10
S4	108	%0,08
K4	16	%0,01
S5	7	%0,00
K5	2	%0,00

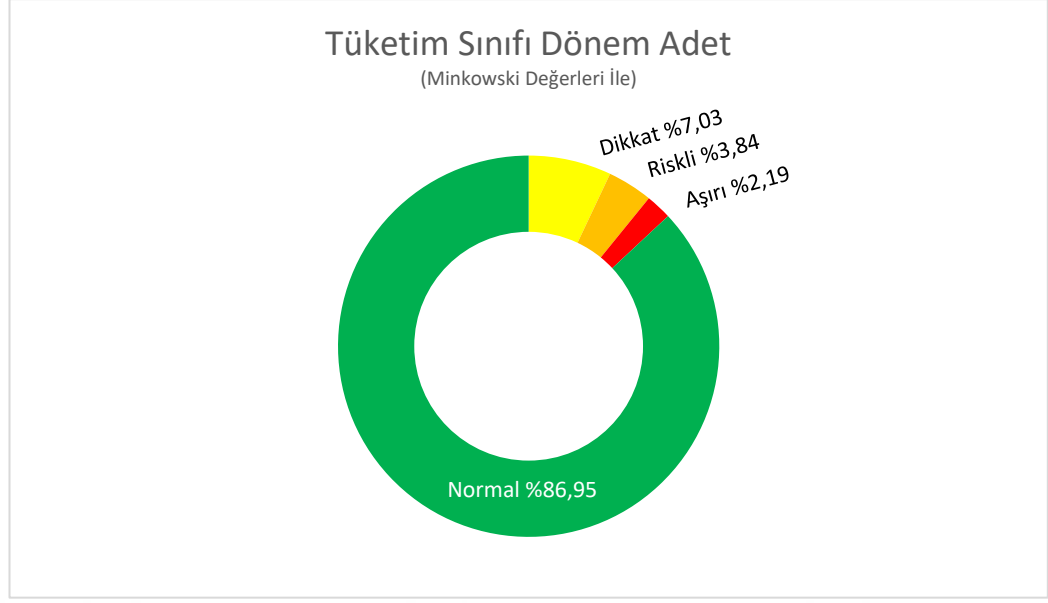
İkinci aşamada hem gerçek hem de minkowski mesafe değerleriyle elde edilen alt verisetler, çalışma kapsamında belirlenen 7 farklı anomali analiz algoritması ile analiz edilmiş, anormal tüketim tespitleri, anomali puanlamaları ve içmesuyu tüketim sınıflandırmaları yapılmıştır.

Toplam 1.325.969 adet içme suyu tüketim gerçek değerleri kullanılarak yapılan analizde, Şekil 5.2’de gösterildiği gibi 1.196.181 adet normal tüketim, 129.788 adet de anormal tüketim tespit edilmiş olup, bütün tüketimlerde %9,79 oranında bir anomali durum olduğu anlaşılmıştır. Anomali tespit yöntemlerindeki eşik değeri ile bu oran yukarı veya aşağı yönlü değiştirilebilir. Anormal tüketimler de kendi içinde “Dikkat” etiketli tüketim sayısı 62.239 (anormal tüketimler içinde %47,95), “Riskli” etiketli tüketim sayısı 30.418 (anormal tüketimler içinde %23,44), “Aşırı” etiketli tüketim sayısı 37.131 (anormal tüketimler içinde %28,61) adet olarak tespit edilmiştir.



Şekil 5.2. Abonelerin gerçek tüketim değerleriyle tespit edilen normal ve anormal tüketim sınıfları grafiği.

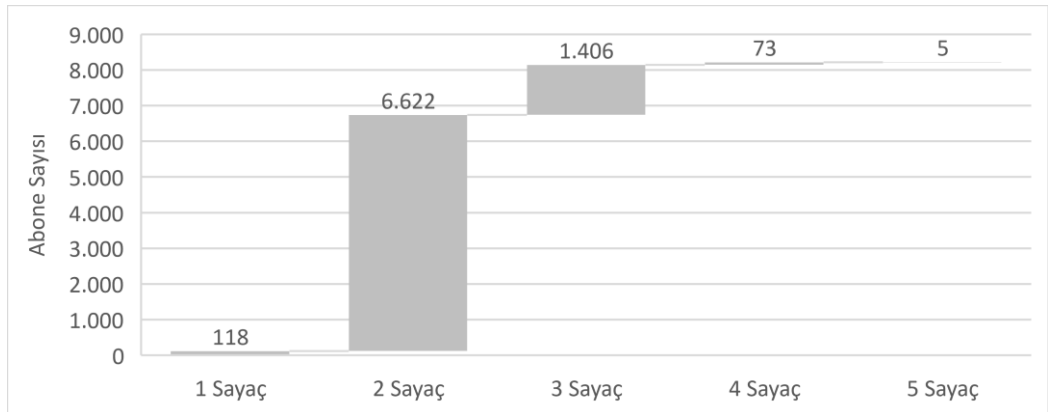
Toplam 1.325.969 adet içme suyu tüketiminin minkowski mesafe değerlerine dönüştürülerek yapılan analizinde, Şekil 5.3’te gösterildiği gibi 1.152.927 adet normal tüketim, 173.042 adet de anormal tüketim olarak tespit edilmiş olup %13,05 oranında bir anomali durum söz konusu olduğu anlaşılmıştır. Ayrıca anormal tüketimlerde kendi içinde “Dikkat” etiketli tüketim sayısı 93.177 (anormal tüketimler içinde %53), “Riskli” etiketli tüketim sayısı 50.856 (anormal tüketimler içinde %29,39), “Aşırı” etiketli tüketim sayısı 29.009 (anormal tüketimler içinde %16,76) adet olarak tespit edilmiştir.



Şekil 5.3. Abonelerin minkowski tüketim değerleriyle tespit edilen normal ve anormal tüketim sınıfları grafiği.

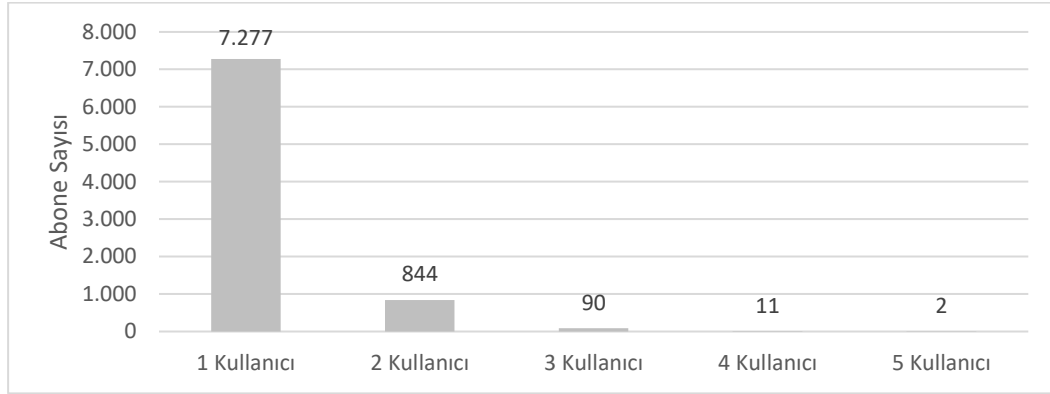
Çalışmada gerçek değerler ve minkowski mesafe değerleri ile yapılan anormal tüketim tespitlerinde, aynı anomali tespit yöntemleri ve eşik değerleri kullanılmasına rağmen farklı sonuçlar elde edilmiştir. Minkowski değerleriyle toplamda %3,26'lık bir oranda daha fazla anormal tüketim tespiti olmasına rağmen, “Aşırı” tüketim sınıf tespitinde %0,61 oranında bir düşüşün olması minkowski değerleri kullanılarak elde edilen tespitlerin gerçek değerlerle tespit edilenlerden daha iyi sonuçlar verdiği şeklinde yorumlanmıştır.

Ayrıca abonelerin en fazla 167 en az 160 dönem sayaç okuması yapılmış olup Şekil 5.4'te gösterildiği gibi %80,52'sinin tüketim süresince 2. sayacı kullandıkları tespit edilmiştir.



Şekil 5.4. Tüketim süresince abonelerin kullandıkları sayaç durum grafiği.

Ayrıca içme suyu abonelerinin Şekil 5.5'te de gösterildiği gibi %88,48'inin tüketimlerinin aynı kullanıcı tarafından yapıldığı tespit edilmiştir.



Şekil 5.5. Tüketim süresince sözleşme yapan kullanıcı durum grafiği.

Abone tüketimlerinin tüketim süresince Tablo 5.4'te gösterildiği gibi yalnızca %0,24'ünün kademeye girdiği, %0,55'inin de ödenmediği tespit edilmiştir.

Tablo 5.4. Sayaç, kullanıcı, kademe ve ödeme durumları özet tablosu.

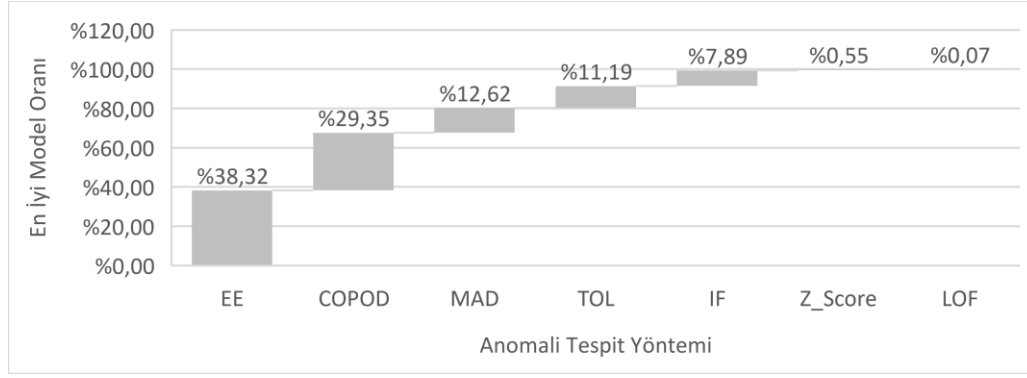
	Kademesiz	Kademeli	Ödenen	Ödenmeyen	Toplam
Abone Sayısı	6.833	1.391	8.224	3.189	8.224
Okuma Sayısı	1.322.781	3.188	1.318.707	7.262	1.325.969

Çalışmada her bir abone için toplamda 112 adet anormal tüketim tespit modeli geliştirilmiş; doğruluk, duyarlılık, kesinlik, hassasiyet, özgüllük, MAE ve MSE performans metrikleriyle modeller karşılaştırılmıştır. Toplamda Tablo 5.5'te gösterildiği gibi 921.088 adet model geliştirilmiş, bunlardan 16.077 adeti (%1,75) en iyi model olarak maksimum doğruluk performansı göstermiştir.

Tablo 5.5. Model kapsamında geliştirilen anomali tespit model sayıları tablosu.

	Abone Sayısı	Veriseti Sayısı	Anomali Algoritma Sayısı	Anomali Tespit Model Sayısı
Gerçek değerler ile	8.224	8	7	460.544
Minkowski değerleri ile	8.224	8	7	460.544
Toplam				921.088

Modellerin performans doğruluk oranları en düşük %43, en yüksek %100 ve ortalama %85 olarak tespit edilmiş olup Şekil 5.6'da gösterildiği gibi EE yöntemiyle elde edilen modellerin çoğunlukla en iyi modeller olduğu, Z Skor ve LOF yöntemleri ile elde edilen modellerin çok düşük oranlarda maksimum doğruluk sağladığı anlaşılmıştır.



Şekil 5.6. Maksimum doğruluk sağlayan anomali analiz yöntem oranları grafiği.

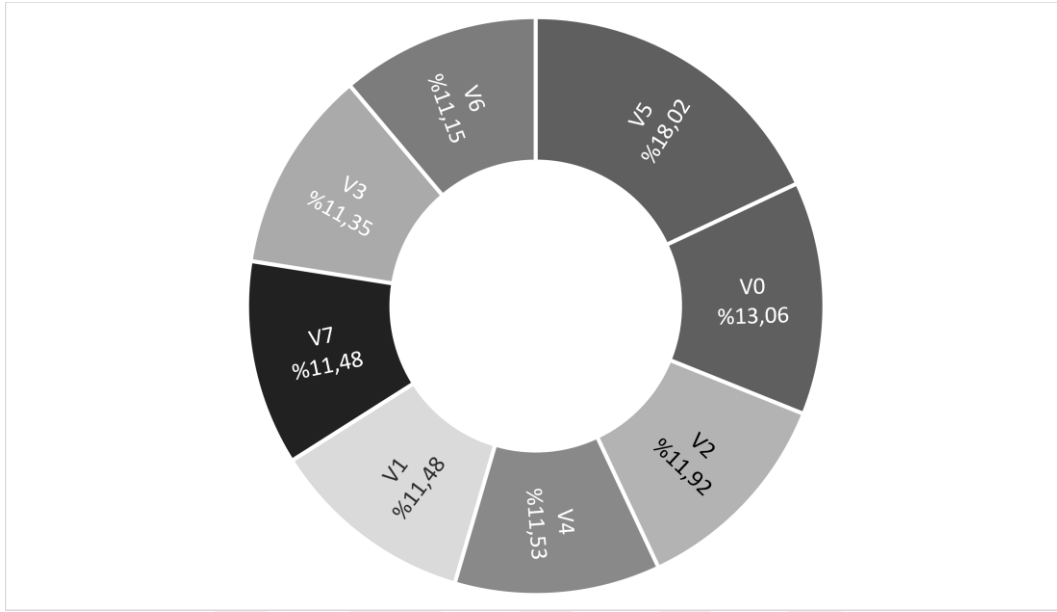
Her bir abone için geliştirilen anormal tüketim tespit modellerinden doğruluk performans metriği %90 ve üzeri olanlar Tablo 5.6’da da gösterildiği gibi sıralanır ve doğruluk oranı en yüksek olanı veya kullanılan anomali analiz yöntemine göre veya kullanılan alt verisetine göre idare tarafından belirlenebilir. Her bir abone için ayrı ayrı olmak üzere belirlenen en iyi anormal tüketim tespit modeli ile bir sonraki okuma döneminde anormal tüketim analizinin yapılması daha az işlem ile ve daha kısa sürede mümkün olacaktır.

Tablo 5.6. Her bir aboneye ait doğruluk performans metriği maksimum olan en iyi model tablosu.

No	Abone No	En iyi Model	Doğruluk
1	12000541	c_V5_COPOD	0,993902439
2	12000950	g_V1_EE	0,993902439
3	10003387	c_V4_COPOD	0,993865031
4	10004150	c_V2_IF	0,993865031
5	4002847	c_V7_COPOD	0,99375
...	...	...	...
8220	12001129	g_V3_IF	0,975308642
8221	4011157	c_V6_Z_score	0,962962963
8222	6009693	g_V1_EE	0,95
8223	216960	c_V2_TOL	0,9
8224	217139	c_V4_TOL	0,9

Çalışmada elde edilen bütün tüketim sınıfı tahmin modellerinden, doğruluk performans metriği %90 ve üzeri olanlar dikkate alındığında, Şekil 5.7’de de gösterildiği üzere örneklem olarak alınan 8.224 abonenin 7.150 tanesi (%86,94) alt verisetlerinden (V1, V2, V3, V4, V5, V6, V7), diğer 1.074 adet abonenin (%13,06) ise temel verisetinden (V0) elde edilebilmektedir. Her bir aboneye ait geliştirilen maksimum doğruluk performanslı model en fazla oranda f skor öznelilik seçim

yöntemi ile elde edilen V5 alt veriseti kullanılarak elde edilmiştir. Diğer alt verisetleri ile de birbirine yakın oranlarda başarılı sonuçlar elde edildiği anlaşılmıştır.



Şekil 5.7. Doğruluk performansı %90 ve üzeri olan tahmin modellerinde kullanılan veriseti oranları grafiği.

Burada işlem sayısı ve süresi azaltılmak istendiğinde her bir abone için alt verisetlerinden en az öznelikli olanı tercih edilebilir.

Model geliştirme sürecinde işlemci hızı 1.1 GHz, RAM kapasitesi 4GB olan bir PC kullanılarak her bir abone için model ön işlemleri ortalama 10 sn, anomali analizleri ortalama 8 sn ve performans ölçüm ve karşılaştırma işlemleri ortalama 3 sn olmak üzere toplamda yaklaşık 96 saat sürmüştür. Çalışma kapsamında analizi yapılan 8.224 abonenin belirlenmiş en iyi modeli kullanılarak anomali tespiti yapılmasında bu süre yaklaşık olarak 51 dk sürmektedir. Böylece önerilen yöntem ile 4 günlük bir analiz ortalama 1 saatten az bir zamanda yapılabilmektedir. Genellikle bütün sular idarelerinde mevcut olan profesyonel donanım altyapısının kullanılması ile çalışma kapsamında önerilen analiz işlem sürelerinin daha da kısaltılması mümkün olacaktır.

Çalışmanın temel amacı olan abonelerin duyarlı hale getirilmesi ile tüketim davranışlarında oluşan anormal tüketimlerinin azaltılması, tüketimlerin normal seviyelere çekilmesi durumunda, tespit edilen ‘Aşırı’ sınıftaki tüketimlerde %20 tasarruf ile 139.427 m³, ‘Riskli’ sınıftaki tüketimlerden %15 tasarruf ile 154.058 m³ ve ‘Dikkat’ sınıftaki tüketimlerden de %10 tasarruf ile 169.402 m³ olmak üzere

toplamda sistemdeki %2,58'lik 462.887 m³ suyun daha verimli kullanımının sağlanacağı öngörülmektedir.

Abone faturalarında ise 'Aşırı' sınıftaki tüketimlerde %20 tasarruf ile 284.304,37 ₺, 'Riskli' sınıftaki tüketimlerden %15 tasarruf ile 288.181,09 ₺ ve 'Dikkat' sınıftaki tüketimlerden de %10 tasarruf ile 349.438,92 ₺ olmak üzere içme suyu fatura toplamalarında %5,15'lik 921.924,38 ₺ tasarrufun sağlanacağı öngörülmektedir.

5.2. ML Analiz Sonuçları

Çalışma sonunda tüketim sınıf etiketlemesi ile gözetimli hale getirilmiş veriseti, DT, NB, KNN, LJR, MLP-NN, RF ve GB olmak üzere 7 farklı ML sınıflandırma algoritmaları kullanılarak içme suyu tüketim sınıfı tahmin modelleri geliştirilmiştir. Tüketim verileri TS verilerden oluştuğu için ilk yıllar eğitim, son yıllar da kontrol verisi olmak üzere %90'a, %10 oranlarında bölünerek ve bölünmeden kullanılmıştır. Geliştirilen modeller Tablo 5.7'de gösterildiği gibi ACC ve R² performans metrikleriyle karşılaştırılmış, her bir içme suyu abonesi için en iyi model seçilmiştir.

Tablo 5.7 İçme suyu tüketim sınıf tahmin modelleri performans tablosu.

ML Model	Veriseti Bölünmeden				%90 Eğitim - %10 Test			
	Minkowski Değerler		Gerçek Değerler		Minkowski Değerler		Gerçek Değerler	
	ACC	R ²	ACC	R ²	ACC	R ²	ACC	R ²
1 DT	1,00	1,00	0,99	0,73	1,00	1,00	1,00	1,00
2 Gaussian NB	0,94	-0,36*	0,69	-2,88*	1,00	1,00	0,69	0,00
3 KNN	1,00	1,00	0,99	0,73	1,00	1,00	0,94	0,00
4 LJR	0,87	0,00	0,86	-0,27*	1,00	1,00	1,00	1,00
5 MLP-NN	0,94	-0,52*	0,89	-0,97*	1,00	1,00	1,00	1,00
6 RF	1,00	1,00	0,99	0,73	1,00	1,00	1,00	1,00
7 GB	1,00	1,00	0,99	0,73	1,00	1,00	1,00	1,00

*R² değerinin negatif olması modelin çok daha kötü olduğunu göstermektedir.

Minkowski mesafe değerleriyle elde edilen modeller, gerçek değerler kullanılarak geliştirilen modellerden daha iyi performans göstermiştir. Modeller, kurumun isteği doğrultusunda veya belli periyotlarda otomatik olarak yeni veriler de dikkate alınarak güncellenebilir ve böylece daha iyi performans sonuçları elde edilebilir.

6. SONUÇ VE ÖNERİLER

Çalışma sonucunda abonelerin sayaç okuma dönemlerine ait abone, sayaç, sayaç okuma, tarife, tahakkuk ve tahsilat verilerinden tüketim davranış modelleri geliştirilmiş ve tüketim sınıflandırma etiketlenmesi yapılmıştır. Her bir aboneye özel olarak geliştirilen tahmin modeli ile her okuma döneminde sayaç okuma esnasında gerçekleşen tüketimin sınıfı belirlenmesi sağlanmıştır.

İçmesuyu abone tüketimlerinden anormal tüketimlerin tespit edilmesi ve tüketim sınıflarının belirlenmesi ile; tüketim alışkanlıklarının su israfına neden olacak düzeyde değişmiş olabileceği, iç tesisatta bir sızıntı olabileceği, kullanılan içme suyu sayacının ölçüm hassasiyetini kaybetmiş olabileceği, ölçüm dışı tüketim veya kaçak kullanımın söz konusu olabileceği, sayaç okuma memurunun endeksi yanlış girmiş veya sayacı yanlış okumuş olabileceği, sayaç endeks okuma döneminin ay aşırı geç veya daha kısa sürelerde yapılmış olabileceği, bahçe, peyzaj veya tarım amaçlı sulama yapılıyor olabileceği, hanenin boş veya tatile çıkmış olabileceği, hane halkı sayısının doğum, ölüm vb. nedenlerle artmış veya azalmış olabileceği, iklim şartlarının klima kullanımı vb. nedenler ile değişmiş olabileceği değerlendirmelerinin daha sağlıklı yapılabilmesi öngörülmektedir.

Ayrıca çalışmada önerilen yöntemle sular idaresinin her okuma döneminde, anormal tüketim analiz raporu ve anormal içmesuyu tüketimi yapan abone listesi oluşturabileceği, sahada yapılan sayaç arıza, fiziki kayıp kaçak gibi içme suyu denetimlerinde bu listelerden istifade edebileceği öngörülmektedir. Böylece denetimlerde genelleme yapılmayarak daha özenli ve itinalı davranılacağı, anomali tespitinin kontrolünün ve geri dönüşlerinin de sağlanmış olacağı öngörülmektedir. Çalışma ile analizi yapılan abonelerin, yaklaşık 16 yıllık tüketim verilerinden, “Aşırı” sınıflı 27.463 adet tüketim dikkate alındığında yaklaşık olarak aylık 137 adet kontrolün gerektiği, günlük 25 adet kontrol yapan bir ekip ile 1 haftaya yakın bir zamanda kontrol edilmesinin mümkün olduğu öngörülmektedir. Bütün abonelerin çalışmada önerilen yöntemle ait analizlerinin yapılmasıyla daha bütüncül bir kontrol listesi elde edilebilir, ekip sayısı ve anormal tüketim sayısı dikkate alınarak bir planlama yapılabilir. Ayrıca

istenirse ekip sayısına ve ekiplerin kontrollerini yapabilecekleri anomali sayısına göre de anomali tespit yöntemlerinin eşik değerleri yukarı veya aşağı yönlü ayarlanarak planlama yapılabilir.

Abonelerin sayaç okumalarının yapılamaması veya okuma dönem sürelerinin ay aşırı günlerde veya daha kısa sürelerde yapılmasının önüne geçilerek, sayaç okumalarının her bir abone için kayıp dönem olmadan her ayın aynı gününde yapılmasının tüketim tahmin model performansını arttıracığı, ayrıca mevsimsel şartların dolayısıyla meteorolojik verilerin hane içindeki tüketimlere etkisi gerçeğinin tek başına yeterli olmayacağı, hanelerde klima gibi hane içindeki iklim şartlarını değiştiren sistemlerin varlığına ait bilgilerin de dikkate alınmasının faydalı olacağı öngörülmektedir.

Ayrıca önerilen yöntem kullanılarak, sayaçların içeride olması ve kullanıcıların evde olmaması durumlarında ya da sayaç arıza dönemlerinde yapılan kıyas tüketimlerin, her aboneye özel olarak tahmin edilmesi veya belirlenmesi daha adilane, daha spesifik ve daha bilimsel bir yaklaşım sağlayacaktır.

Ayrıca çalışmada abone devri, kapama ve açma işlemlerinden dolayı tüketim davranış modellerinin yenilenmesinin uygun olacağı, çalışmada tüketim gün sayısının, kullanıcı değişiminin, sayaç değişiminin, ödeme durumunun dikkate alındığı ancak başkaca parametrelerin de etkili olabileceği gerektiğinde önerilen modele bunlarında eklenebileceği önerilmektedir.

Anomali analizleri sonucunda yapılan tespitler bir öngörü vermektedir ve gerçekliği ancak saha çalışmaları sonucunda gelen geri dönütler ile anlaşılacaktır. Dolayısıyla hangi anomali tespit yöntemi kullanılırsa kullanılsın, abone tüketimlerinin sınıflarının öngörülmesinde tam bir başarı sağlanması arzu edilen bir sonuçtur, fakat tam bir başarı sağlanmazsa dahi önemli olan başarı oranını artıracak adımlar atmak, etkili olduğu anlaşılan başkaca veriler varsa bunları temin etmek ve geri dönütlerle gelen hataları düzelterek ilerlemek olmalıdır.

Ayrıca sular idarelerinin çoğunda yapılan şebeke basınç denetimlerinden elde edilen kayıp kaçak lokasyonları ile bu çalışmadan elde edilen anormal tüketim yapan abonelere ait lokasyonlar eşleştirilerek kayıp kaçak denetimlerinde daha spesifik çözümler elde edilebilir.

Ayrıca abonelere özel tahmin modelleri geliştirildiğinde, bölgelere ayrılmış içmesuyu şebekesinde de ihtiyaç duyulacak aylık, yıllık içme suyu miktarları daha gerçekçi

olarak tahmin edilebilecektir. Böylece arz yönlü değil talep yönlü bir bakış açısıyla su yöntemi sağlanmış olacaktır. Bu bakış açısıyla özellikle suya ulaşımın çok zor olduğu bölgelerde ve kuraklık dönemlerinde daha etkili analizler yapılmasına yardımcı olacak ve sonuçlar alınmasını sağlayacaktır. Bu çalışmadaki taleplerin disipline edilmesi hedefi bu amaca hizmet etmektedir.

Çalışmada her bir abone için modellerin geliştirilmesi ve en iyi modelin doğruluk performans metriği, işlem adeti ve işlem süreleri dikkate alınarak seçilmesi anlık ihtiyaç duyulacak bir durum olmayıp sayaç okuma işleminin dönemsel yapıyor olması avantajı ile ilgili ay içinde analizin yapılmasının yeterli olacağı, dolayısıyla modellerin hız performanslarından çok, veri kalitesi ve model verimliliğinin öncelenebileceği ifade edilebilir. Sahada sayaç okuması esnasında anlık sınıflandırma yapılması daha önce geliştirilmiş modellerin kullanımıyla sağlanacaktır.

Tatil günleri, pandemi gibi etkenlerin sebep olduğu olağanüstü yaşam koşulları tüketimi dönemsel etkileyip davranış modellerini yanıltabileceği öngörülmektedir. Bu husus başka bir çalışmanın konusu olarak da değerlendirilebilir.

Çalışma sonucunda içmesuyu abonelerine ait anormal tüketim dönem sayısının toplam tüketim dönem sayısındaki oranının yaklaşık %2 olduğu düşünülürse verisetinin dengesiz bir yapıda olduğundan bahsedilebilir. Çalışmanın devamı ve tamamlayıcısı olarak tüketim sınıfı tahmin modellerinin geliştirilmesi düşünüldüğünde, elde edilen gözetimli verisetlerinin KNN algoritması temelli SMOTE gibi veri çoğaltma algoritmaları veya normal tüketimlerin anormal tüketim değerleri ile değiştirilmesi yöntemleri ile dengeli hale getirilmesinin tahmin modellerinin performanslarını arttıracığı öngörülmektedir. Çalışmada içme suyu tüketim sınıfı tahmin model performansları yeterli olduğu için buna gerek duyulmamıştır. İhtiyaç duyulması halinde normal tüketim değerleri yerine kullanılacak anormal tüketim değerleri, yaz ayları tüketiminin kış aylarında veya kış ayları tüketimlerinin yaz aylarında yapılmış gibi gösterilmesiyle de elde edilebilir.

Çalışma sonunda elde edilen gözetimli veriseti, tahmin modelleri geliştirilmeden önce tekrardan öz nitelik seçim işlemlerine tabi tutulabileceği, yeterli sayıda faktörün seçilebileceği, analiz hızının, verimin ve performans metrik değerlerinin arttırılabileceği öngörülmektedir.

İçme suyu abone sayaçlarının, ölçüm hassasiyetlerinden dolayı en geç 10 yılda bir ya kalibrasyon ayarlarının yapılması ya da değişiminin yapılması gerekmekte olup, çalışma ile yaklaşık 16 yıldır tek sayaç kullanan 118 abonenin sayaç değişimlerinin yapılmasının, benzer olarak sayaç değişim kayıtlarına ulaşamayan diğer bütün abonelerin endeks kontrolü ile 10 yıllık sayaç denetimlerinin yapılmasının daha doğru tüketim ölçüm değerlerini sağlayacağı öngörülmektedir.

Abonelere sayaç okuma esnasında verilen su bildirim makbuzlarıyla birlikte tüketim alışkanlığının normal veya anormal seyrinde olduğu, anormal ise hangi sınıf bir tüketim yaptıkları bilgisinin verilerek su tüketimi duyarlılığının artırılacağı öngörülmektedir. Ayrıca hane içi su tüketimlerinin genelde hangi amaçlarla yapıldığı, nerelerde yapıldığı, ne kadar yapılması gerektiği ve nerelerde bir tasarruf sağlanabileceği bilgilendirmelerinin veya önerilerinin kurum ve aboneler arasında var olan bütün iletişim kanalları aracılığıyla mesajlar, broşürler kısa filmler vb. yöntemler kullanarak yapılmasının da duyarlılığa katkı sağlayacağı öngörülmektedir.

Tespit edilen anormal tüketimlerde vatandaşın bilgilendirilmesi ve duyarlı hale getirilmesi ile tüketimlerin normal tüketim sınıflarına çekildiği düşünüldüğünde öncelikle bu durumun abone sözleşme sahibinin ekonomisine katkı sağlamasının yanında su idarelerinin su talebini karşılamak için alternatif kaynakların sisteme getirilmesi yatırımları ve arayışlarına destekleyici bir çözüm olacağı düşünülmektedir.

Bu çalışmada sadece konutlardaki içmesuyu tüketimi dikkate alınmış ve modeller geliştirilmiştir. Bununla birlikte önerilen yöntem toplu tüketimlerin söz konusu olduğu diğer bütün sektörlerde de kullanılabilir bir altyığa sahiptir. Aynı yöntem ile elektrik, doğalgaz, internet, mobil hatlar, kredi kartları gibi tüketimlerin bir abonelik sistemi üzerinden dönemsel yapıldığı ve kayıt altına alındığı alanlarda her bir abone için tüketim davranış modeli çıkartılabilir, anormal tüketimler tespit edilip belli sınırlarda tüketimlerin yapılması veya tüketimin disipline edilmesi sağlanabilir, tüketici kitlesinde geri bildirimler yapılarak konuyla ilgili duyarlılıklar geliştirilebilir. Ayrıca bu sektörlerde hizmet veren kurumlar AI ile geliştirilen bu sistemi artı bir hizmet olarak kendi mobil uygulamalarında abonelerine sunarak daha fazla müşteri memnuniyeti sağlamış olacaklardır.

KAYNAKLAR

- Abdel-fattah, M. K., Mokhtar, A., & Abdo, A. I. (2021). Application Of Neural Network And Time Series Modeling To Study The Suitability Of Drain Water Quality For Irrigation : A Case Study From Egypt. *Environmental Science and Pollution Research*, 28(1), 898–914.
- Abro, A. A. (2021). Aykırı Değer Belirleme İçin Melez Bir Zeki Sistem Tasarımı Ve Gerçekleştirimi. İçinde *Dr Tezi*. Ege Üniversitesi.
- Adamowski, J., Chan, H. F., Prasher, S. O., Ozga-zielinski, B., & Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, ANN, and wavelet ANN methods for urban water demand forecasting in Montreal , Canada. *Water Resources Research*, 48(W01528), 1–14. <https://doi.org/10.1029/2010WR009945>
- Adamowski, J. F. (2008). Peak Daily Water Demand Forecast Modeling Using ANN in the city of Ottawa, Canada. *Journal Of Water Resources Planning And Management*, 1(MARCH/APRIL), 119–129.
- Adamowski, J., & Karapataki, C. (2010). Comparison of Multivariate Regression and ANN for Peak Urban Water-Demand Forecasting in Cyprus: Evaluation of Different ANN Learning Algorithms. *Journal of Hydrologic Engineering*, 15(1), 729–743. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0000472](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000472)
- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, 16(2), 270–301. <https://doi.org/10.1177/1094428112470848>
- Ajbar, A. H., & Ali, E. (2012). Water Demand Prediction for Touristic Mecca City in Saudi Arabia using Neural Networks. *World Academy of Science, Engineering and Technology International Journal of Geological and Environmental Engineering*, 6(5), 231–235.
- Akdağ, R. (2014). Yapay Sinir Ağları Yöntemiyle Diyarbakır İli Kent Merkezi İçme Suyu Talep Tahmini Uygulaması. İçinde *Dr Tezi*. Dicle Üniversitesi.
- Akdağ, R. (2016). Yapay Sinir Ağları, Destek Vektör Makineleri ve Box-Jenkins Yöntemleriyle Kentsel İçmesuyu Talebi Tahmini ve Karşılaştırmalı Analizi Urban Water Demand Forecasting and Comparative Analysis by Artificial Neural Networks , Support Vector Machines and Box-J. *Business and Economics Research Journal*, 7(1), 123–138. <https://doi.org/10.20409/berj.2016116808>
- Akpınar, K. O., & Özçelik, İ. (2020). Anomaly Detection on EtherCAT Based Water Level Control Automation. (*UBMK'20*) *5th International Conference on Computer Science and Engineering*, 1(978-1-7281-7565-2/20), 79–82.
- Allawi, M. F., Jaafar, O., Hamzah, F. M., Abdullah, S. M. S., & El-shafie, A. (2018). Review on Applications of Artificial Intelligence Methods for Dam And Reservoir-Hydro-Environment Models. *Environmental Science and Pollution Research*, 25(1), 13446–13469. <https://doi.org/https://doi.org/10.1007/s11356->

- Almalawi, A., Yu, X., Tari, Z., Fahad, A., & Khalil, I. (2014). An Unsupervised Anomaly-Based Detection Approach for Integrity Attacks on SCADA Systems. *Computers & Security*, 46(29 July), 94–110. <https://doi.org/10.1016/j.cose.2014.07.005>
- Altunkaynak, A., Özger, M., & Çakmakçı, M. (2005). Water Consumption Prediction of Istanbul City by Using Fuzzy Logic Approach. *Water Resources Management*, 19(1), 641–654. <https://doi.org/10.1007/s11269-005-7371-1>
- Alvisi, S., Franchini, M., & Marinelli, A. (2007). A Short-Term, Pattern-Based Model for Water-Demand Forecasting. *Journal of Hydroinformatics*, 9(1), 39–50. <https://doi.org/10.2166/hydro.2006.016>
- Arslan, B. Y. (2020). Su Dağıtım Sistemlerindeki Kayıpların Hesaplanması Eskişehir İli Örneği. İçinde *YL Tezi*. Eskişehir Teknik Üniversitesi.
- Arup and Sydney Water. (2015). The Future of Urban Water: Scenarios for Urban Water Utilities in 2040. İçinde *Arup 2015*. <https://www.arup.com/perspectives/publications/research/section/the-future-of-urban-water>
- Athanasiadis, I. N., Montes, A. K., Mitkas, P. A., & Mylopoulos, Y. A. (2005). A Hybrid Agent-Based Model for Estimating Residential Water Demand in Thessaloniki, Greece. *SIMULATION The Society for Modeling and Simulation International*, 81(3), 175–187. <https://doi.org/10.1177/0037549705053172>
- Aydoğdu, M., & Fırat, M. (2015). Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods. *Water Resources Management*, 29(5), 1575–1590. <https://doi.org/10.1007/s11269-014-0895-5>
- Başa, Ş., & Kurt, S. (2017). Su Ve Kanalizasyon İdarelerinde Akıllı Su Yönetimi Uygulamaları: Tekirdağ Örneği. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 22(Kayfor 15 Özel Sayısı), 1519–1532. <https://dergipark.org.tr/tr/pub/sduiibfd/706714>
- Bennett, C., Stewart, R. A., & Beal, C. D. (2013). Expert Systems with Applications ANN-based residential water end-use demand forecasting model. *Expert Systems With Applications*, 40(4), 1014–1023. <https://doi.org/10.1016/j.eswa.2012.08.012>
- Billings, R. B., Horrace, W. C., & Agthe, D. E. (1999). State-Space versus Multiple Regression for Forecasting Urban Water Demand. *Journal of Water Resources Planning and Management*, 125(4(240)), 1–5. [https://doi.org/10.1061/\(ASCE\)0733-9496\(1999\)125](https://doi.org/10.1061/(ASCE)0733-9496(1999)125)
- Billings, R. B., & Jones, C. V. (2008). Forecasting Urban Water Demand (Second Edition). İçinde *Kitap - American Water Works Association (Second)*. www.awwa.org
- Bougadis, J., Adamowski, K., & Diduch, R. (2005). Short-Term Municipal Water Demand Forecasting In Ottawa, Ontario, Canada. *Hydrological Processes*, 19(1), 137–148. <https://doi.org/10.1002/hyp.5763>
- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104.

<https://doi.org/10.1145/342009.335388>

- Budak, H. (2018). Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 22(Özel), 21. <https://doi.org/10.19113/sdufbed.01653>
- Bulut, O. (2020). İçme Suyu Şebekelerinde Oluşan Su Kayıp Ve Kaçakları Belirlemek İçin Kullanılan SCADA Sisteminin Değerlendirilmesi: Erzincan İli Örneği. İçinde *YL Tezi*. Erzincan Binali Yıldırım Üniversitesi.
- Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature Selection in Machine Learning: A New Perspective. *Neurocomputing*, 300, 70–79. <https://doi.org/10.1016/j.neucom.2017.11.077>
- Caiado, J. (2008). Munich Personal RePEc Archive Forecasting water consumption in Spain Using Univariate Time Series Models. *MPRA*, 6610(1), 1–10. <https://mpa.ub.uni-muenchen.de/6610/>
- Campbell, H. E., Johnson, R. M., & Larson, E. H. (2004). Prices, devices, people, or rules: The relative effectiveness of policy instruments in water conservation. *Review of Policy Research*, 21(5), 637–662. <https://doi.org/10.1111/j.1541-1338.2004.00099.x>
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection : A Survey. *ACM Computing Surveys*, 41(3), 1–58. <https://doi.org/10.1145/1541880.1541882>
- Cinal, H. (2009). Basınç Yönetimi ile İçmesuyu Şebeke Kayıplarının Azaltılması: Sakarya Örneği. İçinde *YL Tezi*. Sakarya Üniversitesi.
- Cini, J., Mung, A., & Waughray, D. (2014). Global Agenda Council on Water Security 2012-2014. İçinde *World Economic Forum*. <http://www.weforum.org/content/global-agenda-council-water-security-2012-2014>
- Çıtakoğlu, H., & Özeren, Y. (2021). Sakarya Havzası Su Kalitesi Parametrelerinin Yapay Sinir Ağları ile Modellenmesi Sakarya Basin Water Quality Parameters Modeling With Artificial Neural Networks. *Avrupa Bilim ve Teknoloji Dergisi, Özel Sayı*(24), 10–17. <https://doi.org/10.31590/ejosat.898046>
- Clarke, G. P., Kashti, A., McDonald, A., & Williamson, P. (1997). Estimating Small Area Demand for Water: A New Methodology. *J.CIWEM*, 11(1), 186–192.
- Cousineau, D., & Chartier, S. (2010). Outliers Detection and Treatment: A review. *International Journal of Psychological Research*, 3(1), 58–67.
- Cutore, P., Campisano, A., Kapelan, Z., Modica, C., & Savic, D. (2008). Probabilistic Prediction of Urban Water Consumption Using the SCEM-UA Algorithm. *Urban Water Journal*, 5(2), 125–132. <https://doi.org/10.1080/15730620701754434>
- Denis, D. J. (2021). Univariate, Bivariate, and Multivariate Statistics Using R - Quantitative Tools for Data Analysis and Data Science. İçinde *John Wiley & Sons, Inc*. John Wiley & Sons, Inc. <https://doi.org/10.3138/9781442666009-fm>
- Dogo, E. M., Nwulu, N. I., Twala, B., & Aigbavboa, C. (2021). Accessing Imbalance Learning Using Dynamic Selection Approach in Water Quality Anomaly Detection. *Symmetry*, 13, 818.
- Doorn, N. (2020). Artificial Intelligence in The Water Domain: Opportunities For Responsible Use. *Science of the Total Environment*, 755(January), 1–11.

<https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.142561>

- Düğenci, M. (2007). Arı Algoritması'nın Yapay Sinir Ağı Öğrenmesi İçin Kullanımı ve Atıksu Arıtma Tesis Kontrolü Uygulaması. İçinde *Dr Tezi*. Sakarya Üniversitesi.
- Durivage, M. A. (2014). Practical Engineering, Process, and Reliability Statistics. İçinde *American Society for Quality* (C. 47, Sayı 3). <https://doi.org/10.1080/00224065.2015.11918134>
- Froukh, M. L. (2001). Decision-Support System for Domestic Water Demand Forecasting and Management. *Water Resources Management*, 15(1), 363–382.
- Gerger, R., & Aslan, A. (2019). Şanlıurfa İli İçin İçme Suyu Kayıp ve Kaçaklarının Tespiti. *Harran Üniversitesi Mühendislik Dergisi*, 4(2), 26–35. <http://dergipark.gov.tr/humder>
- Ghalekhondabi, I., Ardjmand, E., Ii, W. A. Y., & Weckman, G. R. (2017). Water Demand Forecasting: Review of Soft Computing Methods. *Environ Monit Assess*, 189(313), 1–13. <https://doi.org/10.1007/s10661-017-6030-3>
- Ghiassi, M., Zimbira, D. K., & Saidane, H. (2008). Urban Water Demand Forecasting with a Dynamic ANN Model. *Journal of Water Resources Planning and Management*, 9496(MARCH/APRIL), 138–146. [https://doi.org/10.1061/\(ASCE\)0733-9496\(2008\)134](https://doi.org/10.1061/(ASCE)0733-9496(2008)134)
- Gökçen, A., & Yeşil, B. (2021). Zero Flow Rate Detection of Ultrasonic Water Meter Using Machine Learning Techniques. *Avrupa Bilim ve Teknoloji Dergisi*, 26, 477–481. <https://doi.org/10.31590/EJOSAT.961090>
- Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. İçinde *Dr Tezi* (Sayı April). The University of Waikato.
- Herrera, M., Torgo, L., Izquierdo, J., & Pérez-Garcia, R. (2010). Predictive models for forecasting hourly urban water demand in Monte Carlo. *Journal of Hydrology*, 387(1), 141–150. <https://doi.org/10.1016/j.jhydrol.2010.04.005>
- Hodge, V. J., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2), 85–126. <https://doi.org/10.1023/B:AIRE.0000045502.10941.a9>
- Institute, P. (2023a). *Water Conflict Chronology*. Pacific Institute. <https://www.worldwater.org/conflict/list/>
- Institute, P. (2023b). *World Water Conflict Map*. Pacific Institute. <https://www.worldwater.org/conflict/map/>
- International Water Association. (2017). The IWA principles for water wise cities - for urban stakeholders to develop a shared vision and act towards sustainable urban water in resilient and liveable cities. *Urban Stakeholders to Develop a Shared Vision and Act towards Sustainable Urban Water in Resilient and Liveable Cities*, 2nd, 1–6. https://iwa-network.org/wp-content/uploads/2016/10/IWA_Brochure_Water_Wise_Communities_SCREEN-1.pdf
- Jain, A., & Ormsbee, Li. E. (2002). Short-Term Water Demand Forecast Modeling Techniques—Conventional Methods Versus AI. *American Water Works Association*, 94(7), 64–72. <https://www.jstor.org/stable/41298416>

- Jain, A., Varshney, A. K., & Joshi, U. C. (2001). Short-Term Water Demand Forecast Modelling at IIT Kanpur Using ANN. *Water Resources Management*, 15(1), 299–321.
- Jalal, D., & Ezzedine, T. (2019). Toward a Smart Real Time Monitoring System for Drinking Water Based on Machine Learning. *2019 27th International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2019*, 1–5. <https://doi.org/10.23919/SOFTCOM.2019.8903866>
- Jin, T., Cai, S., Jiang, D., & Liu, J. (2019). A Data-Driven Model For Real-Time Water Quality Prediction And Early Warning by An Integration Method. *Environmental Science and Pollution Research*, 26(1), 30374–30385.
- Kadinski, L., & Ostfeld, A. (2021). Incorporation Of COVID-19-Inspired Behaviour Into Contamination Responses. *Water*, 2863(13), 1–21.
- Kadılar, G. Ö. (2019). Entropi Nedir? Hangi Alanlarda Kullanılır? İçinde *TÜBİTAK 2237-A Bilimsel Eğitim Etkinliklerini Destekleme Programı*. www.kantitatifekoloji.net/algoritma
- Karakuş, C. B., Yıldız, S., & Cerit, O. (2010). Sivas Kent İçme Suyu Şebekesindeki Su Kayıpları ve Kayıp Oranını Azaltma Çalışmaları. *Selçuk Üniversitesi Mühendislik, Bilim Ve Teknoloji Dergisi*, 25(1), 1–10.
- Kızıllöz, B., & Şişman, E. (2021). Gelir Getirmeyen Su Oranı Tahmin Modelleri Prediction Models For Non-Revenue Water Ratio. *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 10(1), 276–283. <https://doi.org/10.28948/ngmuh.789694>
- Körlü, S. (2019). Yeraltı Suyu Seviye Değişiminin Bulanık Mantık Ve Destek Vektör Makineleri Yöntemleri İle Tahmini. İçinde *YL Tezi*. İskenderun Teknik Üniversitesi.
- Kulanuwat, L., Chantrapornchai, C., Maleewong, M., Wongchaisuwat, P., Wimala, S., Sarinnapakorn, K., & Boonya-Aroonnet, S. (2021). Anomaly Detection Using a Sliding Window Technique and Data Imputation with Machine Learning for Hydrological Time Series. *Water*, 13(1862), 1–20. <https://doi.org/10.3390/w13131862>
- Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020). COPOD: Copula-based outlier detection. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2020-Novem(1), 1118–1123. <https://doi.org/10.1109/ICDM50108.2020.00135>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1(ICDM08), 413–422. <https://cs.nju.edu.cn/zhoush/zhoush.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhoush/zhoush.files/publication/icdm08b.pdf?q=isolation-forest>
- Liu, Z., Wan, J., Ma, Y., & Wang, Y. (2019). Online Prediction of Effluent COD in The Anaerobic Wastewater Treatment System Based on PCA-LSSVM Algorithm. *Environmental Science and Pollution Research*, 26(1), 12828–12841. <https://doi.org/https://doi.org/10.1007/s11356-019-04671-8>
- LOF - Unsupervised Outlier Detection using Local Outlier Factor. (2021). İçinde *Çalışma Notları*.
- Lund, N. S. V., Kirstein, J. K., Madsen, H., Mark, O., Mikkelsen, P. S., & Borup, M.

- (2021). Feasibility Of Using Smart Meter Water Consumption Data And In-Sewer Flow Observations For Sewer System Analysis: A Case Study. *Journal of Hydroinformatics*, 23(4), 795–812. <https://doi.org/10.2166/hydro.2021.166>
- March, H., Morote, Á. F., Rico, A. M., & Saurí, D. (2017). Household Smart Water Metering in Spain: Insights from The Experience of Remote Meter Reading in Alicante. *Sustainability (Switzerland)*, 9(4). <https://doi.org/10.3390/su9040582>
- Mayer, P., Dziegielewski, B., Kiefer, J. C., Lantz, G. L., Opitz, E. M., & Porter, G. a. (2000). Commercial and Institutional End Uses of Water. *Içinde Power*.
- Mayer, P. W., DeOreo, W. B., Opitz, E. M., Kiefer, J. C., Davis, W. Y., Dziegielewski, B., & Nelson, J. O. (1999). Residential End Uses of Water Executive Summary. *Içinde AWWA Research Foundation and American Water Works Association*.
- Morán-Fernández, L., Bolón-Canedo, V., & Alonso-Betanzos, A. (2016). *** iptal *** Data Complexity Measures for Analyzing The Effect of SMOTE Over Microarrays. *ESANN 2016 - 24th European Symposium on Artificial Neural Networks, November*, 289–294.
- Novakovic, J. (2009). Using Information Gain Attribute Evaluation to Classify Sonar Targets. *17 thTelecommunucation Forum, November 24-26*, 1351–1354.
- Önen, F., Aslan, B., & Hamidi, N. (2018). Diyarbakır Kenti İçmesuyu İhtiyacının Genetik İfadeli Programlama İle Modellenmesi. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 9(2), 859–870. <https://dergipark.org.tr/tr/pub/dumf/399323>
- Özmen, A., Er, F., Atlas, M., Aslanargün, A., Peker, K. Ö., Şıklar, E., & Sönmez, H. (2012). İstatistik. *Içinde Kitap-Anadolu Üniversitesi*.
- Rousseeuw, J. P. and K. V. D. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3), 212–223.
- Şen, C. (2018). İçmesuyu Tesislerinde SCADA Sistemlerine Sezgisel Yaklaşımların Uygulanması. *Içinde YL Tezi. Sakarya Üniversitesi*.
- Seo, Y., Kwon, S., & Choi, Y. (2018). Short-Term Water Demand Forecasting Model Combining Variational Mode Decomposition and Extreme Learning Machine. *Hydrology*, 5(54), 1–19. <https://doi.org/10.3390/hydrology5040054>
- Tabachnick, B. G., & Fidell, L. S. (2013). Using Multivariate Statistics. *Içinde Pearson Education, Inc. (C. 6)*. <https://doi.org/10.1037/022267>
- Topuz, M. D. (2014). Makine Öğrenmesi Algoritmaları Ve Anomali Tespiti. *Içinde YL Tezi. Bahçeşehir Üniversitesi*.
- Tukey, J. w. (1977). Exploring Data Analysis. *Içinde Addison-Wesley Publishing Company. ADDISON-WESLEY PUBLISHING COMPANY*. <https://doi.org/10.1525/9780520338210>
- Türkiye Cumhuriyeti Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı. (2020). Sürdürülebilir Kalkınma Amaçları ve Göstergeleri. *Içinde Y. H. G. MÜDÜRLÜĞÜ & B. V. B. Y. D. BAŞKANLIĞI (Ed.), Strateji ve Bütçe Başkanlığı Yayınları: C. Haziran 20 (Haziran 20, ss. 1–42). STRATEJİ VE BÜTÇE BAŞKANLIĞI*. <http://www.surdurulebilirkalkinma.gov.tr/wp-content/uploads/2021/02/SKA-ve-Gostergeleri-Kapak-Birlestirilmis.pdf>
- Uddin, M. T., & Uddin, M. A. (2015). A Guided Random Forest based Feature

- Selection Approach for Aactivity Recognition. *2nd International Conference on Electrical Engineering and Information and Communication Technology, iCEEICT 2015, May*, 1–7. <https://doi.org/10.1109/ICEEICT.2015.7307376>
- Yang, T.-M., Fan, S.-K., Fan, C., & Hsu, N.-S. (2014). Establishment Of Turbidity Forecasting Model And Early-Warning System For Source Water Turbidity Management Using Back-Propagation Artificial Neural Network Algorithm And Probability Analysis. *Environ Monit Assess*, 186(1), 4925–4934. <https://doi.org/10.1007/s10661-014-3748-z>
- Yavuz, M. A. (2009). Kaçak Su Kullanımının Tespitinde Veri Madenciliği Yaklaşımı. İçinde *YL Tezi*. SAKARYA ÜNİVERSİTESİ.
- YILDIZ, D., & ÖZGÜLER, H. (2020). Yapay Zeka ve Su Yönetimi. İçinde *Rapor* (C. 30, Sayı 1).
- Yıldız, E. (2011). İçmesuyu Şebekelerinde Yapay Sinir Ağları Kullanılarak Harcama Eğrisi Tahmini ile Performans Tayini ve Kaçakları En Düşük Seviyede Tutmak İçin Optimum Vana Optimizasyonu. İçinde *Dr Tezi* (Sayı August). Middle East Technical University.
- Yousefi, P. (2020). Integrated Management Plan Of Water Distribution Systems: Forecasting Approach. İçinde *Dr Tezi*. The University Of British Columbia.
- Yüksel, İ., Songur, M., & Demirel, İ. H. (2018). Diyarbakır İçme Suyu Şebekesindeki Su Kayıplarını Önlemek İçin Hidrolik Yöntem Geliştirme. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 9(1), 535–542. <https://dergipark.org.tr/tr/pub/dumf/413066>
- Zhang, Q., Li, Z., Snowling, S., Siam, A., & El-Dakhkhni, W. (2019). Predictive Models For Wastewater Flow Forecasting Based on Time Series Analysis And Artificial Neural Network. *Water Science & Technology*, 80(2), 243–253. <https://doi.org/10.2166/wst.2019.263>
- Zhang, R., Nie, F., Li, X., & Wei, X. (2019). Feature Selection with Multi-view Data: A Survey. *Information Fusion*, 50(March 2019), 158–167. <https://doi.org/10.1016/j.inffus.2018.11.019>
- Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L. B., & Pan, B. (2019). Integrating Water Quality and Operation Into Prediction Of Water Production in Drinking Water Treatment Plants by Genetic Algorithm Enhanced Artificial Neural Network. *Water Research*, 164(114888), 1–12. <https://doi.org/10.1016/j.watres.2019.114888>
- Zhu1, S., Hadzima-Nyarko, M., Gao, A., Wang, F., Wu, J., & Wu, S. (2019). Two Hybrid Data-Driven Models For Modeling Water-Air Temperature Relationship in Rivers. *Environmental Science and Pollution Research*, 26(1), 12622–12630.



EKLER

EK A. Literatürde geçen aykırı değer tanımları tablosu (Aguinis vd., 2013).

EK B. Dünyada su ile ilgili yapılan çatışmaların haritası (Institute, 2023b).



EK A

Tablo A.1. Literatürde geçen aykırı değer tanımları (Aguinis ve ark, 2013).

Metod	Grubu	Açıklama
1 Kutu grafiği	Tek katmanlı teknikler	Bir yapının en küçük değerinin (aykırı değerler hariç), alt çeyreğin (Q1), medyanın (Q2), üst çeyreğin (Q3) ve en büyük değerinin (aykırı değerler hariç) bir özetini gösteren bir grafik. Aykırı değerler, grafiğin bıyıklarının ötesinde kalan noktalar olarak tanımlanabilir (yani, aykırı değerler hariç en küçük ve en büyük değerler).
2 Kök yaprak grafiği	ve Tek katmanlı teknikler	Eş zamanlı olarak nicel verileri sıralayan ve bir dağılımın şekli hakkında bilgi sağlayan bir çizim. Çiftlerin geri kalanından önemli ölçüde uzak olan gövde ve yaprak çiftleri, aykırı değerlerin varlığına işaret eder.
3 Şematik grafik analizi	Tek katmanlı teknikler	Bir kutu grafiğine benzer, ancak özellikle bir meta-analiz bağlamında efekt boyutları için kullanılır.
4 Standart sapma analizi	Tek katmanlı teknikler	Bir veri noktasının standart sapma birimleri cinsinden ortalamadan uzaklığı.
5 Yüzde analizi	Tek katmanlı teknikler	Yüzdelik dilimine göre endekslenen puan dağılımındaki bir veri noktasının göreceli konumu
6 Dağılım grafiği	Çok katmanlı ("mesafe") teknikleri	Biri x ekseninde (genellikle bağımsız değişken) ve diğeri y ekseninde (genellikle bağımlı değişken) olmak üzere iki değişkenin değerlerinin grafiği. Potansiyel bir aykırı değer, verilerin merkezinden çok uzakta bulunan bir veri noktası tarafından tanımlanabilir.
7 q-q grafiği	Çok katmanlı ("mesafe") teknikleri	İki olasılık dağılımını, niceliklerini birbirine karşı çizerek karşılaştıran bir çizim (q nicelik anlamına gelir). Doğrusal olmayan bir eğilim, aykırı değer(ler)in olası varlığını gösterir.
8 p-p grafiği	Çok katmanlı ("mesafe") teknikleri	İki kümülatif dağılım fonksiyonunu birbirine karşı çizerek iki veri setinin (genellikle gözlenen ve beklenen) benzerlik derecesini değerlendiren bir çizim (p olasılığı temsil eder). Doğrusal olmayan bir eğilim, aykırı değer(ler)in olası varlığını gösterir.
9 Standartlaştırılmış artık	Çok katmanlı ("mesafe") teknikleri	i. gözlemin artık değerinin bir standart sapma terimine bölünmesiyle hesaplanan artık değer. Yüksek standartlaştırılmış artık değerlere sahip gözlemlerin aykırı olması muhtemeldir. Bununla birlikte, bir gözlemin standartlaştırılmış artık değeri, bir gözlemin öngörücü değişkenler üzerindeki aykırılığını ölçmez.
10 Öğrenci-leştirilmiş artık (Studentized residual)	Çok katmanlı ("mesafe") teknikleri	Hem gözlemin standartlaştırılmış kalıntı değeri (yani bir tür mesafe) açısından uzaklığını hem de gözlemin tahmin değişkenleri (yani başka bir tür mesafe) üzerindeki uzaklığını ölçen artık değer, öyle ki bir veri noktası her iki tür mesafe açısından da uzak olan, standartlaştırılmış artık değerinden daha büyük bir öğrencileştirilmiş artık değere sahip olacaktır. Öğrencileştirilmiş artık değerleri yüksek olan gözlemlerin aykırı olması muhtemeldir.

Tablo A.1. (Devamı) Literatürde geçen aykırı değer tanımları (Aguinis ve ark, 2013).

Metod	Grubu	Açıklama
11 Standart-laştırılmış silinmiş artık (Standardized deleted residual)	Çok katmanlı ("mesafe") teknikleri	Odak gözlemi için tahmin edilen değer, gözlemin kendisi olmadan hesaplanması dışında, standartlaştırılmış bir artık ile aynı olan bir artık değer. Bu hariç tutma, odak gözlemin artık değeri düşürmesini ve standart sapma terimini şişirmesini engeller; bu tür bir deflasyon ve şişirme, gözlemin herhangi bir aykırılığının varlığını maskeleymektedir. Yüksek standartlaştırılmış silinmiş artık değerlere sahip gözlemlerin aykırı olması muhtemeldir.
12 Öğrencileştirilmiş silinmiş artık (yani, harici olarak öğrencileştirilmiş artık, jackknife residual)	Çok katmanlı ("mesafe") teknikleri	Odak gözlemi için tahmin edilen değer, gözlemin kendisi olmadan hesaplanması dışında, öğrenilen artık ile aynı olan bir artık değer. Bu hariç tutma, odak gözlemin artık değeri düşürmesini ve standart sapma terimini şişirmesini engeller; bu tür bir deflasyon ve şişirme, gözlemin herhangi bir aykırılığının varlığını maskeleymektedir. Silinen kalıntı değerleri yüksek olan gözlemlerin aykırı olması muhtemeldir.
13 Öklid mesafesi	Çok katmanlı ("mesafe") teknikleri	Bir, iki veya n boyutlu uzayda belirtilen iki nokta arasındaki doğru parçasının uzunluğu. İki veri noktası arasındaki büyük bir Öklid mesafesi, iki veri noktasından birinin aykırı değer olduğu anlamına gelebilir.
14 Mahalanobis mesafesi	Çok katmanlı ("mesafe") teknikleri	Öklid uzaklığına benzer, ancak bunda farklı olan Mahalanobis mesafesi, bir veri noktası ile merkez noktası arasındaki çizgi parçasının uzunluğudur (başka bir gözlem yerine), merkez merkezin tüm tahmin edicinin araçlarının kesişiminde oluşturulan nokta olduğu değişkenler. Büyük bir Mahalanobis mesafesi, karşılık gelen gözlemin bir aykırı değer olduğu anlamına gelebilir.
15 K-kümeleme (değiştirilmiş şapka matrisi olsun veya olmasın) veya diğer benzer küme analizi teknikleri	Çok katmanlı ("mesafe") teknikleri	Daha sonra bir veya daha fazla çoklu vaka teşhisi ile değerlendirilmesi gereken farklı aday alt kümeleri verir.
16 Orijinalin ve temel bileşen değişkenlerinin 2 veya 3 boyutlu çizimleri	Çok katmanlı ("mesafe") teknikleri	PCA'nın bir sonucu olarak üretilen iki veya üç boyutlu değişken grafiği. Yalıtılmış bir veri noktası, olası bir aykırı değeri belirtir.
17 Otokorelasyon fonksiyonu grafiği	Çok katmanlı ("mesafe") teknikleri	Değişken zaman gecikmelerinde veri değerleri için otokorelasyonların hesaplanmasıyla oluşturulan bir çizim. Potansiyel aykırı değerler, diğer veri noktalarından uzakta bulunan veri noktaları tarafından tanımlanabilir.
18 Zaman planı	Çok katmanlı ("mesafe") teknikleri	Belirli bir değişken ve zaman arasındaki ilişkinin grafiği. Potansiyel aykırı değerler, diğer veri noktalarından uzakta bulunan veri noktaları tarafından tanımlanabilir.
19 Aşırı öğrencileştirilmiş sapma (yani Grubbs yöntemi)	Çok katmanlı ("mesafe") teknikleri	Bir değişkenin ortalama ve sorgu değeri arasındaki farkın standart sapma değerine bölümü.

Tablo A.1. (Devamı) Literatürde geçen aykırı değer tanımları (Aguinis ve ark, 2013).

	Metod	Grubu	Açıklama
18	Zaman planı	Çok katmanlı ("mesafe") teknikleri	Belirli bir değişken ve zaman arasındaki ilişkinin grafiği. Potansiyel aykırı değerler, diğer veri noktalarından uzakta bulunan veri noktaları tarafından tanımlanabilir.
18	Zaman planı	Çok katmanlı ("mesafe") teknikleri	Belirli bir değişken ve zaman arasındaki ilişkinin grafiği. Potansiyel aykırı değerler, diğer veri noktalarından uzakta bulunan veri noktaları tarafından tanımlanabilir.
19	Aşırı öğrencileştirilmiş sapma (yani Grubbs yöntemi)	Çok katmanlı ("mesafe") teknikleri	Bir değişkenin ortalama ve sorgu değeri arasındaki farkın standart sapma değerine bölümü.
20	Hosmer Lemeshow uyum iyiliği testi	Çok katmanlı ("mesafe") teknikleri	Gözlenen ve beklenen (yani, tahmin edilen) frekanslar tablosundan bir Pearson ki-kare istatistiği.
21	Kaldıraç değerleri (Leverage Values)	Çok katmanlı ("mesafe") teknikleri	Hat (Projeksiyon veya izdüşüm) matrisinin köşegen elemanları olarak da bilinen kaldıraç değerleri, tahmin ediciler alanında gözlemlerin ne ölçüde aykırı olduğunu ölçer.
22	Merkezlenmiş kaldıraç değerleri	Çok katmanlı ("mesafe") teknikleri	Merkezlenmiş kaldıraç değerleri endeksi. Belirli istatistiksel paketler (örneğin, SPSS), normal kaldıraç değerleri yerine merkezli kaldıraç değerleri bildirir.
23	Standartlaştırılmış çok değişkenli kalıntının silinmesi	Çok katmanlı ("mesafe") teknikleri	Çok seviyeli modelleme bağlamında standartlaştırılmış artık terim. Bu, daha yüksek seviyede bir aykırı değer model uyumu üzerindeki etkisinin değerlendirilmesine izin verir. Daha yüksek seviyede bir aykırı değer bulunursa, daha düşük seviyeli birimler araştırılmalıdır.
24	Aşçının Yemeği	Etkime teknikleri	Bir veri noktasının bir bütün olarak tüm regresyon katsayıları üzerindeki etkisini değerlendirir.
25	Modifiye Cook's Di	Etkime teknikleri	Cook's Di'ye benzer, ancak standartlaştırılmış artıklar yerine standartlaştırılmış silinmiş artıkları kullanır.
26	Genelleştirilmiş Cook's Di	Etkime teknikleri	Cook's Di'ye benzer, ancak bir veri noktasının parametre tahminleri üzerindeki etkisini değerlendirmek için yapısal eşitlik modellemesine uygulanır.
27	Uyum farkı, standartlaştırılmış (DFFITSi)	Etkime teknikleri	Tıpkı Cook's Di gibi, bu teknik de bir veri noktasının bir bütün olarak tüm regresyon katsayıları üzerindeki etkisini değerlendirir. İki teknik arasındaki büyük fark, farklı ölçeklerde var olan bilgileri üretmeleridir.
28	Beta farkı, standartlaştırılmış (DFBETAS _{ij})	Etkime teknikleri	Bir i durumunun dahil edilmesinin, tek bir j regresyon katsayısında (yani bir eğim veya kesişim) bir artışa veya azalmaya yol açıp açmadığını gösterir.

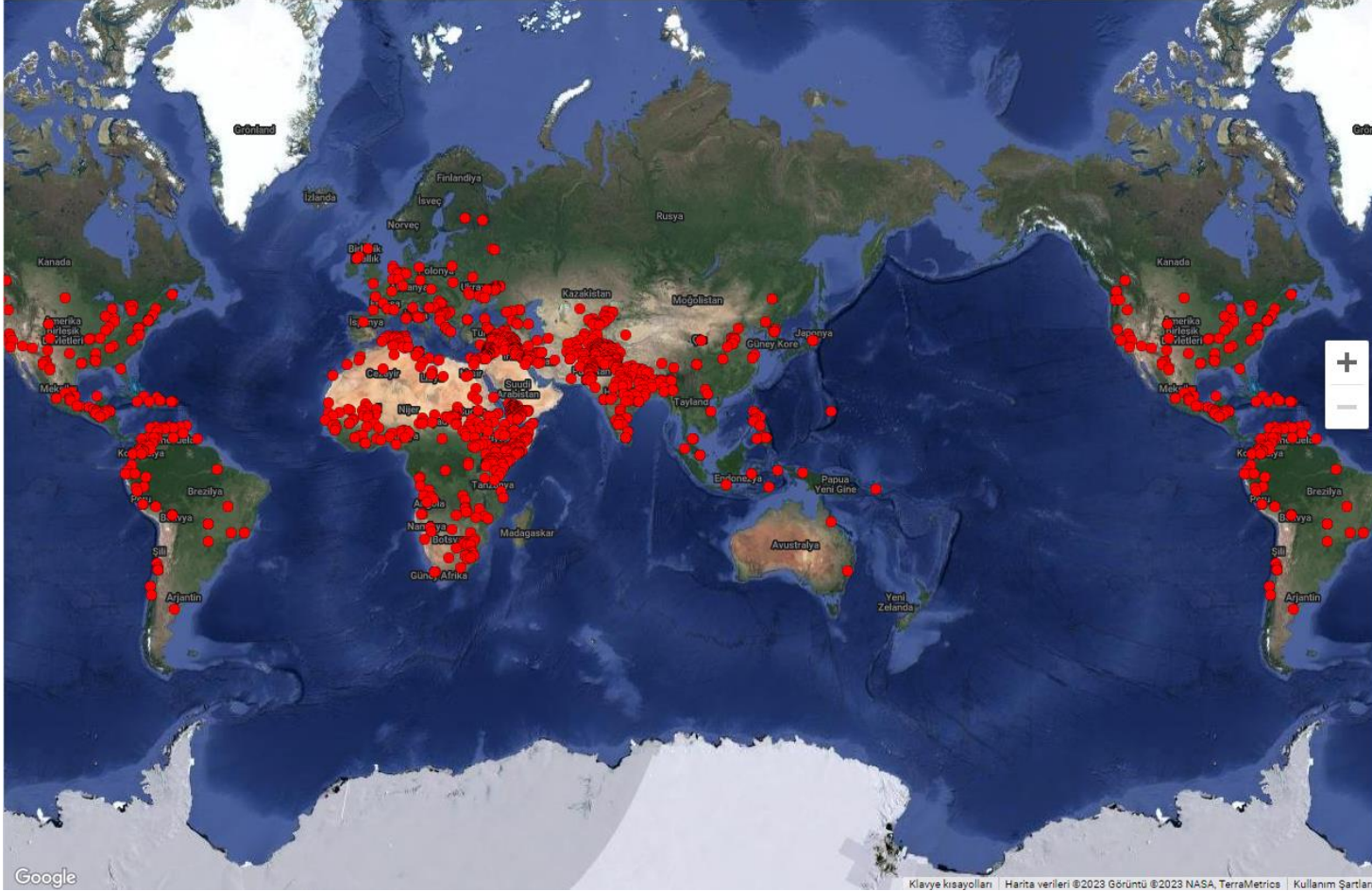
Tablo A.1. (Devamı) Literatürde geçen aykırı değer tanımları (Aguinis ve ark, 2013).

Metod			Grubu	Açıklama
29	Ki-kare testi	fark	Etkime teknikleri	Bu yöntem, yapısal eşitlik modellemesi (SEM) yürüten bir araştırmacının, aykırı değer içeren ve aykırı değer içermeyen iki model arasındaki farkı değerlendirmesini sağlar.
30	Tek parametre etkisi		Etkime teknikleri	DFBETAS _{ij} 'e benzer şekilde, bu tekniği SEM'de, bir aykırı değer modeldeki tüm parametre tahminleri üzerindeki genel etkisinin aksine, bir aykırı değer belirli bir parametre tahmini üzerindeki etkisini değerlendirmek için kullanılır.
31	Ortalama kare sapma tekniği		Etkime teknikleri	Çok seviyeli modelleme yapılırken, Cook'sDi'nin doğrudan bir analogu olan bu yöntem, her grubun sabit ve/veya rastgele parametreler üzerindeki etkisini araştırır ve daha yüksek seviyeli tahmin aykırı değerlerinin tanımlanmasına izin verir. Daha yüksek seviyede bir aykırı değer bulunursa, daha düşük seviyeli birimler araştırılmalıdır.
32	Örnek meta-analitik sapma (SAMD)	ayarlı	Etkime teknikleri	Meta-analizde, bu test istatistiği, her bir birincil seviye etki büyüklüğü tahmininin değeri ile bu etki büyüklüğü olmadan hesaplanan ortalama örnek ağırlıklı katsayı arasındaki farkı alır ve bu etki büyüklüğü olmadan hesaplanan ortalama örnek ağırlıklı katsayı ve ardından fark değerini birincil düzey çalışmanın örneklem büyüklüğüne göre ayarlar. Aykırı değerler, uç SAMD değerleriyle tanımlanır.
33	Aykırı değer içeren, içermeyen analiz gerçekleştirme		Etkime teknikleri	Bu teknik, belirli bir veri noktası olan ve olmayan istatistiksel analizin yürütülmesini ifade eder. Sonuçlar iki analiz arasında farklılık gösteriyorsa, veri noktası bir aykırı değer olarak tanımlanır.
34	En yakın komşu teknikleri	yakın	Etkime teknikleri	Öklid, Mahalanobis gibi çeşitli mesafe metrikleri ile sorgu değerine en yakın değer hesaplanması. Teknikler arasında KNN, optimize edilmiş en yakın komşu, verimli Tip 1 en yakın komşu, Tip 2 en yakın komşu, azaltılmış özelliklere sahip en yakın komşu, ejderha yöntemi, PAM (medoidler etrafında bölümlendirme), CLARANS (rastgele aramaya dayalı büyük uygulamaları kümeleme) ve grafik bağlantı yöntemi.
35	Parametrik olmayan yöntemler		Etkime teknikleri	Veriler hakkında herhangi bir kısıtlayıcı varsayımda bulunmadan düzleştirilmiş bir eğri uydurmaktan oluşur. İlişkide doğrusal olmayan bir eğilim, aykırı değerlerin varlığına işaret eder.
36	Parametrik yöntemler		Etkime teknikleri	Parametrik olmayan yöntemlerden farklı olarak, verilerin doğası hakkında belirli varsayımlarda bulunur. Böyle bir varsayım, verilerin belirli bir olasılık dağılımı türünden (örneğin, normal dağılım) gelmesidir. Aykırı değerler, bu tekniklerle, verilerin doğasına ilişkin beklentilerin dışında kalan veri noktaları olarak tanımlanır. Parametrik yöntemler, dışbükey soyma, elipsoid soyma, yinelenmeli silme, yinelenmeli kırpma, derinlik kırpma, en küçük kareler ortancası, en az kırpılmış kareler ve M-tahmini içerir.

Tablo A.1. (Devamı) Literatürde geçen aykırı değer tanımları (Aguinis ve ark, 2013).

	Metod	Grubu	Açıklama
37	Semiparametrik yöntemler	Etkime teknikleri	Bu yöntemler, tek global dağıtım modeli yerine yerel kümeleri veya çekirdekleri araştırmak için parametrik yöntemlerin hızı ve karmaşıklığını parametrik olmayan yöntemlerin esnekliğiyle birleştirir. Aykırı değerler, düşük yoğunluklu bölgelerde bulunuyor olarak tanımlanır.
38	Yinelemeli aykırı değer tanımlama prosedürü	Etkime teknikleri	Bir dizi adımda, bu prosedür, bir TS analizi yapılırken kullanılan tahmin prosedürüne duyarlı veri noktalarını belirlemek için artık standart sapmanın tahminine izin verir. Bu tür veri noktaları daha sonra aykırı değerler olarak tanımlanır.
39	Bağımsız bileşen analizi (PCA)	Etkime teknikleri	Aralarındaki istatistiksel bağımsızlığı en üst düzeye çıkararak bağımsız bileşenleri ayırmak için kullanılan bir hesaplama yöntemi. Ayrı bağımsız bileşenler, bir TS analizinde bulunduğu aykırı değerler olarak tanımlanır.

EK B



Şekil B.1. Tarihte su ile ilgili yapılmış çatışmaların dünya haritasında gösterimi. (Institute,2023b)

ÖZGEÇMİŞ

Ad-Soyad : İsmail GÜNEY

ÖĞRENİM DURUMU:

- **Lisans** : 2003, Sakarya Üniversitesi / Mühendislik Fakültesi
/ Bilgisayar Mühendisliği
- **Yükseklisans** : Devam ediyor, Sakarya Üniversitesi / Fen Bilimleri Enstitüsü
/ Bilişim Sistemleri Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

- 2019-Halen Maski Genel Müdürlüğü'nde Bilgisayar Mühendisi olarak çalışıyor.
- 2014-2019 yılları arasında Maski Genel Müdürlüğü Abone İşleri Dairesi Başkanı olarak görev yaptı.
- 2013-2014 yılları arasında Malatya Büyükşehir Belediyesi Strateji Geliştirme Müdürü olarak görev yaptı.
- 2009-2013 yılları arasında Malatya Büyükşehir Belediyesi Malatya Kent Bilgi Sistemi (MAKBİS) Proje Koordinatörü olarak görev yaptı.
- 2008-2009 yılları arasında Malatya Büyükşehir Belediyesi Esenlik Ltd. Şti.'de AR-GE ve Bilgi İşlem Müdürü olarak görev yaptı.
- 2004-2008 yılları arasında Aysis Yazılım – Malatya şirketinde iş ortağı olarak görev yaptı.

TEZDEN TÜRETİLEN ESERLER:

- Güney, İ., & Selvi, İ. H. (2023). Makine Öğrenmesi Yöntemleriyle Anormal İçme Suyu Tüketimlerinin Tespit Edilmesi ve Tahmin Modellerinin. Zeki Sistemler Teori ve Uygulamaları Dergisi, 6(2), 159–173.
<https://doi.org/10.38016/jista.1226528>

DİĞER ESERLER: