T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK AMAÇLI MODELLEME ARAÇLARI İLE ATIK SU ARITMA TESİSLERİNİN PERFORMANS DEĞERLENDİRİLMESİ

Faruk DİKMEN

DOKTORA TEZİ

Çevre Mühendisliği Anabilim Dalı

Çevre Mühendisliği Programı

Danışman

Prof. Dr. Ahmet DEMİR

Şubat, 2026

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK AMAÇLI MODELLEME ARAÇLARI İLE ATIK SU ARITMA TESİSLERİNİN PERFORMANS DEĞERLENDİRİLMESİ

Faruk DİKMEN tarafından hazırlanan tez çalışması 16.02.2026 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Çevre Mühendisliği Anabilim Dalı, Çevre Mühendisliği Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Prof. Dr. Ahmet DEMİR
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Ahmet DEMİR, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Bestami ÖZKAYA, Üye
İstinye Üniversitesi _____

Prof. Dr. İsmail TORÖZ, Üye
İstanbul Teknik Üniversitesi _____

Prof. Dr. Mehmet Sinan BİLGİLİ, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Mahmut ALTINBAŞ, Üye
İstanbul Teknik Üniversitesi _____

Danışmanım Prof. Dr. Ahmet Demir sorumluluğunda tarafımca hazırlanan "Çok Amaçlı Modelleme Araçları ile Atık Su Arıtma Tesislerinin Performans Değerlendirilmesi" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Faruk Dikmen

İmza

Aileme

ve

eşime

TEŞEKKÜR

Başta, danışmanım Prof. Dr. Ahmet DEMİR'e; yalnızca doktora ve tez sürecimde değil, çalışma hayatım boyunca, bilgi ve tecrübeleriyle bana yol gösterdiği, sorumluluk almamı teşvik ettiği ve yöneticilik anlayışıyla mesleki gelişimime önemli katkılar sunduğu için en içten teşekkürlerimi sunarım. Akademik rehberliğinin yanı sıra, birlikte çalışma kültürü ve disiplin anlayışı, bu çalışmanın ortaya çıkmasında belirleyici olmuştur.

Tezimin her aşamasında desteğini ve ilgisini esirgemeyen, bu çalışmanın ortaya çıkmasında vazgeçilmez bir role sahip olan Prof. Dr. Bestami ÖZKAYA'ya ayrıca teşekkür ederim. Yalnızca bir akademisyen değil, aynı zamanda güven veren bir dost olarak her konuda yanımda olan hocamın katkıları, bu tezin fikrî ve akademik altyapısının şekillenmesinde büyük rol oynamıştır.

Lisans eğitimimden bu yana, 1990'lı yıllardan itibaren akademik hayatımda emeği bulunan, bilgi birikimi ve akademik duruşuyla her zaman örnek aldığım Prof. Dr. İsmail TORÖZ'a; yıllar boyunca kazandırdığı bakış açısı ve mesleki vizyon için şükranlarımı sunarım.

Tez kapsamında geliştirilen yazılım ve makale çalışmalarında bilgi ve tecrübeleriyle katkı sağlayan Doç. Dr. Cevat REŞİT'e değerli destekleri için teşekkür ederim. Ayrıca, kayıt sürecimde ve derslere katılım aşamasında desteğini hiçbir zaman esirgemeyen, her zaman yanımda olan merhum Salih ASAL'ı saygı ve rahmetle anmak isterim.

Son olarak, bu uzun ve yorucu süreç boyunca gösterdikleri sabır, anlayış ve verdikleri manevi destek için sevgili eşim ve kıymetli çocuklarıma en içten teşekkürlerimi sunarım.

Faruk DİKMEN

İÇİNDEKİLER

SİMGE LİSTESİ x

KISALTMA LİSTESİ xii

ŞEKİL LİSTESİ xv

TABLO LİSTESİ xvii

ÖZET xviii

ABSTRACT xx

1 GİRİŞ 1

1.1 Bağlam 1

1.2 Problemin Tanımı 4

1.3 Tezin Ana Hatları 7

2 LİTERATÜR 9

2.1 Atık Su Arıtmanın Gelişimi 9

2.1.1 Atık Su Arıtmanın Tarih Süreci 9

2.2 Atık Su Arıtmada Tesis Modellemesi: Tarihsel Gelişim 15

2.2.1 Atık Su Arıtma Modellemesine Giriş ve Önemi 15

2.2.2 Atık Su Arıtmada Modelleme Yaklaşımlarının Sınıflandırılması 16

2.2.3 Atık Su Arıtma Modelleme Tarihçesi 17

2.2.4 Mühendislikte Standardizasyon: Alman ATV- DWA A 131 31

2.3 Dijital Simülasyon Araçları ve Literatür Uygulamaları 36

2.3.1 GPS-X (Hydromantis) 37

2.3.2 BioWin (EnviroSim) 38

2.3.3 West (DHI)……39

2.3.4 STOAT (WRc)……..39

2.3.5 ASIM (EAWAG)……..40

2.3.6 SIMBA (IFAK)…….41

2.3.7 SUMO (Dynamita)……42

2.4 Deterministik Modelleme Yaklaşımlarının Sınırları ve Dijitalleşme ile Birlikte Yapay Zekaya Geçiş: IWA Vizyonu 43

2.5 Zekanın Modellenmesinden Makine Öğrenmesine Geçişi: Teoriden Mühendislik Sistemlerine 47

2.5.1 Teorik Temeller: Makineler Düşünebilir mi?.....48

2.5.2 Bağlantıcılık ve Veriden Öğrenme Dönemi……. 48

2.5.3 Metodolojik Genişleme ve Makine Öğrenmesinin Gelişimi…… 49

2.5.4 Büyük Veri ve Derin Öğrenme Ekseninde Makine Öğrenmesindeki Dönüşüm…… 49

2.6 Atık Su Arıtma Süreçlerinde Yapay Zeka ve Makine Öğrenmesi: Literatür Analizi (2020-2025).....52

2.6.1 Bibliyometrik Analiz ve Akademik Trend…… 53

2.6.2 Deşarj Kalite Tahmini İçin Yapılan YZ Uygulamaları…….53

2.6.3 Enerji Tüketim Optimizasyonu ve Karar Destek Mekanizmaları Üzerine Yapılan Çalışmalar…… 54

2.6.4 Membran Teknolojileri ile Kirletici Giderimini Odağa Alan Çalışmalar…… 55

2.6.5 İleri Çevre Teknolojilerini ve Yeni Nesil Kirleticileri Konu Alan YZ Destekli Çalışmalar…… 56

3 VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ TEMELLERİ 58

3.1 Veri Ön İşleme.....59

3.1.1 Eksik Olan Veri Sorununun Yönetimi……..59

3.1.2 Kategorik Verilen Kodlanması…… 62

3.2 Veri Madenciliğinde Özellik Seçimi (Feature Selection 65
3.2.1 Özellik Seçimi Yöntemleri…… 66
3.3 Veri Setini Bölme Stratejileri..... 71
3.3.1 Basit Ayırma Yöntemi (Hold-Out)……... 71
3.3.2 Tabakalı Örnekleme (Stratified Splitting)…… 72
3.4 Makine Öğrenme Algoritmaları..... 72
3.4.1 Gözetimli Öğrenme (Supervised)……. 75
3.4.2 Gözetimsiz Öğrenme…… 91
3.4.3 Pekiştirmeli Öğrenme……... 93
3.4.4 Derin Öğrenme Mimarileri……... 94
3.5 Model Değerlendirme Metrikleri..... 95
3.5.1 Kök Ortalama Kare Hatası (RMSE)…… 96
3.5.2 Ortalama Mutlak Hata (MAE)……..97
3.5.3 Belirleme Katsayısı (R^2)…… 97
4 MATERYAL VE METOD **99**
4.1 Çalışma Alanı ve Kullanılan Veri Seti..... 99
4.1.1 İncelenen Atık Su Arıtma Tesisi (Tesis-A)…… 100
4.1.2 Veri Seti ve Özellikleri……... 104
4.2 Önerilen Metodolojik Yaklaşımın Genel Çerçevesi 106
4.3 Deneysel Ortam ve Çalışmadaki Yazılım Araçları..... 108
4.4 Eksik Veri Giderme 109
4.5 Kategorik Verilen Kodlanması 110
4.6 Özellik Seçimi-Boyut İndirgeme 111
4.6.1 SelectKbest Yönteminin Seçim Gerekçesi 111
4.6.2 Karşılıklı Bilgi Yönteminin Seçim Gerekçesi…… 112
4.6.3 RFE-Model Tabanlı Seçimin Gerekçesi…… 112

4.7 Veri Seti Bölme....113
4.8 Makine Öğrenme Algoritmaları....113
4.8.1 Ağaç Tabanlı Topluluk Öğrenme Modellerinin Seçim Gerekçesi.114
4.8.2 Diğer Makine Öğrenmesi Modellerinin Seçilmeme Gerekçeleri... 115
4.8.3 Derin Öğrenme Yönteminin Seçilmeme Gerekçesi……116
4.9 Model Değerlendirme Metriklerinin Seçimi....117
4.9.1 Kök Ortalama Kare Hatası (RMSE) Seçilme Gerekçesi…… 117
4.9.2 Ortalama Mutlak Hata (MAE) Seçilme Gerekçesi ……117
4.9.3 Belirlilik Katsayısı (R^2) Seçilme Gerekçesi……118
4.9.4 Diğer Metriklerin Seçilmeme Nedenleri……118
5 BULGULAR 120
5.1 Veri Seti Analizi ve Ön İşleme 120
5.2 Özellik Seçimi Senaryoları Bulguları 123
5.2.1 SelectKBest Yöntemi Sonuçları (Senaryo I)…… 124
5.2.2 Karlışıklı Bilgi (MI) Yöntemi Sonuçları (Senaryo II)…… 125
5.2.3 RFE Yöntemi Sonuçları (Senaryo III)…… 127
5.3 Veri Setini Bölme İşlemi....128
5.4 MÖ Model Uygulamaları ve Değerleme Metrikleri Sonucu 130
5.4.1 Senaryo-1 Analizi: SelectKBest Özellik Seçimi Sonuçları…… 132
5.4.2 Senaryo-2 Analizi MI Özellik Seçimi Sonuçları…… 137
5.4.3 Senaryo-3 Analizi: RFE Özellik Seçimi Sonuçları…… 142
5.5 Senaryoların Çapraz Karılaştırılmalı Analizi....146
6.2 Özellik Seçimi Senaryolarının Değerlendirilmesi 149
6.3 Makine Öğrenmesi Algoritmalarının Tartışılması....152
6.4 Metodolojik Yaklaşımın Güçlü Yönleri 154
6.5 Çalışmanın Kısıtları, Zorluklar ve Gelecek Perspektifi....155

6.6 Tezin Literatüre Katkısı ve Çıkarımlar 156

7 SONUÇ **157**

KAYNAKÇA **160**

A PYTHON KOD SATIRLARI **175**

B LABORATUVARDAKİ ANALİZLERİN DENEY YÖNTEMİ VE REFERANS METOTLARI **179**

TEZDEN ÜRETİLMİŞ YAYINLAR **180**

SİMGE LİSTESİ

$X_{B,H}$	Aktif Heterotrof Biyokütle
$X_{B,A}$	Aktif Ototrof Biyokütle
S_{ALK}	Alkalinite
k_a	Amonifikasyon Hız Sabiti
K_{NH}	Amonyak Azotu İçin Yarı Doygunluk Sabiti
S_{NH}	Amonyak Azotu Konsantrasyonu
η	Anoksik Büyüme İçin Düzeltme Faktörü
η_h	Anoksik Hidroliz İçin Düzeltme Faktörü
X_P	Biyokütle Bozunumundan Kaynaklanan Partiküler Ürünler
Y	Biyokütle Verim Katsayısı
K_O	Çözünmüş Oksijen İçin Yarı Doygunluk Sabiti
S_O	Çözünmüş Oksijen Konsantrasyonu
b_H	Heterotrof Biyokütle Bozulma Hızı
Y_H	Heterotrof Biyokütle Verim Katsayısı
k_h	Hidroliz Hız Sabiti
b	İçsel Solunum Hızı
S_I	İnert Çözünmüş Organik Madde
X_I	İnert Partiküler Organik Madde
S_S	Kolay Ayrışabilen Çözünmüş Organik Madde
μ_{max}	Maksimum Spesifik Büyüme Hızı
k	Maksimum Substrat Kullanım Hızı
K_{NO}	Nitrat Azotu İçin Yarı Doygunluk Sabiti
S_{NO}	Nitrit ve Nitrat Azotu Konsantrasyonu
b_A	Ototrof Biyokütle Bozulma Hızı
Y_A	Ototrof Biyokütle Verim Katsayısı

K_X	Partiküler Substrat Hidrolizi İçin Yarı Doygunluk Sabiti
μ	Spesifik Büyüme Hızı
K_S	Yarı Doygunluk Sabiti
X_S	Yavaş Ayrışabilen Partiküler Organik Madde

KISALTMA LİSTESİ

AAT	Atık Su Arıtma Tesisi
AÇ	Aktif Çamur
AÇS	Aktif Çamur Sistemleri
ADM	Anaerobik Çamur Çürütme Modeli
AKM	Askıda Katı Madde
ANN	Yapay Sinir Ağları
ASM	Aktif Çamur Modeli
ATV	Alman Atık Su Teknolojisi Birliği
BOİ	Biyolojik Oksijen İhtiyacı
BİO-P	Biyolojik Fosfor Giderimi
BNR	Biyolojik Nutrient Giderimi
CNN	Evrişimsel Sinir Ağları
CPU	Merkezi İşlem Birimi
DBP	Dezenfeksiyon Yan Ürünleri
DWA	Alman Su, Atık Su ve Atık derneği
DÖ	Derin Öğrenme
DSV	Seyreltilmiş Çamur Hacmi
DT	Karar Ağaçları
EBPR	Aşırı Biyolojik Fosfor Giderimi
F/M	Substrat/Biyokütle
GBM	Gradyan Artırma Makinesi
GPU	Grafik İşlem Birimi
IAWQ	Uluslararası Su Kalitesi Derneği
IWA	Uluslararası Su Birliği
IWAPR	Uluslararası Su Kirliliği ve Kontrolü Birliği

IWSA	Uluslararası Su Temini Birliği
KOİ	Kimyasal Oksijen İhtiyacı
LCA	Yaşam Döngüsü Değerlendirmesi
LSTM	Uzun Kısa Süreli Bellek
MAE	Ortalama Mutlak Hata
MAPE	Ortalama Mutlak Yüzde Hata
MBE	Ortalama Sapma Hatası
MBR	Membran Biyoreaktör
MDP	Markov Karar Süreçleri
MedAE	Medyan Mutlak Hata
MIT	Massachusetts Teknoloji Enstitüsü
MÖ	Makina Öğrenmesi
MLR	Çoklu Lineer Regresyon
MLSS	Askıda KAtı Madde
MLVSS	Askıda Uçucu KAtı Madde
PAO	Fosfor Biriktiren Mikroorganizmalar
PCA	Temel Bileşen Analizi
PFAS	Polifloroalkil Maddeler
PHA	Polihidroksialkanoatlar
PP	Polifosfat
PPCP	İlaçlar ve Kişisel Bakım Ürünleri
RF	Rastgele Orman
RFE	Özyinelemeli Özellik Eleme
RMSE	Kök Ortalama Kare Hata
RSM	Yanıt Yüzey Metodolojisi
SGB	Stokastik Gradyan Artırma
SRT	Çamur Yaşı
SVR	Destek Vektör Regresyonu
SVM	Destek Vektör Makineleri
TN	Toplam Azot

TP	Toplam Fosfor
UAKM	Uçucu Askıda Katı Madde
UCT	Cape Town Üniversitesi
YSA	Yapay Sinir Ağları
YZ	Yapay Zeka

ŞEKİL LİSTESİ

Şekil 2.1 Arden ve Lockett tarafından geliştirilen ilk AÇ prosesi.....**11**
Şekil 2.2 Monod kinetik modeli **19**
Şekil 2.3 ASM1 modeline göre karbonlu bileşenlerin fraksiyonu **22**
Şekil 2.4 ASM1 modeline göre azotlu bileşenlerin fraksiyonu.....**23**
Şekil 2.5 GPS-X kullanıcı arayüzü.....**37**
Şekil 2.6 BioWin kullanıcı arayüzü.....**38**
Şekil 2.7 WEST kullanıcı arayüzü **39**
Şekil 2.8 STOAT kullanıcı arayüzü **40**
Şekil 2.9 ASIM kullanıcı arayüzü **41**
Şekil 2.10 SIMBA kullanıcı arayüzü.....**42**
Şekil 2.11 SUMO kullanıcı arayüzü.....**43**
Şekil 2.12 Yapay zekanın evrimi.....**51**
Ş**ekil 3.1** Eksik veri giderme teknikleri **59**
Şekil 3.2 Etiket kodlaması.....**63**
Şekil 3.3 Vektörel kodlama **64**
Şekil 3.4 Hedef odaklı kodlama **65**
Şekil 3.5 Özellik seçimi yöntemleri **70**
Şekil 3.6 Yapay zeka'nın bileşenleri **74**
Şekil 3.7 Öğrenme şekillerine göre makine öğrenmesi algoritmalarının yapısal gösterimi……..…….....**76**
Şekil 3.8 Doğrusal ve polinomsal regresyon.....**77**
Şekil 3.9 Destek vektör regresyonu (SVR) **78**
Şekil 3.10 Sınıflandırma modelleri düşünme şekli.....**79**
Şekil 3.11 Lojistik regresyon algoritması.....**79**
Şekil 3.12 Destek vektör makine algoritması.....**80**
Şekil 3.13 Karar ağaçları algoritması **81**
Şekil 3.14 En yakın komşu algoritması **82**
Şekil 3.15 Topluluk öğrenmesinde paralel ve sıralı öğrenme yöntemleri **83**
Şekil 3.16 Rastgele orman algoritması.....**85**

Şekil 3.17 GBM-XGBoost ile LightGBM modellerinin ağaç büyüme düzeni **.....90**
Şekil 4.1 Tesis-A'ya ait tesis proses akış diyagramı **.....103**
Şekil 4.2 Metodoloji akış şeması **.....107**
Şekil 5.1 Giriş parametreleri ve işletme parametrelerinin istatiksel dağılımı **.....122**
Şekil 5.2 Etiket kodlama ile işlenen hava durumu ve giriş parametrelerinin pearson korelasyon ısı haritası **.....123**
Şekil 5.3 SelectKBest yöntemi ile seçilen etkili 10 adet özellik ve F-skorları**....125**
Şekil 5.4 Karşılıklı bilgi(MI) ile seçilen en etkili 10 adet özellik skorları **.....126**
Şekil 5.5 RFE ile seçilen en etkili 10 adet özellik ve önem skorları **..... 127**
Şekil 5.6 Dört ayrı veri bölümlemesi ile çalışıldığını gösterir grafik. **.....129**
Şekil 5.7 Tez kapsamında yapılan çalışmalarda ulaşılan sonuçlar **.....131**
Şekil 5.8 Senaryo-1'de SeelctKBest ile seçilmiş özellikler**.....132**
Şekil 5.9 Senaryo-2'de MI ile seçilmiş özellikler **.....138**
Şekil 5.10 Senaryo-3'de RFE ile seçilmiş özellikler**.....142**

TABLO LİSTESİ

Tablo 2.1 ASM1 modeline göre proses stokiyometrisi ve kinetiği **26**
Tablo 2.2 ASM1 ve ASM3 modellerindeki biyokimyasal prosesler **27**
Tablo 2.3 ASM2 ve ASM2d modellerindeki biyokimyasal prosesler **30**
Tablo 2.4 ASM modelleri ile ATV standartlarının karşılaştırılması **33**
Tablo 2.5 IWA'nın dijitalleşme ve YZ kronolojisi **46**
Tablo 3.1 Bagging ve rastgele alt uzay yöntemlerinin karşılaştırılması **84**
Tablo 3.2 Gradyan artırma logaritması **87**
Tablo 3.3 XGBoost'ta eklenen matematiksel ifadeler **89**
Tablo 3.4 LightGBM'de eklenen matematiksel ifadeler **91**
Tablo 4.1 Tesis-A'ya ait tasarım kriterleri ve tabi olduğu deşarj limitleri **100**
Tablo 4.2 Tezde kullanılan verilerin kategorizasyonu **105**
Tablo 4.3 Veri setine ait bazı temel tanımlayıcı istatistikler **106**
Tablo 5.1 Eksikliği giderilmiş veri setinin istatistiği **121**
Tablo 5.2 GBM-çıkış KOİ tahmin başarısının veri bölme oranlarına göre durumu **130**
Tablo 5.3 Senaryoların kurgu yapısı **132**
Tablo 5.4 Senaryo-I'e göre MÖ model ve değerleme sonuçları **133**
Tablo 5.5 Senaryo-1'e ait en iyiler tablosu **135**
Tablo 5.6 Senaryo-I'e göre MÖ model ve değerleme sonuçları **139**
Tablo 5.7 Senaryo-2'ye ait en iyiler tablosu **140**
Tablo 5.8 Senaryo-3'e göre MÖ model ve değerleme sonuçları **144**
Tablo 5.9 Senaryo-3'ye ait en iyiler tablosu **145**
Tablo 5.10 Hedef değişkenler ve özellik seçimleri karşılaştırılması **147**
Tablo 5.11 Hedef değişkenler ve MÖ algoritmalarının kıyaslanması **148**

ÖZET

Çok Amaçlı Modelleme Araçları ile Atık Su Arıtma Tesislerinin Performans Değerlendirmesi

Faruk DİKMEN

Çevre Mühendisliği Anabilim Dalı

Doktora Tezi

Danışman: Prof. Dr. Ahmet DEMİR

Dünyamızdaki su kaynakları sınırlıdır ve su çevrimi kapalı bir sistemdir. Tüm yaşamın sürdürülebilirliği açasından bu kaynağın korunması, artan nüfus, sanayileşmenin ve iklim değişikliğinin ağır baskısı altında gittikçe zorlaşmaktadır. Bu sebeplerle atık su arıtma tesislerinin (AAT) etkin bir şekilde işletilmesi büyük önem arz etmektedir. Bu tesisler, fiziksel, kimyasal ve biyolojik proseslerin bir arada yürütüldüğü; doğrusal ve doğrusal olmayan ilişkilerin bulunduğu yüksek karmaşıklık içeren sistemlerdir. Geleneksel deterministik tasarım ve modelleme yaklaşımları, kararlı hal işletme stratejileri, ani pik yüklenmeleri, endüstriyel deşarjlar ve değişen iklim baskıları sebebi ile yetersiz kalmaktadır. Bu durum özellikle küçük-orta boyutlu tesislerde yeterli uzman personel bulunmamasından kaynaklı sorunlara yol açmaktadır.

Bu tez kapsamında, IWA'nın Dijital Su Vizyonu ve Endüstri 4.0 prensipleri çerçevesinde atık su arıtma tesislerinin performansının değerlendirilmesi ve çıkış suyu kalitesinin tahmin edilmesi kapsamında veriye dayalı Makine Öğrenmesi yöntemlerinin uygulanabilirliği denenmiştir. Çalışmada büyük ölçekli bir kentsel atık su arıtma tesisinin (Tesis-A) 1073 günlük verileri öncelikle veri madenciliği

araçları ile işlenmiş, özellik seçimi uygulamaları ile indirgenmiş ve makine öğrenme algoritmalarıyla modellenmiştir. Hedef değişkenler olarak seçilen tesis çıkış parametreleri (KOİ, BOİ, AKM, TN ve TP) model ile tahmin edilmiş ve değerlendirme metrikleri ile kıyaslanmışlardır. Bu kullanılan yöntemlerle 6 adımlık bir uygulamanın bütünsel bir yaklaşımı olarak, metodolojik bütünlük sağlanmaya çalışılmıştır. Model yalnızca algoritmaların başarısına değil, Özellik Seçimi yöntemlerini merkeze alarak ilerlemiştir. Elde edilen bulgular, özellik seçimi açısından literatürdeki yaygın kanının aksini göstermiş, hesaplama maliyeti yüksek olan Özyinelemeli Özellik Seçim (RFE) yönteminin atık su arıtma tesisi verilerinde yüksek korelasyon ilişkisi ve veri gürültüsü sayesinde başarısız olduğunu göstermiştir. Buna karşılık, Filtreleme yöntemlerinden olan SelectKBest'in doğrusal ilişkileri yakalamada, Karşılıklı Bilgi Yönetiminin de doğrusal olmayan ilişkileri kavramada daha iyi performans gösterdiği ortaya konmuştur.

Algoritmalar açısından ise; gürültülü ve pik yüklemelere sahip olan AAT verilerinde Karar Ağaçları algoritması gibi tekil modellerin başarısız olduğu, Topluluk Öğrenmesi algoritmalarının daha başarılı olduğu sonucuna ulaşılmıştır. Bunların arasından özellikle XGBoost ve GBM yüksek doğruluk oranları sergileyerek (AKM için, R^2: %97), sanal sensör potansiyeli olabileceğini otaya koymuştur. TN ve TP parametreleri içinde uygulanan model sonuçlarının tahmin hata değerleri (MAE), deşarj limitlerinin %10'un altında kalmış, bu yönü ile de erken uyarı sistemi olarak kullanılabileceğini göstermiştir.

Sonuç olarak bu tez çalışması; AAT'lerin işletilmesinde deterministik modelleri dışlamadan, onları tamamlayıcı nitelikte, kolay ve erişilebilir, veri odaklı ve düşük maliyetli uzmanlık gerektirmeyen bir dijital karar destek mekanizmasının uygulanabilirliğini sunmaktadır.

Anahtar Kelimeler: Atık su arıtma, özellik seçimi, makine öğrenmesi, karar destek sistemleri, Endüstri 4.0

ABSTRACT

Performance Evaluatıon Of Wastewater Treatmentplants Using Multi-Objective Modelling Tools

Faruk DİKMEN

Department of Environmental Engineering
Doctor of Philosophy Thesis

Supervisor: Prof. Dr. Ahmet DEMİR

Water resources on our planet are limited, and the water cycle is a closed system. he protection of this vital resource, essential for the sustainability of all life, is becoming increasingly difficult under the severe pressures of a growing population, industrialization, and climate change. For these reasons, the efficient operation of wastewater treatment plants (WWTPs) is of critical importance. These plants are highly complex systems that integrate physical, chemical, and biological processes, characterized by a mix of linear and non-linear relationships. Traditional deterministic design and modeling approaches, based on steady-state operating strategies, are proving inadequate in the face of sudden peak loads, industrial discharges, and the pressures of a changing climate. This situation leads to significant problems, particularly in small-to-medium-sized facilities, due to a lack of specialized personnel.

In this thesis, the applicability of data-driven Machine Learning methods for evaluating the performance of wastewater treatment plants and predicting effluent quality was tested within the framework of IWA's Digital Water Vision and Industry 4.0 principles. In the study, 1073 days of data from a large-scale municipal wastewater treatment plant (Plant-A) were first processed using data mining tools,

reduced through feature selection applications, and modeled using machine learning algorithms. The plant effluent parameters selected as target variables (COD, BOD, TSS, TN, and TP) were predicted by the models and compared using evaluation metrics. A methodological integrity was aimed for by using a holistic 6-step application approach with these methods. The modeling process did not focus solely on the success of the algorithms but progressed by centering on Feature Selection methods. The findings contradicted the common consensus in the literature regarding feature selection, demonstrating that the computationally expensive Recursive Feature Elimination (RFE) method failed on wastewater treatment plant data due to high correlation relationships and data noise. In contrast, it was revealed that Filter methods, specifically SelectKBest, performed better in capturing linear relationships, while Mutual Information (MI) performed better in grasping nonlinear relationships.

Regarding the algorithms, it was concluded that single models, such as Decision Trees, failed on WWTP data characterized by noise and peak loads, whereas Ensemble Learning algorithms were more successful. Among these, XGBoost and GBM exhibited high accuracy rates (for TSS, R^2: 97%), demonstrating their potential as "soft sensors." The prediction error values (MAE) of the model results applied for TN and TP parameters remained below 10% of the discharge limits, indicating that they could serve as an early warning system.

In conclusion, this thesis presents the applicability of an accessible, data-driven, low-cost digital decision support mechanism that does not require high-level expertise, serving as a complementary tool to deterministic models rather than excluding them in the operation of WWTPs.

Keywords: Wastewater treatment, feature selection, machine learning, decision support system, Industry 4.0

1
GİRİŞ

1.1 Bağlam

Gezegenimizde hayat su ile başladı ve onun yokluğunda yaşamın varlığı mümkün değildir. Günümüzde suya olan bağımlılığımız artık kişisel ihtiyaçlarımızın çok ötesindedir. Endüstriyel üretimlerimiz, geliştirdiğimiz teknolojilerimiz ve gelecek amaçlarımız hep suya ihtiyaç duymaktadır. Dünya yüzeyinin %71'i sudan oluşur. Ancak, bunun sadece %0,3'ü kullanılabilir su kısmını oluşturmaktadır. Suya olan bu ihtiyacımız yeraltı suları, nehirler ve göller gibi tatlı su kaynaklarından karşılanır [1]. Dünyamızdaki, kullanageldiğimiz tüm doğal kaynaklarımız sınırlıdır ve kapalı bir çevrim sistemindedir. Dünya dışı bir sistemden kaynak girişi bulunmamaktadır. Bu nedenle suyumuz sınırlı bir kaynaktır. Milyarlarca yıl öncesinde, evrenin oluşum sürecinde, dünyamızda meydana gelen su, bugün de kullanmakta olduğumuz sudur. Yaşama kaynaklık eden su, tatsız, kokusuz ve renksizdir. Kimyasal yönü ile inceleyecek olursak, çok iyi bir çözücüdür. Bu özelliğinden dolayı da canlıların ihtiyacı olan mineraller sudan karşılanmaktadır [2]. Ancak yine bu özelliğinden dolayı da çok kolay kirlenebilmekte, faydalı kullanımı zarar görmektedir. Eskiden bazı doğal olaylar sonucu kirlenebilen su, günümüzde insan faaliyetlerinde kullanılan, metaller, toksik kimyasallar ve zirai ilaçlar gibi kimyasalları çözerek bünyesine alması ile kirlenmektedir. Artık yararlı kullanımını kaybeden su tarafımızdan atık su olarak tanımlanır.

Oluşan bu atık suların, doğal su çevriminden ayrılması, imkansız olduğu gibi etkilerinin tamamen ortadan kaldırılması da mümkün değildir. Artan nüfus, düzensiz kentleşme politikaları, önceliklenen sanayi kalkınma hamleleri gibi gelişme ve büyümeyi ön plana alan yaklaşımlar, çevresel sürdürülebilirlik açısından çok kritik konular haline gelmiştir. Bu ve benzeri birçok sebeple; atık suların

arıtılması, kirleticilerin suyun doğal döngüsünden uzaklaştırılması ve mevcut suyun kalitesinin korunması bakımından çok önemlidir [3]. Ayrıca gelişen teknoloji ve endüstri beraberinde yeni tanımlı kirletici türlerini de getirmiştir. İlaçlar ve kişisel bakım ürünleri (PPCP), polifloroalkil maddeler (PFAS) ve dezenfeksiyon yan ürünleri (DBP) gibi yeni tip kirleticilerin de su kaynaklarından uzaklaştırılması zorunlu hale gelmiştir [4,5]. Atık suların arıtılması konusu sadece çevresel etkileri bakımından değil, sosyal, ekonomik ve sürdürülebilirlik açısından da önem arz etmektedir. Gelişmekte olan ülkeler çevrenin korunmasını önceliklemeyip, büyümeye odaklandıklarından dolayı, endüstriyel, evsel ve zirai faaliyetler sonucunda kirlenen su kaynakları sebebi ile suya erişim sorunlarının artması ile karşı karşıya kalmaktadırlar [6]. Kısıtlanan kaynak, sürdürülebilirliği engellediğinden atık suların arıtılması mecburiyete dönüşmektedir. Ancak geleneksel atık su arıtma prosesleri aşırı enerji tüketimi, kimyasal kullanım gereksinimi ve oluşan çamurun bertarafındaki zorluklar nedeni ile sürdürülebilirliği olumsuz etkilemektedir [7].

Atık su arıtma tesisleri (AAT), fiziksel, kimyasal ve mikrobiyolojik reaksiyonların kullanıldığı, enerji yoğun, karmaşık ve doğrusal olmayan ilişkileri barındıran bir dizi işlemler zinciridir. Günümüzde giderek katılaşan alıcı ortama deşarj limitleri ve artan işletme maaliyetlerini düşürme ihtiyacı, bu zincir işlemler için gelişmiş kontrol sistemlerine olan ilgiyi arttırmıştır. Bu tesislerdeki enerji tüketimin büyük bölümü mikrobiyolojik faaliyetlerin ihtiyacı olan oksijeni sağlayan havalandırma sistemlerinden kaynaklanmaktadır. Sistemin oksijen ihtiyacını düzenlemek, karmaşık ve öngörülemeyen biyolojik süreçler nedeni ile zordur [8]. Tesis profesyonelleri ve operatörler biyoreaktörlerin performansını etkileyen değişiklikler karşısında sistemi dengeye ulaştırabilmek için ani kararlar almak ve operasyonel değişiklikler yapmak zorunda kalabilmektedirler [9]. AAT'ler işletilirken ki süreçte, enerji tasarrufu ile kirletici giderimi arasında bir denge kurulmasını gerektirmektedir. Bu sebeple, hem kaliteli bir çıkış suyu sağlamak, hem de enerji tasarrufu sağlamak için, sisteme etki eden kritik parametrelerin öncelikle belirlenmesi ve sonrasında tahmin edilmesi mühendislik açısından hayati öneme sahiptir [10]. AAT'lerin proseslerine etki eden faktörlerin ve kinetiğinin hesaplanabilirliği için ilk uygulamaya konulduğu 1913 yılından bu yana sürekli bir inceleme ve geliştirme konusu olmuştur.

Eski adıyla IWAPR, şimdiki adı ile Uluslararası Su Birliği (IWA); biyolojik atık su artıma tesislerinin hem işletilmesinde hem de projelendirilmesinde kullanılabilecek işlevsel modeller geliştirmek için 1983 yılında bir proje çalışma grubu kurmuştur. Bu grup, yürüttüğü kapsamlı çalışmalar sayesinde, aktif çamur sistemlerindeki karbon oksidasyonunun, nitrifikasyonun ve denitrifikasyonun performanslarını gerçeğe en yakın şekilde simüle eden Aktif Çamur Modeli No.1 (ASM1)'i 1987 senesinde ortaya koymuştur. Çalışma grubuna ait diğer modeller ASM2(1995), ASM2d(1999) ve ASM3(1999) devam eden yıllarda yayınlanmıştır [11]. ASM'ler nutrient giderimine yönelik aktif çamur proseslerinin tasarımı, işletilmesi ve iyileştirilmesi için yaygın olarak kullanılan ve kabul görmüş metodolojiler olarak öne çıkmaktadırlar [12]. Bu modeller matematiksel olarak hesaplanmış çok yönlü kinetik katsayılarla tariflenmişlerdir. Sistem dinamiklerini doğru şekilde temsil etme kapasiteleri olmalarına rağmen atık su giriş karakterizasyonunun doğru bir şekilde gerçekleştirilmesine, dinamik modelin kalibrasyonuna ve validasyonuna ihtiyaç duyarlar [13]. AAT'lerin anlık değişimlere karşı davranışlarını görebilmek için her defasında giriş atık su karakterizasyonu ile hesaplamaları yeniden yapıp işletme şartlarında yorumlama ihtiyacı olmaktadır. Bunun hesaplanabilirliği için bilgisayar yazılım ortamında simülasyon programları geliştirilmiştir. Bu modellerin birçok ticari isimleri ve farklı geliştiricileri vardır; Gps-x, West, Aquasim, Biowin, Simba, Sumo vd gibi. Bu uygulamalar başarılı bir şekilde AAT'leri simüle edebilseler de kullanımları çok ciddi profesyonellik ve bir eğitim gerektirmektedir.

AAT kendisi ve çevresi veri üreten bir süreçtir. Tesislerin mekaniksel işletme verileri (bakım, onarım ve arıza), biyolojik ve kimyasal süreçlerin karmaşık verileri, AAT'lerin bulunduğu çevresinden, havzasından, ve ikliminden kaynaklanan gelen verileri çok çeşitlidir. Bu kadar geniş ve sistemi etkileyen karmaşık verilerin yönetilmesi de profesyonellik gerektirmektedir. Bu karmaşıklığı yönetmek, günümüzde Endüstri 4.0 olarak adlandıran 4.sanayi devrimi ile akıllandırılmış tesisler ile mümkün kılınabilmektedir. Endüstri 4.0; üretkenliği ve esnekliği artırmak için, Nesnelerin İnterneti (IoT), yapay zeka (YZ) ve robot bilimi gibi en son teknolojilerin endüstriyel süreçlere dahil edilmesiyle karakterize edilir [14]. Endüstri 4.0 bu bahsi geçen bileşenleri entegre bir şekilde kullanarak yüksek düzeyde otomasyon, arttırılmış verimlilik ve sürekli optimizasyon ile karakterize edilen akıllı tesislerdeki üretim süreçlerinin bütününü ifade etmektedir. Bu

kapsamda akıllandırılmış tesisler, verimlilik ve üretkenliği arttırmak amacı ile dijital teknolojileri ve veri odaklı çözümleri sistematik olarak kullanan tesisler olarak tanımlanmaktadır [15]. AAT'lerde bu sistemin kullanılması, üretilen verilerin nesnelerin interneti (IoT) ile sahadan alınması, veri madenciliği metotları ile işlenmesi sonrasında bir karar merkezince işlenmesi ve sistemin optimize edilmesidir. Burada karar merkezi, artık günümüzde nümerik modellemeden çok YZ uygulamalarına doğru yönelmektedir. YZ uygulamalarından derin öğrenme (DÖ) veya makine öğrenmesi (MÖ), sistemin sürekli olarak bulut tabanlı veri kümesinden gelen verileri işleyerek, sistemini güncel duruma adapte edip ve yeni durumlara göre kendi karar mekanizmasını revize ederek çalışırlar.

Bu tez kapsamında, Endüstri4.0 in uygulamalarından YZ'nin bir uygulaması olan (MÖ) ile AAT çıkış suyu kalite parametreleri tahmin edilemeye çalışılacaktır. MÖ ile sistemi etkileyen parametrelerin öncelikleri belirlenecek, sonrasında MÖ algoritmalarından bazıları ile sistem karakterize edilmeye çalışılacaktır. Buradaki amaç, işletme profesyonellerine karar destek mekanizması oluşturmak, olabilir ise bir erken uyarı protokolü geliştirmek olacaktır. Askıda Katı Madde (AKM), Kimyasal Oksijen İhtiyacı (KOİ), Biyokimyasal Oksijen İhtiyacı (BOİ), Toplam Azot (TN) ve Toplam Fosfor (TP) parametrelerinin çıkış konsantrasyonlarını tahmine yönelik bir model oluşturulmaya çalışılacaktır.

1.2 Problemin Tanımı

AAT tasarımı için yaygın kullanılan ASM'lerin bir versiyonu olan Alman Atık Su Mühendisliği Derneği Standardı (ATV)-131E ülkemizdeki su ve atık sı idarelerince yaygın olarak kullanılmaktadır. Bu standarda göre en büyük şehirlerimizden, en küçük ölçekli yerleşim birimlerimize kadar birçok alanda inşa edilmiş AAT'ler mevcuttur. Ve bu tesislerin işletilmesinde yeteri kadar AAT işletme profesyonelleri lokasyonları sebebi ile istihdam edilememektedir. Bu standardın yaygın kullanılmasının olası bazı olumsuz yönleri vardır;

- Bu standart, yerelde bazı kabullerle (atık su karakterizasyonu) doğrudan projelere uygulandığında, öngörülen hacimlerin çok fazla olduğu görülmektedir [16]. Bu hacim fazlalığı işletme kolaylığı sağlasa da, inşaat maliyetleri gibi, enerji tüketimlerini arttırmaktadır. Örneğin; Ülkemizde incelenen bazı tesislerde

kişi başı özgül enerji tüketimi yaklaşık 42 kW-saat/kişi.yıl hesaplanmış iken bu değer Almanya da ortalama 23-30 kW-saat/kişi.yıl değerleri arasındadır [17].

- Orijinal tasarım değerleri önerilen aralıklara tam uymadığından, bu yöntemin sonuçları ile gerçek durumdaki sonuçlar farklı olmaktadır. Büyük ve küçük AAT'ler arasında tasarım ve işletme farklılıkları vardır ve bu farklılıklar büyük tesislerde daha belirgindir [18].
- Yenilikçi uygulamalara çok uygun olmayabilmektedir. Tesislerin teknolojik dönüşümüne karşı biraz dirençlidir.

AAT'lerin işletme koşulları hem kontrol edilebilir hem de kontrol edilmesi mümkün olmayan şartlar altındadır. Bu değişkenlikler; ani yük dalgalanmalı, endüstriyel kaçak su deşarjları, mekanik ve elektriksel arızalar, mevsimsel farklılıkların artık her yıl öngörülemez şekilde değişmesi vb. sayabiliriz. Küresel ısınmanın iklim değişikliklerine etkisi ve artıma tesisi işletilmesine etkileri de belki incelenmesi gereken bir konudur. Bu tür koşulların deterministik hesaplamalarla inşa edilmiş AAT'lerin işletilmesinde, deterministik modellerle simüle edilmesi oldukça zordur. Bundan dolayıdır ki, klasik kontrol sistemlerinin veya sabit işletme stratejilerinin ötesinde, her şarta uyarlanabilir ve tahmin yöntemlerine dayalı bir yaklaşım gerekmektedir.

Matematik modellere dayalı simülasyon uygulamaları da Endüstri 4.0 içerisinde kendine yer bulabilmektedir. Nesnelerin interneti ile makinelerden gelen her veri, laboratuvarlarda üretilen her değer, saha ve saha dışından derlenen anlık bilgiler, bu simülasyon programlarında değerlendirilerek işletme operatörlerine bilgi vermektedir. Ancak bu simülasyon uygulamaları, olası deşarj limit aşımlarını tahmin etmekten ziyade var olan durum ile ilgili bilgi vermekte, bu var olan durumlar ile limit aşımı tespiti kullanıcı becerisine kalmaktadır. Bu yazılımların kullanımı da belirli bir uzmanlık gerektirmektedir. Ayrıca değişen her işletme durumuna göre (hacimsel ayarlamalar) yeniden simülasyon ayarlamaları yapılması gerektirmektedir. Bu da yine uzmanlık kısıtı ile sisteminin kısıtlanmasına sebep olmaktadır.

Dijital ikiz (digital twin) uygulamaları da Endüstri 4.0 içerisinde yer bulmaktadır. Dijital ikiz kavramı, fiziksel bir sistemin sanal bir modelini oluşturarak, AAT'den gelen online verilerin ve sisteme işlenen diğer dijital verilerin (laboratuvar ve saha)

sürekli optimizasyon ve analiz yapılmasına imkan veren, bir yazılım ve uygulamalar dizisidir. Dijital ikiz ülkemizde, belirli büyük tesislerde, yeterli yetişmiş uzman tarafından uygulanabilir bir yöntemdir. Ayrıca dijital ikiz sistemine entegre edilmiş bir YZ karar destek uygulaması gelecekte kaçınılmaz olarak yönelinecek bir mecradır. Ancak bizim çalışmamızın odağı daha çok, işletme uzmanlığının tam olarak sağlanamadığı, tesislerin verilerinin çıkış parametrelerinin tahminini yapan bir YZ modeli geliştirmektir.

YZ ve MÖ entegrasyonu, işletme profesyonellerinin her an süreçleri değerlendirdiği, veri tabanlı bir destek ortamı yaratabilir. Yukarıda sayılan olumsuzluklar, MÖ tabanlı yaklaşımlarla aşılabilir. MÖ modelleri çok yönlü, uyarlanabilir, yeni şartları öğrenip adapte olabilir ve kendi kendini geliştirebilir yapılardır. Bu modeller AAT'lerdeki farklı parametreler arasındaki karmaşık etkileşimleri tanımlayabilir ve uzun dönemde işletme süreçlerinde daha esnek çözümler üretebilir. Bununla birlikte, makine öğrenmesi yaklaşımlarının güvenilir ve başarılı sonuçlar üretebilmesi, kullanılan verilerin hem niceliksel yeterliliğine hem de niteliksel doğruluğuna doğrudan bağlıdır. Bu doğrultuda, söz konusu çalışmada uzun dönemli ve geniş kapsamlı veri setine sahip bir tesis ele alınmıştır. Elde edilen veriler kullanılarak bir tahmin modeli geliştirilecek; belirlenen giriş değişkenleri temelinde, deşarj limitlerinin aşılma olasılıklarının öngörülmesi hedeflenmektedir. Bu süreçte, modelleme aşamasına geçilmeden önce, veri madenciliği teknikleri aracılığıyla verilerin derlenmesi, ön işleme tabi tutulması ve analize uygun hâle getirilmesi sağlanacaktır. Bu aşamanın ardından, özellik seçimi (feature selection) yöntemleri kullanılarak modelleme sürecine en yüksek katkıyı sağlayan değişkenler belirlenecek; takiben veri seti eğitim ve test alt kümelerine ayrılacaktır. Derlenen veriler, beş farklı makine öğrenmesi algoritması kullanılarak modellenecek ve söz konusu modeller aracılığıyla çıktı değişkenlerinin tahmin edilmesi amaçlanacaktır. Son aşamada ise üç farklı performans metriği üzerinden modellerin başarı düzeyleri karşılaştırmalı olarak değerlendirilecektir.

Bu doğrultuda, elde edilecek bulgular aracılığıyla, hangi makine öğrenmesi modelinin hangi parametrelerin tahmininde daha yüksek başarı sergilediğinin ve kullanılan doğrulayıcı metriklerden hangilerinin sonuçlara daha güçlü biçimde yakınsadığının ortaya konulması hedeflenmektedir. Bu yönüyle çalışmanın, ilgili literatüre yöntemsel ve karşılaştırmalı bir katkı sunacağı değerlendirilmektedir.

Yapay zekâ uygulamalarının, atık suların oluşumundan itibaren havza ölçeğinde izlenmesi ve yönetilmesi süreçlerinde, atık suların toplama sistemleri aracılığıyla arıtma tesislerine ulaştırılmasında, arıtma tesislerinin tasarım aşamalarında ve proseslerin işletilmesi ile optimizasyonunda gelecekte yaygın ve etkin bir biçimde kullanılacağı öngörülmektedir. Bu bağlamda, yapay zekâ temelli yaklaşımların, karar destek mekanizmalarının güçlendirilmesine, işletme verimliliğinin artırılmasına ve çevresel sürdürülebilirliğin sağlanmasına önemli katkılar sunacağı değerlendirilmektedir Son yıllarda hızla gelişen büyük veri (Big Data) uygulamaları, verilerin bulut ortamlarında depolanması ve açık kaynaklı veri ile uygulama paylaşımlarının yaygınlaşması, yapay zekâ teknolojilerine bugünden tam olarak kestirilemeyecek bir gelişim alanı sunmaktadır. Gelecekte, atık su arıtma tesislerinin; atık suyun oluşum anında, tüketicilerin su kullanım miktarlarının anlık olarak ölçülmesinden elde edilen verileri, kanalizasyon şebekelerinin SCADA sistemleri üzerinden izlenmesiyle sağlanan bilgileri ve meteorolojik tahmin verilerini birlikte işleyerek, atık sular henüz arıtma tesisine ulaşmadan önce operasyonel müdahalelerde bulunabilecek tam otomasyonlu yapılar hâline gelmesi öngörülebilir. Bu kapsamda yürütülen bu çalışma, literatürde yer alan benzer araştırmalarla birlikte, söz konusu dönüşüm sürecinin ve geleceğe uzanan gelişim basamaklarının ilk adımlarını oluşturmaktadır.

1.3 Tezin Ana Hatları

Bu çalışma ile, AAT'lerin çıkış parametrelerini doğru tahmin eden bir MÖ modeli geliştirilerek bir karar destek sisteminin geliştirilmesi amaçlanmaktadır. Çalışmamız beş ana bölümden oluşmaktadır.

Birinci bölümde, araştırmanın genel çerçevesi ortaya koyulmaktadır. Bu bölümde çalışmanın bağlamı, araştırmanın amacı ve önemi ile tezin ana hatları sunulmuştur. Ayrıca klasik tasarım ile yapılmış AAT'lerde işletme yaklaşımının kısıtlılıkları, dijital dönüşümün gerekliliği vurgulanıp, YZ uygulamalarından bahisle MÖ'ne dayalı modellerin potansiyeline değinilmiştir.

İkinci bölümde mevcut literatür anlatılmıştır. Atık su arıtmanın tarihçesi, kinetik ve stokiyometrik yaklaşımlar hakkında bilgi verilmiştir. Gelişen teknoloji ile yeni yaklaşımlardan olan Endüstri 4.0, Big Data, IoT gibi uygulamaların çevre mühendisliğinde kullanım alanlarına değinilmiştir. Sonrasında da MÖ konusundaki

mevcut çalışmalar ve modelleme yaklaşımları hakkında bilgi verilmiştir. Yine bu bölümde bizim çalışmamızın literatüre katkısına değinilmiştir.

Üçündü bölümde, araştırmamızın metodolojisinin oluşturulduğu çatı anlatılmaktadır. Burada ilk önce hangi yazılım dilinin neden tercih edildiği, kısaca verilerin elde edilişi, elde edilen veri setlerinin özellikleri, kullanılan özellik mühendisliği teknikleri ve MÖ algoritmaları anlatılmış. Sonrasında ortaya konulan modellerin tahmin yaklaşımları değerlendiren metodolojiler anlatılmıştır.

Dördüncü bölümde çalışmamızdaki uygulamalarımız ve bulgularımız anlatılmıştır. Bu bölümde geliştirdiğimiz modellerin performansları ve birbirleri ile kıyaslamalarına yer verilmiştir. Ortaya çıkan bulgular uygulanabilirlik açısından değerlendirilmiştir.

Son bölün olan beşinci bölümde ise sonuçlar değerlendirilmiş ve öneriler sunulmuştur. Denenen algoritmaların çevre mühendisliğinin AAT işletilmesi konusundaki sınırlılıkları, potansiyeli ve gelecek perspektifine ait öneriler sunulmuştur. Literatüre katkısı değerlendirilmiştir.

Yukarıda anlatılan yapı çerçevesinde hazırlanan bu doktora tezimiz, geleneksel mühendislik sınırları dışında, çevre mühendisliği alanında YZ ve MÖ uygulamalarının AAT işletmelerine entegrasyonu konusunda bilimsel ve uygulamalı bir temel oluşturmayı hedeflemiştir.

2
LİTERATÜR

2.1 Atık Su Arıtmanın Gelişimi

Bu başlık altında, atık su arıtmasının tanımı, ilk yaklaşımları günümüze kadar geçirdiği süreçler anlatılacaktır.

2.1.1 Atık Su Arıtmanın Tarih Süreci

Atık suyun tarihçesi, insanlık tarihinin yaşadığı çevre ile olan ilişkisinin bir yansımasıdır:

- Antik Dönem : Uzaklaştır
- Orta Çağ : Geriye dönüş ve hijyenin çöküşü
- Sanayi Devrimi : Oluşan Krizi görmeden gelmeye çalış veya kaçın.
- 20. Yüzyıl : Kirleticiyi gider, çevresel etkiyi sınırlandır.
- 21 Yüzyıl : Geri kazanımla değer oluştur.

Antik Dönem (Uzaklaştırma- MÖ3000-MS 500); Bu dönemde arıtma diye bir kavram yoktu, öncelik yaşanılan yerden hızlıca uzaklaştırmaydı. Bu da genel bir uygulama değildi, sadece ilk medeni şehirlerde görülmüştür. İlk uygulama MÖ 2600'lü yıllarda İndus Vadisi uygarlığında (Mohenjo-Daro ve Harappa) evlerden gelen atık suların pişmiş tuğlalardan yapılmış kapalı kanallarla toplanıp nehre deşarj edilmesidir. Kanalların tıkanmasını önlemek amacıyla basit ızgaralar da kullanılmıştır [19]. Antik Roma'da (MS 500–400) "Cloaca Maxima" (Büyük Kanal) gibi anıtsal bir kanalizasyon sistemi inşa edilmiştir. Bu sistemlerle şehirde kullanılmış sular Tiber Nehri'ne taşınırken aynı zamanda şehirdeki bataklık alanların kurutulması da sağlanmıştır. Bu durum halk sağlığı açısından olumlu bir yaklaşım olsa da söz konusu dönemde kirliliğin giderilmesi konusu hiç gündeme gelmemiştir [20].

Orta Çağ (Altyapısızlık ve Sağlık Krizi- MS 500-1700); Roma'nın çöküşünden sonra Avrupa, kamu hizmetlerinin ve mühendislik uygulamalarının büyük bir gerilemesine sahne oldu. Roma döneminden kalan yapılar kullanılmaz duruma gelirken, mevcut olanlar korunmadığı gibi yeni gelişmeler de yaşanmadı. Orta Çağ Avrupası'nda şehirlerin sokakları dar ve düzensizdi; evlerde oluşan atık sular doğrudan sokaklara bırakılırdı. Sokaklar zamanla açık kanallara dönüştü. 14. yüzyılda Londra sokaklarından atık suları uzaklaştırmak için açık kanallar ve hendekler yapılsa da bunlar sık sık tıkandığından bizzat kirlilik kaynağı hâline geldi. 1300'lü yıllarda Fleet Nehri'ndeki kirlilik o kadar artmıştı ki parlamento onu "pislik kanalı" olarak adlandırdı [21]. Dönemin hijyen anlayışı koku ile sınırlı kaldı. Hastalıkların yayılmasının sebebini bu koku olarak gördükleri için karşılaştıkları durumu "Koku Teorisi (Miasma Theory)" adını verdikleri bir yaklaşımla çözmeye çalıştılar. Atık suyu arıtmayı değil, kokuyu uzaklaştırmayı önceliklediler. Bu ise gerçek hastalık taşıyıcılarının yayılmasını engelleyemedi. Avrupa nüfusunun yaklaşık %30'u kolera, tifo ve veba gibi hastalıklar sebebiyle öldü [22].

Aynı dönemde İslam coğrafyasında standartlar Roma dönemine daha yakındı. Bağdat, Kurtuba ve Kahire gibi şehirlerde kanalizasyon sistemi vardı. Temizliğin dini önemi ve tuvalet kültürü bu çağlarda İslam coğrafyasını salgınlardan büyük ölçüde korumuştur. Örneğin 14. ve 15. yüzyılda Endülüs'te büyük şehirlerde gelişmiş kanalizasyon şebekesi olduğu gibi, kırsal ve az nüfuslu mahallelerde fosseptik uygulamaları vardı [23].

Sanayi Devrimi (18. Ve 19. Yüzyıllar); Sanayi tesislerinin kurulmasıyla artan istihdam ihtiyacı, şehirlere yönelik büyük göç dalgalarını başlattı. Çarpık kentleşme, halihazırda tarihi altyapısı yetersiz olan şehirlere daha da büyük yükler getirdi. Ayrıca bu dönemde, atık suların kirlilik içeriğine yeni bir tanım eklendi: "Endüstriyel Kirlilik". Ham madde kullanımındaki verimsizlik ve üretim sistemlerinin denetimsizliği; fabrika atıklarının, bugünkünden çok daha yoğun bir kirlilik yüküyle ve kontrolsüzce alıcı ortamlara deşarj edilmesine yol açtı. Kirliliği ve sonuçlarını bilimsel temellerle açıklama noktasında, o dönemin fen bilimleri ve mühendislik becerileri henüz yeterli düzeyde değildi. Nitekim Thames Nehri o kadar kirlenmişti ki, bu durum tarihe "Büyük Koku (The Great Stink)" olarak geçti ve 1858 yılında parlamento çalışmalarının dahi aksamasına sebep oldu [24]. 1854 yılında Londra'daki kolera salgınının su kaynaklı olduğunun anlaşılması ve

ardından edebi eserlere dahi konu olan "Büyük Koku" hadisesinin yaşanmasıyla birlikte bir dizi önlem alınmaya başlandı. Bu kapsamda, mekanik arıtmanın ilk örnekleri sayılabilecek kum tutucular ve çökeltme tankları geliştirildi. Özellikle kireç kullanımıyla kimyasal çöktürme yöntemleri denendi. 19. yüzyılın sonlarına doğru ise doğal süreçleri taklit eden damlatmalı filtreler (trickling filters) ve çamur çürütme sistemleri gibi biyolojik arıtmanın öncü uygulamaları literatüre girdi [25].

20. Yüzyıl (Biyolojik Arıtma-1900-1970); Bu dönemde devrim niteliğinde bir gelişme yaşandı. İngiltere Manchester yakınlarındaki bir AAT(Şekil2.1) de çalışan Edward Arden ve W.T. Lockett, atık suyun havalandırıldıktan bir müddet sonra mikrobiyolojik olarak aktivasyonunun arttığını ve organik maddelerin daha hızlı ayrıştığını gözlemledi. Bu ayrışma sonrasında bir biyokütle oluştuğunu da fark etiler. Bu mikrobiyal topluluğa da "Aktif Çamur (AÇ)" adını verdiler.



Şekil 2.1 Arden ve Lockett tarafından geliştirilen ilk AÇ prosesi (Kaynak:[27])

AÇ, organik ve inorganik maddeler içeren atık su ile hem canlı hem de ölmüş mikroorganizmaların bir karışımıdır. AÇ prosesi de, bu karışımda bulunan canlı mikroorganizmaların, aerobik ortamda oksijeni kullanarak atık sudaki organik maddeyi ayrıştırmasına dayalı olarak geliştirilen bir atık su arıtma sistemidir [28]. AÇ günümüzde de askıda büyüyen sistem olarak atık suyun biyolojik olarak arıtılmasında kullanılan en yaygın prosestir [29].

İkinci Dünya Savaşı'na kadar geçen süreçte, Aktif Çamur (AÇ) sistemlerinin tasarımı mühendislik açısından standartlaşmaya başlamış; havalandırma havuzları, çamur geri devri ve çöktürme üniteleri gibi günümüzde hâlâ kullanılan temel prosesler geliştirilmiştir. Buna paralel olarak, büyük metropollerde yüksek kapasiteli tesisler işletmeye alınmaya başlanmıştır [30]. Ancak savaş, atık su arıtımı alanındaki teknik ilerlemeleri durma noktasına getirmiştir. 1948'e kadar bu alanda

kayda değer bir çalışma yapılamamış; birçok tesisin zarar görmesi ve işletme süreçlerinin aksaması nedeniyle alıcı ortamlardaki kirlilik yükü de artmıştır [31].

Savaş sonrası gözlenen nüfus artışı, yeniden inşa süreçleri, kimyasal tüketimindeki artış ve ağır sanayi hamleleri, deşarj standartlarının sıkılaştırılmasını beraberinde getirmiştir. 1950-1960'lı yıllarda AÇ verimliliğini artırmak için yüksek yüklü aktif çamur, damlatmalı filtreler ve biyodisk prosesleri geliştirilmiştir [25]. Ayrıca bu dönemde, mikrobiyolojik büyümenin modellenmesinde ilk adım olan "Monod Denklemi", 1942 yılında J. Monod tarafından literatüre kazandırılmıştır [32]. 1960'ların sonuna gelindiğinde, dünyadaki özgürlük hareketlerinin etkisiyle çevre bilinci de artmış; Amerika Birleşik Devletleri (ABD)'nde "Temiz Su Hareketleri" güçlenerek 1970'lerde yürürlüğe giren pek çok çevresel düzenlemenin zeminini hazırlamıştır. Böylelikle 1970'lere kadar geçen süreç, atık su arıtma teknolojilerinin hem bilimsel ve mühendislik hem de politik açıdan önemli gelişmelere sahne olmasını sağlamıştır [33].

20. yüzyılın son çeyreğinde; AÇS yalnızca karbon odaklı organik madde giderimi için değil, aynı zamanda azot ve fosfor gibi nütrientlerin (besin maddelerinin) de giderilmesi için geliştirilmeye başlanmıştır. Nitrifikasyon-denitrifikasyon prosesleri AÇS'lerde standart hale gelmiştir. A2/O, Bardenpho ve UCT gibi ileri biyolojik atık su arıtma teknolojileri 2000'li yıllardan önce geliştirilmiştir [11]. 1990'lardan sonraki süreçte; enerji maliyetlerindeki düşüş ve nanoteknolojinin gelişmesine bağlı olarak membran imalatının kolaylaşması ile Membran Biyoreaktörler (MBR) tercih edilir olmaya başlanmıştır. MBR sistemleri; düşük alan ihtiyacı, yüksek çıkış suyu kalitesi ve yüksek biyokütle konsantrasyonlarında çalışabilme avantajları sebebiyle kentsel ve endüstriyel atık suların arıtılmasında yaygın olarak benimsenmiştir [34].

Yine bu yıllarda kirliliğin global ölçekli etkileri nedeniyle, su kaynaklarının korunmasına yönelik bütüncül yaklaşımlar önem kazanmıştır. Bazı global ölçekli kuruluşlar birleşerek tek çatı altında daha büyük bir yapıya dönüşmüştür. 1947 yılında kurulan Uluslararası Su Kirliliği Araştırmaları Birliği (IAWPRC), atık su arıtımı, su kirliliği kontrolü ve çevresel kimya üzerine küresel ölçekte yapılan araştırmaları paylaşmaktaydı. Bir diğer kurum da 1972 yılında kurulmuş olan Uluslararası Su Temini Birliği (IWSA) idi. Bu iki kurumun iş birliği, politikalar ve uygulama alanlarındaki bağı güçlendirmek amacıyla zamanla artmıştır. Bu stratejik

sürecin sonunda kurumlar, 1999 yılında birleşerek Uluslararası Su Birliği (IWA) adını almıştır [35]. Bu birleşme, bütünleşik yaklaşımların kurumsal düzeyde benimsenmesine öncülük ederek IWA'nın günümüzdeki küresel yapısını meydana getirmiştir. Bu dönüşüm sonrası uzmanlık grupları; teknik altyapılarını ve bölgesel ağlarını güçlendirerek, özellikle gelişmekte olan ülkelerde birçok proje yürütmeye başlamıştır [36]. Birlik ayrıca; Water Science and Technology, Water Research ve Journal of Water and Health gibi prestijli dergilerle akademik gelişmeleri de desteklemektedir.

21.Yüzyıl (Yeni Teknolojiler ve Geri Kazanım); karbon ayak izi ve yaşam döngüsü değerlendirmesi (LCA) gibi kavramlar ortaya çıkmıştır. Çevresel duyarlılık, iklim anlaşmaları, kirliliğin ve çevresel etkilerinin artması, yeni teknolojilerin geliştirilmesini de zorunlu kılmıştır. Atık çamur oluşumunun düşük olmasını sağlayan ve yer ihtiyacını azaltan teknolojilere yönelim artmıştır. Bu kapsamda; aerobik, anaerobik ve anoksik zonları tek bir hacimde birleştirebilen, daha az çamur oluşumu sağlayan, yüksek biyokütle konsantrasyonunda çalışabilen ve daha az alana ihtiyaç duyan aerobik granül çamur teknolojisi geliştirilmiştir. Bu sistem Avrupa'da yaygınlaşmaktadır [37]. Yeni tip kirleticiler de bu dönemde tanımlanmıştır. İlaç kalıntıları ve endokrin bozucular gibi mikro kirleticilerin giderimi için ise ozonlama ve granül aktif karbon adsorpsiyonu gibi ileri arıtma yöntemleri birçok büyük ölçekli tesise entegre edilmiştir [38]. Granül aktif karbon, mikro kirleticileri adsorbe etmek için yaygın olarak kullanılmaktadır. Ozonlama ve ileri oksidasyon prosesleri, bu yeni tanımlanmış inatçı organik bileşikleri parçalamak ve daha zararsız hale getirmek amacıyla İsviçre'deki büyük tesislerde kamu gücü ile zorunlu olarak kullanılmaktadır [39].

20. yüzyılın son çeyreğinde nutrient giderimi teknik açıdan çözümlenirken, 21. yüzyılda artık giderilen nutrientlerin geri kazanımına odaklanılmıştır. Fosfor, dünyamızda sınırlı bir rezerve sahiptir. Canlılık için enerji transferindeki önemi sebebiyle, tarımsal faaliyetler aracılığıyla fosfor döngüsüne hızla geri kazandırılmalıdır. Atık sudaki fosforun %90'ı çamurda yoğunlaşmaktadır. Ancak bu çamurun doğrudan tarımda kullanımı; içerdiği patojenler, ağır metaller ve toksik organik bileşikler sebebiyle hem toprak kirliliğine yol açması hem de hijyen riskleri nedeniyle mümkün olmamaktadır [40]. Özellikle çamur çürütme tanklarında meydana gelen bir fosfat tuzu olan struvit, kontrollü çöktürme metodu ile sistemden

uzaklaştırılarak yüksek kaliteli bir gübre olarak kullanılmaktadır. Ayrıca AirPrex, Pearl ve Ostara gibi ticari tescilli prosesler ile doğrudan geri akımlar veya atık su üzerinden fosfor kazanılıp yine gübre formuna getirilebilmektedir [41].

21. yüzyılda artan nüfus ve azalan kaynaklar, küresel su kıtlığı sorununu ortaya çıkarmıştır. Bu sebeple atık su, artık bir kaynak olarak kabul edilmeye başlanmıştır. Arıtılmış atık suyun yeraltı sularını beslemek maksadıyla bir rezervuara deşarj edilmesi ve doğal bir bariyerden geçerek yeniden kullanılması, uzun süredir uygulanagelen bir yöntemdir. Ancak 21. yüzyıldaki en cesur yaklaşım; MBR, ters ozmos ve ultraviyole ile dezenfeksiyon aşamalarından sonra, doğrudan atık suyun içme suyu olarak kullanımıdır. Bu uygulama Singapur ve Namibya'da kısıtlı olsa da kullanılmaktadır [42].

AAT'ler, tarihsel süreçte yoğun enerji tüketen tesisler olarak işletilmiştir. Günümüzde ise temel hedef; bu tesislerin enerji tüketimini minimize etmek, hatta enerji pozitif bir işletme haline getirmektir. Yaygın kullanılan mezofilik çürütmenin dışında, daha yüksek biyogaz üretimi sağlayan termofilik çamur çürütme prosesleri yaygınlaşmaya başlamıştır [43]. Bu teknolojinin bir üst aşaması olarak, çamuru yüksek basınç (6-12 bar) ve yüksek sıcaklıkta (150-170 °C) işleyen termal hidroliz yöntemi de geliştirilmiştir. Bu prosesle biyogaz üretimi %30-50 oranında artarken, bertaraf edilmesi gereken çamur miktarı da %25-35 oranında azalmaktadır [44]. Üretilen biyogaz; ısı elde edilmesi için yakma motorlarında, elektrik üretmek için ise kojenerasyon ünitelerinde yakılmaktadır. Günümüzde bunların dışında, yeni teknolojilerle geliştirilen mikrotürbinler kullanılarak daha verimli enerji üretimi sağlanmıştır. Ayrıca, elde edilen biyogazdan hidrojen elde edilerek yakıt hücresi ile elektrik üretimi de gerçekleştirilmektedir. Bu prensiple işletilen bazı tesisler de mevcuttur.

21. yüzyılın en önemli gelişim alanlarından biri de hiç kuşkusuz dijital dönüşüm olmuştur. İlk aşamada klasik SCADA sistemleri ve gelişmiş sensörler yardımıyla proses parametreleri izlenerek sistem kontrol altında tutulmaya başlanmıştır. Son yıllarda ise IoT, büyük veri, veri madenciliği ve YZ gibi kavramlar AAT işletmeciliğinde kullanılmaya başlanmıştır. Geliştirilen veri odaklı ve deterministik modellerin ulaştığı en ileri nokta dijital ikiz (digital twin) kavramıdır. Bu sistemlerle tesisin sanal bir kopyası oluşturularak, simülasyon modunda tesis işletimine olanak sağlanmaktadır [45]. Günümüzde birçok ticari isim altında benzer

çalışmalar mevcuttur. Güncel süreçte artık dijital ikiz kavramının ötesine geçilerek, tamamen kendi kendine işleyen yapay zeka destekli tesis yaklaşımları oluşturulmak istenmektedir. AAT işletilmesi ekosisteminde, oluşturulan muazzam miktardaki veri işlenerek anlamlandırılmayı beklemektedir. Birçok güncel araştırmada; yapay zeka, makine öğrenmesi ve derin öğrenme uygulamaları kullanılarak AAT'lerin işletilmesindeki karmaşıklığı gidermek amacıyla yeni algoritmalar geliştirilmektedir [46].

2.2 Atık Su Arıtmada Tesis Modellemesi: Tarihsel Gelişim

AAT'ler; hızlı kentleşme ve sanayinin gelişmesiyle birlikte, su kaynaklarının korunması ve sürdürülebilirliği açısından kritik bir öneme sahip olmuştur. AAT'lerin tasarımı ile işletme ve enerji yönünden optimize edilerek kontrol altında tutulması hem deşarj standartlarına uyum sağlanması hem de işletme maliyetlerinin düşürülmesi açısından büyük önem arz etmektedir. Tesislerin tasarımı ve modellenmesi, birçok karmaşık biyolojik ve kimyasal sürecin çözümlenmesini gerektirmektedir. Bu doğrultuda; biyologlar, kimyagerler ve mühendisler, AÇ sistemlerinin bu karmaşık doğasını anlamlandırabilmek için yıllar içerisinde ileri düzey matematiksel modeller geliştirmişlerdir.

2.2.1 Atık Su Arıtma Modellemesine Giriş ve Önemi

ATT modelleme çalışmaları, tesiste meydana gelen tüm reaksiyonlara ve hidrolik olaylara matematiksel denklemlerle bir anlam getirmeye yönelmektir. Bu modeller sadece teorik çalışmalara bir altlık teşkil etmemiş, pratik uygulamalara da yönelmiştir.

Literatür araştırmalarında modelleme yaklaşımlarının önemine şu yönlerden değinilir:

- Tesis Tasarımı ve Ölçeklendirme : Yeni inşa edilecek tesislerin doğru bir şekilde yapılması veya mevcut tesislerin revizyonunun inşasının bilimsel bir yaklaşımla yapılması açısından modelleme gereklidir [47].
- Optimizasyon: Enerji ve arıtma verimliliğinin maksimize edilmesi için, işletme parametrelerinin (sisteme verilmesi gereken hava miktarı, geri devir oranları vd.) değişen tesis giriş değerlerine göre ayarlanması [48].

- Otomatik Kontrol: Modelleme yaklaşımları, SCADA sistemi bağlamında ve simülasyonlarla proseslerin otomatik olarak kontrol strateji geliştirilebilmesinde de önemlidir (örneğin çözünmüş oksijen kontrolu) [49].
- Araştırma ve Teknoloji Geliştirme: Her yeni modelleme çalışmaları bir sonraki çalışmalara altlık teşkil etmekte ve sürekli bir güncelleme sağlanmaktadır. Modelleme yaklaşımlarının aslın önemi aslında buradadır. Sistemin karmaşıklığı, gelişen teknolojinin iştahını buraya yönlendirmektedir.

2.2.2 Atık Su Arıtmada Modelleme Yaklaşımlarının Sınıflandırılması

AAT modelleme yaklaşımları temelde, neyi amaçladıklarına ve yaklaşımlarına göre sınıflandırılırlar. Bunlar:

- Mekanik Prensiplere Dayalı Kinetik Modeller (Beyaz Kutu): Sistemin fiziksel olarak nasıl işlediğini, kütle korunumunu, enerji korunumunu ve akışkanlar hidroliğini gibi konuları açıklar. Bunlara bağlı olarak da proseslerin hızını ve işleyişini nelerin kontrol ettiğini, reaksiyonların hızını vs. tanımlamaya odaklanır. Beyaz kutu modelleri yeni bir AAT'nin projelendirilmesinde ve konfigürasyonunda kullanılır. Tamamen ölçülebilir, hesaplanabilir matematiksel modellerdir [50]. IWA tarafından tarihsel süreç içerisinde ortaya konan ASM'ler bunlara en güzel örneklerdir. Bunların haricinde, çamur çökelme modelleri, anaerobik çürütme modelleri (ADM), oksijen transfer modelleri ve hidrolik modeller gibi deterministik modeller bu gruba girer.
- Veri Tabanlı Modeller (Kara Kutu): Bu modeller, sürecin nasıl işlediğini irdelemeden sadece giriş ve çıkış değerleri arasında bir korelasyon ile çıkışı tahmin etmeye çalışır. Belirli bir alanda uzmanlık gerektirmeden sonucu hızlıca bulmaya çalışan modellerdir [51]. Çoğu AAT inşa edilirken, yüksek düzeyde belirsizlik ve doğrusal olmayan ilişkileri içeren AÇ sistemi tercih edilmektedir [52]. Kara kutu modelleme birden fazla değişkenin etkilediği, aralarında doğrusal ilişki olmayan süreçleri modellemede oldukça etkilidir ve bu onu AAT'leri gibi oldukça karmaşık yapıları analiz etmede başarılı bir model yapar [53]. En yaygın kullanılan kara kutu modeller; çıkış parametrelerini, birden fazla giriş parametresinin lineer kombinasyonu olarak tahmin eden çoklu lineer regresyon (MLR) ve çok sayıda ilişkili değişkeni birbirleri ile ilişkisiz daha az temel bileşene indirgeyen temel bileşen analizi (PCA) gibi istatiksel modeller ile

YZ/MÖ modelleridir [54,55]. En yaygın olarak karşılaşılan yapay zeka uygulamaları; yapay sinir ağları (ANN), destek vektör makineleri (SVM), rastgele orman (Random Forest), uzun kısa süreli bellek (LSTM) vb. modellerdir [56].

- Hibrid Modeller (Gri Kutu): Gri kutu modeller, beya kutu ve kara kutu modellerin dezavantajlarını minimize ederek maksimum avantaj sağlamayı hedefleyen bir modelleme yaklaşımını benimser. Bu modelin ana felsefesi; fiziksel ve hesaplanabilir kısımları hesapla, fakat karmaşık kısımları, veriye dayalı tahminlerle çözümlemek üzerine kuruludur. Kurulan model tamamen şeffaf veya tamamen kapalı değildir [57]. Burada şu şekilde uygulamalar olabilir. Bir beyaz kutu modelin çıktıları, bir kara kutu modelin çıktıları olabilir veya beyaz kutu model ve kara kutu model ayrı ayrı çalışır ve sonrasında tahminler birleştirilerek bir MÖ ile nihai sonuç oluşturulur. Bir başka hibrit model oluşturulma şeklide, fiziksel bir modelin kalbi olan verilerin anlık ölçülmesi veya kesinliği zor olan parametreler bir kara kutu modeli ile tahmin edilerek beyaz kutu modelini çalıştırır [58]. Bu şekilde en gerçek değere yaklaşılmaya çalışılır. Bu en yaygın kullanılan gri kutu modelidir. Gri kutu modelleme daha yüksek tahmin doğruluğu sunabilir. Kara kutuya göre daha az geriye ihtiyaç duyar ve mühendislik hesaplamalarının doğrultusuna uygun tahmin yapmaya çalışır. Bu sebeple açıklanabilirliği daha iyidir. Ancak hem deterministik hem de YZ modellemesinde uzmanlık gerektirir. Bu yönü ile aslında dijital ikizin geliştirilmesi noktasında çığır açıcıdır. Ancak literatürümüz henüz, MÖ ile yeteri kadar doygunluk ve kabul edilebilir yöntemler geliştirebilmiş değildir. Bunlar gerçekleştiğinde gri kutu modelleme dijital ikizin ana omurgasını oluşturacaktır. Bu çalışmada da söz konusu bütünleşik yapının kurulması amaçlanmaktadır.

2.2.3 Atık Su Arıtma Modelleme Tarihçesi

AAT modelleme çalışmaları, 20. yüzyılın ikinci çeyreğinden itibaren çevre mühendisliği alanında akademik düzeyde başlamıştır. Bu ilk dönem yaklaşımları genellikle ampirik veya yarı ampirik metotlara dayalı olarak yürütülmekteydi. Temelde, tesiste meydana gelen karmaşık reaksiyonların tamamı tam olarak anlaşılamadığından ve alıcı ortam kirliliğine etki eden temel faktör olarak sadece karbon görüldüğünden; modelleme çalışmaları Biyokimyasal Oksijen İhtiyacı

(BOİ) veya Kimyasal Oksijen İhtiyacı (KOİ) üzerinden gerçekleştirilmekteydi. Bu ilk modelleme çalışmaları da, söz konusu iki parametrenin giderilmesini basit kütle dengesi ve değişmeyen kinetik sabitler üzerinden hesaplamaktaydı.

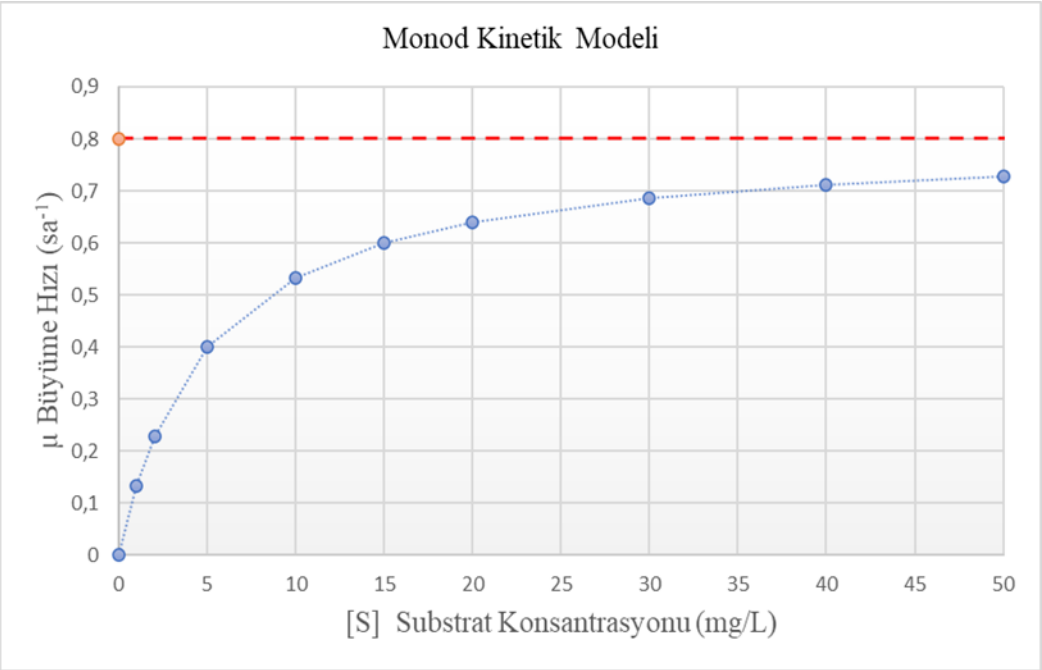
2.2.3.1 Kinetik Yaklaşımlar ve İlk Stokiyometrik Modeller (…-1970'ler)

AÇ prosesinin matematik modeli, ilk kez 1949 yılında Fransız mikrobiyolog Lacques Monod tarafından ortaya konmuştur [59]. Monod, bakteri büyüme hızı ile AÇ'deki besin miktarına (substrat) dayanan bir denklem (2.1) ve bu denkleme bağlı olarak, Şekil 2.1'de gösterilen bir doygunluk eğrisi ile çalışmasını neticelendirmiştir.

$$\mu = \mu_{max}\left(\frac{S}{K_s + S}\right) \tag{2.1}$$

μ : spesifik büyüme hızı
μ_{max}: maksimum büyüme hızı
S : substrat konsantrasyonu
K_S : Yarı doygunluk sabiti

Monod'un bu çalışması mikrobiyal büyümenin matematiksel modellemesinin başlangıç noktasını oluşturmaktadır. Bu yaklaşım; atık su arıtımı, biyoreaktör tasarımı ve fermantasyon mühendisliği gibi birçok alanın günümüzdeki teorik temelini teşkil etmektedir.



Şekil 2.2 Monod kinetik modeli

Monod sonrasında Amerikalı akademisyen W.Wesley Eckenfelder 1950'lerin sonunda gerçekleştirdiği bir dizi çalışma ile AÇ kinetiğini geliştirmiş; oluşturduğu matematiksel modellerle bu süreci mühendislik tasarımına uyarlamıştır [60]. Eckenfelder substrat giderim hızının, ortamdaki organik madde konsantrasyonu ile açıklanabileceğini ortaya koymuştur. Bu doğrultuda, Monod kinetiği ile uyumlu olan substrat giderim hızını (2.2) numaralı denklemdeki şekilde tanımlamıştır. Bu matematiksel model ile mühendislik uygulamalarında, hedeflenen bir BOİ konsantrasyonu için reaktör hacminin hesaplanmasına olanak sağlanmıştır [61].

$$\frac{dS}{dt} = -k\frac{XS}{K_S + S} \tag{2.2}$$

S : Substrat konsantrasyonu (BOİ, mg/L)

X : Askıda Katı Madde (AKM) konsantrasyonu (mg/L)

k : Maksimum substrat kullanım kızı

K_S : Yarı doygunluk sabiti

Eckenfelder çamur yaşı kavramının tasarımdaki önemini vurgulamış; çamur yaşı, BOİ giderim verimi ve çamur üretimi arasındaki ilişkiyi modellemiştir [61].

Bu dönemde modeller, tek bir kirletici konsantrasyonu üzerinden ele alınan; reaktöre giriş-çıkış miktarları ve reaksiyon hızı üzerinden hesaplama yapılan basit kütle dengelerinden ibaretti. Burada idealize edilmiş sürekli tam karışımlı bir tank kabulü yapılırken kinetik sabitler de buna göre belirleniyordu. Buradaki μ ve K_S gibi kinetik parametreler deneysel yöntemlerle ya laboratuvarlarda ya da pilot tesislerde hesaplanmaktaydı. Tesisin gerçek şartları ve değişken pH değerleri gibi dinamik durumlar ise göz önüne alınamıyordu.

Daha sonraki yıllarda, arıtma profesyonelleri yoğun bir şekilde modelleme konusuna yönelmişlerdir. Lawrence ve McCarty 1970 yılında, çamur yaşının AÇ tesislerinin tasarım ve işletilmesinde birleştirici bir parametre olduğunu ortaya koyan bir çalışma yayınlamışlardır. Bu çalışmanın özel amacı, Aç tasarımını kinetik denklemlerle birleştirerek hedeflenen çıkış suyu kalitesi için gerekli çamur yaşını ortaya koymaktı [62]. Yine bu dönemde modelleme ve tasarım bilgileri standart mühendislik ders kitaplarında artık yerini bulmaya başlamıştır.1972 yılında Metcalf&Eddy tarafından yayınlanan kitapta, atık su mühendisliğinin pratiklerinin ve kütle denge prensiplerinin, kinetik ile nasıl birleştirilebileceği detaylı olarak anlatılmıştır [63].

Lawrence ve McCarty in modeli kararlı durumu yeteri kadar karşılasa da, dinamik durumdaki tesisi davranışlarını karşılamıyordu. 1974 yılında John F. Andrews kararlı durum yaklaşımından sıyrılarak, tesisin operasyonel gerçekliğini yansıtan anlık, dinamik durumuna odaklanmıştır [64]. Andrews kurguladığı dinamik modelle, sadece substrat ile biyokütle arasındaki ilişkiyi değil, pH, sıcaklık ve çözünmüş oksijen gibi değişken parametrelerin de etkilerini reaksiyon hızına entegre etmiştir. Bu çalışma literatüre modelleme çalışmalarının sadece tesis boyutlandırmada ile ilgili bir konu olmayıp işletme operasyonları üzerinde de uygulanabilirliğine bir kapı aralamıştır.

Dold, Ekama ve Marais, 1980 yılında gelindiğinde bir başka önemli akademik gelişmeyi kaydetmişlerdir [65]. Bu makale kurucu bir makale niteliğinde bir çalışmadır. Karbon oksidasyonu, nitrifikasyon ve denitrifikasyonun tek bir AÇ sisteminde gerçekleşebileceğini ortaya koymuştur. Burada aerobik ve anoksik şartlardaki, heterotrofik ve ototrofik bakterilerin büyüme hızının monod denklemine dayalı kinetiği çıkarılmıştır. Bu çalışmada da kararlı durum değil dinamik duruma göre modelleme yaklaşımı benimsenmiştir. Dinamik şartlar

altında; substrat, biyokütle ve azot formları tek tek kütle denkliği denklemleri oluşturulmuştur. 1980 yılında geliştirilen bu model, "Cape Town Üniversitesi (UCT)" prosesi olarak anılmaya başlanmıştır. Bu çalışma daha sonraki yıllarda -aşağıda da anlatacağımız- IWA tarafından hazırlanacak olan ASM modellerine doğrudan temel olarak alınacaktır. Günümüzde kullanılan deterministik modellerin ilk temel taşıdır. Yine benzer bir çalışmayı 1976 yılında yapan Henze, nitrifikasyon bakterilerinin büyüme hızını modellemiştir [66]. Ancak Henze'yi literatürde asıl önemli kılan unsur; KOİ'yi çözünmüş/hızlı parçalanan, partiküler/yavaş parçalanan ve parçalanamaz olarak fraksiyonlara ayırma fikridir. Ayrıca yine 1980 yılında yapılan bir çalışma ile nitrifikasyon hızının atık sudaki çözünmüş oksijen konsantrasyonu ile olan bağımlılığı ortaya konulmuş ve kinetiği hesaplanmıştır [67].

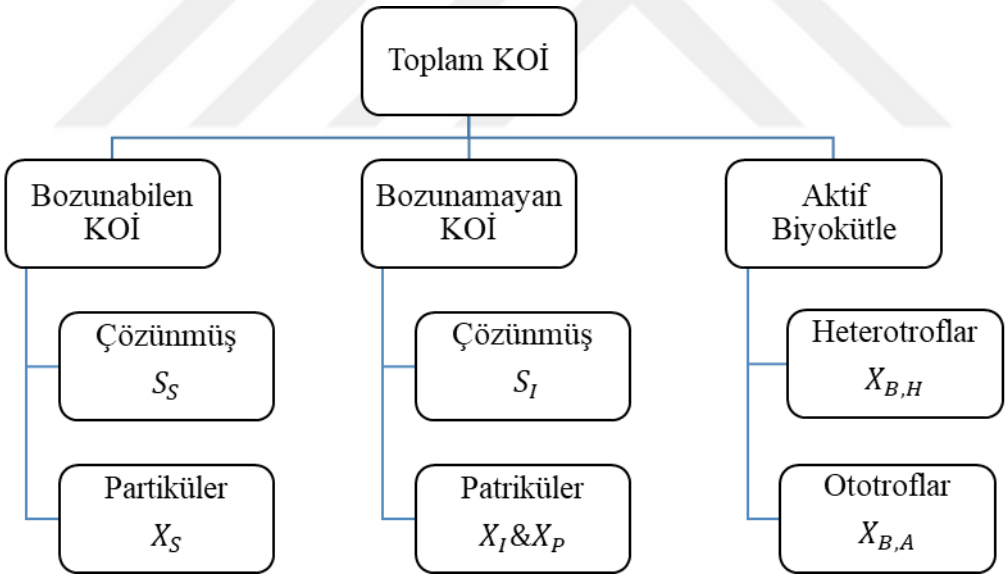
Literatürde, yukarıda temel yapı taşları olarak sunulan, Monod Kinetiği, Çamur Yaşı, Çevresel Dinamikler ve KOİ'yi fraksiyonlama gibi çalışmalar sonraki yıllarda IWA tarafından yapılacak olan deterministik modellemelere kaynaklık eden araştırmalar olmuşlardır.

2.2.3.2 Aktif Çamur Modellemede IWA ve ASM1-ASM3

1983 yılında Uluslararası Su Kalitesi Derneği (IAWQ) (şimdiki adı ile IWA), AÇ sistemlerinin tasarımında mühendislik açısından bir akademik standarda ulaşmak ve AAT işletilmesinin dinamiklerinin genel geçer şekilde anlaşılması için, yapılmış tüm çalışmaları özetleyen ve çerçeve oluşturmak amacıyla bir çalışma grubu oluşturdu. Bir diğer amaç ise karbon oksidasyonu, nitrifikasyon ve denitrifikasyon yapan tek çamurlu arıtma proseslerinin performansını tahmin etmek için en gerçekçi, ve en basit matematik modeli oluşturmak ve fikir birliğine varmak idi [68]. Bu çalışma grubu 1987 tarihinde, 5 yıl süren çalışmaları sonunda Aktif Çamur Model No.1 (ASM1)'i literatüre kazandırdı [69]. Model, karbon oksidasyonu, nitrifikasyon ve denitrifikasyonu ele almış ve oluşan toplam çamur miktarını tespit etmeyi öngörmüştür [70]. ASM1 de ele alınan iki önemli hipotez, "bisubstrat hipotezi" ve "ölüm-yenileme hipotezi (lizis)" UCT modelinden alınmış; hem heterotrofik hem de ototrofik bakterilerin büyüme hızları için de Monod kinetiği kullanılmıştır. Organik substrattaki elektron eşdeğeri, biyokütle ve kullanılan oksijen arasındaki bağlantı, kütle dengesi için KOİ seçilmiştir. Bu modelde KOİ de yapılan deneysel çalışmalar neticesinde hızlı-kolay ve yavaş-zor ayrışabilen KOİ

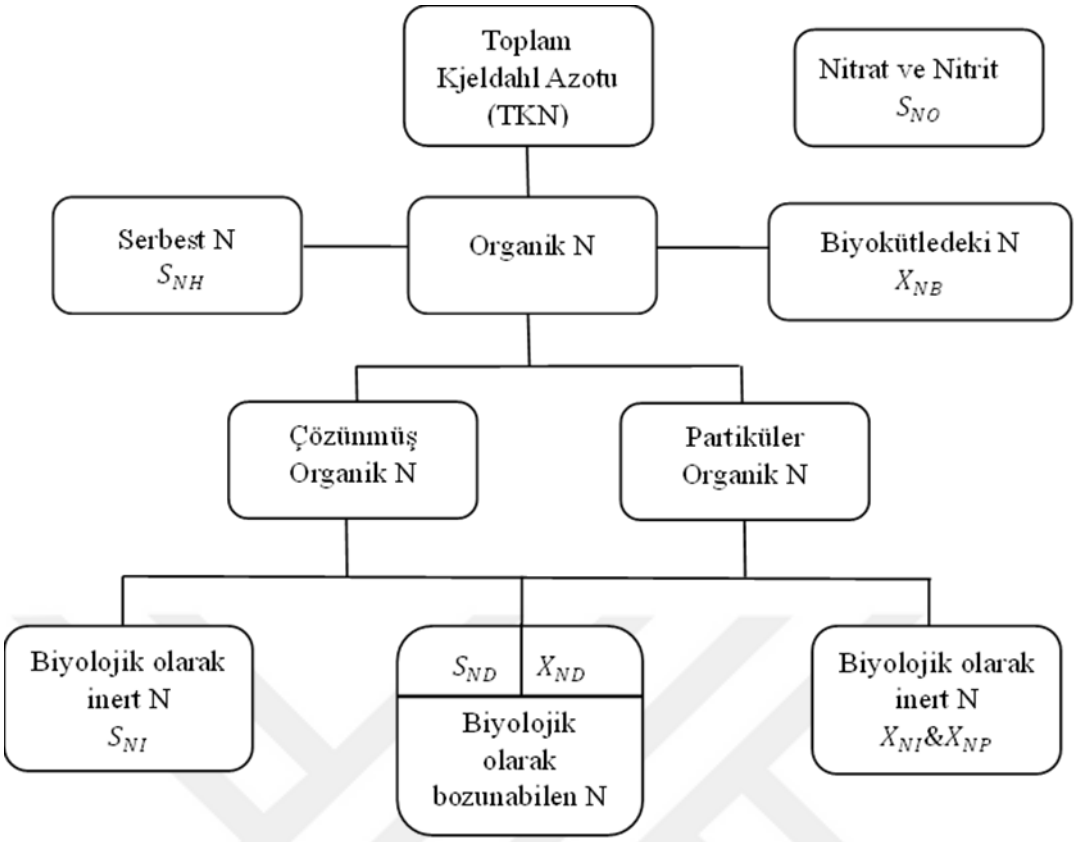
fraksiyonu olarak ikiye ayrılarak modele katılmıştır. Hücre duvarından kolayca geçebilen, çözünmüş substratın basit moleküllerden oluştuğunu ve daha hızlı olarak sentez reaksiyonlarına girdiği öngörülürken, daha büyük partikül halindeki moleküllerin de hücre dışı enzimlerle hidroliz reaksiyonları sayesinde besin olarak kullanılabildiği ortaya konulmuştur [68]. ASM1' de hem aerobik hem de anoksik şartlar altında tek bir lizis tanımlandı. Hem karbon hem de azot yönünden tek bir ölüm yenileme tanımlanması yapılamasının iki sebebi vardı: i) lizis denilen, ölen hücrelerin yeniden biyokütle oluşumunda kullanılmasının deneysel sonuçlara, besin- çamur dengesi açısından en iyi şeklide uyması, ii) heterotrof ve ototrofları ayırmamanın hesaplama süresini kısaltması idi [71].

ASM1 de karbon; substrat olarak biyolojik olarak bozunabilen (biodegradable), bozunamayan (non-biodegradable) ve aktif biyokütle olarak 3 kısımda incelenir Şekil 2.2) . Burada atık sudaki bozunamayan(inert) KOİ'nin çözünmüş kısmı (SI) sistemden değişmeden çıkarken, partikül kısmı (XI) aktif biyokütle ile sistemden uzaklaştırılmaktadır.



Şekil 2.3 ASM1 modeline göre karbonlu bileşenlerin fraksiyonu

ASM'de azot, Toplam Kjeldahl Azotu (TKN) olarak tanımlanmaktadır. Karbon bileşenleri gibi azotta, bozunabilen ve bozunamayan olarak Şekil 2.4'deki gösterildiği şekli ile sınıflandırılmaktadır.



Şekil 2.4 ASM1 modeline göre azotlu bileşenlerin fraksiyonu

Bu yukarıda gösterilen model bileşenlerinden başka, Çözünmüş oksijen konsantrasyonu (S_O) ve alkalinite (S_{alk}) de vardır. Bu şekilde ASM1 modelinde toplamda 13 bileşen 8 proses kinetiği ve 19 parametre içermektedir. Bunların tamamı tek bir tablo halinde Tablo 2.1 de verilmiştir [72].

1987 yılında yayımlanan bu ilk makaleden günümüze kadar geçen süre içerisinde akademik dergilerde bu çalışmaya yaklaşık 800 ila 900 defa atıfta bulunulmuştur. Yayınlandığı ilk yıllarda da aşağıda belirtilen kabul ve kısıtlar nedeniyle çeşitli eleştiriler almıştır:

- Sistemde sabit sıcaklıkta çalışıldığı varsayılmıştır.
- pH değeri sabit ve nötre yakın olduğu kabul edilmiştir.
- Organik maddenin yapısındaki değişimler dikkate alınmamıştır.
- Organik madde giderimi ve hücre büyümesi üzerine, azot, fosfor ve diğer nütrientlerin kısıtlılığı nedeniyle meydana gelebilecek etkiler göz ardı edilmiştir.
- Nitrifikasyon için belirlenen katsayıların, ortam şartlarına göre değişmediği varsayılmıştır.

- Heterotrofik hücrelerin reaktör içinde homojen dağıldığı kabul edilmiş ve zamanla bakteri türlerinin değişmediği de bu kabule eklenmiştir.
- Partiküler halde bulunan substratın çamurda tutulmasının anlık olduğu kabulü ile tamamı hesaba katılmıştır.

1999 yılına gelindiğinde yine IWA tarafından ASM1 in revizesi yapılmış ve ASM3 modeli kamuoyu ile paylaşılmıştır. Bu çalışmada biyokütle bozunum kinetiği ve substrat tüketimindeki basitleştirmeleri giderme amacı güdülmüştür [73]. ASM1'e benzer şekilde karbon oksidasyonu, nitrifikasyon ve denitrifikasyon prosesi ele alınmış, çamur üretim miktarı hesabı katılmıştır. ASM3, ASM1'den temel düzeyde iki farklı yaklaşımı sunar:

- ASM1 AÇ'daki biyokütle azalmasını ve bozunmasını, tek bir sistem olan "ölüm-yenileme (lizis)" ile açıklamıştır. Ancak burada biyokütle içerisinde depolanan substratın kullanılması ile hücrenin fiziksel olarak parçalanmasını ayırt edememektedir. Bu durum, sistem uzun çamur yaşlarında çalışırken çamur üretim veriminde hatalara sebep olur. Ancak ASM3 de endojen(içsel) solunum tanımlaması ile bunun önüne geçilmeye çalışılmıştır. Burada modele dahil edilen konu; düşük substrat konsantrasyonlarında biyokütle depolanmış substratı kullanarak hayatta kalır [73].
- ASM1 de, çözünür KOİ, biyokütle tarafından doğrudan büyüme için kullanılır kabulü var iken, ASM3'te substratın iki aşamalı olarak tüketildiği kabul edilir. Biyokütle substratı önce hızlıca biyokütle içine alıp depolamakta, sonrasında büyüme için kullanmaktadır. Bu yaklaşımın; deneysel değerlerle hesaplanan kinetik dinamiklerin ASM3 ile daha iyi temsil edildiğini ortaya koyduğu görülmüştür [74].

ASM1 ve ASM3 modelleri, atık su arıtımında karbon ve azot giderimi üzerine kurulmuştur. Burada organik maddelerin oksidasyonu ve Nitrifikasyon-Denitrifikasyon (N-DN) prosesleri vardır. Bu modeller temelde iki aşamalı olan nitrifikasyon sürecini ($NH_4^+ \rightarrow NO_2^- \rightarrow NO_3^-$) basitleştirilmiş matematiksel işlemler için tek aşamalı olarak gösterir [75]. Tablo 2.2 de ASM1 ve ASM3 modellerindeki çözünmüş, partiküler formdaki durumlar için biyokimyasal süreçler verilmiştir.

ASM3; özellikle AÇ sisteminde bolluk-kıtlık durumlarında yaşanan dinamik yükleme farklılıklarında daha iyi tahmin yapmak üzere geliştirildi. Bu teorik olarak çok doğru çalışsa da kalibrasyon karmaşıklığını artırmıştır. Literatürde yapılan en yaygın eleştiri bu yapısal tanımlanabilir olma sorunudur. Şöyle ki; ASM3 de substratın önce bakteriler tarafından hücre içinde depolandığı, sonrasında büyümede kullanıldığı varsayımı üzerine modelleme yapılmıştır. Bu doğru bir yaklaşım olsa da Gernaey ayrı ayrı iki çalışmada bu durumun yani, depolama verimi ile büyüme veriminin ayırt edilerek model kalibrasyonuna eklenmesinin belirsizleştiğini ortaya koymuştur [76]. Bir diğer hususta ASM3 işlem kolaylığı için nitrifikasyonu tek basamaklı olarak açıklamış olduğunu belirtmiştik. Modeldeki bu kolaylık, NO_2 birikimini ve sera gazı olan N_2O'nin üretimini hiç hesaba katmamaktadır [77]. Ancak bu artık günümüzde en önemli sera gazı kaynaklarından biridir. Her ne kadar 2020 yılınsa yapılan bir çalışmada N_2O gazı tahmin etme hususunda başarılı bir çalışma yapılmışsa da bu defa kritik parametre sayısı nerede ise iki katına çıkmıştır [78].

ASM1 ve ASM3'ün en büyük eksiliği ve hesap dışına bıraktığı şey ise biyolojik fosfor giderimi (Bio-P) eksikliğidir ve hiç değinilmemiştir. ASM3'e özellikle fosfor giderimini eklemlemeye çalışan bir çalışmada, model yapısının çok karmaşık bir hale geldiği ve pH değişimini modellemek için ise birçok yeni kinetiklere gerek duyulduğunu ortaya koymuştur [79].

Tablo 2.1 ASM1 modeline göre karbon oksidasyonu, nitrifikasyon ve denitrifikasyon stokiyometrisi ve proses kinetiği [72]

Bileşen → i j Proses ↓	1 S_1	2 S_S	3 X_1	4 X_S	5 $X_{B,H}$	6 $X_{B,A}$	7 X_P	8 S_O	9 S_{NO}	10 S_{NH}	11 S_{ND}	12 X_{ND}	13 S_{ALK}	Proses Hızı ρ_i [$ML^{-1}T^{-1}$]
1 Heterotrofların aerobik üremesi		$-\frac{1}{Y_H}$			1			$-\frac{1-Y_H}{Y_H}$		$-i_{XB}$			$-\frac{i_{XB}}{14}$	$\mu_H\left(\frac{S_S}{K_S+S_S}\right)\left(\frac{K_{O,H}}{K_{O,H}+S_O}\right)X_{B,H}$
2 Heterotrofların anoksik üremesi		$-\frac{1}{Y_H}$			1				$-\frac{1-Y_H}{2.86Y_H}$	$-i_{XB}$			$\frac{1-Y_H}{14-2.86Y_H}-{}^{i_{XB}}/_{14}$	$\mu_H\left(\frac{S_S}{K_S+S_S}\right)\left(\frac{K_{O,H}}{K_{O,H}+S_O}\right)\times\left(\frac{S_{NO}}{K_{NO}+S_{NO}}\right)\eta_B X_{B,H}$
3 Ototrofların aerobik üremesi						1		$-\frac{4.57-Y_A}{Y_A}$		$-i_{XB}-\frac{1}{Y_A}$			$-\frac{i_{XB}}{14}-\frac{1}{7Y_A}$	$\mu_A\left(\frac{S_{NH}}{K_{NH}+S_{NH}}\right)\left(\frac{S_O}{K_{O,A}+S_O}\right)X_{B,A}$
4 Heterotrofların bozulması				$1-f_P$	-1		f_P		$\frac{1}{Y_A}$			$i_{XB}-f_P i_{XP}$		$b_H X_{B,H}$
5 Ototrofların bozulması				$1-f_P$		-1	f_P					$i_{XB}-f_P i_{XP}$		$b_A X_{B,A}$
6 Çözünmüş organik azotun amonifikasyonu										1	-1		$\frac{1}{14}$	$k_s S_{ND} X_{B,H}$
7 Tutunmuş organiklerin hidrolizi		1		-1										$k_h\frac{X_S/X_{B,H}}{K_X+(X_S/X_{B,H})}\left[\left(\frac{S_O}{K_{O,H}+S_O}\right)+\eta_h\left(\frac{K_{O,H}}{K_{O,H}+S_O}\right)\left(\frac{S_{NO}}{K_{NO}+S_{NO}}\right)\right]X_{B,H}$
8 Tutunmuş organik azotun hidrolizi											1	-1		$\rho_7(X_{ND}/X_S)$
Gözlenen Dönüşüm Hızı	$r_i=\sum_J r_{ij}=\sum_j \upsilon_{ij}\rho_i$													

Tablo 2.2 ASM1 ve ASM3 modellerindeki biyokimyasal prosesler

No	Proses	ASM1	ASM3
1	Tutulmuş organiklerin hidrolizi	X	X
2	Tutunmuş organik N hidrolizi		
Heterotrofik organizmalar, aerobik ve denitrifikasyon aktivitesi			
3	Biyolojik hızlı parçalanabilir (SS) substratın aerobik depolanması		X
4	SS'in anoksik depolanması		X
5	Aerobik büyüme	X	X
6	Anoksik büyüme (Denitrifikasyon)	X	X
7	Aerobik içsel solunum		X
8	Anoksik içsel solunum		X
9	Heterotrofik organizmalarca aerobik solunumla hücre içi depolama		
10	Heterotrofik organizmalarca anoksik solunumla hücre içi depolama		X
11	Heterotrof organizmaların ölümü	X	
Ototrofik organizmalar, nitrifikasyon aktivitesi			
12	Nitrifikasyon bakterilerinin aerobik büyümesi	X	X
13	Aerobik içsel sonunum		X
14	Anoksik içsel solunum		X
15	Ototrof organizmaların ölümü	X	
16	Çözünebilir organik N'in amonifikasyonu	X	

2.2.3.3 Aktif Çamur Modellemede IWA ve ASM2-ASM2d

ASM1, C ve N'u başarı ile modellemesine rağmen, alıcı ortamdaki ötrofikasyonun temel sebeplerinden biri olan fosforu modellememiştir. 1990'lı yılların başında deşarj standartlarının sıkılaştırılması ve Aşırı Biyolojik Fosfor Giderimi (EBPR) proseslerinin yaygınlaştırılmaya çalışılmasından sebeple yeni bir modelleme standardına ihtiyaç duyulmuştur. IAWQ tarafından 1995 yılında ASM2 modeli açıklanmadan önce de birçok biyolojik fosfor giderimi üzerine çalışmalar yapılmıştır. İlk dönem çalışmalar, prosesin biyokimyasal çözümlenmesi ve proses konfigürasyonu hakkında yapılan çalışmalar olmuştur. 1965 yılında ilk olarak mikroorganizmaların ihtiyaçlarının çok üzerindeki fosforu bünyelerine alabildiklerini ortaya koyan bir çalışma yapılmıştır. Burada, bu duruma lüks tüketim anlamına gelen (Luxury Uptake) bir tabir kullanılmıştır[80]. 1975 yılına gelindiğinde Fuhs ve Chen biyolojik fosfor giderimi için sisteme mutlak anaerobik bir hacim eklenmesi gerektiğini belirtip, prosesten sorumlu ana bakteri topluluğunun ise Acinetobacter olduğunu ortaya koymuşlardır [81]. Bu çalışmadan bir yıl sonra James Barnard, Bio-P için, atık suyun tesise ilk girdiğinde öncelikle anaerobik bir havuza alınması gerektiğini ortaya koyarak Bardenpho ve Phoredox proseslerini geliştirmiştir. Bio-P giderimini laboratuvar ortamından gerçek ölçekteki tesislere geçişin adımı atılmış olmuştur [82]. UCT grubu, 1983 yılında Bio-P deki anaerobik hacimde, fosfor salınımı ile KOİ alımı arasındaki stokiyometrik temelleri ortaya koymaya çalışmışlardır [83]. Yine aynı grup; 1989 yılında, fosfor biriktiren organizmaların (PAO) büyüme, fosfor depolama ve ölüm kinetiklerini Monod benzeri denklemlerle ifade ederek, ASM2'nin ilk adımını atmış oldular [84].

ASM2, ASM1'in Bio-P prosesi eklenmiş bir versiyonu olarak 1995 yılında IAWQ çalışma grubu tarafından yayınlanmış bir modelleme çalışmasıdır [85]. Bu çalışma ile AÇ modellemesine heterotrofik ve ototrofik biyokütlenin yanına PAO'larda eklenmiştir. Ayrıca, ASM2 de anaerobik koşullardaki fermantasyonda sürece dahil edilmiştir. Burada hızlı ayrışabilen substrat (S_F), fermantasyon sonundaki asetat vb. gibi ürünlere (S_A) dönüşür. Bu adım ile PAO'lar anaerobik koşullarda beslenebilirler. Modele göre, Bio-P prosesi şu şekilde işlemektedir: Anaerobik şartlarda PAO'lar hücre içindeki polifosfatlarını (X_{PP}) tüketerek enerji üretir ve bu ürettikleri enerji ile ortamdaki fermantasyon ürünü olan asetatı hücre içine alıp

depolarlar (X_{PHA}). Daha sonra aerobik koşullara alınan PAO'lar depoladıkları X_{PHA} oksitleyerek büyürken, atık su ortamında bulunan orto-fosfatları (S_{PO_4}), hücre içine alarak polifosfat olarak depolarlar. Bu şekilde atık su ortamındaki fosforu gidermiş olurlar.

ASM2'deki bir diğer geliştirme de sadece Bio-P prosesi ile fosfor giderimi modellenmemiş, ayrıca kimyasal olarak metal tuzları ile fosfor giderimi de modellenmiştir [85].

ASM2 de PAO'ların sadece aerobik koşullarda fosfor depolayabildikleri modellenmişti. Oysa ki yapılan bazı tam ölçek-laboratuvar çalışmalarında araştırmacılar oksijen tüketimi ve çamur üretiminin model uymadığını ve ölüm hızlarının modeldeki gibi sabit olmadığını ve daha yavaş gerçekleştiğini ortaya koydular [86,87]. Bu durum 1993 yılında, yani model çalışması yapılmadan iki yıl önce yapılan çalışmalarla ortaya konulmuştu. Kuba ve Kern anoksik koşullar altında, bazı PAO türlerinin oksijen yerine nitrat kullanarak fosfor depolayabildiklerini ispatlamışlardı [88,89]. IAWQ çalışma grubu bunun model üzerinde çok büyük etkisi olmayacağını düşünerek dışarıda bırakmışlardı. Ancak ortaya çıkan durum bir düzeltme gerektiriyordu.

Çalışma grubu ASM2 de bilinçli bir şekilde dışarıda bırakılan denitrifikasyon yapabilen PAO'ları da hesaba katan yeni bir çalışmayı 1999 yılında yayınladırlar. Modele de denitrifikasyona atfen ASM2d ismini verdiler [90]. ASM2d modeline şu iki evre: i) PAO'lar oksijen yerine nitratı kullanarak bünyelerindeki polihidroksialkanoatları (PHA) depolayabilmeleri, ii) PAO'ların anoksik koşullarda ortamdaki fosfatı hücre içine alıp polifostat olarak depolamaları, eklemlenmiş oldu. Bu eklemlemelerle AAT'lerdeki fosfor giderimi, gerçek tesis verilerine oldukça yaklaşan bir model ile ifade edilmiş oldu. ASM2 ve ASM2d ye ait biyokimyasal süreçler Tablo 2.3'de verilmiştir.

Tablo 2.3 ASM2 ve ASM2d modellerindeki biyokimyasal prosesler

No	Proses	ASM2	ASM2d
1	Aerobik hidroliz	x	x
2	Anoksik hidroliz	x	x
3	Anaerobik hidroliz	x	x
Heterotrofik organizmalar: X_H			
4	Kolay parçalanabilir substratın fermantatif büyümesi (S_F)	x	x
5	Asetatın büyümesi (S_A)	x	x
6	S_F ile denitrifikasyon	x	x
7	S_A ile denitrifikasyon	x	x
8	Fermantasyon	x	x
9	Lizis	x	x
Fosfor Biriktiren Organizmalar (PAO): X_{PAO}			
10	PHA'ların depolanması (X_{PHA})	x	x
11	Polifostatların aerobik depolanması	x	x
12	Polifosfatların (PP) anoksik depolanması		x
13	Anoksik büyüme		x
14	Aerobik büyüme	x	x
15	X_{PAO} lizisi	x	x
16	X_{PP} lizisi	x	x
17	X_{PHA} lizisi	x	x
Ototrofik (Nitrifikasyon) Organizmalar: X_{AUT}			
18	Aerobik büyüme	x	x
19	Lizis	x	x
Fosforun metal tuzları ile eş zamanlı çökelmesi			
20	Çökelme	x	x
21	Yeniden Çözünme	x	x

ASM2d, 2004 yılında yapılan bir çalışmada AÇ modelleme için "altın standart" olarak değerlendirilmiştir [91]. Ancak bu modelde ciddi itirazlara maruz kalmıştır. Bu itiraz sebepleri genelde bir takım biyokimyasal olayları fazla basitleştirmesi yahut tamamen yok sayılması etrafında toplanmaktadır. Şöyle ki:

- Glikojen biriktiren organizmaların (GAO) model dışı bırakılması; GAO'lar da PAO'lar gibi uçucu yağ asitlerini (VFA), besin kaynağı olarak kullanırlar. Atık su ortamında bu iki bakteriyolojik kültür rekabet halindedir. Oysa model sadece PAO'nun asetat tükettiğini varsaymıştır. Eğer ortamda GAO'lar baskın ise, model tesisin fosfor giderim verimini gerçekte olduğundan çok daha yüksek tahmin etmektedir [92].
- Hücre içi PHA ve glikojen depolama: PAO'lar anaerobik fazda hücre içine aldıkları VFA'yı PHA y dönüştürmek için enerji olarak hücre içerisinde depoladıkları glikojeni kullanırlar. Aerobik faza geçtiklerinde de PHA yı yakar, glikojen deposunu yeniler. Yani glikojen deposu dinamik bir döngü içerisindedir. Ancak ASM2d modeli, bunu tek bir seviyede kabul eder. Bu durum, dinamik yük değişimlerinde modelin tepkisini hatalı kılmaktadır [93].
- Sıcaklığa bağlı rekabet: Literatürde birçok çalışmada, 20^{o} C üzerindeki sıcaklıklarda GAO'ların PAO'lara karşı üstünlük sağladıkları kanıtlanmıştır [94]. ASM2d'nin bu sıcaklık etkisini yeterince yansıtmaması, özellikle değişken iklim koşullarında modelin güvenilirliğini kısıtlamaktadır.

2.2.4 Mühendislikte Standardizasyon: Alman ATV- DWA A 131

AAT'lerin güvenilir bir şekilde inşa edilmesi, verimli ve sürdürülebilir bir şekilde çalıştırılması ve kamu otoriteleri tarafından hem yapım hem de işletim koşullarının denetlenebilir olması için belirli standartlara uyulması zorunluluğu vardır. Özellikle biyolojik süreçlerin modellenmesinden sonra bu modele uygun mühendislik uygulamalarında standartlaştırılması elzemdir. Atık su artma tesisi yapımı işini yüklenecek ekosistemin; imalat sanayinin, teknik ve uygulamacı birimlerin uyacağı bir kriterler olmalıdır. İşte bu zorlamalar altında Almanya merkezli Alman Su, Atık Su ve Atık Derneği (DWA) tarafından yayımlanan ATV-A 131 standardı öncelikle yerel sonrasında evrensel bir standart olarak sunulmuştur [95].

Atık su arıtmanın başlangıç yıllarında, mühendislik uygulamaları yerel bilgiye, kısıtlı ampirik bilgilerin hacimsel hesaplamalara dönüştürülmesine ve geçmiş

tecrübelere dayanıyordu. ASM modelleri, biyolojik reaksiyonların ayrıntılı olarak analizlerini, kinetiğini modellese de reaktörlerin boyutlandırılması ihtiyacı devam etmiştir [96].

ATV standartlarının temel yaklaşımı karmaşık biyokimyasal reaksiyonlara dayalı kinetik süreçleri basit anlaşılabilir ampirik formüllere ve emniyet katsayılarına indirgeyerek ifade etmektir. Bu yaklaşım, standartlara göre inşa edilmiş tesislerin en kötü şartların oluştuğu senaryolarda dahi "kararlı durum (steady-state)"'u koruyarak deşarj limitlerini sağlamayı amaçlamaktadır. Yani; ASM'ler proseslerin "anlaşılmasına" imkan verirken, ATV standartları "proseslerin "inşa edilmesine" katkı sağlamaktadır. ATV standartları da üç ayrı tarihsel süreçten geçmiştir:

- ATV-A 131 (1991): 1991 yılında yayımlanan yönerge genel olarak karbon giderimine odaklansa da azot giderimi için, nitrifikasyon-denitrifikasyon süreçlerinde genel kabuller yapmıştır [97]. Uygulamaya konulduğu yıllarda AÇ mekanizması yeteri kadar çözümlenemediği için geniş ve büyük emniyet katsayıları kullanılmış; bu durum inşa edilen tesislerdeki havuzların hacminin gereğinden çok büyük olmasına sebebiyet vermiştir. Ne kadar geniş, o kadar yüksek emniyetli durum ve kolay sağlanabilir deşarj limitleri düşüncesi hakimdir.
- ATV-DVWK-A131 (2000) : Bu aslında genel dünyadaki diğer standartlar açısından da bir dönüm noktası olmuştur [98]. Bu standartta, ASM mantığı ile uyum sağlamak için hesaplamalara KOİ fraksiyonları da dahil edilmiştir. Ayrıca denitrifikasyon hacminin belirlenebilmesi için (V_D/V_{AT}) önceki versiyona göre çok daha kapsamlı ampirik hesaplamalar geliştirmiştir. Sıcaklık değişimlerinin nitrifikasyon kinetiği üzerindeki etkileri de irdelenmiştir.
- DWA-A 131(2016): Bu son ve güncel versiyonda enerji verimliği ön planda tutulmuş, fosfor giderimi ve geri kazanım detaylandırılarak incelenmiştir [99]. Ayrıca burada dinamik simülasyon araçlarının, statik olan bu hesaplamaların doğrulanmasında kullanılması önemli bir şekilde belirtilmiştir.

ASM'lerin dinamik değişimleri modellediğini, ATV'lerin de denge durumunu esas aldığını aşağıdaki Tablo 2.4 daha anlaşılır göstermek için hazırlanmıştır.

Tablo 2.4 ASM modelleri ile ATV standartlarının karşılaştırılması

Karşılaştırma Kriteri	ASM Modelleri	ATV(DWA) Standartları
Matematiksel Temel	Diferansiyel Denklemler	Cebirsel Formüller
Zaman Boyutu	Dinamik (Değişken)	Statik (Kararlı Durum)
Tasarım Mantığı	Prosesi Modellemek-Anlamak	Tesis İnşası – Deşarj Garantisi
Veri İhtiyacı	Yüksek (Kinetik katsayılar, detaylı atık su karakterizasyonu	Düşük (Kabuller ile belirlenmiş tasarım parametreleri)
Kısıtlar	Değişken Giriş Yükü – Sürekli Akış	Kritik Sıcaklık (T_{min}) ve Pik Yük (Q_{max})

2.2.4.1 DVWK-A 'nın Yenilikçi Yaklaşımları

ATV-A 131E'nin getirdiği en büyük yenilik, AAT'ni oluşturan her bir birimin (biyolojik rektörler, ön ve son çökeltim tankları vb.) bağımsız ünitiler olarak ele almayıp, bütünleşik bir sistem olarak ele almasıdır [30]. Eski metedolojilerde biyolojik reaktörler F/M oranı üzerinden tekil olarak hesaba konu edilip boyutlandırılırken, ATV, son çöktürme tankının katı madde yükleme oranı üzerinden hesap ederek sistemi geriye doğru tasarlar [96]. ATV'nin kendine has yeni tasarımsal yaklaşımını şu aşağıdaki gibi üç ana başlık altında toplayabiliriz.

1) Substrat/Biyokütle'den (F/M)'den Çamur Yaşı'na (SRT) Geçiş: Klasik tasarım yöntemlerinde havyz hacimleri, F/M oranı ile hesaplanıyordu. Ancak azot giderimi olan tesislerde bu yöntem biyolojik büyümeyi işin içine dahil etmede yetersiz kalmıştır. Karbon gideren bakteriler hızlı büyürken, azot gideren bakteriler yavaş büyürler bu sebeple hesaplara tam olarak dahil edilememiş olurlar. ATV ise nitrifikasyon bakterilerinin sistemden büyümeden atılmasını önlemek için bu bakterilerin maksimum büyüme hızını çamur yaşı limiti olarak kabul ederek giriş substrat yüküne göre sistemde tutulması gereken mikro organizma süresini hesaplayarak boyutlandırma yapar [30, 95].

2) Kritik Büyüme Hızı(μ_{max}) ve Güvenlik Faktörü (SF): ATv burada çamur yaşını şu aşağıdaki formül ile hesaplar (2.3):

$$SRT_{aerobik} = SF.\frac{1}{\mu_{max,T} - b_T} \quad (2.3)$$

Burada; $\mu_{max,T}$ sıcaklığa bağlı olarak ototrofik bakteriler maksimum büyüme hızıdır. b_T ise bakterilerin içsel solunum hızını ifade eder [95]. ATV ileride değineceğimiz gibi emniyet faktörünü biraz yüksek (SF=1,5-2,5) bir aralıkta tutarak güvenli alanda kalmayı tercih eder.

3) Biyolojik reaktör ve Son Çöktürme Tankının Entegrasyonu: ATV son çöktürme havuzlarını sadece hidrolik olarak ele almayıp biyolojik prosesin hacim belirleyicisi olarak değerlendirir. ATV, biyolojik reaktördeki konsantrasyonu, son çöktürme havuzunun alanı ve çamur hacim endeksi (SVI) ile ilişkilendirir [100]. Eğer çamurun çökelme karakteristiği kötü ise reaktördeki biyolojik konsantrasyonu düşürerek havuz hacmini arttırır. Bu şekildeki tasarım prensibi olası AKM kaçaklarının önüne geçmiş olur.

2.2.4.2 DVWK-A'nın Standartlarının Zayıf Yönleri ve Kısıtları

ATV (DWA) standartları, deşarj standartları açısından güvenli bölgede kalması ve uygulama ve hesap kolaylığı açısından AAT tasarım ve inşasında yaygın kullanılmaktadır. Ülkemizde, Su Kirliliği Kontrol Yönetmeliği ve AAT Teknik Usuller Tebliği gibi yasal mevzuatlarda ATV-A131E standardı kabul görülmüş mühendislik uygulamaları olarak geçmektedir. Ayrıca Altyapı İhale ve Yapım Yönetmeliklerinde "tesis tasarımı ATV-A 131 veya eşdeğer bir standarda uygun olmalıdır" gibi ibareler bulunmaktadır. Bizim bu çalışma kapsamında incelediğimiz AAT, bu standarda göre inşa edilmiştir. İleri ki bölümlerde detaylıca incelenecektir.

Literatürde ATV-A 131E ve benzeri statik tasarım standartlarına getirilen eleştiriler, aşırı boyutlandırma ve parametre hassasiyetleri ile ilgili olmaktadır. Bunu 4 alt başlık halinde şöyle özetleyebiliriz.

1) AÇ'in Dinamik Yapısını Yok Sayma: Biyolojik bir süreç olan aktif çamur hesaplamaları doğası gereği, dinamiktir ve küçük durum değişimlerinin etkilerini yansıtmaktadırlar. Ancak ATV tasarımı "kararlı durum (steady-state) varsayımına dayanarak yapılmaktadır. Tesiste meydana gelen hidrolik ve

organik yük dalgalanmalarının anlık etkilerini kabul edilebilir kılmak için "pik faktörler" kullansa da, bu karmaşık biyokimyasal süreçlerin doğasını basitleştirmiş olmaktadır [101]. Ayrıca nitrifikasyon ve denitrifikasyon arasındaki dinamik bağlamı; ATV, sadece hacimsel oran (V_D/V_{AT}) ile ifade etmiş, oksijen transferi ve alkalinitedeki değişimleri göz ardı etmiştir. 2010 yılında Makina, bu durumu, "kara kutu yaklaşımı" olarak tanımlayarak, statik standartların, tesisin limit koşullardaki durumlarındaki davranışını tahmin etmede başarısız olduğunu simülasyon programları ile göstermiştir [102].

2) Yüksek Güvenlik Faktörleri (SF) ve Aşırı Boyutlandırma (Oversizing): ATV, belirsizliklerin etkisi gidermek için yüksek SF kullanır. Literatür bu yaklaşım sebebi ile, tesislerin gereğinden daha büyük inşa edildiğini ve yatarım maliyetlerinin gereksiz arttırıldığını vurgulamaktadır [96]. Reaktör hacimlerinin, bu pik güvenlik faktörleri sebebi ile , yılın sadece çok kısıtlı zamanlarda düşük sıcaklık veya çok yüksek yük alacağı hesaba katılarak boyutlandırılmasının, tesisin çok düşük F/M oranları da verimsiz bir şekilde çalışmasına sebep olmaktadır. 2014 yılında IWA tarafından hazırlanan teknik raporda, kararlı duruma göre tasarlanan AAT'lerin, dinamik kontrol stratejileri ile işletilen tesislere göre %20-30 oranında daha fazla inşaat maliyetine katlanmak durumunda kaldıklarını Benchmark Simülasyon Modelleri ile ortaya konulmuştur [103].

3) Kinetik Katsayıların Yereli Yansıtmaması: ATV standartları mikroorganizmaların büyümesini ($Y_{H,max}$) ve çürümesini (b_H) karakterize eden katsayıları literatürden, varsayılan (default) değer olarak alır. Fakat tüm atık sular havzasının özelliklerini sergilediğinden farklıdır. İstanbul da yapılan bir çalışmada tüm havzalardaki farklı tesislere gelen giriş atık suyu karakterizasyonun farklılığı ortaya konmuştur [104]. Ayrıca 2012 yılında IWA tarafından yayınlanan teknik raporda da standart katsayılar kullanılarak yapılan AAT tasarımlarının, gerçek atık su karakterizasyonuna göre yapılan modellere kıyasla olması gereken durumdan önemli derecede saptığı vurgulanmıştır [105].

4) Enerji Verimliliği ve İşletme Optimizasyonu Kısıtı: Günümüz AAT işletmeciliği artık sadece kirliliği önlemeyi değil, enerjiyi verimli kullanmayı, karbon ayak izini, geri kazanımı öncelikleyen bir yaklaşıma doğru evrilmektedir. Ancak ATV gibi statik tasarım standartları, öncelik olarak limit aşımını engellemeyi kabul

ettiği için, enerji optimizasyonunu arka plana atmaktadırlar [106]. Bir diğer hususta, bu tür tasarımlarda havalandırma için yapılan blower kapasite hesaplarında pik yükler hesaba katıldığından, ihtiyaçtan fazla enerji tüketen bir havalandırma sistemi inşa edilmiş olmaktadır [107].

Nihayetinde; literatür ATV yaklaşımındaki statik kararlı durum tasarım standartlarının "tesis boyutlandırılması için güvenli bir rehber" olduğunu kabul etmekle birlikte, tesisin işletilmesi için bu deterministik kuralların ötesine geçilerek dinamik simülasyon, yapay zeka destekleri gibi araçların kullanılmasını önermektedir [108]. Sistemin doğrusal olmayan karmaşık yapısını öğrenebilen Yapay Zeka ve Makine Öğrenmesi algoritmaları bir sonraki işletim araçları olarak öne çıkmaktadır.

2.3 Dijital Simülasyon Araçları ve Literatür Uygulamaları

ASM modelleri biyolojik süreçlerin matematik olarak ifade edilmesinde çok önemli olsa da başlangıçta bu modellerin mühendislik uygulamalarına aktarılması pek mümkün olmamıştır. Bunun sebebi; bu modellerin çok sayıda bileşenden oluşması ve birbirine bağlı non-lineer diferansiyel denklem setlerini içermesidir. ASM'lerin dinamik durumunun zorluklarından kurtulmak için, pik faktör kabulleri ile kararlı durum varsayımı altında geleneksel tasarım standartları geliştirilmiştir. Ancak işletim koşulları yine dinamik şartlar altındadır. Tesis içindeki hidrolik geri devirler, biyokimyasal mekanizmalar ve AÇ'nin çökelme süreçleri dinamik koşullar altındaki karmaşık etkileşimlerimi manuel olarak hesaplayıp anlık sonuçlara ulaşmak çok büyük bir hesaplama yükü oluşturur [109].

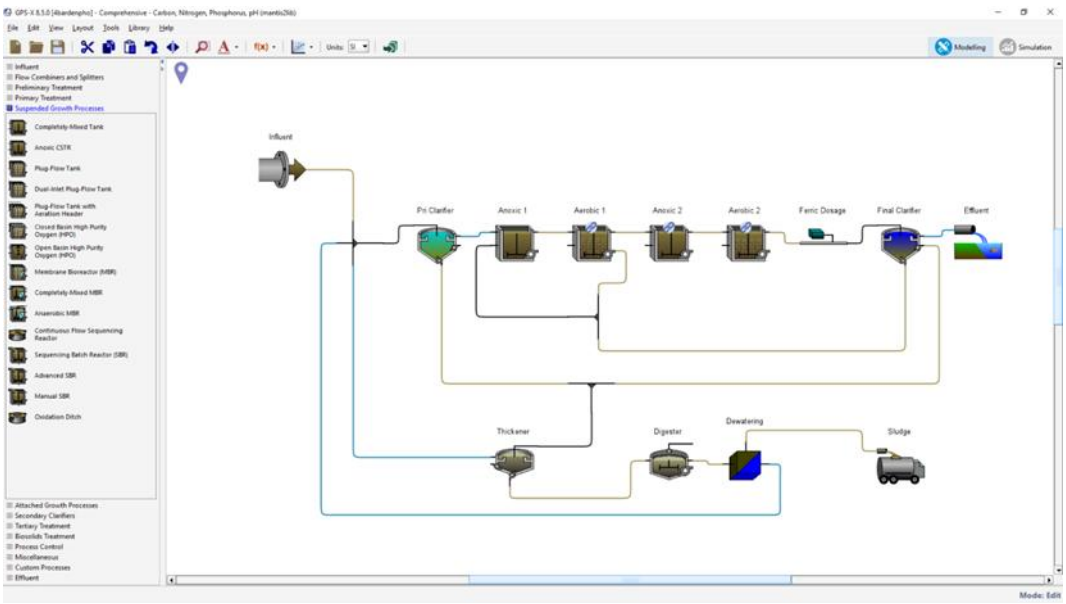
Bu sınırlamaları aşmak amacı ile deterministik modelleri çözümleyen, hidrolik ve biyokimyasal süreçleri bütünleşik olarak ele alan bilgisayarın işlem gücünden yararlanan kararlı durumu dinamik olarak simüle eden araçlar geliştirilmiştir. Veri işleme biliminin gelişimi ile birlikte; Endüstri 4.0 ve IoT teknolojilerinin tesislere uygulanması ile geliştirilen bu yazılımlar; tesisin mevcut işletim şartlarını, kapasitesine göre durumunu, işletme optimizasyonu için yapılması gerekenleri ve "ne olursa ne olur (what-if)" senaryolarını kolaylıkla yapmaya odaklanmışlardır. Duarte ve arkadaşları 2024 yılında yaptıkları bir çalışmada bu simülatörlerin başarısını vurgulamakla birlikte; yüksek kalibrasyon gereksinimlerinin araştırmacıları yapay zeka gibi veriye dayalı yeni yaklaşımlara yönelttiğini de

vurgulamışlardır [110]. Bu çalışma kapsamında, söz konusu deterministik modellerin temel yapısına kısaca değinilecektir.

2.3.1 GPS-X (Hydromantis)

GPS-X simülasyon programı; 30 yıla yakın geçmişi ile sektörün en eski programlarında biridir. Program, farklı işletme senaryoları ile tesisin işletilmesini ekonomik olarak optimum şartlara getirirken, deşarj standartlarının da korunmasını sağlayabilmektedir [111]. Program kendi kütüphanesindeki sürükle bırak proses kısımları ve gerçek tesisi değerlerinin girilmesi ile tesisin modelini oluşturabilmektedir. GPS-X; ASM1, ASM2/2d, ASM3 ile modelleme çalışmasına imkan verirken, ayrıca firma tarafından geliştirilen Mantis Aktif Çamur Modelini de barındırmaktadır. Bu program, what-if senaryolarına olanak tanıyarak, gerçekleşmemiş senaryoların olası çıktılarını da gösterdiğinden, manuel optimizasyon süreçlerini de desteklemektedir. Kullanıcı kolaylığı sağlayan ara yüzü Şekil 2.5 de görülmektedir [111].

2020 yılında yapılan bir çalışmada; tam ölçekli bir AAT'de havalandırma stratejileri üzerine yürütülen senaryo analizleri ile çıkış suyu kalitesinde limit aşımı yaşanmadan, enerji tüketiminde önemli düzeyde optimizasyon sağlandığı ortaya konulmuştur [112].

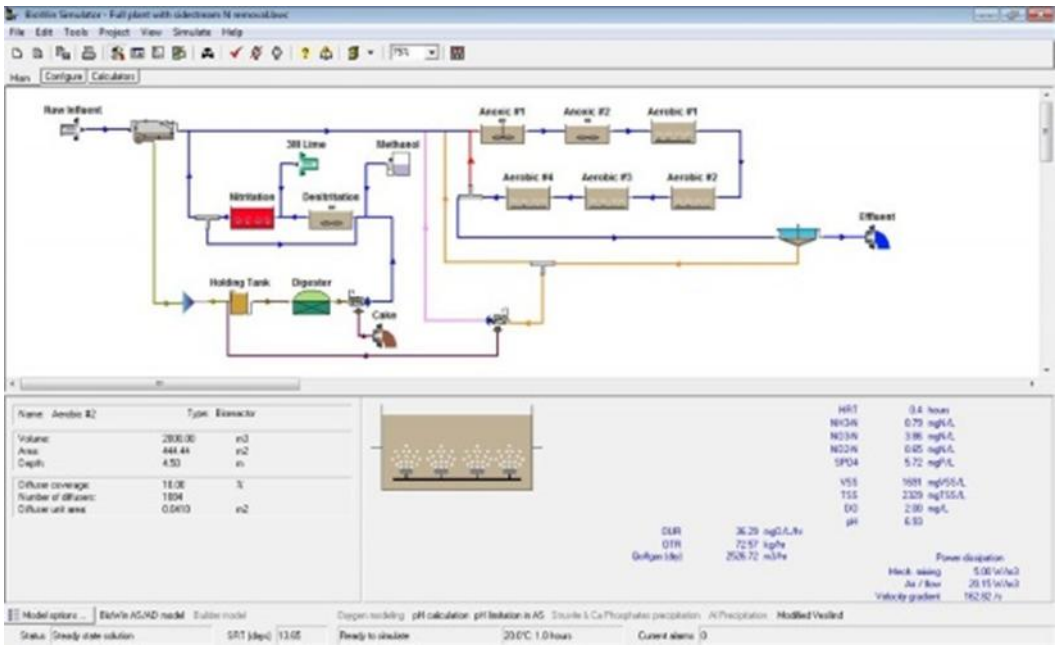


Şekil 2.5 GPS-X kullanıcı arayüzü (Kaynak:[111])

2.3.2 BioWin (EnviroSim)

BioWin; pH ve alkalinite gibi fizikokimyasal parametreleri, fiziksel ve biyolojik proseslerle (fosfor giderimi ve anaerobik çamur çürütmeyi de dahil ederek) birlikte başarılı şekilde değerlendirebilen bir simülatördür. Yazılım, "Aktif Çamur/Anaerobik Çürütme Modeli (ASDM)" yapısını kullanmaktadır. Ayrıca ASM1 ve ASM3'ü de kütüphanesinde bulundurmaktadır. Özellikle de Kuzey Amerika ülkelerinde bir simülatör uygulama standardı olarak kabul edilmekte ve yaygın olarak kullanılmaktadır [113]. Model arayüzüne ait gösterim Şekil 2.6'da verilmiştir.

2022 yılında yapılan bir çalışmada, pilot ölçekli olarak kurulan bir anaerobik reaktörün performans tahmini BioWin kullanılarak gerçekleştirilmiştir. Söz konusu çalışmada anaerobik reaktör; AÇ sistemlerine kıyasla daha az enerji tüketimi ve daha az çamur üretimi sağlaması nedeniyle tercih edilmiştir. Çalışmada yapılan kalibrasyonlar sonucunda, BioWin'in; KOİ ve AKM gibi temel parametrelerin giderim verimini tahmin etmede tutarlı olduğu ve sistemin tasarımında, analizinde güvenilir bir performans sunduğu belirtilmiştir [114].

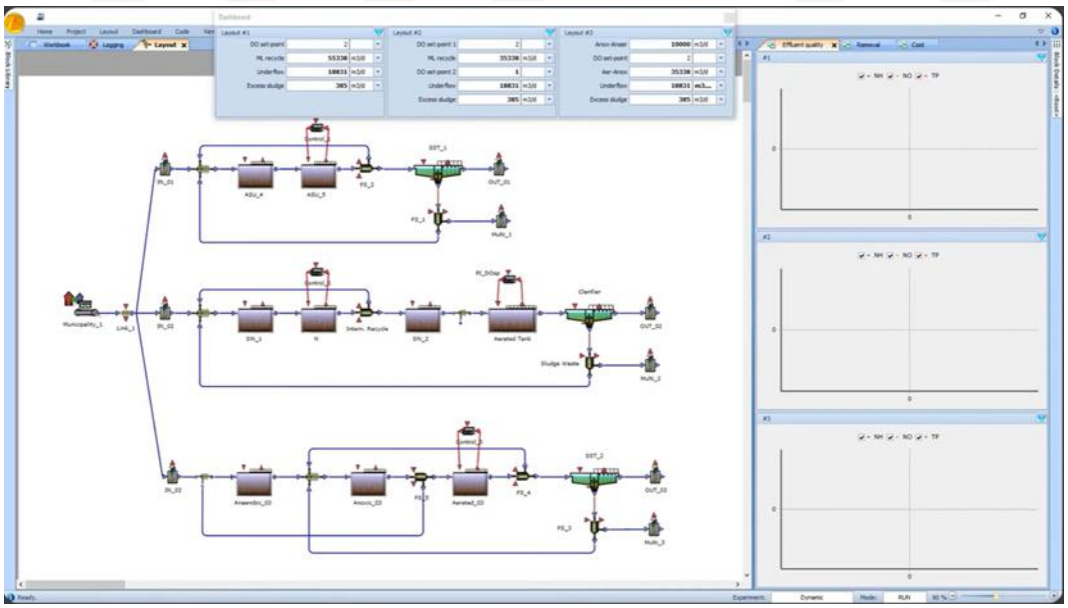


Şekil 2. 6 BioWin kullanıcı arayüzü (Kaynak:[113])

2.3.3 West (DHI)

WEST (Worldwide Engine fır Simulation, Training and Automation), model yapısının esnekliği sayesinde kullanıcısına geniş özgürlük alanı sağlayan AAT simülasyon yazılımlarından biridir. WEST kütüphanesinde standart olarak sadece ASM1 olmasına rağmen bulundurduğu "Model Specification Language (MSL)" tabanlı yapısı sayesinde istenilen kinetik denklemler sisteme tanımlanabilmektedir. Bu esneklik; yeni araştırma ve proses yapılarının kurgulandığı akademik çalışmalarda yazılımın daha çok tercih edilmesini sağlamaktadır [115]. WEST'e ait kullanıcı arayüzü görseli Şekil 2.7 de verilmiştir.

Danimarka da 2020 yılında yapılan çalışmada WEST tam ölçekli, ASM2d modeli taban alınarak tasarlanan bir AAT'de, simülatör olarak kullanılmıştır. Simülatöre dinamik istatistiksel denklemler de eklenerek, AAT'nin anoksik ve aerobik tankları ile son çöktürmeler tek bir hat olarak modellenmiştir. Sıcaklık, pH ve toplam AKM değerleri, değişen operasyonel şartlar altında 3 günlük sürekli izleme verileri ile uyumlu olarak tahmin edilebilmiştir. Sonuç olarak; simülatörün dijital ikiz olarak kullanılabilme ve dinamik şartlardaki belirsizlikleri giderme noktasındaki üstünlüğü ortaya konmuştur [116].

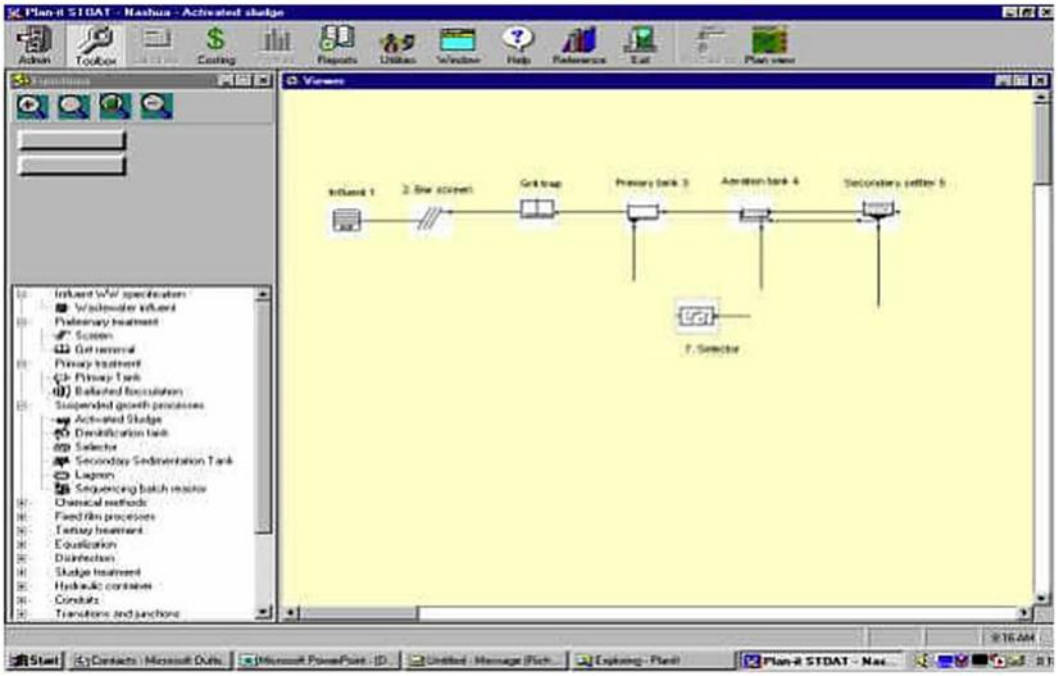


Şekil 2.7 WEST kullanıcı arayüzü (Kaynak:[115])

2.3.4 STOAT (WRc)

İngiltere'de WRc Grubu tarafından geliştirilen bu simülatör, barındırdığı avantajlar nedeniyle dünyada yaygın kullanılan bir yazılımdır. Özellikle AAT'leri sadece

proses bazlı modellemekle kalmayıp, AAT'ye gelen atık suların kaynağı olan havzadaki kanalizasyon şebekelerini de modelleme sürecine dahil eden entegre bir yapı barındırmaktadır.



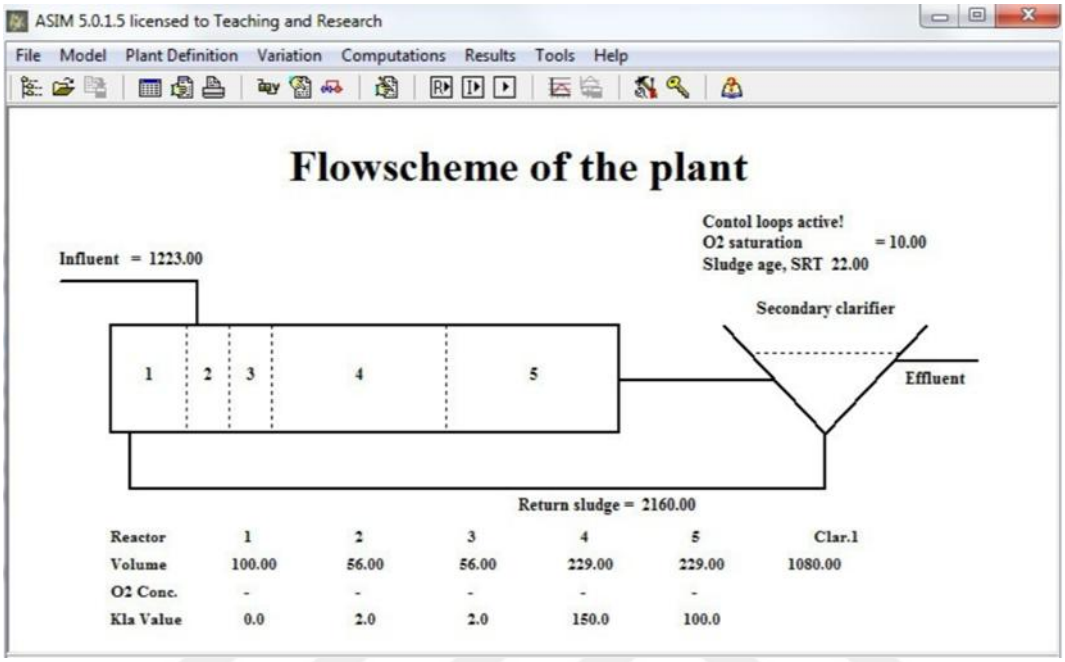
Şekil 2. 8 STOAT kullanıcı arayüzü (Kaynak:[117])

Tesislerin yağışlı havalarda değişen dinamik yüklerini ve taşkın olaylarını tesis hidroliği ile birlikte ele alıp modele dahil edebilen güçlü bir araçtır [117]. Bu bağlamda 2016 yılında yapılan bir çalışmada STOAT kullanılarak kanalizasyon şebekesi ile tesis hidroliği arasındaki etkileşimi modellenmiş ve taşkın durumlarının artıma verimine olan etkisi simüle edilmiştir [118]. Şekil 2.8 de kullanıcı arayüzü görseli verilmiştir.

2.3.5 ASIM (EAWAG)

ASIM İsviçre'deki EAWAG (Su Araştırmaları Enstitüsü) da görev yapan Prof.Dr.Willi Gujer ve ekibi tarafından 1980 li yılalrda geliştirilmeye başlnamıştır. IWA ekibi tarafından da kullanılan simülatör, ASM modellerinin ilk uygulandığı yazılımdır. Matris tabanlı model düzeni olan bu ilk uygulama aracı, tam ölçekli tesislerden ziyade, biyokinetik modellerin matris yapısının öğretilmesi ve model kalibrasyon yaklaşımının benimsetilmesi amacı ile genellikle akademik çevrelerce eğitim aracı olarak kullanılmaktadır. Açık kaynak kodludur, içerdiği kaynak kodlarına ekleme yapılarak model bileşenlerinin değiştirilmesine izin veriri. Diğer

ticari model yazımlarının aksine arayüzü sadece ve daha basit görülebilir. Ancak çok iyi bir model öğrenme aracıdır. Kinetik katsayıların etkilerini öğrencilerin anlaması daha kolaydır [119]. ASIM'e ait bu basit ara yüz Şekil 2.9 da görülmektedir.

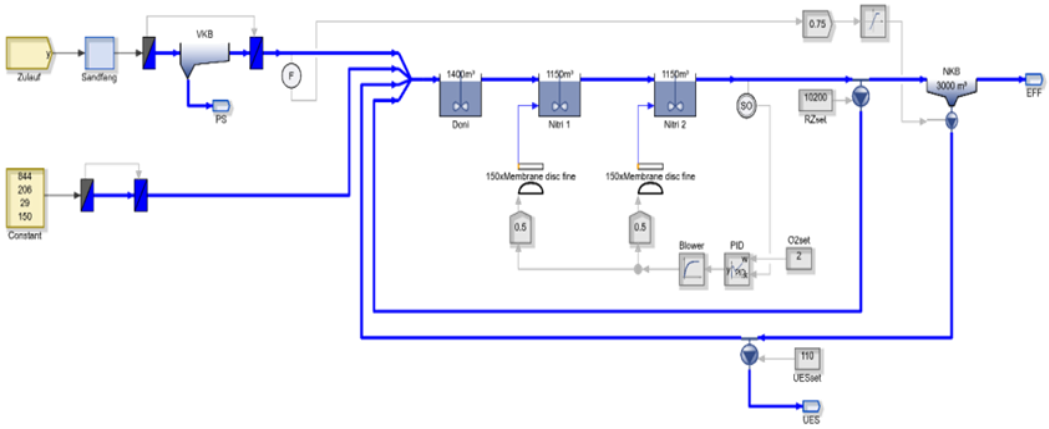


Şekil 2. 9 ASIM kullanıcı arayüzü (Kaynak:[119])

Gujer, bu matris tabanlı programın 1995 yılında yayımlanan bir makale ile akademik çevre tarafından kolayca kullanılabileceğini ve ASM'lerin modellenmesindeki güçlü hesap kabiliyetini belirtmiştir [120].

2.3.6 SIMBA (IFAK)

SIMBA; Alman IFAK Enstitüsü tarafından 1994 yılında MATLAB/Simulink eklentisi şeklinde geliştirilmiştir. 2013 yılına gelindiğinde ise artık bağımsız bir platformda simülasyon yazılımı olarak şekillendirilmiştir. Ancak geliştirildiği MATLAB mimarisi devam ettiği için SIMBA; sadece klasik bir biyolojik proses simülatörü değil, aynı zamanda gelişmiş kontrol ve otomasyon sistemlerinin test edilebildiği bir laboratuvar niteliğindedir. Bünyesinde SCADA arayüzünü de barındırmaktadır. Bu simülasyon modeli ile havzadaki kanalizasyon hatları ile birlikte alıcı ortam şartları da izlenebilmektedir. En büyük dezavantajı; MATLAB altyapısı gerektirdiğinden daha fazla profesyonellik ve daha yüksek lisans ücreti gerektirmesidir. SIMBA'ya ait kullanıcı arayüzü Şekil 2.10'da verilmiştir [121].

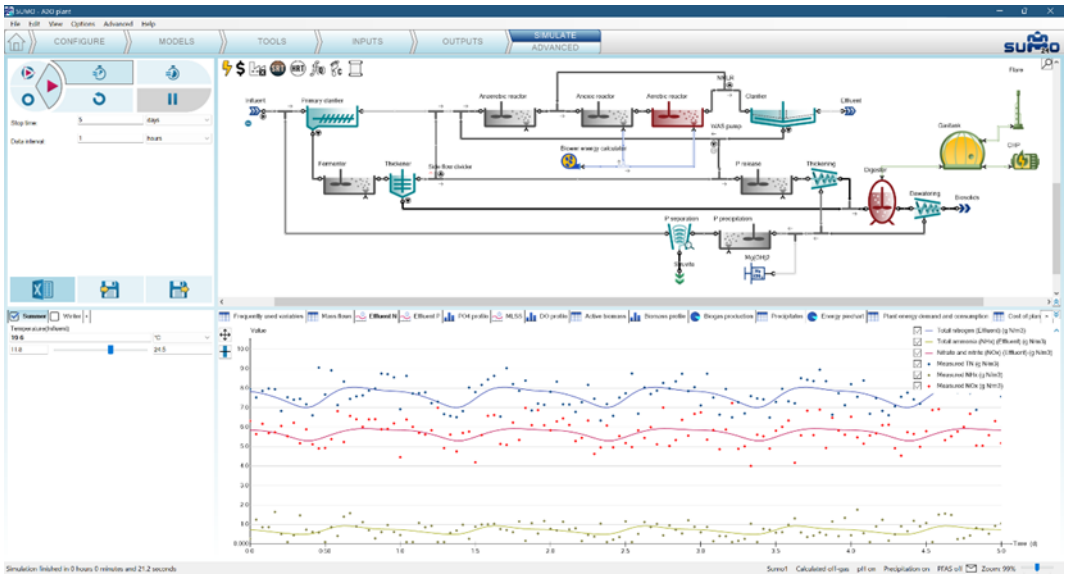


Şekil 2.10 SIMBA kullanıcı arayüzü (Kaynak:[121])

2004 yılında yapılan bir çalışmada; kanalizasyon şebekesi ve arıtma tesisi arasındaki etkileşimi gerçek zamanlı olarak simüle etme yeteneği ortaya konulmuştur. Bu bağlamda, SIMBA'nın otomasyon stratejilerini geliştirmede önemli bir araç olduğu öne sürülmektedir [122].

2.3.7 SUMO (Dynamita)

SUMO; İsviçre merkezli bir ticari firma olan Dynamita tarafından 2000'li yılların başında geliştirilmeye başlanmıştır. Yazılım, özellikle 2010'lu yıllardan sonra açık kaynak kodlu olarak sunulmaya başlanmıştır. Açık kaynak kodlu yapısı, yazılım dili olan Python ile desteklenmektedir. Diğer simülasyon programlarına göre daha hızlı çözümleme yapmayı amaçlayarak piyasaya çıkmıştır. Açık kaynak kodlu olması, yeni gelişen proseslerin SUMO'ya kullanıcı tarafından uyarlanmasını ve modelin genişletilmesini sağlamaktadır. Bu özelliği ile hem atık su profesyonellerinin hem de araştırmacıların ilgisini çekmektedir. Programa ait kullanıcı arayüzü Şekil 2.11'de verilmiştir [123].



Şekil 2.11 SUMO kullanıcı arayüzü (Kaynak:[123])

2022 yılında İnsel ve diğerlerinin yaptığı bir çalışmada; tam ölçekli, ATV-A 131E tasarım kriterlerine göre inşa edilmiş bir AAT'de, gerçek operasyonel veriler ile ASM yaklaşımından sapmaları tespit etmek için bütüncül bir yaklaşımla SUMO simülasyonu kullanılarak tesis izlenmiştir. Bu çalışma sonucunda tesisin karbon (C), azot (N) ve fosfor (P) giderimi ile biyogaz üretimi performansları değerlendirilerek literatürle karşılaştırılmıştır. Bu değerlendirme kapsamında; laboratuvar çalışmaları ile öncelikle atık suyun karakterizasyonu ortaya konulmuş, sonrasında tesise ait kinetik katsayılar SUMO programına aktarılmıştır. Çalışma neticesinde, söz konusu tesisin nitrifikasyon hızının literatüre kıyasla %30, anaerobik hidrolizinin ise %70 daha düşük olduğu SUMO simülasyonu ile ortaya konulmuştur [124].

2.4 Deterministik Modelleme Yaklaşımlarının Sınırları ve Dijitalleşme ile Birlikte Yapay Zekaya Geçiş: IWA Vizyonu

AAT'lerin tasarımı, işletmesel kontrolü ve optimizasyonu; IWA tarafından geliştirilen ASM tabanlı modeller ile bu modeller baz alınarak hazırlanan simülasyon programları (GPS-X, SUMO, STOAT vb.) kullanılarak yapılmaktadır. Bu yaklaşımların tamamı "beyaz kutu" modeller olup; AÇ sistemlerindeki biyokimyasal reaksiyonları stokiyometrik ve kinetik denklemlerle açıklayarak, her bir prosesteki mekanizmaları fizikokimyasal olarak modellemek için bir dizi matris çözümlemesini kullanır. Bunların tamamı; AAT tasarımı, akademik araştırmaların

sürdürülmesi ve mekanizmaların anlaşılabilmesi için vazgeçilmez bir değer taşımaktadır. Ancak modern AAT'lerin dinamik yapıları ve Endüstri 4.0 dönüşüm gereklilikleri karşısında bazı sınırlamaları bulunmaktadır. Bu sınırlamalar aşağıda maddeler halinde şöyle özetlenebilir:

1) Karmaşık parametreler ve kalibrasyon sorunu: Deterministik modeller, her tesis için hassas kalibrasyon gerektiren çok sayıda parametre içerir. Literatürde "aşırı parametreleştirme" (over-parameterization) denilen bu durum; yaklaşımın teorik olarak başarılı olmasına rağmen, sahadaki uygulamalarda bariz sapmalara sebep olabilmektedir. Her tesis için gerekli olan fraksiyon analizleri hem maliyet hem zaman kaybı yaratmakta hem de her şartta sürdürülebilir olmamaktadır. Özellikle küçük tesisler için bunu yapmak verimsizdir [103].
2) Lineer ve stabil olmayan durumlar ve belirsizlikler: AAT işletme şartları; sıcaklık değişimleri, öngörülemeyen yağış rejimi farklılıkları (iklim değişikliği baskısı), debi dalgalanmaları ve havzadaki endüstriyel deşarj kaynaklı yük öngörülemezliği sebebiyle doğrusal olmayan ve kaosa yatkın şartlardır. Bilinen simülasyon uygulamaları "kararlı durum" (steady-state) varsayımı altında çalışmakta; tesiste meydana gelen filamentli bakteriler sebebiyle çamur kabarması gibi biyolojik anormal durumlarda doğru tahmin yeteneğini kaybedebilmektedir [108].
3) Uzmanlık Gereksinimi: Deterministik modellerin tamamı, AAT işletme profesyonellerinin özel bir eğitimle öğrenebilecekleri uygulamalardır. Her tesiste bu uygulamaları kullanacak personel yetiştirmek hem zahmetli hem ekonomik olmamaktadır. Saha operatörlerinin bu eğitimi alabilmeleri için akademik yetkinlikleri her zaman mümkün olmadığından (özellikle küçük tesislerde) "uygulama boşluğu" meydana gelmektedir [125].

Yukarıda bahsedilen bu metodolojik sıkışmışlık, sadece akademik bir yaklaşımın tanımlaması değildir. IWA'nın kendisi de bu durumu belirleyerek ele alınması gereken bir zorunluluk olarak ortaya koymuştur. Son on yıla yakın süredir bu konuda bir perspektif sunmaya çalışmaktadır. Tablo 2.5 de IWA'nın sürece yaklaşım şekli ve tarihçesi verilmiştir.

Çalışmamızın yürütülme sebeplerinden biri de olan, IWA'nın 2020 yılında yayımlamış olduğu "Su Sektörü İçin Yapay Zeka Çözümleri" deklarasyonu,

sektördeki dönüşümü açıkça göstermektedir [129]. Bu rapor, 2019 yılında ortaya konulan "Dijital Su" kavramının devamı niteliğindedir [128]. Bu belgede IWA; yukarıda sıraladığımız deterministik yöntemlerin açmazlarını aşmak için Yapay Zeka (YZ) ve Makine Öğrenmesini (MÖ) kullanmayı bir paradigma değişimi olarak tanımlamaktadır. Bu yeni vizyon; deterministik modellerin reaktif tasarım yaklaşımını kullanıp, bir sonraki adım olarak YZ destekli "ön kestirim" yapan, tahmine dayalı karar destek sistemlerinin geliştirilmesini içermektedir [129].

Yine raporda YZ destekli yaklaşımlarının atık su hatlarındaki tıkanıklıkları ve taşmaları oluşmadan önleme kabiliyeti olduğundan bahsetmektedir. Ayrıca, Aquasuite gibi YZ destekli mimarilerin, tahmine dayalı öngörü kabiliyeti sayesinde tesisin ihtiyacı olan havayı sağlayan blowerların enerji tüketiminde %15'e yaklaşan tasarruf sağladığına da değinilmiştir [129]. Bu geleneksel modellerin yüksek hesaplama yüküne karşılık, YZ'nın üstünlüğünüm bir göstergesidir. IWA; YZ ile veri setlerindeki gürültüyü azaltarak yanlış alarmların önüne geçilip, gerçek zamanlı doğru müdahale edilebildiğini vurgulamıştır.

Yukarıda vurgulanan geleneksel, teorik modellerin getirdiği yüksek maliyetli kalibrasyon ve uzmanlık gerekliliği, bir alternatif olarak "kara kutu" ve veri odaklı yaklaşımların gelişimini zorunlu kılmaktadır [131].

Tablo 2.5 IWA'nın dijitalleşme ve YZ kronolojisi

Yıl	IWA'daki Gelişmeler	Akademik Anlam	IWA'nın Çalışması
2015-2016	Ön Tartışma Dönemi	Büyük Veri, sensörler ve otomasyon konuları ele alındı. Henüz dijitalleşme yoktu.	IWA konferans bildirgeleri
2017	Farkındalık Oluşturma	Dijital su diye bir kavram geçmez ama akıllı sistemler tartışmalarda tematik olarak geçer [126].	IWA yıllık rapor 2017-2018
2018	Programın Kuramsal Olarak Başlatılması	İlk defa "Dijital Water Programme" yi resmen yapılandırmıştır [127].	IWA yıllık rapor 2018-2019
2019	Çerçevenin Kurulması	"Digital Water" raporu yayınlayarak, IoT, Endsütri 4.0, veri yönetimi, akıllı sistemler ilk defa bütüncül bir yaklaşımla ele alınan bir çerçeveye oturtulmuştur [128].	IWA Water Report 2019
2020	Yapay Zekaya Yönelim	Bu yıl, bir yenilikle ilk defa Yapay Zeka ve veri analizini konu alan White Paper yayınlamıştır [129].	Digitial Water: AI Solution for the Water Sector
2020-2021	Genişleme	Artık rutin olarak White Paper serileri yayınlanmaya başlanmış. Dijital ikiz, veri kalitesi ve güvenirliği, karar destek sistemleri ele alınmaya başlanmıştır [130].	Digitial Water White Paper Series
2022	Küresel Uygulama ve Paylaşım	İlk IWA Digitial Water Summit düzenlenmiş, kürelsel ölçekte dijitalleşme ve Yapay Zeka konuları örnekleri ile birlikte tartışılmıştır.	IWA Digitial Water Summit
2023-…..	Olgunlaşma Dönemi	Yapay Zeka, dijital ikiz, veri madenciliği gibi konular artık IWA da temel söylemler arasında yerini alarak, kitaplar ve yeni White Paper'lar yayınlanmaya devam etmiştir.	Digital Water Book (2022+)

Yürütülen bu tez çalışması ile yukarıda özetlenen deterministik modelleme kısıtlarına ve IWA'nın "Dijitalleşen Su" vizyonuna metodolojik bir yanıt verilmeye çalışılmıştır. Çalışmamız; ham verilerin işlenmesinden, özellik mühendisliği (feature engineering) uygulamalarından ve katmanlı veri bölme gibi veri ön işleme tekniklerinden yararlanılarak, sahada uygulanabilir bir YZ modeli önermektedir. Bu yaklaşımdaki temel amaç; işletmesel olarak daha erişilebilir, gerçek zamanlı olaylarda daha hızlı tepki verebilir ve minimum uzmanlık desteğine ihtiyaç duyarak yüksek doğrulukta senaryo tahminleri yapabilir bir karar destek mekanizması oluşturmaktır. Dolayısıyla; bu tezde geliştirilen MÖ yaklaşımı sadece akademik bir çerçeve değil, aynı zamanda IWA'nın veri odaklı dönüşüm stratejisi içerisinde sahada uygulanabilirliği yüksek bir çözüm önerisi olarak konumlandırılmaktadır.

Peki, nedir bu YZ? Bu sorunun cevabını bir alt başlıkta ayrıntılı olarak sunacağız.

2.5 Zekanın Modellenmesinden Makine Öğrenmesine Geçişi: Teoriden Mühendislik Sistemlerine

1956 yılında ilk defa bilimsel bir terim olarak ortaya çıkan "Yapay Zeka" (Artificial Intelligence); insanın bilgiyi işleyebilme yeteneğinin algoritmalar ve hesaplamalar aracılığıyla matematiksel ve istatistiksel olarak taklit edilmesi süreci olup, modern bilimin en canlı alanlarından biridir [132]. YZ'nin gelişim süreci; sadece bilgisayar yazılımındaki yeni bir çerçeve değil; felsefe, matematik, biyoloji ve mühendisliğin hem kesişim hem de bileşim kümesinde meydana gelen bir paradigma değişimidir. Bu olgu, 20. yüzyılın ortalarında doğmuş; günümüz endüstrisinde operasyonel yaklaşımların merkezine yerleşerek radikal bir dönüşüm sürecini başlatmıştır. Salt hesaplama becerilerinin ötesine geçmekle kalmayıp; karmaşık, stokastik ve doğrusal olmayan (non-lineer) sistemlerin içsel mekanizmalarını anlamaya yönelmiştir.

YZ'nin kronolojik tarihçesi; sadece teknik bir araştırmalar kronolojisi değil, aynı zamanda bilginin oluşturulması, tutulması ve işlenmesine ait felsefi bir paradigmadan bir diğerine geçiştir. Bilgi; saklanması ve düzenli kaydı tutulması gereken, kaydı tutulurken işlenen bir meta mıdır? Yoksa ham olarak düzensiz depolanıp, kullanıcının istediğini seçip amaçladığı doğrultuda yorumlaması mıdır?

2.5.1 Teorik Temeller: Makineler Düşünebilir mi?

Literatürde modern YZ fikrinin kronolojik olarak ortaya çıktığı ilk tarih 1950 yılı olmuştur. Alan Turing, "Makineler Düşünebilir mi?" sorusunu sorduğu makalesinde; zekayı fiziksel bir yapıdan ziyade, işlevsel bir çıktı olarak tanımlamıştır [133]. Turing; hâlâ mihenk taşı olarak kabul edilen ve kendi adıyla anılan bir test ile (Turing Testi) makinelerin zekasını ölçmeyi önermiştir [133].

1956 yılındaki dönüm noktası, aslında bir başlangıç noktası olmuş ve John McCarthy tarafından Dartmouth Konferansı'nda YZ terimi ilk defa telaffuz edilmiştir. "Sembolik YZ" olarak adlandırılan ve bilginin semboller ile mantık dizileri şeklinde temsil edildiği yeni bir yaklaşım başlamış oldu [132]. Burada klasik mantığın "Yukarıdan-Aşağıya" yaklaşımı benimsenerek bir düşünme modeli geliştirildi. Ancak bu yaklaşımın doğası, deterministik yaklaşımın çok dışına çıkamamış ve gerçek dünya olaylarındaki belirsizlikleri tanımlamada yetersiz kalmıştır. Bu dönemdeki yaklaşım kapalı sistemlerde başarılı olsa da atık su arıtma gibi doğrusal olmayan (non-lineer) ve dinamik yapıların doğasını kavramaktan uzak kalmış; eksik bilgi ve yüksek veri boyutluluğu ile başa çıkamamıştır. Bu dönemdeki sistemler "kapalı dünya" varsayımı altında çalışmaktaydı ve sistemin tanımlı evreni dışındaki her bilgi yanlış olarak kabul ediliyordu [134].

2.5.2 Bağlantıcılık ve Veriden Öğrenme Dönemi

1960'ların sonunda Massachusetts Teknoloji Enstitüsü (MIT) araştırmacıları M. Minsky ve S. Papert; geliştirilen tek katmanlı bu YZ yaklaşımının "XOR" (Özel Veya) problemini çözemediğini ispat etmiş ve YZ'nin ilk kışını (AI Winter) başlatmışlardır [134]. Yaklaşık 15 yıl süren bu durgunluk döneminin ardından; 1980'li yıllarda "Aşağıdan-Yukarıya" öğrenen bağlantıcı (connectionist) modellerin doğuşu ile YZ'nin ikinci devri başlamıştır. 1986 yılına gelindiğinde; Rumelhart, Hinton ve Williams tarafından ortaya konulan "geri yayılım" (backpropagation) algoritması sayesinde, çok katmanlı olarak örgülenmiş Yapay Sinir Ağları'nın (YSA) verilerdeki gizli örüntüleri matematiksel olarak optimize edebilmesi sağlanmıştır [135].

Bu dönüşümle birlikte bilginin kaynağı köklü bir şekilde değişmiştir. Geleneksel yaklaşımlarda bilginin kaynağı; uzmanlarca işlenmiş, programlanmış, sınıflandırılmış ve bir şablona dönüştürülmüş kurallar dizisiyken; yeni dönemde

bilginin kaynağı doğrudan ham veri setlerinin içindeki istatistiksel örüntüler olmuştur.

2.5.3 Metodolojik Genişleme ve Makine Öğrenmesinin Gelişimi

Bilgi ve veri yaklaşımındaki gelişmeler ile makinelerin önceden tanımlanmış kuralları takip etmesi yerine, geçmiş örneklemlerden hareketle bilinmeyen durumlar hakkında genelleme yapabilme yeteneği kazanmasına evrilmiştir. V. N. Vapnik geliştirdiği istatistiksel öğrenme teorisi ile, model karmaşıklığı ve genelleme hataları arasındaki ilişkiyi matematiksel bir çerçeveye indirgeyerek MÖ'nün temelini oluşturmuştur [136]. Bu yaklaşım ile öğrenmek, bilişsel bir iş olmaktan ziyade olasılıksal çıkarım ve optimizasyon problemi olarak değerlendirilmeye başlandığından, YZ alanında köklü bir anlayış değişime sebep olmuştur [137].

Sonraki yıllarda MÖ, YZ'den bağımsız bir alan olarak gelişmiş ve sahada uygulama odaklı güçlü algoritmalar geliştirilmiştir. Örneğin; Karnel yöntemleri ile non-lineer problemlerin etkin bir şekilde çözülebildiği ortaya konulmuş, Destek Vektör Makineleri (SVM) yardımı ile yüksek boyutlu veri kümelerinde maksimum marjı öncelikleyerek doğru genellemeler yapabilme yeteneği sunulmuştur [138]. Yine aynı dönemlerde, bagging, boosting ve Rastgele Orman (Random Forest) gibi topluluk (ensemble) öğrenme yöntemleri, tekil olarak kurulan modellerin kararsızlık hatalarını azaltarak daha güvenilir ve doğru tahminler ortaya koyabilmişlerdir [139,140].

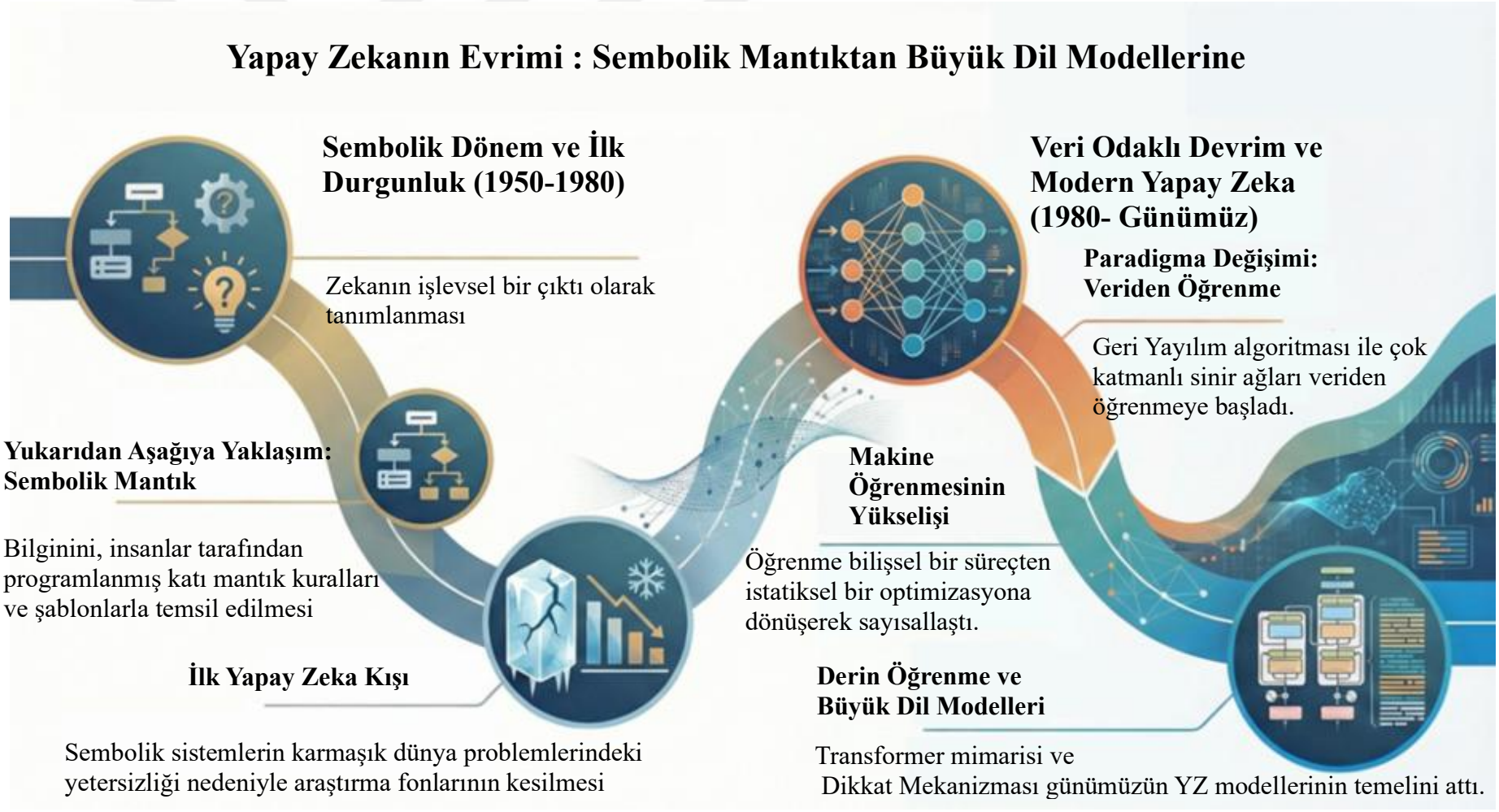
2.5.4 Büyük Veri ve Derin Öğrenme Ekseninde Makine Öğrenmesindeki Dönüşüm

2000'li yılların ortalarına gelindiğinde MÖ; kuramsal altyapısındaki hesaplama becerisinin gelişimi ve veri ölçeğindeki büyüme ile birlikte başka bir döneme girmiştir. İnternetin yaygın kullanılmaya başlanması, her kullanıcıyı bir veri üreticisi haline dönüştüren bulut ortamları, sensör teknolojilerindeki gelişmelerle beraber maliyetlerin düşmesi ve fiziki veri depolama alanlarının kapasitelerinin artması "büyük veri" (big data) olgusunun gelişmesine sebep olmuş; bu da doğal olarak öğrenme algoritmalarının performansının artmasına yol açmıştır. Şöyle ki; Hadoop ve MapReduce gibi dağınık haldeki veriyi kolayca işleyebilen yazılımların gelişmesi ile bu kadar büyük hacimli veri işlenebilmiş ve bunlar üzerinde

istatistiksel öğrenme metodolojileri uygulanabilmiştir. Bu süreçte, "daha karmaşık model" kurmak yerine, "daha fazla veri ile yeterli model" kurma kavramı gelişmiştir [141, 142]. Veri madenciliğindeki bu yükseliş, MÖ'nün endüstriyel alanlarda kullanımını artırmıştır. Ayrıca büyük veride çalışan algoritmaların geliştirilmesi ile de akademik çalışmalar ölçeklenebilirlik üzerinde yoğunlaşmaya başlamıştır.

2010'lu yıllara gelindiğinde MÖ alanındaki çalışmalar, Derin Öğrenme (DÖ) paradigmasının doğmasına ve gelişmesine zemin hazırlamıştır. DÖ; MÖ'nün veri özelliklerini algılayabilmek için ihtiyaç duyduğu manuel özellik mühendisliğine (feature engineering) gerek duymaksızın, çok katmanlı yapısı ile istenilen özellikleri ham veriden kendisi çıkarabilmektedir [143]. Teorik gelişmelerin yanı sıra hesaplama hızlarındaki artış, büyük veriye erişmedeki kolaylık ve bunların oluşturduğu ortak sinerji, YZ'de hızlı gelişmeleri başlatmıştır [143]. Örneğin; 2012 yılında geliştirilen AlexNet mimarisinin görüntü tanıma alanında ortaya koyduğu performans, derin öğrenmenin gelişim kapısını aralamıştır [144]. Bahdanau ve ekibi, "Dikkat Mekanizması"nı (Attention Mechanism) geliştirerek YZ'nin hatırlama ve odaklanma kapasitesinde devrim yaratmıştır [145]. 2016 yılında önerilen ResNet mimarisi, 100+ katmanlı aşırı derin ağların eğitilmesini mümkün kılmıştır [146]. Vaswani ise Transformer mimarisini geliştirerek öz-dikkat (self-attention) mekanizmasına dayanan, doğal dil işleme görevlerinde çok önemli bir yaklaşım ortaya koymuştur [147].

2020 yılından sonrası, ölçek yasaları ve büyük dil modellerinin (LLMs) ortaya çıkışı ile YZ'nin kitleler tarafından hem tanınmasını hem de kullanılmasını sağlamıştır. YZ, ilk kez akademik çevrenin ve endüstrinin kapılarını aralayarak dışarı çıkmıştır. Radford ve arkadaşları; 2018-2021 yılları arasında yapmış oldukları çalışma dizisi ile Üretken Ön Eğitim (GPT) ve çok modlu öğrenme paradigmalarını geliştirerek, YZ'nin ham veriden evrensel temsiller öğrenmesi konusunda modern literatürün ana omurgasını oluşturmuşlardır [148].



Şekil 2.12 Yapay zekanın evrimi (Kaynak: Yazarın kendi metni üzerinden NotebookLM aracılığıyla oluşturulmuştur.)

Biz bu çalışmamızda, YZ uygulamalarından MÖ yöntemini seçtik. Yukarıda detaylıca anlatıldığı üzere; YZ'nin tarihsel evrimi, DÖ mimarilerinin gelişimi üzerinden devam etmektedir. DÖ; özellikle karmaşık ve büyük veri setlerinde (görüntü, ses, doğal dil işleme vb.) oldukça başarılıdır. Ancak bu tez kapsamında incelenen AAT verileri; yapılandırılmış, sınıflandırılmış ve belirli sınırlara sahip verilerdir. DÖ'nün devasa veri ihtiyacı ve "kara kutu" doğası, bu tür veri setlerinde aşırı öğrenmeye (overfitting) sebebiyet verebileceğinden, çalışmanın amacına ulaşmasını zorlaştırabilecektir. Literatürde belirtildiği üzere; veri setinin sınırlı olduğu, değişkenler arasındaki hem doğrusal (lineer) hem de doğrusal olmayan (non-lineer) ilişkilerin istatistiksel olarak açıklanabilmesi gereken durumlarda; MÖ algoritmaları (Random Forest, XGBoost vb.), modelin yorumlanabilirliği açısından çok daha etkin sonuçlar verebilmektedir. Bu bağlamda biz de tezin metodolojik kurgusunda; veri yapımızın karakteristiklerine daha uygun olan, aşırı öğrenme ihtimalinin hem daha az olduğu hem de daha iyi yönetilebildiği, ayrıca proses parametrelerimizi daha şeffaf şekilde analiz etmemizi sağlayan MÖ algoritmalarının esnekliğini ve erişilebilirliğini tercih ettik.

2.6 Atık Su Arıtma Süreçlerinde Yapay Zeka ve Makine Öğrenmesi: Literatür Analizi (2020-2025)

İnsanlığın tarihsel gelişiminde meydana gelen teknolojik devrimler, çoğunlukla çevre teknolojileri alanına en son ve en yavaş şekilde nüfuz etmiştir. Sanayi devriminden sonra başlayan çevreyi ikinci plana atma ve kıt kaynakları sınırsız ve sorumsuzca kullanma alışkanlıkları, geliştirilen yeni teknolojinin bilimsel ve endüstriyel alanlarda dağılımında da kendini göstermiştir. Yeni geliştirilen teknolojik yaklaşımlar, ürünler ve yöntemler, öncelikle savunma, iletişim, ticaret veya eğlence sektörlerinde kendini göstermekte; ticari değerini kaybetmeye başladığı veya teknoloji üreticilerinin pazarının azaldığı evrede nihai uygulama alanı olarak çevre teknolojilerine doğru yönelmektedirler. Günümüzde yaşanan küresel ölçekteki çevre sorunları göz önüne alındığında, bu durumun sürdürülebilir ekolojik denge hedefleri ile bağdaşmadığı, artık değişmesi gereken bir paradigmayı teşmil ettiği aşikardır.

Çevre mühendisliği alanındaki geliştirilen ileri teknolojilerin akademik literatürden, endüstriyel uyarlamalara geçişi; sadece teknik uygunluğa değil, karar

verici mekanizmaların, endüstriyel paydaşların ve çevre sorunları ile karşı karşıya kalan toplumun öncelikleri ve farkındalıklarına bağlıdır. Bilim insanlarının bu farkındalığı arttırabilmesindeki en önemli aracı, bilimsel çalışmaların hem nicelik hem de nitelik olarak etkinliğini arttırarak, paydaşlara ulaşabilirliğini maksimize etmek olabilir. Bu bağlamda, güncel teknolojik trend olan yapay zeka gibi dijital dönüşüm uygulamaları konusunda, çevre mühendisliğinin tüm alt disiplinlerinde daha fazla ve daha nitelikli araştırma yapılması elzemdir.

Bu çalışma ile yukarıda zikredilen zorunluluğa katkı sağlamak amaçlanmıştır. Çevre mühendisliği konularından biri olan atık su arıtma alanında yapılan bizim çalışmamıza geçmeden önce, 2020-2025 yılları arasındaki küresel literatür; yapay zeka, makine öğrenmesi ve derin öğrenme uygulamaları ekseninde sistematik olarak incelenmiştir. Bu alt başlıktaki derleme ile, mevcut durum ortaya konmakta, bilgi birikiminin haritası çıkarılmaya çalışılmakta ve gelecek perspektifine ait bir pencere aralanmaktadır.

2.6.1 Bibliyometrik Analiz ve Akademik Trend

AAT'lerin operasyonel verimliliğini arttırmayı hedefleyen, YZ'nin atık su arıtma alanında kullanımı hakkında yapılan çalışmalar 2000-2023 yılına kadar toplamda 3460 adettir ve özellikle 2018 yılından sonra eksponansiyel bir artış göstermektedir [149]. 2020 yılından sonra eğilimin LSTM ve Evrişimsel Sinir Ağları (CNN) gibi DÖ modellerine doğru yönelmektedir [150].

Çin 342, ABD 209 adet bilimsel yayın ile "Akıllı su Yönetimi" alanını domine etmektedirler [151]. Günümüzde YZ modelleri ile yapılan kirletici giderim verimi tahminlerinin operasyonel kolaylığı ile %10-%23 arasında mali tasarruf sağlandığı ortaya konulmuştur. Özellikle veriye dayalı modeller ile fiziksel tabanlı modellerin birleşimi olan hibrit tabanlı yaklaşımların 2012-2022 yılları arasında diğer konulara nazaran daha yoğun bir şekilde incelenmeye konu olmuşlardır [150].

2.6.2 Deşarj Kalite Tahmini İçin Yapılan YZ Uygulamaları

2025 yılında yapılan bir çalışmada alıcı ortam su kalitesini izlemede MÖ algoritmalarını kullanarak hiper-parametre optimizasyonunu gerçekleştirilmiş ve çalışma %99,84 lük bir tahmin doğruluğu ile neticelenmiştir. Bu literatürdeki en yüksek başarı oranıdır [152].

İsveç'teki bir başka çıkış suyu tahmin modelleme çalışmasında, AAT'den alınan verilerle MÖ uygulamalarından XGBoost algoritmasını kullanarak fosfor giderim verimini tahmin etmeye çalışmışlardır. Çalışma sonunda, geleneksel yapay sini r ağlarına göre, topluluk öğrenmelerinin daha başarılı sonuçlar elde ettiğini ortaya koymuşlardır [153].

2024 yılında Bagheri ve ark. Gerçek zamanlı sensör verilerini kullanarak çözünmüş oksijeni tahmin etmek için CNN-LSTM hibrit modelini deneyimlemişlerdir. 13 ayrı parametre üzerinde yaptıkları çalışmada CNN model katmanının verideki uzamsal özellikleri, LSTM modelinin ise uzun süreli bağımlılıkları iyi yakalayabildiğini vurgulayarak, bu hibrit yapının hata oranının çok düşük olduğunu göstermişlerdir [154].

Bir başka LSTM algoritmasının denendiği çalışmada, su kalitesi marametrelerinin mevsimsel ve iklimsel değişimleri üzerinde durulmuş, yine LSTM'nin uzun süreli bağımlılıkları öğrenme yeteneği kullanılarak, ani değişen su kirliliği durumları, klasik regresyon modellerine göre daha iyi tahmin ortaya koymuştur [155].

Su potabilitesi tahmini amacı ile yapılan bir çalışmada MÖ algoritmalarından Gradient Boosting modelleri kullanılmıştır. Çalışma sonuçlarına göre %99,50 gibi yüksek bir doğruluk oranı ile su sınıflandırılması yapılabildiği gözlemlenmiştir. Ayrıca bu modelin düşük maliyetli sensörler sisteme eklenerek gerçek saha uygulamalarında kolaylıkla kullanılabileceğini vurgulamışlardır [156].

2.6.3 Enerji Tüketim Optimizasyonu ve Karar Destek Mekanizmaları Üzerine Yapılan Çalışmalar

Bir grup araştırmacı MÖ algoritmalarından LightGBM algoritması ile AAT'deki enerji tüketim miktarlarını tahmin etmeye çalışmıştır. Ayrıca AAT'nin karbon ayak izini azaltmak için işletmesel parametrelerin (debi, havalandırma süresi vb.) hangisinin daha yüksek etkiye sahip olduğunu ortaya koyan bir yaklaşım sergilemeye çalışmışlardır [157].

2022 yılındaki, havalandırma tanklarındaki çözünmüş oksijen optimizasyonu ile ilgili çalışmada yine YZ uygulamaları kullanılarak %23 oranında enerji tasarrufu sağlandığı raporlanmıştır [158].

Karar Destek sistemi olarak bir AAT'ye uygulanan dijital ikiz teknolojisinin entegrasyonu incelenmiştir. Çalışmada MÖ ile desteklenen dijital ikizin gerçek

zamanlı verilerle çalışarak, operatörlerin tesis üzerindeki kontrol operasyonlarını geliştirdiğini vurgulamıştır [159].

Otonom kontrol sistemlerinin, atık su kimyasına adaptasyonunu başarılı bir şekilde uygulayan bir çalışma Mohammadi ve ark. tarafından 2024 yılında gerçekleştirilmiştir. Burada fosfor giderimindeki zamansal gecikmeleri Soft Actor-Critic (Takviyeli Öğrenme) algoritmasını başarıyla uygulamışlardır [160].

2.6.4 Membran Teknolojileri ile Kirletici Giderimini Odağa Alan Çalışmalar

Membran gözenek tıkanması (fouling) olgusunun tahmin edilmesini YZ destekli olarak araştıran tüm çalışmaların kapsamlı bir derlemesi 2025 yılında yayımlanmıştır. Buna göre; YZ ile fouling mekanizmalarının erken teşhis edilmesi sayesinde, membran ömrünün %20 oranında uzatıldığı ve buna bağlı olarak işletme maliyetlerinin de %15 düşürülebildiği belirtilmiştir [161].

Bir diğer çalışmada; membran proseslerinde YZ kullanarak "kaynak geri kazanımı" konusu üzerine yoğunlaşılmıştır. Burada da fouling tahmininin sadece sistemi korumak ve maliyetleri düşürmek gibi bir etkisinin bulunmadığı; ayrıca geri kazanım sürecine de olumlu katkılar sağladığı vurgulanmıştır [162].

Geri kazanım ölçeğinde bakıldığında; 2024 yılına tarihlenen bir çalışmada, boya endüstrisine ait bir AAT'de koagülasyon süreci MÖ ile modellenmiştir. Bu modelleme sonucunda, %30 oranında kimyasal tasarrufu sağlanmıştır [163].

Bir başka çalışmada; atık sudan kurşun kirleticisinin giderimi hem DÖ hem de Yanıt Yüzey Metodolojisi (RSM) uygulanarak birbirleriyle kıyaslanmıştır. DÖ'nün değişkenler arasındaki yüksek boyutlu ve doğrusal olmayan (non-lineer) ilişkileri daha iyi tespit etmesi nedeniyle, çok daha başarılı tahminlerde bulunduğu sonucuna ulaşılmıştır [164].

2021 yılında yapılan bir çalışmada, YSA modeli ile nanopartikül destekli ağır metal giderimi modellenmiştir. Çalışma, uygun bimetalik (Fe/Cu) dozajının giderim verimi üzerindeki en etkili bileşen olduğunu duyarlılık analizleri aracılığıyla göstermiştir [165].

2.6.5 İleri Çevre Teknolojilerini ve Yeni Nesil Kirleticileri Konu Alan YZ Destekli Çalışmalar

Polifloroalkil (PFAS) gibi kalıcı kirleticilerin taşınımı ve giderimini araştıran bir çalışmada; MÖ algoritmaları, veri setindeki eksikliklere rağmen %85 oranında doğru ve etkili sonuçlar vermiştir [166].

Leng v eark. Biokömür üretimini ve bunun arıtmadaki verimini MÖ algoritmalarını kullanarak modellemiş, piroliz koşullarının yüzey alanı ve azot içeriği üzerindeki etkilerini de %90 oranında doğru tahmin etmişlerdir [166].

YZ ve IoT'un, "Akıllı Su Yönetimi" sistemine entegrasyonunun tartışıldığı bir çalışmada, sahada geniş bir şekilde yayılmış olan sensör ağlarından gelen anlık verilerin, YZ ile işlenmesinden sebeple; tesisin olası durumlara tepki sürelerinin kısaldığını belirtmişlerdir [167].

Bu başlık altına dahil edilebilecek, elektrokimyasal arıtma teknolojileri konusundaki YZ uygulamaları hakkında bir derleme 2025 yılında sunulmuştur. Bu derleme makalede; akım yoğunluğu ve pH etkileşimlerinin YSA ve ANFIS modelleri ile yaklaşık 0,99 R^2 skorlu doğruluk oranı ile optimize edilebileceği ortaya konmuştur [168].

Sun ve arkadaşlarının (2025) yaptıkları derleme çalışma, gelecek vizyonunun açıklanması bakımından dikkat çekici olarak değerlendirilebilir. Çalışmada, yukarıdaki tüm alt başlıklar altındaki çalışmalardan bahsedilmiş ve YZ uygulamalarının genelleşmesi için veri bilimi ve sensör teknolojilerindeki kısıtlara değinilmiştir. Burada; operatör için uzmanlık geliştirici ve gerçek bir yardımcı paydaş olarak "Açıklanabilir Yapay Zeka" (XAI) önerilmektedir.Ayrıca "Kenarda Yapay Zeka" (Edge AI) ile gerçek zamanlı karar verme mekanizmaları öne çıkarılmaktadır. Ancak çalışmanın sonunda; hâlâ veri, XAI, Kenar YZ ve teknoloji uyumu gibi konularda akademik boşluklar olduğu vurgulanmaktadır [169].

AAT'lerin iklim değişikliklerine bağlı değişen yağış rejimlerinde nasıl davranması gerektiğini YZ destekli olarak ortaya koyan ve 2025 yılında Hocaoğlu ve ark. tarafından yürütülen çalışma; IWA Publishing tarafından yayımlanan ve IWA ile ilişkili "Journal of Water and Climate Change" dergisinde yayımlanmıştır. Çalışmada; ani yağış anlarında AAT'nin bypass yapmasına gerek kalmayacak şekilde adaptif kontrol stratejileri sunulmaktadır [170].

Yukarıda alt başlıklar halinde 2020-2025 yılları arasında yapılmış olan çalışmaları ve güncel eğilimleri inceledik. Mevcut durumda yapılan çalışmaların genelde YZ ve MÖ uygulamalarının sınırlı sayıda parametre ve standart veri setleri üzerinde yoğunlaştığı görülmektedir. Yürüttüğümüz bu tez çalışmasında; tüm temel deşarj kalite standartlarında adı geçen KOİ, BOİ, AKM, TN ve TP'yi kapsayan parametrelerin tamamını inceledik. Ayrıca bunları incelerken, tesisteki tüm parametreler üzerinden meydana gelen değişiklikleri etkileyen faktörlerin tespitinde; daha önce bu kapsamda uygulanmayan bir yaklaşımla "Karşılıklı Özellik Seçimi" (Feature Selection) tekniklerini kullandık. Bu sayede tezimizi sadece akademik bir çalışma olmanın ötesine taşıyarak, gerçek zamanlı tesis yönetimi için bir basamak haline getirmeyi amaçladık.

3
VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ TEMELLERİ

AAT'lerden elde edilen ham verilerin analitik değerlere evrilmesi ve anlamlandırılan verilerden etkin tahminler yapılması için sistematik bir veri işleme ve MÖ metodolojisi gerekir. Bu noktada; dağınık ve işlenmemiş veri setleri içerisinden anlamlı ve kullanılabilir verilerin elde edilmesinde "Veri Madenciliği" devreye girer [171].Ham verinin doğrudan modele verilmesi; "çöp girer, çöp çıkar" (garbage in, garbage out) kuralınca modelin tahmin ve genelleme verimini düşürdüğünden dolayı veri işleme, MÖ algoritmaları ile bilgi üretme sürecinin en önemli adımını oluşturur [172]. Sonrasında kurulacak MÖ algoritmalarının tahmin doğruluklarındaki yanılsamaların nedenleri, veri kalitesinden ziyade başka durumlarda aranmalıdır.

Bu bölümde; AAT verileri ile kurulacak olan MÖ algoritmalarının yüksek doğrulukla ve etkin bir şekilde tahmin yapabilmesi için gerekli teorik altyapı sunulmaktadır. Verilerin temizlenmesi ve hazırlanmasının ardından; özellik seçimi ve mühendisliği ile en belirleyici parametrelerin hangi yöntemlerle belirlenebileceği aktarılacak; çeşitli MÖ algoritmaları ile modelleme yapılıp nihayetinde model başarısı kapsamlı metriklerle değerlendirilecektir. Bu metodolojik çerçeve ile tesis verileri aracılığıyla çıkış suyu kalitesinin tahmin edilmesi amaçlanmaktadır.

Bu bölüm; çevre mühendisliği bilimi çerçevesinde araştırma yapanların güncel veri bilimi araçlarını kavrayabilmesi ve tezde kullanılan metodolojinin kavramsal altyapısının açık bir şekilde anlaşılabilmesi amacıyla, disiplinler arası bir bakış açısı doğrultusunda detaylandırılmıştır. Bu kapsamlı anlatım ile bu alanda çalışma yapacak araştırmacılar için yol gösterici bir rehber olması hedeflenmiştir.

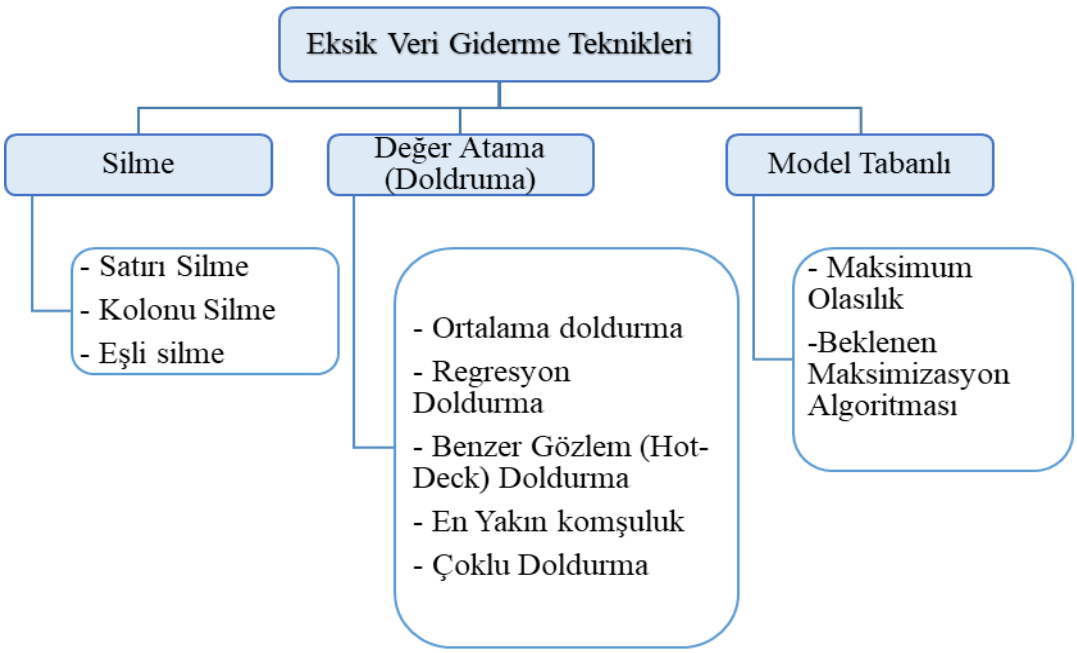
3.1 Veri Ön İşleme

Ham veriler; aykırı değerler, eksik veriler ve ölçek farklılıkları gibi sorunları doğaları gereği bünyesinde bulundurur. Bu sorunları gidermek için veri ön işleme, modelleme çalışmalarının ilk adımını oluşturur [139]. AAT'lerdeki sensör verileri; SCADA sistemleri üzerinden belirli aralıklarla sürekli olarak kaydedilen mekanik veri toplama süreçleridir. Bunlarda meydana gelen arızalar veya bakım dönemlerinde sistemin devre dışında bırakılması; veri setlerinde eksik verilerin oluşmasına sebep olmaktadır. Ayrıca meteorolojik şartlar ve proses değişkenleri gibi verilerin sayısallaştırılmaya (encoding) tabi tutulması gerekebilir [173].

Tezin bu bölümünde; kurulacak MÖ modellerinin tahmin yeteneğini artırmak amacıyla veri kalitesini yükseltmek için kullanılabilecek temel veri ön işleme teknikleri yeterli ölçüde açıklanmaya çalışılacaktır.

3.1.1 Eksik Olan Veri Sorununun Yönetimi

Literatürde genelde veri madenciliğinde eksik verilerle başa çıkmak; üç ayrı başlık altında sınıflandırılmıştır [171]. Bunlar; kaydı silme (görmezden gelme), doldurma (değer atama) ve tahmine dayalı değer atamadır. Her bir yönteme ait alt teknikler Şekil 3.1'de verilmiştir.



Şekil 3.1 Eksik veri giderme teknikleri (Kaynak: [174])

1) Kaydı Silme: Eksik verinin sadece kendisinin değil; veri setindeki tüm satırın, tüm sütunun veya veriler iki indisli veri çifti şeklinde ise her ikisinin de silinmesi şeklinde yapılır. Eğer veri seti çok büyükse ve veriler rastgele dağılmışsa veri silme işlemi uygulanabilir. Ancak veriler yeterli miktarda değilse veri setinin küçülmesi, kurulacak modelin tahmin gücünü zayıflatacaktır [173]. Ayrıca bu işlem çalışmada bir yanlılık (bias) sorunu oluşturabilir. Hangi verinin neden eksik olduğunun cevabı bilinmediği gibi; silinmiş veri sütunu veya satırı gibi çoklu değişkenlerin anlam ve etki değerinin ne olduğu gibi soruları yansız olarak cevaplamak oldukça güçtür. Bu sebeple, çok özel durumlar gerektirmedikçe literatüre göre silme işleminden ziyade doldurma (imputation) yöntemleri tercih edilmektedir.

2) Değer Atama (Doldurma): Bu yöntemde eksik olan veri hücresine tek bir değer atanır. Uygulama kolaylığından dolayı yaygındır. Azda olsa işlenmiş veri setlerinde etkin olarak kullanılabilir.

   Ortalama-Mod-Medyan şeklinde doldurma en yayın kullanım şeklidir. Sayısal verilerdeki eksik hücresel değer ilgili parametrenin değerlerinin zaman serisine göre ortalaması veya medyanı ile doldurulurken, sayısal olmayan eksik değerler ile mod (en sık gözlemlenen durum) ile doldurulur. Bunun bir tek sakıncası standart hatanın küçük olarak değerlendirilmesi olabilir [175].

   Regresyon temelli doldurma şeklinde ise; eksikliğin bulunduğu değişken ile veri setindeki diğer değişkenler arasında bir regresyon modeli kullanılarak eksiklik tahmin edilmeye çalışılmaktadır. Bu yöntemde ise şu sakınca vardır: Eksik verilere yapay bir keskinlik katar ve belirsiz ilişkiler hesaba katılmamış olur. Bu da kurulacak model sonuçlarına yansıyabilmektedir [176].

   Eksik olan veriyi belirleyen diğer parametrelerin değerleri ile veri setinde benzerlik gösteren veri ile eksikliği giderme işlemi "Benzer Gözlem (Hot-Deck)" ile doldurulmaktadır. Benzerlik, yardımcı değişkenlere göre belirlenir. Burada gerçek gözlem değerleri kullanıldığı için veri setinin orijinal dağılım karakterini iyi yansıtabilmektedir [177].

   En yakın komşuluk (k-NN) yönteminde, eksik olan veriye benzer şartlarda olan "k" adet en yakın benzer komşularını bulur; eksik değer bunların ortalaması veya modu ile yerine konulmaktadır. Bu yönü ile aslında Hot-Deck metodunun daha

sistematik halidir denilebilir. Komşularının uzaklığının ağırlıklandırılması, belirli katsayılar ile ilişkilendirilerek de yapılabilmektedir. Bu teknikte; eksik veriyi yerine koymadaki tahmin gücü aslında "k" sayısının ve uzaklık ölçütlerinin doğru hesaplanmasına bağlıdır [178].

Eksik veriyi yerine koyma ve doldurma tekniklerinden bir diğeri de "Çoklu Doldurma" işlemidir. Geleneksel yöntemlerden farkı; eksik değerlerin her biri için tek bir değer değil, birden çok değer üretilmesi ve bu üretilen değerlerle bir tahmin kümesi oluşturulmasıdır. Eksik verilerin tamamı için her bir eksikliğe "m" adet rastgele tahmin yapılır. Böylelikle tüm eksiklikler için "m" adet küme oluşturulur. Sonrasında istatistiksel regresyon çalıştırılarak "m" adet regresyon katsayısı elde edilir. En nihayetinde "Rubin Kuralları" adı verilen tahminleme formülleri kullanılarak nihai bir sonuca ulaşılmaya çalışılır. Bu yöntemin amacı; veri boşluklarındaki belirsizlikleri de hesaba katarak daha doğru ve güvenilir tahminlerle eksik verilere çözüm bulmaktır [179].

3) Model Tabanlı Eksik Giderme: Eksik veri sorununu çözmenin bir diğer yöntemi de mevcut verilerin meydana geliş şeklinin istatistiksel yapısını modelleyerek tahminde bulunmaktır. Burada bir değer atamaktan önce, mevcut veriler arasındaki ilişkiyi modellemek önceliklidir. Bu yaklaşımın iki yöntemi "Maksimum Olasılık" ve "Beklenen Maksimizasyon" algoritmalarıdır ve bunlar aslında iç içe geçmiş yapıdadır [176].

   Maksimum olasılık yönteminin temel yaklaşımı; öncelikle mevcut gözlemlenen verileri oluşturan modeli kurgulamaktır. Eksik veriler, mevcut kurgulanmış ve anlamlandırılmış verilerin modellenmesinden elde edilir. Buradaki yöntemde ilk önce, verilerin çok değişkenli normal dağılım formunda olduğu varsayımı yapılır; sonrasında bu varsayılan dağılımın istatistiksel parametrelerinin (ortalama vektörü ve kovaryans matrisi), elimizdeki mevcut veriye en iyi uyacak şekilde bulunmasına çalışılır. Bu tahminle bulunan parametreler kullanılarak eksik veriler tahmin edilir [176]. Bu yaklaşımın en önemli avantajı; eksik verilerin doğrudan analiz modeli ile birleştirilebilmesi ve eksik veri için ayrı bir doldurma aşaması gerektirmemesidir. Böylelikle parametre tahmini ile eksik veri işleme süreci aynı anda gerçekleşmiş olur [180].

Beklenen Maksimizasyon algoritması; maksimum olasılık tahminlerini hesaplamak için geliştirilmiş bir sayısal optimizasyon mimarisidir. Bu sebeple bir önceki paragrafta maksimum olasılık anlatılırken iç içe geçmiş iki yöntem denilmiştir. Bu yöntemin uygulaması aslında adında kullanılan beklenti ve maksimizasyonun bir döngü halinde ilerlemesine dayanmaktadır. Şöyle ki; beklenti aşamasında, eksik verilerin yerine onların koşullu beklenen değeri geçici olarak konulur ve sanal bir eksiksiz veri seti oluşturulur. Sonrasında, oluşturulan bu sanal eksiksiz veri setinden, sanki bunlar gerçek veri imiş gibi daha iyi parametre tahminleri yapılır. Yani tamamlanmış bu veri setine en iyi uyan parametreler güncellenir. Bu iki adım durmadan tekrarlanır ve nihai parametre tahminleri, bir önceki tahmin ile arasında anlamlı bir fark kalmayıncaya kadar devam ettirilir. [181].

Bu bölümde veri setimizde bulunan eksik verilere ne şekilde yaklaşabileceğimizin literatürdeki yeri verilmiştir. Metodolojimizin anlatıldığı bir sonraki bölümde, buradaki yöntemlerden hangisini neden seçtiğimizi anlatıp detaylandırılacaktır.

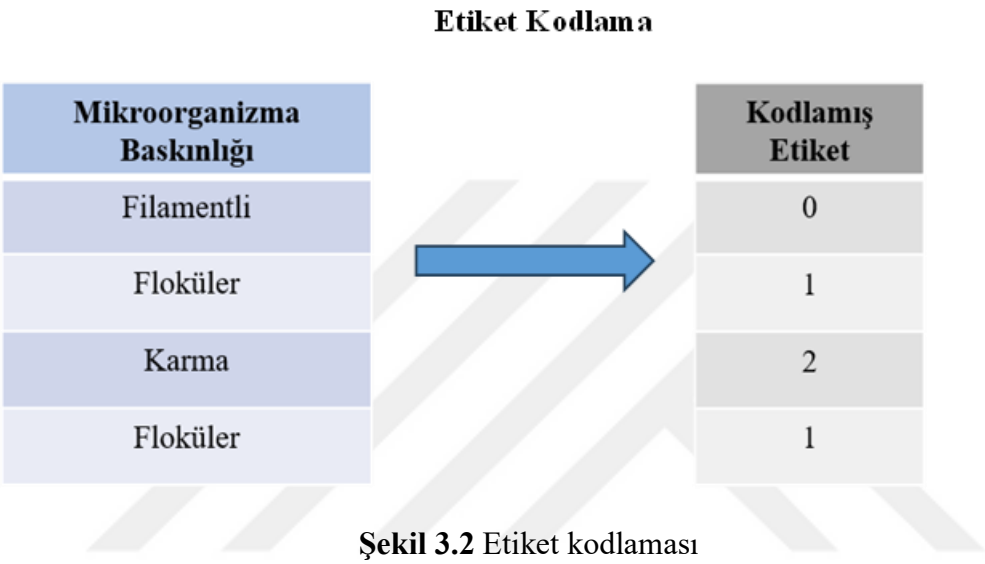
3.1.2 Kategorik Verilen Kodlanması

Sayısal olmayan (kategorik) veriler MÖ algoritmalarınca doğrudan kullanılamazlar. Çünkü MÖ'de kullanılan denklemler sayısal verileri işleyebilirler. AAT'lerde; KOİ, pH, sıcaklık vb. gibi sayısal verilerin yanı sıra; sözel veya kategorik olan havuz tipi, arıza durumu, hava durumu, ekipman durumu gibi veriler de vardır. Bu kategorik verilerin sayısal verilere dönüştürülmesi, veri işleme sürecinin en kritik adımlarından biridir. Literatür incelendiğinde bu amaçla oluşturulmuş çok sayıda yöntem bulunduğu görülmektedir. Ancak bunlar, oluşturuluş şekillerine göre üç ana başlıkta gruplandırılabilir: Etiket Kodlama (Label Encoding), Vektörel Kodlama (One-Hot Encoding) ve Anlamsal Kodlama (Target Encoding) [182]. Bu bölümde bu yöntemlere kısa teorik yaklaşımlarına değinilecektir.

3.1.2.1 Etiket Kodlama (Label Encoding)

MÖ algoritmaları matematiksel denklemlere dayandığı için sayısal olmayan kategorik (sözel) değişkenleri işleyemezler. Etiket kodlama, sayısal olmayan bu verileri sayısal formata dönüştüren yöntemlerden birisidir. Etiket kodlama, basit bir dönüşüm yöntemi olsa da dikkatli kullanılmalıdır [183].

Bu yöntemde, kategorik değişkenin değer kümesinden tam sayı kümesine birebir eşleme yapan bir fonksiyon tanımlanır. Bu eşleme, alfabetik sıraya veya görülme sıklığına göre yapılabilir. Şöyle ki, kategorik değişkeni C ile tanımlar isek, C'nin k tane benzersiz (aynı olmayan) değer kümesi $\{c_1,\ c_2\ ,\ \ldots\ldots\ldots,c_k\}$ olur, kodlama fonksiyonu ise $f: \{c_1, c_2, \ldots\ldots\ldots, c_k\} \rightarrow \{0,1,2, \ldots., k-1\}$ şeklinde tanımlanır [184]. Burada metodolojik olarak önemli olan nokta, bu atamaların tutarlı ve sürekli olmasıdır. Şekil 3.2'de etiket kodlamanın şematik şekli görülmektedir.



Şekil 3.2 Etiket kodlaması

Bu kodlama türü şu durumlarda kullanılması uygundur:

- Sıralı değişkenler varsa, yani kategoriler doğal bir sıraya sahipse (örneğin, eğitim seviyesi, memnuniyet derecesi vb.)...
- MÖ algoritmalarında özellikle karar ağaçları (DT), rastgele orman (RF), gradient boosting gibi ağaç tabanlı algoritmalar kullanılacaksa bu kodlama türü kullanılabilir [185].
- DÖ de, gömme katmanları ile birlikte bu yöntem çok yüksek boyutlu verilerde kullanılabilir [186].

3.1.2.2 Vektörel Kodlama (One-Hot Encoding)

Bu yöntemde k adet kategorik değişkenin her biri, 0 ve 1'lerden oluşan k adet yeni ikili değişken oluşturularak ifade edilir. Her bir kategorik değişken için bir sütun oluşturulur ve gözlenen değere ait kategori ise 1, değilse 0 şeklinde kodlanır [187]. Yönteme ait şematik gösterim Şekil 3.3'te verilmiştir.

Vektörel Kodlama

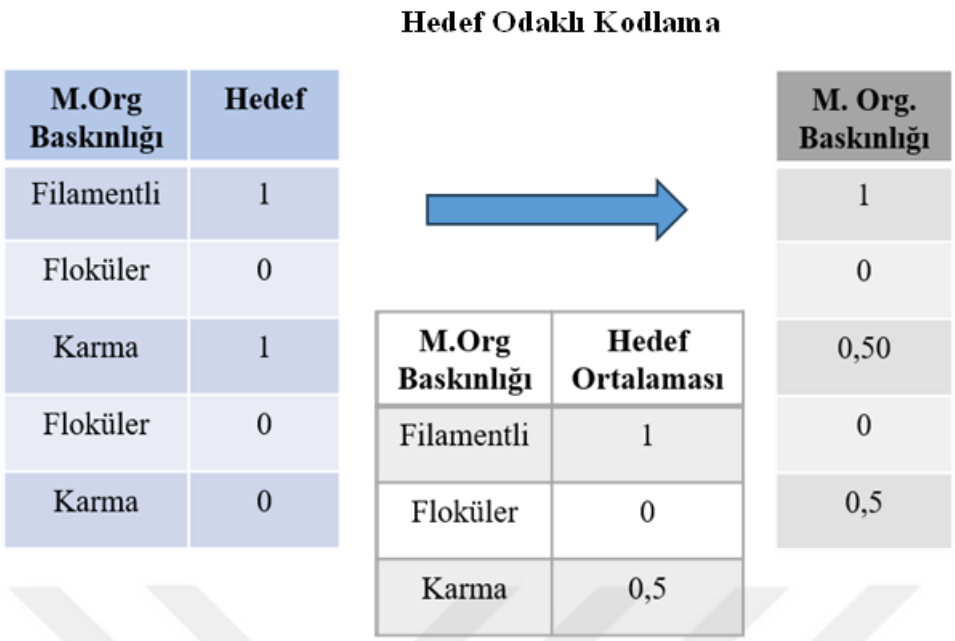
Mikroorganizma Baskınlığı	Filametli M.Org.	Floküler M.Org.	Karma M.Org.
Filamentli	1	0	0
Floküler	0	1	0
Karma	0	0	1
Floküler	0	1	0

Şekil 3.3 Vektörel kodlama

3.1.2.3 Hedef Odaklı Kodlama (Target Encoding)

Çok sayıda kategoriye sahip (yüzlerce farklı istasyon adı gibi) değişkenler için kullanılabilen istatistiksel bir yöntemdir ve kategorilerin hedef değişkene göre ilişkisine dayandırılarak bir sayısallaştırma işlemi yapılır. Bu yaklaşımda kategorik değişkenin her bir sınıfı, o kategoriye ait gözlemlerin hedef değişkeninin ortalama değeri ile değiştirilerek kodlanır. Ne tür hesaplamalar yapıldığı ve yeni kodların yazılımı, Şekil 3.4'te gösterilmiştir.

Bu yöntemin avantajı, kategorik değişken ile hedef değişken arasındaki ilişkiyi doğrudan kodlamasıdır. Tek seferde kodlama (one-hot encoding) yönteminin aksine çok büyük veri setlerinde kullanılabilir ve bellek sorunu oluşturmaz. Ancak yöntemin başarısı hedef değişkene bağımlı olmasıdır ve her yeni veri oluşumu için sistemin yeniden ortalama alarak hesap yapması nedeniyle sürekli güncellenme ve yeniden hesap zorunluluğu vardır.



Şekil 3.4 Hedef odaklı kodlama

3.2 Veri Madenciliğinde Özellik Seçimi (Feature Selection)

Veri madenciliği ve MÖ uygulamalarında, veri toplama teknolojilerinin gelişmesi ve bununla birlikte çok büyük boyutlu verilerin oluşması öğrenme algoritmalarının performanslarını düşürebilmekte, modellerin genellenebilirlik yeteneğini zayıflatıp aşırı öğrenme sorununa yol açabilmektedir [171]. Burada özellik seçimi devreye girmektedir. Özellik seçimi (feature selection), fazla sayıda özellik içeren büyük veri setindeki özelliklerin, en ayırt edici ve gerçek veri setini en iyi yansıtan alt kümesinin oluşturulması sürecidir. Gereksiz ve kurulacak modelin sonucunu değiştirmeyecek olan verilerin elenmesi süreci olarak da tanımlanabilir [188].

Aslında bu metodolojinin temel amacı sadece boyutsal küçültme değil, aynı zamanda MÖ'nün sınıflandırma veya regresyon modellerinin tahmin doğruluğunu korumak ve artırmaktır. Literatürde verilerin özelliklerinin sınıflandırılmasında eğer bir özellik, istenilen hedefin belirlenmesinde istatistiksel olarak bir katkı sağlıyorsa buna "alakalı (relevant)", sağlamıyorsa "alakasız (irrelevant)" olarak nitelendirilir. Eğer bir özellik, yüksek bir korelasyonla başka özellikler tarafından temsil edilebiliyorsa, modelin karmaşıklığını azaltmak amacı ile elenir ve "fazlalık (redundant)" olarak adlandırılır [189].

3.2.1 Özellik Seçimi Yöntemleri

Literatürde özellik seçimi yöntemleri, temel yaklaşımlarına ve bir öğrenme algoritması kullanıp kullanmadıklarına göre şu üç ana başlık altında sınıflandırılır:

- Filtre Yöntemleri (Filter)
- Sarmalayıcı Yöntemler (Wrapper)
- Gömülü Yöntemler (Embedded)

Bu yöntemlerin her biri farklı avantajlar, kısıtlar ve uygulama senaryoları sunar. Hangisinin seçileceği uygulama amacına göre değişir [190].

3.2.1.1 Filtre Yöntemleri (Filter)

Özellik seçimini MÖ algoritmalarından bağımsız olarak, istatistiksel tekniklerle gerçekleştiren yaklaşımlara filtre yöntemleri denir. Bu yöntemler, mevcut verinin özellikleri ile hedef değişken arasındaki ilişkilerin gücü ölçülerek, tek değişkenli veya çok değişkenli testlere dayandırılarak oluşturulur. Filtre yöntemlerinin belirgin avantajları hızlı olmaları ve MÖ algoritmalarından bağımsız çalıştıkları için aşırı öğrenme riski barındırmamalarıdır [191]. Aşağıdaki türleri vardır:

1) Pearson Korelasyon Katsayısı (r): Pearson korelasyon katsayısı, iki sürekli değişken (özellik ve hedef değişken) arasındaki doğrusal ilişkinin gücünü ve yönünü ölçen istatistiksel bir yöntemdir. Karl Pearson tarafından geliştirilmiştir. Özellikle denetimli öğrenme algoritmalarında yaygın olarak kullanılır. Katsayı, iki değişkenin kovaryanslarının standart sapmalarının çarpımına bölünmesiyle bulunur. $-1 \leq r \leq +1$arasında değer alır. (+1) mükemmel pozitif doğrusal ilişki, (-1) mükemmel negatif doğrusal ilişki, (0) ise değişkenler arasında hiç ilişki olmadığını gösterir. Kolay uygulanabilir bir yöntemdir ancak sadece doğrusal ilişkileri yakalayabilir [171].
2) SelectKBest Algoritması: Çok yüksek boyutlu veri setlerinde tek değişkenli ve filtre tabanlı olan bir yöntemdir. Doğrusallık şartı olmadan her bir değişkeni bağımsız olarak ele alıp, hedef değişken ile aralarındaki istatiksel gücü test ederek puanlar. Buradaki puanlamaları sıralayarak en iyi k adet özelliği seçerek diğerlerini eler. Buradaki k adet, model geliştirici tarafından seçilebilmektedir [192]. Bu metodoloji yaptığımız tez çalışmasında kullanılan bir yöntem olduğundan metodoloji kısmında daha detaylı anlatılacaktır.

3) Karşılıklı Bilgi (Mutual Information): Bu yöntemde tezimizde kullanılmıştır. Karşılık bilgi entropi kavramı üzerine kurulmuş bir yöntem olup, istatiksel bağımsızlığı temel alarak bir özelliğin hedeflenen sınırı ne kadar açıkladığını ölçer. Entropi kavramı Claude Shannon tarafından geliştirilmiştir. Bu kavram, rastgele bir değişkenin taşıdığı belirsizliği veya belirliliği ölçer. Bu yöntemin temel avantajı da ilişki türünden bağımsız olması ve doğrusal olmayan sinüzoidal, kuadratik ve karmaşık ilişkileri dahi işleyebilmesidir [193].

4) Varyans Eşikleme: Herhangi bir ayırt edici bilgi taşımayan, varyasyonu ihmal edilebilir olan özellikleri veri setinden eleyerek çıkarmaya dayanan bir yöntemdir. Hedef değişkenden bağımsız olarak çalışan bir filtreleme tekniğidir. Burada tüm örneklemde birbirine yakın veya aynı değerleri alan özellikler düşük varyans olarak kabul edilerek kurulacak MÖ algoritmasının öğrenme aşamasına hiçbir etkisi olmayacağı kabulü ile veri setinden çıkarılır [172]. Veri seti yüksek boyutlu ve birbiri ile korelasyonlu ve benzer özellikleri çok içeriyorsa bu metot, gereksiz boyut artışını azaltacağı için tercih edilebilir bir yöntemdir.

3.2.1.2 Sarmalayıcı Yöntemler (Wrapper)

Sarmalayıcı yöntemler özellik seçimi işini önceden belirlenmiş olan bir MÖ algoritmasını kullanarak yapan yöntemlerdir. Filtre yöntemindeki verinin istatistiksel dinamikleri değil, MÖ algoritmasının tahmin performansı önemlidir. Bu yöntemde özellik seçimi bir "kara kutu" olarak ele alınarak MÖ ile modellenir ve her özellik alt kümesi için model eğitilir. Modelin başarı oranı o modelin uygunluk değeri olarak atanır [171]. Bu gruptaki yöntemler özellik seçimi sürecini "arama problemi" olarak ele alır. En yaygın kullanılan sarmalayıcı yöntemler aşağıdaki şekildedir:

1) İleriye Doğru Seçim (Forward Selection): "Aşağıdan-Yukarıya" bir yaklaşımla çalışan, çalışmaya ilk önce boş bir küme seçerek başlayan iterasyona dayalı bir yöntemdir. Algoritma, seçtiği her yeni özelliği bu boş kümeye ekleyerek modeli eğitir ve sonucu bir doğruluk skoru ile değerlendirir. Mevcut kümeye eklendiğinde modelin performansını en yüksek doğruluk oranı ile artıran özelliği kalıcı olarak seçer. Avantajı diğer sarmalayıcı modellere göre kolay hesaplanabilir olmasına rağmen iç içe geçme sorunu da en büyük dezavantajıdır. Yani ilk başlarda kümeye alınan düşük performanslı değişkenler yeteri kadar

özellik yok ise kümenin içerisinde göreceli olarak yüksek performansından dolayı kalabilir [194].

2) Geriye Doğru Eleme (Backward Elimination): İleriye doğru seçimin aksine, yöntem burada "Yukarıdan-Aşağıya" bir yaklaşım izleyerek, boş kümeden değil tam kümeden başlayarak özellik değerlendirmesi yapar. İterasyon şu şekilde çalışır; tam kümeden özellikler tek tek çıkarılarak modelin performansı ölçülür. Performansı en az etkileyen özellik kalıcı olarak elenerek kümeden çıkartılır. Bu yöntemin ileriye doğru seçilim yöntemine göre avantajı, karmaşık etkileşimleri yakalama kapasitesinin daha yüksek olmasıdır. Ancak tam veri seti ile modellenmeye başlandığı için çok büyük setleri için uygun değildir [139].

3) Özyinelemeli Özellik Eleme Yöntemi (Recursive Feature Elimination-RFE): Bu tez kapsamında da kullanmış olduğumuz RFE yöntemi, ardışık geriye eleme mantığını kullanarak çok daha gelişmiş bir hale getirilen bir optimizasyon yöntemidir. RFE her adımda tüm özellik kümelerini ve kombinasyonlarını denemek yerine, modelin içsel dinamiklerini kullanarak özellikleri eler. Şu aşağıdaki adımları gerçekleştirerek bunu yerine getirir:

- Model Eğitimi: Tüm özellikler kullanılarak temel bir tahmin edici algoritma eğitilir (karar ağacı, rastgele orman vb.)
- Önem Hesaplama: Her bir özellik için model tarafından üretilen ağırlık katsayıları veya özellik önem skorları (seçilen modele göre değişir) hesaplanır.
- Eleme: Belirlenen bir adım sayısına göre, en düşük önem skoruna sahip özellikler kümeden çıkarılarak elenir.
- Yineleme: Kalan özelliklerle model yeniden çalıştırılır ve her adımda yeni bir özellik, istenilen özellik sayısına ulaşılıncaya kadar elenir.

RFE'nin diğer sarmalayıcı yöntemlere nazaran en büyük üstünlüğü her bir eleme adımından sonra modelin tekrar çalıştırılması ile özelliklerin önem skorlarının yeniden hesaplanması ve dolayısıyla modelin kararlılığının sürekli artırılmasıdır [194,195].

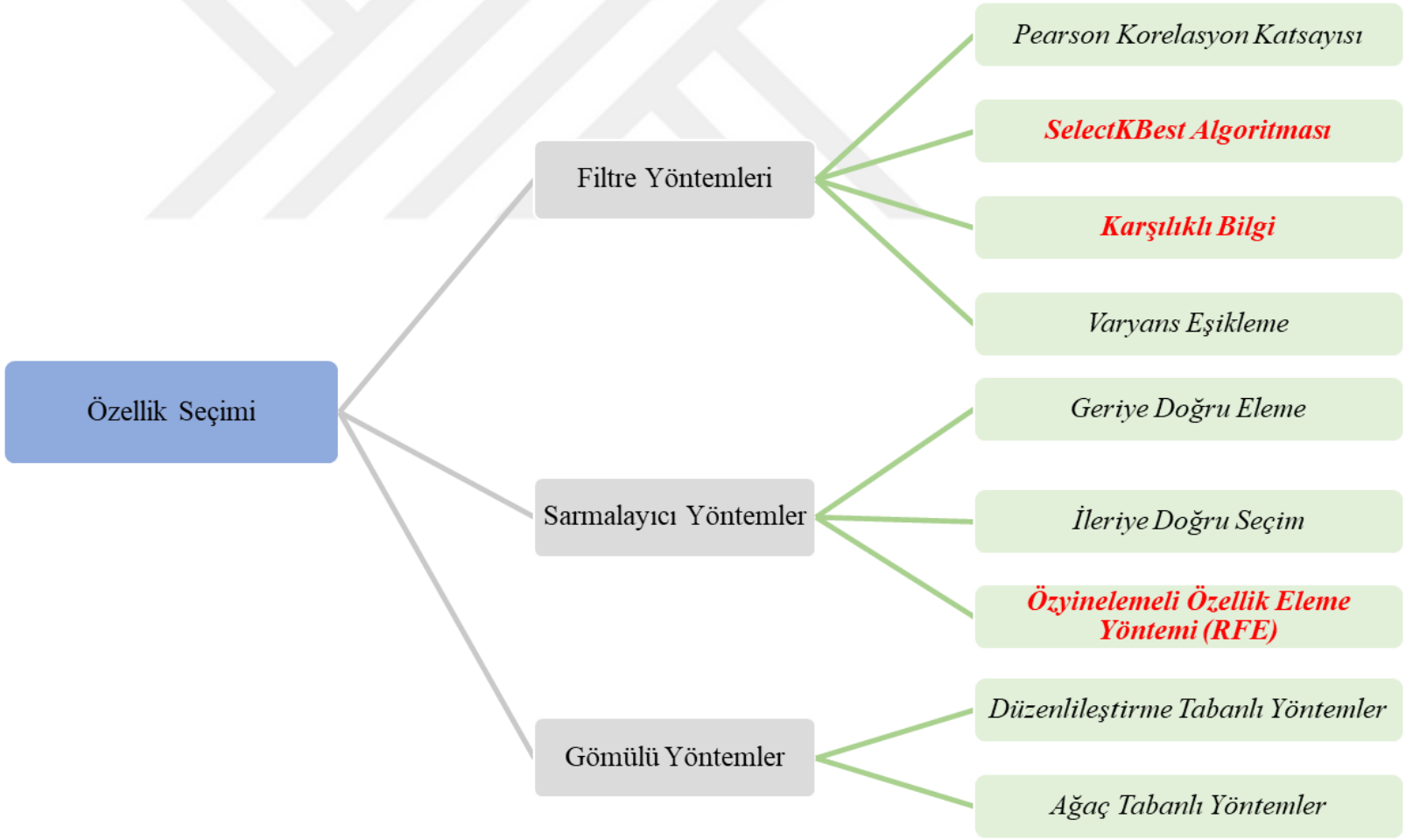
3.2.1.3 Gömülü Yöntemler (Embedded)

Gömülü yöntemler filtre yöntemlerinin hesaplama verimliliğini, sarmalayıcı yöntemlerin de performans üstünlüğünün birleştirildiği hibrit yaklaşımlardır. Şöyle ki: özellik seçim sürecini, doğrudan MÖ modelinin eğitim aşamasına entegre

ederek bunu yapar. Özellik seçimi, geliştirilecek MÖ modelinden ayrı bir ön işlem adımı değil, modelin eğitim sürecinin bir parçası olarak ele alınır. Kurulacak MÖ modeli eğitilirken, hangi özelliklerin tahminleme gücüne en çok katkı sağladığı veya sağlamadığı belirlenir ve buna göre özelliklere etki ağırlığı atanır. Yani algoritmanın doğal yapısı içerisinde gerçekleştirilir. Bu yöntemlerde şu şekilde sınıflandırılabilir:

1) Düzenlileştirme Tabanlı Yöntemler (Regularization): Bu yöntem özellikle doğrusal ilişkilerin modellendiği yapılarda daha çok tercih edilir. Modelin karmaşıklığını kontrol edebilmek için modelin katsayılarına bir ceza puanı eklenmesine dayanır. L1 (LASSO) denilen regresyon uygulama alt yönteminde bu ceza puanı sıfıra eşitlenerek modelden çıkarılırken, L2 (Ridge) denilen başka bir regresyon çeşidinde puan asla sıfıra yaklaştırılmaz, böylelikle kesin bir özellik seçimi yerine sürekli bir küçültme sağlanır. Bir başka yöntem olan Elastik ağ regresyon ise bu L1 ve L2 regresyonunun birleştirilmiş halidir [139].
2) Ağaç Tabanlı Yöntemler: MÖ algoritmalarından Karar Ağaçları veya bunların topluluk versiyonları olan Rastgele Orman veya Gradient Boosting makineleri gibi algoritmalar, özellik seçimi işlemini düğüm bölünmeleri esnasında içsel olarak gerçekleştirir. Her düğümde modeli en iyi ayrıştıran özellik seçilir. Kurulan modelin eğitimi tamamlanmış olduğunda, özelliklerin ağaç yapısındaki kullanım sıklığı ve sonuca katkısı ağacın veya ağaçların doğal yapısı gereği ortaya çıkar ve bir "Özellik Önem Skoru" elde edilir [139]. Gömülü yöntemlerin en büyük avantajı, özellik seçimi işlemlerini kurulacak olan MÖ algoritmasının içerisinde halletmesidir. Özellik seçimi için ayrıca bir model kurulmasına ve bu modelin tekrar tekrar eğitilmesine gerek yoktur [190].

Yukarıda anlatılan özellik seçimi yöntemlerinin toplu olarak gösterimi Şekil 3.5'te verilmiş olup, tez kapsamında kullanılan yöntemler kırmızı ve bold puntolarla belirtilmiştir.



Şekil 3.5 Özellik seçimi yöntemleri

3.3 Veri Setini Bölme Stratejileri

MÖ algoritmaları ile model geliştirmedeki ana hedef, model oluşturulurken kullanılan veri seti örneklerini hedeflenen sonuca en iyi yaklaştırılması değil, veri seti dışında yeni verilerin, yeni olay sonuçlarının karşısında göstermiş olduğu gerçek hedef değerler karşısında yaptığı genelleme yeteneğidir. Literatürde MÖ teorisi kapsamında bahsedildiği üzere; bir modelin başarısının değerlendirilmesi, eğitim verileri üzerindeki hatası ile değil, modelden bağımsız şekilde oluşturulmuş bir veri kümesi üzerinden ölçülen test hatası sonuçları ile değerlendirilmedir [139]. Bu yaklaşım doğrultusunda; veri setinin, modelin kurgulanmasından önce Eğitim (Training) ve Test (Testing) kümeleri olmak üzere iki ayrık alt kümeye bölünmesi literatürde kabul görmüş bir doğrulama prensibidir. Literatürde veri kümesini eğitim verisi ve test verisi olarak bölme işlemi, model doğrulama ile birlikte iç içe değerlendirilen bir süreçtir. Ancak bu tez çalışmasında MÖ algoritmaları uygulandıktan sonra model başarısı ayrı bir şekilde bazı matematiksel metriklerle değerlendirildiği için, bu kısımda sadece veri setini bölme kısmını anlatacağız. Modelin başarı performansı ayrı bir başlık altında (bkz. Bölüm 3.5) yapılandırılmıştır.

3.3.1 Basit Ayırma Yöntemi (Hold-Out)

Basit ayırma yöntemi, veri seti kümesinin (D) önceden belirlenen bir oranda (%70-80; %30-20) herhangi bir kural tanımlanmaksızın rastgele olarak iki parçaya ayrılmasıdır.

- Eğitim Kümesi (D_{train}): Model algoritmasının parametrik yapısı olan ağırlık değerlerinin öğrenilmesi ve içsel optimizasyonunun yapılması amacı ile eğitim esnasında kullanılan veri setinin büyük alt kümesidir. Genellikle verinin tamamının %70-80'lik kısmı bu bölüm için ayrılır.
- Test Kümesi (D_{test}): Model kurulduktan sonra modelin gerçek verilerle test edildiği ve modelin daha önce hiç görmediği veriler ile performansının değerlendirildiği tüm veri setinin küçük alt kümesidir. Veri setinden geriye kalan %30-20'lik kısımdır.

Bu yöntem hesaplama karmaşasını azalttığı için en etkin yaklaşım olmakla birlikte, verinin tamamen rastgele kuralsız bir şekilde bölünmesi, özellikle sınıf dağılımının

veya hedef değişken aralıklarının dengeli olmadığı durumlarda "örnekleme yanlılığına" yol açabildiğinden bu yöntem kullanılmadan önce bu durumlara dikkat edilmelidir. Rastgele seçilimden sebeple eğitim kümesindeki veri dağılımı ile test kümesindeki veri dağılımının istatistiksel özellikleri birbirleri ile aynı olmaması modelin öğrenme sürecini ve başarısını olumsuz etkiler [173].

3.3.2 Tabakalı Örnekleme (Stratified Splitting)

Basit ayırma yönteminin kısıtı olan rastgele bölmenin getireceği dağıtımsal bozulma riskini ortadan kaldırmak için tabakalı örnekleme yöntemi geliştirilmiştir. Literatürde bu yöntem standart hale gelmiştir. Han, Kamber ve Pei'nin tanımladığı üzere veri seti ayrılırken eğitim ve test kümelerinin istatistiksel veri dağılım özelliği korunmaya çalışılır [171].

Bu matematiksel olarak şöyle tanımlanabilir; N elemanlı bir veri kümesinde C_i sınıfının görülme olasılığı $P(C_i)$ ise, eğitim (N_{train})ve test (N_{test}) kümelerinde de oluşturulacak bu sınıflara ait örneklemin görülme oranının da yaklaşık $P(C_i)'ye$ eşit olması temin edilmelidir.

$$\frac{N_{train}^{(i)}}{N_{train}} \approx \frac{N_{test}^{i}}{N_{Test}} \approx P(C_i) \tag{3.1}$$

Denklemde N^i, ilgili alt kümedeki i. Sınıfa ait örnek sayısını ifade etmektedir. Bu veri seti bölme yaklaşımı veri setinin içsel yapısını koruyarak eğitim ve test kümlerinin modeldeki temsil gücünü arttırmaktadır. Böylece hedef değişkenlerdeki uç değerlerin sadece ettim veya sadece test kümesinde toplanmasını engelleyerek, modelin daha kararlı ve dağa gerçeğe yakın doğruluk üretmesini sağlar [172].

3.4 Makine Öğrenme Algoritmaları

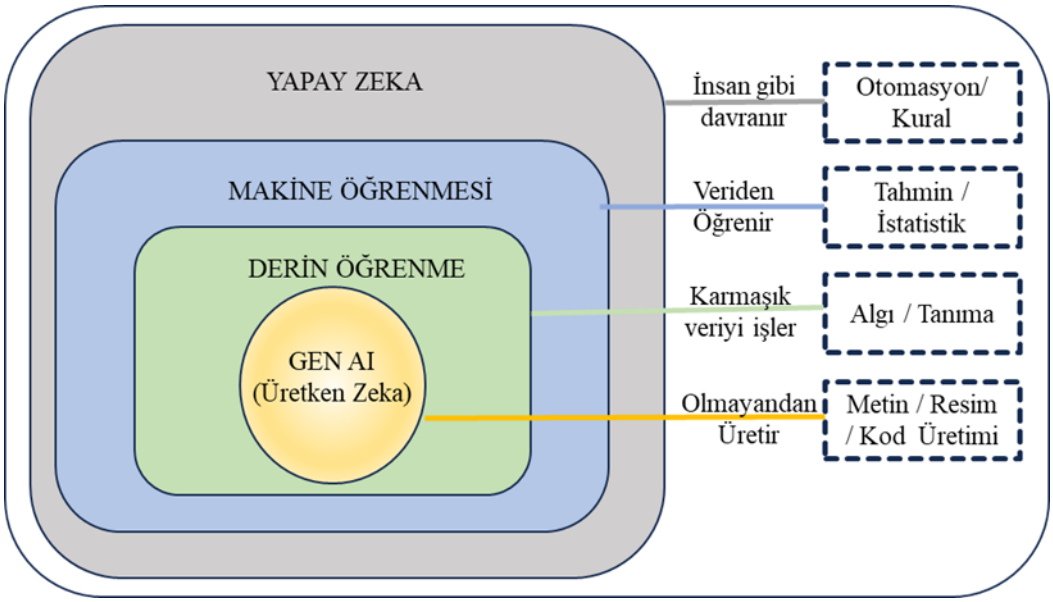
Makine öğrenmesi, açık ve kural tabanlı programlama ihtiyacı duymadan, veri içerisindeki örüntüleri kullanarak performanslarını zamanla artırarak sonuç odaklı tahminler için kendi kendini eğitebilen disiplinler arası bir alandır. 1959 yılında Artur Samuel tarafından "bilgisayarlara, açıkça kural tanımlamadan, öğrenme yeteneği kazandıran çalışma alanı" olarak tanımlanmıştır [196]. Mitchell'in yaptığı daha formal bir tanım ise; "bir bilgisayar programı, T görevler sınıfındaki tahmin performansını P ölçüsü ile E tecrübesine dayanarak artırıyorsa, bu programın

öğrendiği kabul edilir" [137]. Bu tanımlar doğrultusunda güncel ve anlaşılır şöyle bir tanım yapılabilir; MÖ, insanların öğrenme yaklaşımına benzeyerek, veriler arasındaki ilişki ve model algoritmalarına dayanarak, hedeflenen sonuçlara dair yaptığı tahminleri kademeli olarak arttırabilen bir YZ ve bilgisayar programlama bilimi alt dalıdır. Geleneksel modellemeler değişkenler arasındaki ilişkilere odaklanırken, MÖ sonuca odaklanmaktadır. MÖ'nün daha iyi anlaşılabilmesi için, şu şekilde farklılıklar üzerinden açıklanabilir:

- Geleneksel Programlama; Programı oluşturan karşılaşılan problemi analiz eder, ilişkileri anlar ve bu ilişkilere dayanarak algoritmayı kodlar. Sonrasında veriler bu kurallardan geçirilerek çıktı üretilir. Yani; "Girdi + Kural = Çıktı" üretilir. Bu yöntem tüm olasılıkların önceden tanımlanabildiği deterministik problemlerin çözümü için uygundur.
- Makine Öğrenmesi; Seçilen bir MÖ model algoritması, giriş verilerini ve hedeflenen çıktıları alır, bu ikisi arasındaki ilişkinin kurallarını kendisi sürekli denemeler yaparak keşfeder. Yani; Girdi + Çıktı = Kural keşfedilir. Bu yaklaşım, yüz tanıma, doğal dil işleme gibi, kuralların elle yazılamayacak kadar karmaşık olduğu stokastik problemlerin çözümü için kullanılır [197].

MÖ, YZ'nin bir alt bileşenidir. DÖ ise MÖ'nün daha gelişmiş ve daha karmaşık büyük boyutlu veri kümelerine uygulanan gelişmiş şeklidir. Bunların dışında günümüzde artık karşılaşmış olduğu problemler karşısında kendi algoritmalarını kodlayabilen Üretken Yapay Zeka (ÜYZ) vardır. ÜYZ, büyük veri kümelerinden öğrenerek yeni ve benzersiz içerikler üretebilen yapay zeka yaklaşımıdır. Bunlar arasındaki ilişki Şekil 3.6 da verilmiştir.

MÖ, daha önceki bölümlerde anlatılan veri madenciliği uygulamalarını da içinde barındıran bir yaklaşımdır. Veriyi işleme mantığı, mevcut veri setindeki durum ve ilişkiler için de çalıştırılabilen algoritmaları barındırır.



Şekil 3.6 Yapay zeka'nın bileşenleri

MÖ modellemesinin öğrenme süreci, matematiksel olarak istatistiksel öğrenme teorisi çerçevesinde, bir optimizasyon problemi olarak ifade edilir [136]. Öğrenme süreci aşağıdaki üç ana bileşenin etkileşiminden oluşur:

1) Karar Fonksiyonu (Model): Bu aşamada, seçilen algoritma girdi verilerinden yararlanarak, veri içerisindeki ilişkiler hakkında bir tahmin modeli oluşturur. Bu süreç, verinin eğitilmesi olarak da adlandırılır. Eğitim veri setindeki veriler kullanılarak seçilen algoritma eğitilerek öğrenme süreci gerçekleştirilir ve probleme özgü bir model yapılandırılır. Bu aşamaya kısaca girdiyi çıktıya eşleyen hipotez fonksiyonu da diyebiliriz.

2) Amaç/Hata Fonksiyonu (Loss Function): Oluşturulan modelin çıktılarının doğruluğunu değerlendirmek amacı ile kullanılan fonksiyondur. Hata fonksiyonu, gerçek değerler ile modelin tahmin ettiği değerler arasındaki farkı nicel olarak ifade ederek, modelin başarısını ölçer. Kısa tabir ile, modelin tahmini ile gerçek değer arasındaki farkı ölçen bir metriktir.

3) Optimizasyon Algoritması Süreci: Model parametrelerinin, eğitim veri setindeki değerlere daha çok yaklaşacak şekilde modelin güncellenmesidir. Modelin tahminleri ile gerçek değerler arasındaki farkı minimize etmek amacı ile modeldeki parametreler arasındaki ilişkilerin ağırlık katsayıları iteratif olarak ayarlanır. Bu ayarlama, hata fonksiyonunun istenilen doğruluk değerine

yakınsamasına kadar devam ettirilir. Bununla da modelin genelleme ve tahmin yeteneği artırılmış olur.

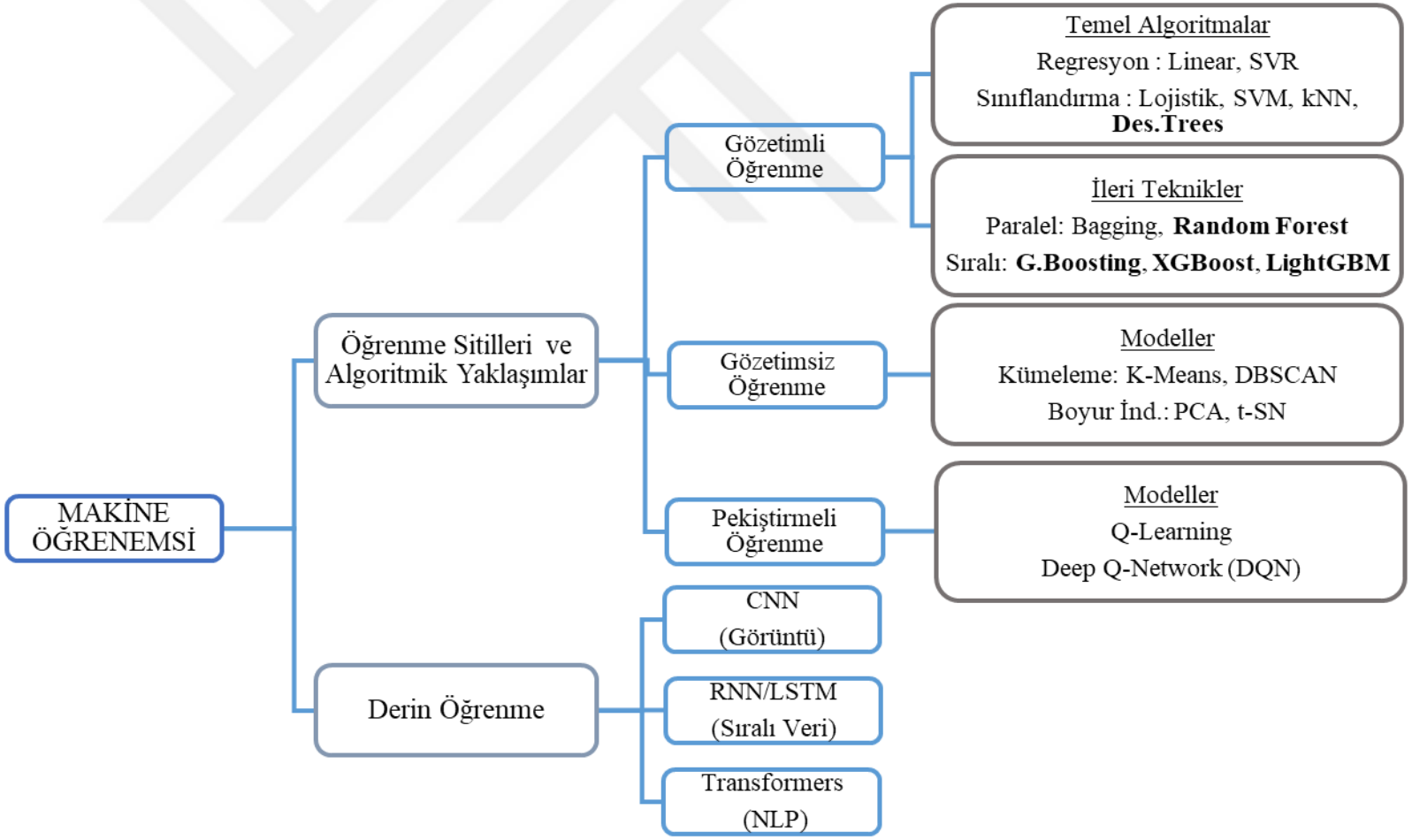
MÖ, verilerin etiketli olup olmaması, geri bildirim mekanizması ve kullandığı tekniklerin derinliğine göre sınıflandırılır. Paradigma seviyesinde farklı, uygulanan teknik genişlik bakımından dolayı farklı bir sınıflandırma çatısı uygulanabilir. Bu çalışma kapsamında yapılan sınıflandırma değerlendirilmesi Şekil 3.7'de gösterilmiştir.

3.4.1 Gözetimli Öğrenme (Supervised)

Gözetimli öğrenme, kurulacak model algoritmasını eğitmek için etiketli veri kullanan bir MÖ yaklaşımıdır [137]. Model etiketli verileri kullanarak eğitilir, sonrasında test verileri ile tahmin yaklaşımının doğruluk oranına göre döngü devam ettirilir veya kesilir. Bu yaklaşım daha yüksek doğruluk oranlarına ulaşabilmesine rağmen veri ön işleme, veri etiketleme gibi işlemler gerektirir. İlk geliştirilen temel algoritmalar ve sık kullanılan modeller bu sınıfta olmuş; veri işleme, aşırı öğrenme sorunu ve algoritma bazlı sorunların çözümü için ileri tekniklerle geliştirilmiştir. Gözetimli öğrenme aşağıdaki gibi sınıflandırılır.

3.4.1.1 Temel Algoritmalar ve Modeller

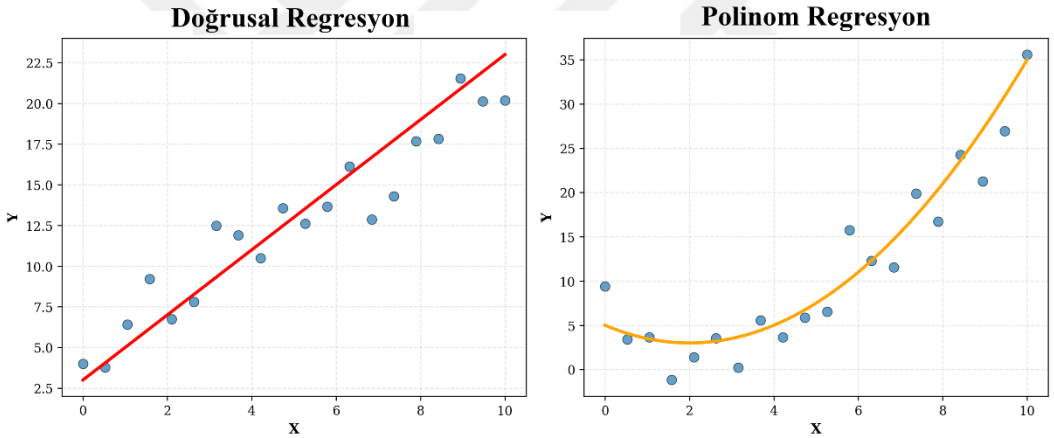
Bu gruptaki yöntemlerde genellikle tekil MÖ modelleri kurulmakta, elde veri setinde eğitilerek tahminleme amacı ile kullanılırlar [137]. Bu gruptaki algoritmalarda model, girdi ile çıktı arasındaki ilişkiyi doğrudan öğrenmeyi hedeflemekte, herhangi bir başka model algoritmasından yardım almamakta, ya da aşamalı bir yapı içermemektedir. Temel algoritmalar, yüksek yorumlanabilirlik özellikleri, göreceli düşük maliyetli olarak kullanılabilmeleri ve karşılaştırma çalışmalarında akademik kolaylık sağladıkları için yaygındırlar [198]. Bu gruptaki algoritmalar, regresyon ve sınıflandırma modelleri olarak iki ayrı başlık altında toplanmışlardır.



Şekil 3.7 Öğrenme şekillerine göre makine öğrenmesi algoritmalarının yapısal gösterimi

1) Regresyon Modelleri: Bu modeller, sürekli nicel çıktı değişkeninin tahmini için kullanılan gözetimli MÖ algoritmalarıdır. Regresyon modellemeleri mühendislik uygulamalarında ve özellikle çevresel sistemlerin modellenmesinde kullanılırlar [198].

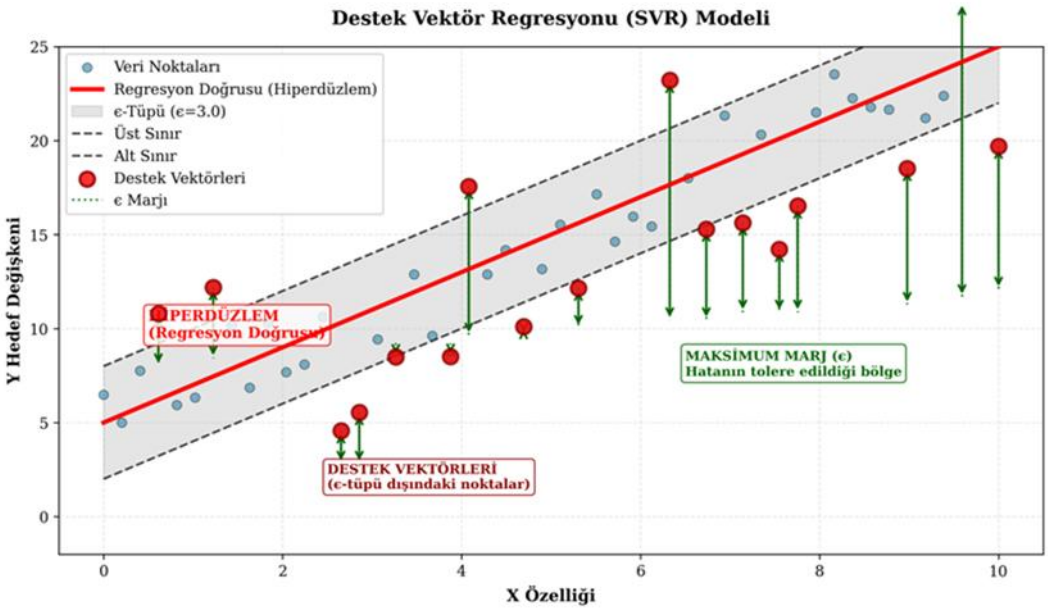
a. Doğrusal ve Polinomsal Regresyon: Değişkenler arasındaki ilişkiyi doğrusal bir denklem ile veya n. dereceden polinomlarla modeller. Genel olarak $y = wx + b$ şeklinde ifade edilmektedir. Doğrusal denklemin, modeli tanımlamakta yetersiz kaldığı durumlarda veriler arasındaki ilişki birden çok doğrusal ilişki ile tanımlanmaya çalışılarak polinomsal regresyon uygulanır. Bu tam doğrusal olmayan ilişkilerin yakalanması için uygun bir yol açmadır [139]. Grafiksel gösterimi Şekil 3.8'de gösterilmektedir.



Şekil 3.8 Doğrusal ve polinomsal regresyon

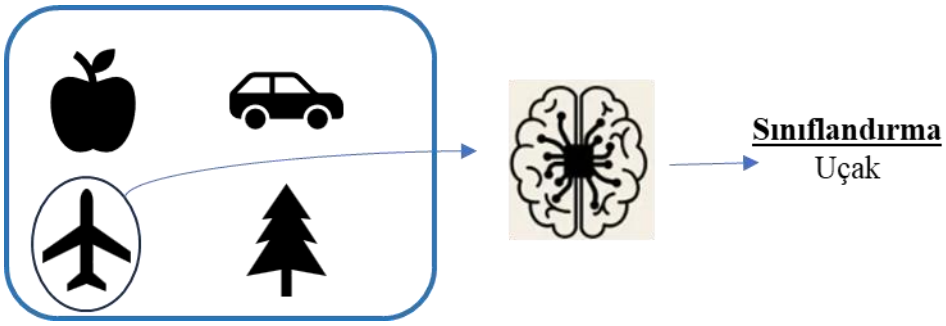
b. Destek Vektör Regresyonu (SVR): Destek vektör makinelerinin prensiplerini regresyon problemlerine uyarlayan bir yöntemdir. SVR yaklaşımında amaç, model uygulandıktan sonra tahmin sonucu ile gerçek değerler arasındaki farkın, önceden tanımlanan bir € tolerans aralığı içerisinde kalmasına çalışan ve model karmaşıklığını azaltan fonksiyonun elde edilmesidir. Yani hem tahmin hatasını küçültmek hem de modelin genelleme yeteneğini arttıran bir optimizasyon kurma işlemidir [199]. Şekil 3.9'da, yöntemin mantığı şematik olarak gösterilmiştir. Burada; € tüpü, tahmin çizgisinin aşağı ve yukarısında tolerans değeri kadar (€) oluşturulan bant arlığıdır. Eğer bir veri bu bant içerisinde kalıyorsa model için sıfır hata olarak kabul edilir, bant dışında kalan noktalar

hata olarak kabul edilir ve model bu noktalara göre kendini düzeltir [184]. Bandın dışında kalan ve modelin öğrenmesine destek veren bu noktalara, destek vektörleri denir [200]. Modelin şeklini veren (doğrunun eğimi ve konumunu) bu az sayıdaki noktalardır. Bant içinde kalan noktalar modelin eğitilmesinde herhangi bir etkiye sahip değildir. Bu yöntem küçük ve orta boyutlu veri setlerindeki doğrusal olmayan ilişkileri modellemede güçlü bir performans sunar.



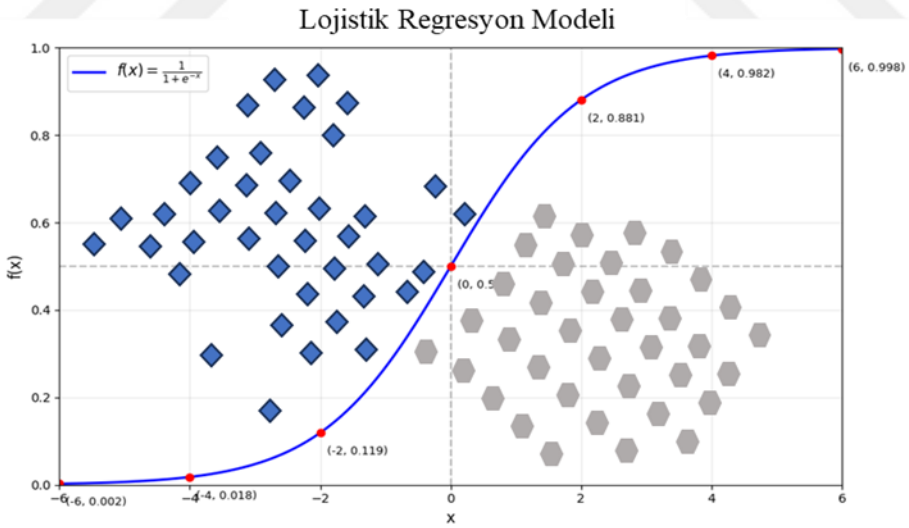
Şekil 3.9 Destek vektör regresyonu (SVR)

2) Sınıflandırma Modelleri: Çıktı değişkeninin sonlu sayıda sınıf etiketi aldığı MÖ yöntemidir. Bu modellerin amacı; girdi değişkenlerine bağlı olarak, gözlemlerin hangi sınıfa ait olduklarının belirlenmesidir. Sınıflandırma modelleri, karar verme süreçlerinde, örüntü tanımada ve tahmin oluşturmada yaygın kullanılan bir tür etiketleme yöntemidir [198]. Model grubunun çalışma mantığı Şekil 3.10'da şematik olarak gösterilmiştir. Alt başlıklarda aşağıda verildiği şekilde birçok algoritması vardır.



Şekil 3.10 Sınıflandırma modelleri düşünme şekli

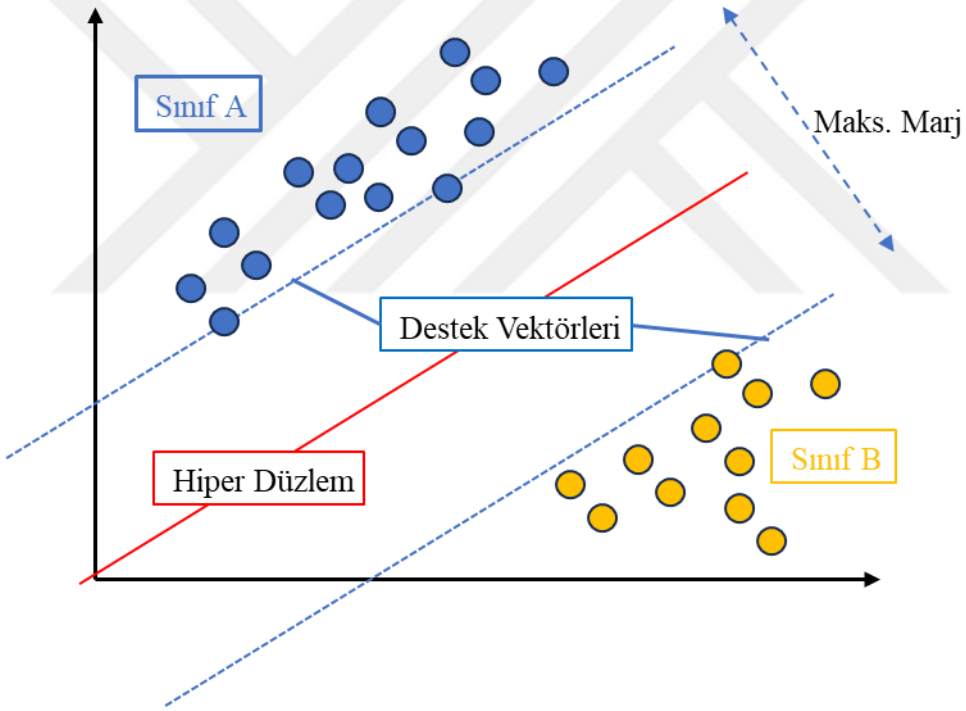
a- Lojistik Regresyon: Sınıflandırma problemlerinde en yaygın kullanılan modeldir ve özellikle ikili sınıflandırmaya özgü düzenlenmiştir. Bu algoritmada, modelin çıktısı, doğrusal bir kombinasyonun lojistik fonksiyon aracı ile olasılık değerine dönüştürülmesi ile elde edilir. $F(x) = 1/(1 + e^{-x})$ lojistik fonksiyonu bağımsız bir değişkeni 0 ile 1 arasındaki bir değere çevirir. Genelde eşik değer 0,5 olarak kabul edilir. Model çalıştıktan sonra üretilen 0-1 arasındaki nümerik değerin büyüklüğüne, yani eşik değerin aşılıp aşılmadığına göre onaylandı veya onaylanmadı gibi bir sınıflama yapılır [200]. Şekil 3.11'de görüldüğü gibi lojistik(sigmoid) grafiğindeki eşik değer çizgisine göre sınıfsal gruplama yapılır.



Şekil 3.11 Lojistik regresyon algoritması

b- Destek Vektör Makineleri (SVM): Bu algoritmada veriler, yüksek boyutlu bir uzayda noktalar olarak ele alınır ve noktalar en uygun şekilde sınıflandırılmaya çalışılır. SVR'den farklı olarak burada sürekli bir değer tahmini ile uğraşılmaz; veri noktalarını birbirinden ayıran en güvenli sınır çizilmeye çalışılır. Algoritmanın temel görevi pozitif örnekleri (+1), negatif örneklerden (-1) ayıran

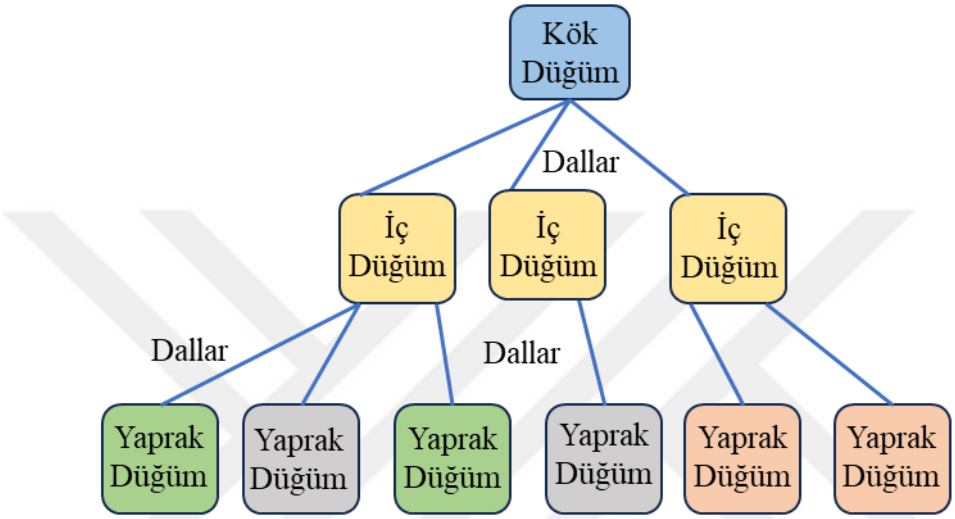
bir sınır çizgisi -teknik terim olarak, “hiper düzlem”- çizmektir. Burada karar sınırına en yakın olan veri noktalarına “destek vektörleri”, karar sınırı ile bu vektörler arasındaki mesafeye de “marj” denir. SVM sadece veri setindeki sınıfları oluşturmaya çalışmaz, marjı maksimum yapmaya çalışır. Marjı maksimum yapmanın mantığı, genelleme yeteneğinin artırmaktadır. Örneğin yeni bir sensörden gelen verileri marj ne kadar geniş olursa o kadar doğru sınıflandırabilir [200]. Şekil 3.12’de sistemin mantığı anlatılmaktadır. Şekilde görülen hiper düzlemi oluşturan vektörler arasındaki mesafeyi iki ülke arasında çekilen tampon bölge gibi düşünebiliriz Bu tampon bölgenin genişliği, yani vektörler arasındaki marjın büyüklüğü ne kadar fazla ise iki ülkenin birbirinden ayrımı o kadar kesindir; hangi bölgenin toprağının nereye ait olduğu o kadar tartışmasızdır.



Şekil 3.12 Destek vektör makine algoritması (Kaynak: [200])

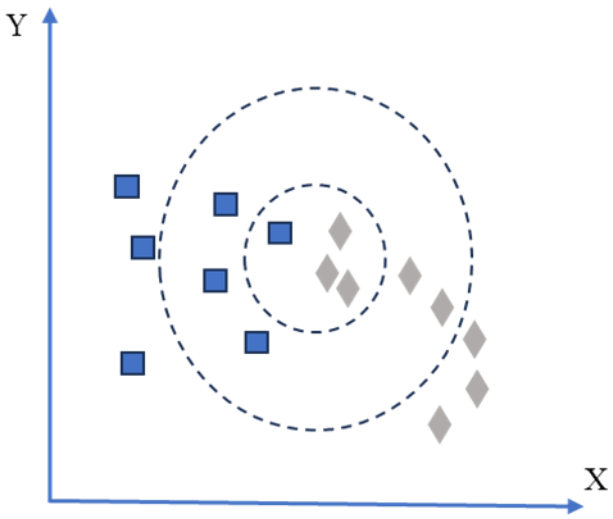
c- Karar Ağaçları (Decision Tree): Karar ağaçları algoritması; girdi olarak özellik vektörlerini kullanıp bir dizi karar kuralları aracılığı ile veriyi alt kümelere ayıran gözetimli bir MÖ’dir. Model; kök düğümden başlayarak iç düğümler ve yaprak düğümlerden oluşan bir ağaç formundadır. Modelde iç düğümler her bir özellik üzerindeki testi, dallar bu testin olası sonuçlarını, yapraklar ise nihai tahmin edilen sınıf etiketini temsil etmektedir [198]. Ağacın büyümesi, her adımda

veriyi en iyi şekilde ayıran özniteliğin seçilmesi ile gerçekleşir. Karar ağaçları parametrik olmayan modellerdir; bu sebeple eğitim veri setlerine aşırı uyum sağlamaya yatkındırlar. Bunun önüne geçmek için ağaç derinliği sınırlandırılabilir veya ağaç tam olarak büyüdükten sonra modele katkı sağlamayan dallar budanır. Karar ağaçları algoritmasının şematik gösterimi Şekil 3.13'te sunulmuştur. Bu yöntem tezde kullanılmış olup metodoloji kısmında daha detaylı bir şekilde değinilecektir.



Şekil 3.13 Karar ağaçları algoritması

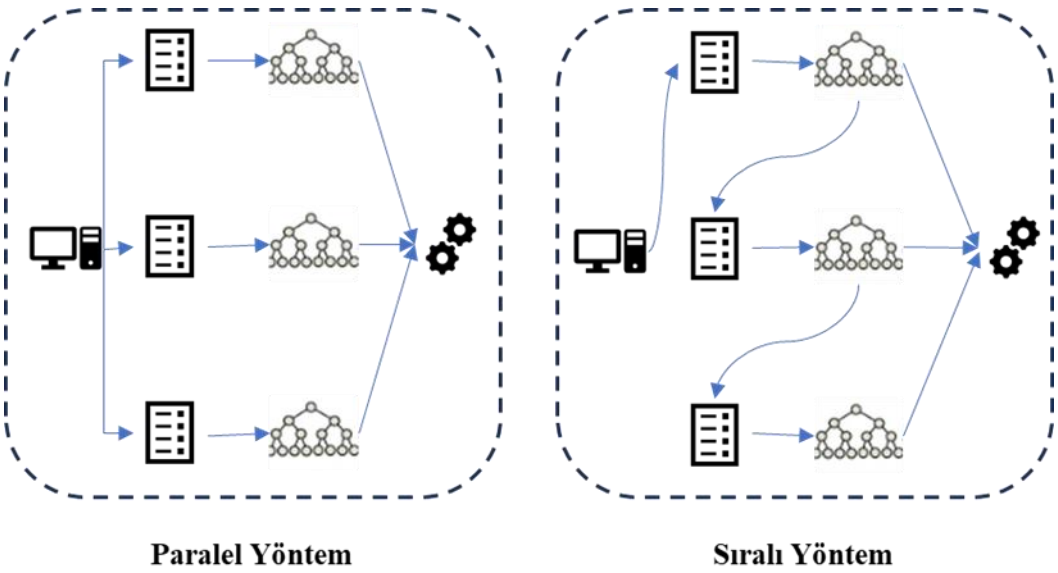
d- En Yakın Komşu Algoritması (k-NN): Yöntem; örnek tabanlı veya tembel öğrenme yaklaşımı ile çalışan, gözetimli bir sınıflandırma algoritmasıdır. Eğitim aşamasında tanımlı bir model veya karar fonksiyonu olmayıp, eğitim veri kümesindeki tüm verilerin gözlem değerlerini özellik uzayında bir referans kümesi şeklinde saklar [139]. Veri noktaları, uzayda birbirine yakın veya uzak olarak kümelenir. Sınıflandırma aşamasında; yeni bir veri bu uzaydaki konumuna göre değerlendirilir ve bu örneğe en yakın "k" adet komşu belirlenir (Şekil 3.14). Her bir "k" komşu arasındaki uzaklık Öklid mesafesi ile hesaplanır.



Şekil 3.14 En yakın komşu algoritması

3.4.1.2 İleri Teknikler-Topluluk Yöntemleri

Topluluk öğrenmesi (ensemble learning); tek bir iyi tahmin oluşturan model kurmak yerine, birden çok zayıf algoritmanın tahminlerini birleştirerek daha yüksek doğruluğu olan bir "meta-model" oluşturmayı hedefleyen gelişmiş, ileri teknik bir MÖ paradigmasıdır [200]. Bu yaklaşımın literatürdeki bir diğer terimsel tanımı "kalabalıkların bilgeliği" ifadesidir [184]. Bu yöntemler; özellikle karar ağaçlarında olduğu gibi varyansı veya yanlılığı azaltmak için kullanılırlar. Bu yeni nesil MÖ uygulamaları literatürde, özellikle karmaşık veri setlerinde DÖ'ye güçlü bir alternatif olarak önerilmektedir [200]. Buradaki varyans ifadesi modelin, eğitim verisindeki gürültüye karşı aşırı duyarlı olmasını (aşırı öğrenme) tanımlarken; yanlılık ifadesi de verideki karmaşık ilişkileri yakalayamamasını (eksik öğrenme) belirtmektedir. Topluluk öğrenmesi; bu amaçları doğrultusunda paralel ve sıralı olmak üzere iki ayrı şekilde sınıflandırılır (Şekil 3.15). Paralel topluluk öğrenmesi sınıfı varyansı azaltmaya odaklanırken, sıralı topluluk öğrenmesi sınıfı da yanlılığı azaltmayı amaçlar. Bunlar da kendi aralarında yaklaşım farklılıklarına göre sınıflandırılırlar.



Şekil 3.15 Topluluk öğrenmesinde paralel ve sıralı öğrenme yöntemleri

Bu tez kapmasında dört adet topluluk yöntemi kullanılmıştır. Aşağıdaki şekilde alt başlıklarda incelendiği şekilde sınıflandırılmaktadır.

1) Paralel Yöntemler: Paralel yöntemlerde bireysel öğrenici modeller birbirinden bağımsız ve eş zamanlı olarak eğitilir. Burada; bireysel ve bağımsız hataları olan modellerin birleştiğinde daha az hata ve daha yüksek doğrulukla güvenilir çıktılar ürettiği gözlenmektedir [200]. Paralel yöntemler kendi içinde iki farklı yaklaşım sergilerler. Bagging yöntemi veri örneklerini rastgeleleştirirken; rastgele uzay yöntemi, modeli çeşitlendirmek için özellikleri rastgeleleştirir.

a- Bagging (Bootstrap Aggregating): Bagging yöntemi, paralel topluluk öğrenmesinin en temel örneklerinden biridir. Bu yöntemde; eğitim veri setinde yerine koymalı örnekleme ile birden fazla alt veri eğitim kümesi oluşturulur. Oluşturulan bu veri alt kümeleri için aynı MÖ algoritması uygulanır ve modeller birbirinden bağımsız olarak eğitilir. Eğitim süreçleri paralel olarak yürütülür; tahmin aşamasında ise her bağımsız tahmin sonucu birleştirilerek tek bir tahmin değeri üretilir [200].

b- Rastgele Alt Uzay Yöntemi (Random Subspace Method): Bu yöntem, bireysel modeller arasındaki korelasyonu azaltarak modelin genelleme yeteneğini iyileştirmeyi amaçlar. Bagging'den farklı olarak aynı veri setinde birden fazla farklı eğitim algoritması çalıştırır. Başka bir ifade ile aynı gözlemler sabit tutulurken, her model için kullanılan öznitelikler farklılaştırılır. Bu şekilde her

temel model, problemin başka bir kısmına veya farklı bir özellik kombinasyonuna göre eğitilir. Bu iki yaklaşım arasındaki fark aşağıdaki Tablo 3.1'de verilmiştir.

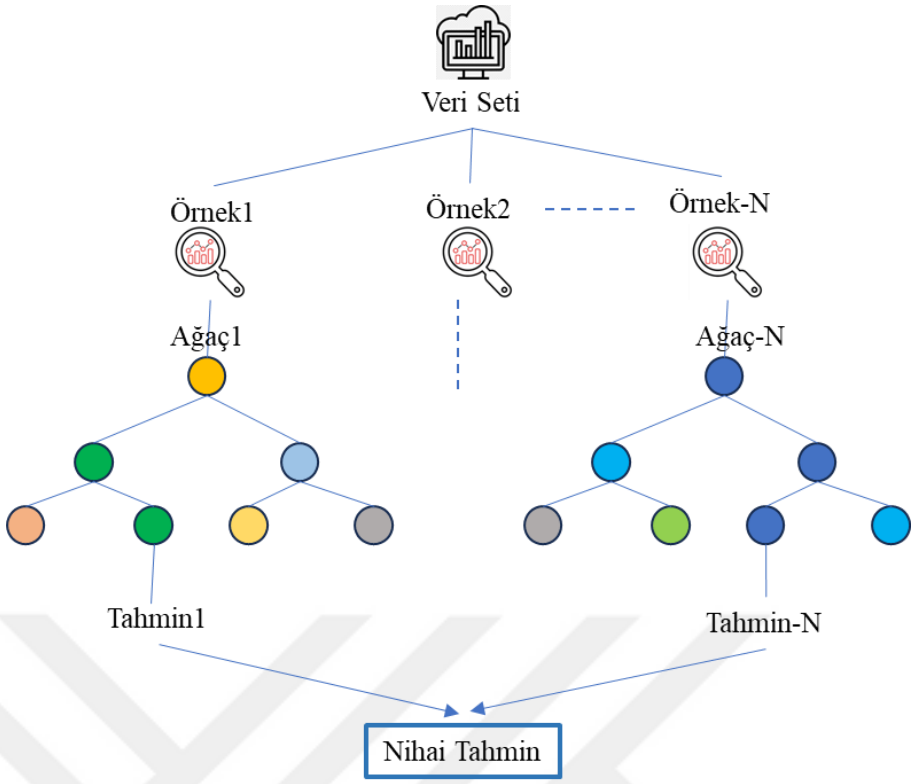
Tablo 3.1 Bagging ve rastgele alt uzay yöntemlerinin karşılaştırılması

Kriter	Bagging	Rastgele Alt Uzay
Rastgelelik	Örnekler (satırlar)	Özellikler (sütunlar)
Veri örnekleri	Farklı	Aynı
Özellik Seti	Aynı	Farklı
Amaç	Varyansı azaltmak	Benzerlik Azaltımı
Tipik Model	Karar Ağacı	kNN

Paralel yöntemlerin her iki alt sınıfının özelliklerini (Bagging ve Random Subspace) kullanan en bilindik örnek algoritma, Rastgele Orman (Random Forest) algoritmasıdır. Rastgele Orman modeli; bagging ve rastgele alt uzay yaklaşımlarının rastgeleliklerini kullanan hibrit bir algoritmadır.

- Rastgele Orman Algoritması (Random Forest): Bu algoritma, bagging yaklaşımını temel alan, karar ağaçlarının topluluk öğrenmesini şeklinde geliştirilmiş ve rastgele alt uzay yönteminin de rastgeleliğini kullanan bir paralel yöntem algoritmasıdır [201]. Bu yöntemde, her bir karar ağacı algoritması farklı bir veri alt kümesinde eğitilir (bagging özelliği); ağacın büyüme sürecinde ise rastgele belirlenen bir özellik alt kümesi içerisinden en uygun bölünme noktası aranır (subspace özelliği). Bu yaklaşım; özellikle veri setinde en önemli özelliklerin baskınlığının önüne geçip, tüm farklı ağaçlarda benzer bölünmelere gidilmesini engellemektedir [200].

  Bu yöntemi, bir konu hakkındaki bir kitabı 100 farklı kişiye (kişi=algoritma) verip, farklı bölümlerini okutup, onalar ortak bir soru sorup, kısıtlanmış kelime grubu ile cevap vermelerini istemeye benzetebiliriz. Böylece her bir birey problemin farklı bir detayında uzmanlaşır ve grubun ortak kararı tek bir kişinin gözünden kaçabilecek tüm detayları da kapsamış olur. Tez çalışmamıza bu algoritma, bahsedilen özelliklerinden dolayı dahil edilmiştir. Bu sebeple burada bir alt başlık açılarak konu detaylandırılmıştır.



Şekil 3.16 Rastgele orman algoritması

2) Sıralı Yöntemler (Boosting): Bu yöntemin tek temsilcisi "Artırma (Boosting)" yaklaşımıdır. Bu yaklaşımın ana yapısı; birden fazla MÖ algoritmasının, birbirine bağımlı ve sıralı bir şekilde peş peşe eğitildiği topluluk öğrenmesidir. Artırma yöntemlerinde, topluluktaki her yeni model, kendinden önceki model veya modellerin hatalarından eğitilir ve tahmin performansını iteratif bir şekilde arttırır [139]. Yani her yeni model önceki modelin hatalı tahmin ettiği örneklere daha fazla ağırlık vererek öğrenme sürecini sürdürür. Bu yönü ile bagging yaklaşımından farklı olarak varyansı değil yanlılığı azaltmaya çalışır. Birçok geliştirilmiş model uygulamaları ve bazı modellerinde geliştirilmiş ileri alt uygulamaları vardır.

a- Adaptif Artırma (AdaBoost): 1997 yılında geliştirilen boosting yaklaşımının ilk ve pratik uygulamasıdır. Temel prensibi; yanlış sınıflandırılan örneklere daha fazla ağırlık vererek sonraki zayıf MÖ algoritmalarının bu örneklere odaklanmasını sağlamaktır. Modelde, her iterasyon adımında tekil MÖ algoritmasının ağırlığı tahmin hatasına bağlı olarak hesaplanır. Nihai modelin tahmin doğruluğu, tüm öğrencilerin kombinasyonuna göre belirlenir. Bu sebeple "adaptif" olarak isimlendirilmiştir [184]. Adaptif artırmanın, düşük karmaşıklığa

sahip karar ağaçları ile birlikte kullanıldığında daha yüksek performansla tahminler yapabildiği literatürde belirtilmiştir. En olumsuz yönü ise gürültülü ve aykırı değerlerin yüksek olduğu veri setlerinde yöntemin hassasiyetini kaybetmesidir [200].

b- Gradyan Artırma (Gradient Boosting): Boosting yaklaşımını daha genel ve matematiksel olarak ele alan bir yaklaşımdır. Bu yaklaşımın temel fikri; öğrenme problemini bir optimizasyon işi olarak ele alarak her iterasyonda modelin hatalarının gradyan (türev) olarak minimize edilmesini sağlamaktır [199]. Adaptif artırmada her yeni zayıf öğrenci önceki öğrenicilerin en çok hata yaptığı örneklere odaklanarak ağırlıklarını değiştirirken; gradyan artırmada her yeni öğrenici önceki modelin yaptığı kalıntı hataları tahmin etmeye çalışarak, gradyan inişiyle hatayı minimize ederek optimizasyon yapmaya çalışır [184]. Gradyan artırma daha gelişmiş bir tekniktir ve karmaşık problemlere daha kolay uygulanır. Adaptif artırma daha tek seviyeli karar ağaçları (stumps) kullanırken; gradyan artırmada daha derin ağaçlar kullanılır. Gradyan artırma hem bir yöntem hem de bir algoritmadır. Ayrıca bu modelin geliştirilmiş daha güncel algoritmaları da vardır ve alt başlıklar altında incelenmektedir. Bunlar, bu tez çalışması kapsamında MÖ algoritması olarak kullanılmışlardır.

i) Gradyan Artırma Makineleri (GBM): Gradyan artırma makineleri (GBM) olarak doğrudan model olarak uygulanan algoritma, Friedman tarafından geliştirilmiş klasik bir artırma modelidir [139]. Model; adım adım ve artırımlı olarak oluşturulan etkili bir MÖ algoritmasıdır. Sıralı modeller eğitilme sürecinde, kendinden bir önceki modelin hatalarını düzeltmeyi öğrenir ve böylelikle kademeli olarak genel tahmin doğruluğu gelişmiş olur. Gradyan artırma genellikle karar ağaçları başta olmak üzere zayıf tahmin modellerini birleştirerek güçlü bir topluluk öğrenme modeli oluşturma şeklinde çalışır. Tablo 3.2'de, gradyan artırmanın matematiksel adımları gösterilmektedir. Burada görüleceği üzere ilk adımda tahmin modeli oluşturulur. Veri kümesindeki her örnek, bu model tarafından üretilen ilk tahminlere sahiptir. İkinci adımda bu modellerin artık (residual) hataları hesaplanır. Beklenen değer ile gerçekleşen değer arasındaki fark artık değerdir. Üçüncü adımda ise önceki ağaçların yaptıkları hatalar düzeltilerek öğrenme devam eder; her yeni ağaç, önceki ağacın artıklarını tahmin etmek üzere eğitilir. Dördüncü adımda, kayıp fonksiyonunu (hatayı) en

aza indirgemek için modelin parametreleri değiştirilir. Bu minimizasyon için kayıp fonksiyonunun negatif gradyanına yeni ağaçlar uyarlanır ve kademeli artırma sağlanır. Son olarak her ağaçtan elde edilen ağırlıklı tahminlerden genel ve son tahmin elde edilir [202].

GBM, genelde yüksek tahmin doğrulukları ortaya koymasına rağmen; ağaç derinliği, öğrenme oranı ve iterasyon sayısı uygun seçilmediğinde aşırı öğrenme riski ortaya çıkmakta ve tahmin doğruluğu azalmaktadır [184].

Tablo 3.2 Gradyan artırma logaritması (Kaynak:[202])

1. **Input :** Eğitim veri seti $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$, Temel Öğrenen h(x), Öğrenme Oranı η, Ağaç Sayısı T
2. **Output :** Son Tahmin Modeli FT (x)
3. **Adım 1:** Modeli Başlat
4. Başlangıç tahminini hedef değerlerin ortalaması olarak ayarlayın:

$$F_0(x) = \arg\min_c \sum_{i=1}^{N} L(Y_i, c) \tag{3.2}$$

5. **Adım 2:** t = 1'den T'ye kadar yineleme yapın
6. Kalıntıları Hesapla (Kayıp Fonksiyonunun Negatif Gradyanı):

$$r_i^{(t)} = -\frac{\partial L(Y_i, F_{t-1}(X_i))}{\partial F} \quad . \tag{3.3}$$

7. Zayıf bir öğreneni ht(x) kalıntıları tahmin edecek şekilde eğitin:

$$h_t(x) = \arg\min_h \sum_{i=1}^{N} (r_i^{(t)} - h(X_i))^2 \tag{3.4}$$

8. En uygun adım boyutunu γt hesaplayın:

$$\gamma_t = \arg\min_\gamma \sum_{i=1}^{N} L(Y_i, F_{t-1}(X_i) + \gamma h_t(X_i)) \tag{3.5}$$

9. Modeli güncelleyin:

$$F_t(x) = F_{t-1}(x) + \eta\gamma_t h_t(x) \tag{3.6}$$

10. **Adım 3:** Son Modele Geri Dön FT (x)

ii) Gelişmiş Gradyan Artırma-XGBoost: XGBoost algoritması; GBM'in çeşitli düzenlemelerle optimize edilmiş, yüksek performanslı şeklidir. 2016 yılında geliştirilmiş olan bu model; GBM'in kısıtı olan aşırı öğrenmenin önüne geçmekte, boş verileri yönetebilmekte ve tüm bunları daha hızlı yapabilmektedir [203]. Klasik GBM artırma algoritması yalnızca hata kayıp fonksiyonunun birinci dereceden türevine bağlı olarak hata kalıntılarını öğrenirken; XGBoost, GBM'in bu yapısına düzenlileştirilmiş bir amaç fonksiyonu ve ikinci dereceden türev bilgisi ekleyerek, modeli daha kararlı ve hesaplama açısından da daha verimli bir öğrenme sürecine yönlendirmektedir. Yukarıda sunulan Tablo 3.2'deki GBM'e ait algoritma yapısına eklenen ve XGBoost algoritmasını oluşturan değişiklikler Tablo 3.3'te verilmiştir. Tablo 3.3'te tüm formülasyon değil, sadece değişiklik yapılan kısımlar madde numarasına göre eklenmiştir. XGBoost; başarısının temelinde bulunan, yukarıda ve Tablo 3.3'te anlatılan farklılıklar sayesinde klasik GBM'e kıyasla literatürdeki birçok anlatımda daha üstün performans sergilediği görülmektedir [203]. Bu tez kapsamında klasik GBM de XGBoost da kullanılmıştır. Literatüre bu yönüyle de katkı sağlanmış olacaktır.

iii) Hafif Gradyan Artırmalı Makine (LightGBM): Microsoft firması tarafından geliştirilen LightGBM, gradyan artırma makinelerinin evriminin son halkalarından biridir. GBM ve XGBoost'ta karşılaşılan maliyet ve bellek kısıtlarını aşmak amacı ile geliştirilmiş yüksek performanslı bir modeldir [204]. Bu model ile getirilen yenilik, "verinin tamamını eğitime dahil etmeli miyiz?" sorusuna verilen iki ayrı cevaptır. Sorudan da anlaşılacağı üzere sonuca daha kısa yoldan gitme amacı taşımaktadır. LightGBM'in bu soruya ilk yaklaşımı; gradyan hatası büyük olan veri noktalarının öğrenme sürecine daha fazla bilgi kattığı varsayımı ile büyük gradyanlı verilerin tamamını tutarken, küçük gradyanlı verilerin bir kısmını atıp kalan kısmını ise bir katsayı ile ağırlıklandırarak dağılımın bozulmasını engellemektir. İkinci cevap yaklaşımı ise; yüksek boyutlu, seyrek veri setlerinde birbirini dışlayan özellikleri birleştirerek özellik sayısını azaltmaktır. Bu iki kolaylaştırıcı hamle ile LightGBM; XGBoost kadar yüksek doğruluğa ulaşan tahminler yapabilmekte ve eğitim süresini kısaltarak bellek kullanımını anlamlı şekilde düşürmektedir [204]. Aslında LightGBM'i ayıran temel fark; diğer boosting metodolojisindeki

modeller genelde ağacı yatay olarak (level-wise) büyütürken, LightGBM'in yaprak odaklı (leaf-wise) olarak ağacı büyüterek hatayı daha belirgin şekilde minimize etmesidir. Bu tez çalışması kapsamında, bu özelliklerinden dolayı çalışmaya dahil edilmiştir.

Tablo 3.3 XGBoost'ta eklenen matematiksel ifadeler

1. Input: Bu kısma λve γ düzenlileştirme parametreleri eklenir.

6. Kalıntıları Hesapla (Kayıp Fonksiyonunun Negatif Gradyanı):
Aynı formül kullanılır ancak XGBoost ikinci derece türevi (Hessian) de kullanılır.

$$r_i^{(t)} = -\frac{\partial L(Y_i, F_{t-1}(X_i))}{\partial F} \tag{3.3}$$

Ayrıca XGBoost'ta g_i ve h_i şu şekilde tanımlanır:

$$g_i = \frac{\partial L}{\partial F} \quad ; \quad h_i = \frac{\partial^2 L}{\partial F^2} \tag{3.7}$$

7. XGBoost'ta ağaç büyüme kare kayıp + düzenlileştirme terimi ile yapılır:

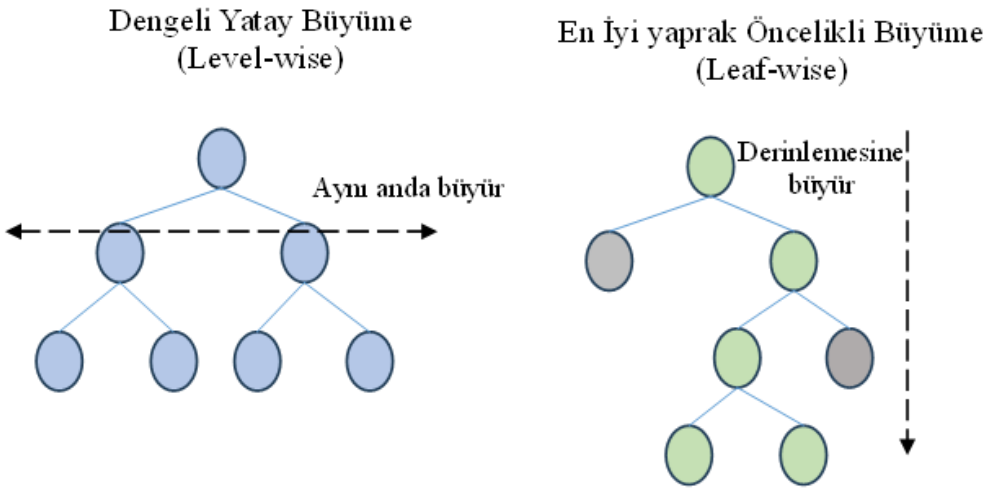
$$Amaç: min \sum_{i=1}^{N} \left[g_i h_t(X_i) + \frac{1}{2} h_i h_t(X_i)^2 \right] + \lambda \|w\|^2 + \alpha \|w\|_1 \tag{3.8}$$

8. Bu adımda XGBoost'ta λ_t ağırlık olarak her yaprak için ayrı ayrı hesaplanır ve düzenlileştirme parametrelerini de içerir.

$$\gamma_{tj} = -\frac{\sum_{i \epsilon I_j} g_i}{\sum_{i \epsilon I_j} h_i + \lambda} \tag{3.9}$$

9. Aynı formül kullanılır ancak; λ_t yukarıdaki hesaplamalardan dolayı artık düzenlileştirmeli olarak hesaplanmış olur.

Tablo 3.4'te LightGBM algoritmasını GBM'den ayıran formül yapısı verilmiştir. En belirgin değişiklikler girdi (input) kısmındadır. Histogram hızı, yaprak ayırma kısıtı, örnekleme ayırımı, özellik ve veri seçimi manuel olarak girilir. Diğer adımlarda, değişen yapı gereği ağacın yatay değil derinlemesine (yaprak odaklı) büyümesini gerektiren matematiksel ifadeler eklenmiştir. Hata hesaplamaları ve döngüler de buna göre değişmiştir. Modelin, GBM ve XGBoost'a göre ağaç oluşum düzeni Şekil 3.17'de verilmiştir.



Şekil 3.17 GBM-XGBoost ile LightGBM modellerinin ağaç büyüme düzeni

c) Kategorisel Artırma (CatBoost): 017 yılında geliştirilen bu yöntem; özellikle kategorik (sözel) değişkenlerin yoğun olarak bulunduğu veri setlerinde, ön işleme gerek duymadan modelleme yaparak yüksek tahmin başarısı gösterebilmektedir [205]. Temel yeniliği; veriyi rastgele permütasyonlarla sıralayarak, her bir örneğin hedef istatistiğini hesaplarken yalnızca kendisinden önceki verilere bakmasıdır. Böylelikle tahmin kayması (prediction shift) problemini çözmeye çalışır. Ayrıca bu yöntem, simetrik ağaç yapısı kurarak modelin tahmin süresini hızlandırır. CatBoost; karmaşık verilerde ön işleme gerektirmeden yüksek performans sunabilmektedir. Ancak genelde, sayısal verilerin çok olduğu mühendislik uygulamalarında tercih edilmemektedir.

d) Stokastik Gradyan Artırma (SGB): 2002 yılında J. H. Friedman tarafından önerilen bu yöntem; klasik GBM modelinin rastgelelik (stokastiklik) eklenerek genişletilmiş bir versiyonudur [206]. Algoritma; her iterasyonda hatayı hesaplamak için tüm veriyi kullanmak yerine, eğitim verisinin rastgele seçilmiş bir alt kümesini kullanır. Bu yaklaşım modelin varyansını azaltarak, modelin aşırı öğrenme riskini düşürür ve eğitim hızını artırır. SGB; sınıflandırma ve regresyon problemlerinde yaygın olarak kullanılan, teorik olarak güçlü ve esnek olan bir gradyan artırma algoritmasıdır.

Tablo 3.4 LightGBM'de eklenen matematiksel ifadeler

1**. Input**: Bu kısma şu aşağıdaki parametreler eklenmiştir:

Histogram bin sayısı: max_bin
Yaprak Sayısı: mum-_leaves
Büyük Gradyanlıların oranı: top_rate
Küçük Gradyanlıların oranı: other_rate
Ağaç başı veri kullanım oranı: bagging_fraciton
Ağaç başı özellik oranı: feature_fraction

7. Zayıf Öğreneni Eğit:
max_bin ile grup oluştur, top_rate ve other_rate ile büyük ve küçük gradyanlı örnekleri böl. Sonrasında aşağıdaki Gain formülü ile en yüksek gain'e sahip yaprağı böl. Gain formülü:

$$Gain = \frac{\left(\sum_{I_L} g_i\right)^2}{\sum_{I_L} h_i + \lambda} + \frac{\left(\sum_{I_R} g_i\right)^2}{\sum_{I_R} h_i + \lambda} - \frac{(\sum_I g_i)^2}{\sum_I h_i + \lambda} - \gamma \tag{3.10}$$

8. En uygun adım boyutunu hesapla:

GBM'de tüm ağaç için bir γ_t var iken LightGBM'de her yaprak için ayrı ağırlık w_i^* hesaplanır.

$$w_i^* = -\frac{\sum_{i \epsilon I_j} g_i}{\sum_{i \epsilon I_j} h_i + \lambda} \tag{3.11}$$

9. Modeli Güncelle:

GBM de tüm ağaç aynı γ_t eklenirken, LightGBM'de x hangi yaprağa denk gelirse o yaprağın ağırlığı w_i^* ile güncellenir;

$$F_t(x) = F_{t-1}(x) + \eta . w_{i(x)}^* \tag{3.12}$$

3.4.2 Gözetimsiz Öğrenme

Gözetimsiz öğrenme yaklaşımları; etiketlenmemiş veri kümeleri üzerindeki örüntüleri, gizli yapısal ilişkileri ve benzerlikleri ortaya çıkarmaya çalışan MÖ yaklaşımlarıdır [199]. Burada model; dışarıdan herhangi bir müdahale olmadan girdi verilerini analiz ederek, yani herhangi bir hedef değişken tanımı olmadan veriyi tanımlayan bir yapı oluşturur. Genellikle üç ana kategori altında incelenir. Bunlar kısaca aşağıdaki başlıklar altında verilmiştir:

3.4.2.1 Kümeleme Analizi (Clustering)

Veri setindeki her bir veriyi benzer özelliklerine göre homojen olarak gruplandırmak, kümeleme işlemidir. Bu gruptaki gözetimsiz öğrenme algoritmaları; her bir kümedeki benzerlikleri maksimize etmeye çalışırken, kümeler arasındaki farklılıkları da maksimize etmeye çalışırlar. En yaygın olarak kullanılan algoritma; merkez tabanlı bir yaklaşım olan k-ortalama (K-Means) algoritmasıdır. K-Means; veriyi önceden belirlenen, merkeze uzaklıkları baz alınan k sayıdaki gruba ayırarak bir sınıflandırma gerçekleştirir. Kümeleme analizi yöntemi kullanan diğer algoritma grupları ise; yoğunluk tabanlı kümeleme (DBSCAN), dağılım tabanlı kümeleme (GMM) ve hiyerarşik kümelemedir [207].

3.4.2.2 Boyut İndirgeme Öğrenimi

Çok geniş ve yüksek boyutlu veri setlerindeki verilerin içerdiği bilgileri aynı tutmaya çalışarak değişkenlerin sayısını azaltmaya dayanan bir yöntemdir. Yapılan bu işlemdeki amaç; hesaplama verimliliğini artırarak model karmaşıklığını azaltmak ve bazen de "boyutluluğun laneti"ni hafifleterek performansı iyileştirmektir [208]. Bu yöntem, iki ana kategoriye ayrılır:

1) Doğrusal Yöntemler: Bu yöntemin en yaygın kullanılan algoritması, Temel Bileşenler Analizi (PCA)'dır. PCA; birbirleri ile ilişkili değişkenleri; birbirleri ile ilişkisiz, dik ve doğrusal bileşenlere dönüştürür. Sonra da bu bileşenleri, verideki varyansı maksimum düzeyde açıklayacak şekilde sıralar [208].
2) Doğrusal olmayan Yöntemleri: Bu yöntemler; genelde yüksek boyutlu veri setlerindeki veriler arasındaki benzerlikleri, daha düşük boyutlu bir haritalamada korumayı amaçlarlar. Bu yöntemler, özellikle veri görselleştirme için etkilidir [209].

3.4.2.3 Birliktelik Kuralı Öğrenimi

Büyük veri setlerindeki verilerin "birlikte görülme" ilişkilerini niceliksel olarak ortaya çıkarmak için kullanılır. Apriori algoritması, bu yöntemin en bilindik temsilcilerinden biridir. Market alışveriş sepeti ve satış fişleri gibi verilerde; "X alındıysa Y de satın alınır" gibi ilişkileri ortaya çıkarmaya çalışır. Yani; "herhangi bir sık öge kümesinin tüm alt kümeleri de sık olmalıdır" ilkesi ile çalışır. Algoritma şu şekilde çalışır: Önce belirli bir destek eşiğini aşan kombinasyonları belirleyerek

sık öge kümelerini bulur; sonra bu sık ögelerden minimum güven değerlerini sağlayan ilişki kurallarını çıkarır [210].

Bu yöntem, genelde müşteri davranışları analizi için kullanılan bir yöntemdir. Kampanya tasarımı, raf yerleşimi optimizasyonu gibi alanlarda yaygın olarak kullanılır.

Bu tez çalışmasında; yukarıda teorik bilgisi kısıtlı olarak verilen gözetimsiz MÖ algoritmaları, AAT çıkış suyu kalitesi tahmininde model oluşturma amaçlı tercih edilmemiştir. Bunun sebeplerinden birkaçı; veri setinin etiketli yapıda olması, değişkenlerin niteliği ve çözmeye çalıştığımız problemin bir regresyon/sınıflandırma problemi olmasıdır. İleride, metodoloji kısmında her bir sebep ayrıca açıklanacaktır.

3.4.3 Pekiştirmeli Öğrenme

Bu öğrenme yöntemi; bir ajanın belirli bir ortamda etkileşime girerek çevresindeki dinamiklere göre, deneme yanılma yoluyla, ödül ve ceza sinyallerine göre karar alma mekanizması geliştirip ödülü maksimize etmeye çalıştığı bir MÖ yöntemidir [211]. Gözetimli öğrenmenin aksine, burada ajana ne yapması gerektiği hakkında doğrudan bir direktif verilmez. Bunun yerine ajanın aldığı eylemlerin uzun vadeli sonuçlarını ödül ve cezaya göre değerlendirerek öğrenmesi beklenir.

Pekiştirmeli öğrenme, Markov Karar Süreçleri (MDP) çerçevesinde tanımlanmaktadır. Bu süreç;

- Durum: Ortamın ajan tarafından gözlemlenen hali
- Eylem: Ajanın bulunduğu durumda seçebileceği hareket.
- Ödül: Eylemden sonra alınan geri bildirim
- Politika: ajanın hangi durumda hangi eylemi seçeceğini belirleyen strateji

Pekiştirmeli öğrenme yaklaşımları genelde; değer tabanlı, politika tabanlı ve aktör-eleştirmen yöntemler olarak sınıflandırılmaktadırlar. En bilindik yöntemleri:

1) Q-Learning: Ajanın; hangi durumda, hangi eylemi yapması gerektiğini belirleyen bir kalite tablosu (Q-Table) oluşturarak öğrenme gerçekleştirir. Burada doğrudan politika fonksiyonu tahmin edilmeye çalışılmaktadır [212].

2) Derin Pekiştirmeli Öğrenme: Q-değerini tahmin etmek için derin sinir ağlarının kullanıldığı bir yöntemdir. Özellikle video oyunları ve gerçek dünya görüntü analizi gibi alanlarda kullanılır [213].

Pekiştirmeli öğrenme; yüksek hesaplama maliyetleri, uzun eğitim süreleri ve karmaşık ortam modelleme gereksinimleri nedeni ile her zaman tercih edilebilir olmamaktadır. Özellikle tam kara kutu olmayan, mekanizması nümerik modellerle ortaya konulabilmiş olan durumlar ve statik veri kümeleri üzerindeki tahmin problemleri için çözüm odaklı sonuçlar üretmeyebilmektedir [211].

Bu çalışmada ele aldığımız problem; çevrim içi karar verme, ajan-ortam etkileşimine dayalı bir yapı barındırmaması ve durum-eylem-ödül döngüsü içermediğinden dolayı, pekiştirmeli MÖ algoritmalarından biri tercih edilmemiştir. Ancak tezin yöntemsel kavramı çerçevesinde bu yaklaşımlar literatür düzeyinde incelenmiş ve "alternatif bir yöntem olabilir mi?" sorusuna cevap aranmıştır.

3.4.4 Derin Öğrenme Mimarileri

Derin öğrenme (DÖ); yapay sinir ağlarının çok sayıda gizli katman içerecek şekilde genişletilmesiyle elde edilen ve verinin hiyerarşik temsillerini öğrenebilen gelişmiş bir MÖ alt dalıdır. Klasik MÖ'de öznitelik çıkarımı uzmanlık gerektiren manuel bir işlem iken; DÖ modellerinde bu süreç otomatikleştirilerek ham veriden anlamlı özellikler doğrudan öğrenilebilmektedir [143]. İnsan beyninin nörolojik çalışma prensiplerinden ve nöronlar arasındaki sinaptik bağlardan ilham alan bu modeller; özellikle büyük veri setlerinde (big data), geleneksel MÖ algoritmalarına kıyasla üstün performans göstermektedir. DÖ, sahip olduğu kendi kendine öğrenebilme yeteneği ile özellikle ses, metin ve zaman serileri gibi yapılandırılmamış veya yüksek boyutlu verilerin analizinde devrimsel başarılar elde etmiştir. DÖ mimarileri, çözülmesi istenilen problemin veri yapısına ve çözüm hedeflerine bağlı olarak farklı şekillerde ortaya konmaktadır. Literatürde yaygın olarak kullanılan mimariler şunlardır:

1) İleri Beslemeli Sinir Ağları: DÖ'nün en temel mimarisidir. Burada bilgi, girdi katmanından çıktı katmanına doğru tek yönlü olarak ilerler; ağ yapısında herhangi bir döngü işlemi bulunmaz. Çok katmanlı algılayıcılar (MLP) olarak da adlandırılan bu yapılar, özellikle tablosal veri setlerinde sınıflandırma ve regresyon görevleri için kullanılır [197].

2) Evrişimli Sinir Ağları (Convolutional Neurol Network-CNN): CNN; barındırdığı matematiksel konvolüsyon (evreşim) işlemine dayanan bu mimari ile verinin mekânsal ilişkilerini koruyarak özellik çıkarımı yapar. Bu yapı, ızgara (grid) tabanlı veri setlerinde yerel örüntülerin öğrenilmesini sağlar. Bu özelliği sayesinde; görüntü işleme, video analizi, medikal görüntüleme ve uzamsal veri analizine dayalı uygulamalardaki başarılarından dolayı yaygın olarak tercih edilir [197].

3) Tekrarlayan Sinir Ağları (RNN) ve Uzun-Kısa Vadeli Bellek (LSTM): RNN; geçmiş zaman adımlarındaki bilgiyi hafızasında tutabilen döngüsel yapıya sahip bir yöntemdir. Bu özelliği sayesinde zaman serileri, doğal dil işleme ve ses tanıma gibi sıralı verilerin analizinde kullanılabilmektedir. Ancak klasik RNN yapıları, uzun veri dizilerinde "kaybolan gradyan" (vanishing gradient) sorunu nedeni ile sınırlı bir performans göstermektedir. Bu sorunun çözümü için de LSTM ağları geliştirilmiştir. LSTM, uzun dönemli bağımlılıkları öğrenmede daha başarılıdır [214].

DÖ modelleri; yüksek öğrenme becerilerine ve kendi kendilerine özellik seçimi yapabilmelerine rağmen; çok büyük verilerde yüksek hesaplama maliyetleri ve sınırlı yorumlanabilirlikleri sebebi ile tercih edilmeyebilmektedir. Mühendislik uygulamalarında; veri seti özellikleri, yapısı ve ele alınan problemin çözümündeki hedefler doğrultusunda DÖ'nün uygunluğu dikkatle irdelenmelidir. Bu tez çalışmasındaki veri seti; düşük-orta büyüklükte, homojen ve iyi yapılandırılmış bir tablo verisinden oluştuğundan ve daha kolay çözümlemeler yapabilen gözetimli MÖ uygulamaları bulunduğundan dolayı DÖ tercih edilmemiştir.

3.5 Model Değerlendirme Metrikleri

MÖ modellerinin geliştirilmesi sürecinde, modellerin tahmin performanslarını nicel olarak ölçmek ve farklı modellerin birbirleri ile kıyaslanmasını sağlamak kritik öneme sahiptir. Model değerlendirme metrikleri; modellerin tahmin performanslarını nesnel olarak ölçmek, farklı modeller arasında karşılaştırma yapabilmek ve en uygun modelin seçilmesine bilimsel bir altlık oluşturmak için geliştirilmiş matematiksel ölçütlerdir [215]. Metrikler; modelin gerçek veriyi ne ölçüde temsil ettiğini, tahmin hatalarının büyüklüğünü ve modelin genelleme yeteneğini ortaya koymaktadır. Model seçimi ile kurulan modelin ilgilenilen

probleme ne denli uyum sağlayarak başarı gösterebildiğini ölçebildikleri için modelleme sürecindeki en kritik rolü üstlenmektedirler.

Literatürde değerlendirme metrikleri, ele alınan problemin türüne göre sınıflandırma veya regresyon olarak iki farklı grupta incelenmektedir. Sınıflandırma problemleri için yapılan model çalışmalarının performansını değerlendiren metrikler; Doğruluk (Accuracy), Hassasiyet (Precision), Duyarlılık (Recall) ve F-1 Skoru gibi yöntemlerdir. Regresyon problemleri için ise Kök Ortalama Kare Hata (RMSE), Ortalama Mutlak Hata (MAE) ve Belirleme Katsayısı (R^2) gibi metrikler öne çıkmaktadır [215]. Değerlendirme metrikleri her geçen gün gelişmekte ve yeni yaklaşımlar literatüre dahil olmaktadır.

Bu tez kapsamında ele alınan problem, AAT çıkış suyu parametrelerinin nümerik olarak tahmin edilmesi olduğundan bir regresyon yaklaşımıdır. Bu nedenle, tez kapsamında model değerlendirme metriği olarak sınıflandırma metrikleri kapsam dışı bırakılmış; burada sadece teorik çerçevede var olduklarına değinilmekle yetinilmiştir. Regresyon analizlerinde literatürde Ortalama Mutlak Yüzde Hata (MAPE), Medyan Mutlak Hata (MedAE) veya Ortalama Sapma Hatası (MBE) gibi metrikler de kullanılmaktadır. Ancak bu çalışmada kullanılan yöntemler aşağıda sırasıyla anlatılmıştır.

3.5.1 Kök Ortalama Kare Hatası (RMSE)

Kök ortalama kare hatası (RMSE), tahmin edilen değer ile gerçek değerler arasındaki farkların karelerinin ortalamasının karekökü olarak tanımlanır [216]. Matematiksel ifadesi aşağıdaki gibidir.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.13}$$

Denklem 3.13'de n örnek sayısını, y_i gerçek değeri $\hat{y}_i$ tahmin edilen değeri ifade etmektedir.

RMSE'nin en önemli özelliği hataları kareleri ile hesaplamasıdır. Bu özelliği sayesinde büyük tekil hatalar, küçük tekil hatalara göre daha büyük katsayılarla ağırlıklandırılır (cezalandırılır) ve metriğin değerini büyük ölçüde artırır. Özellikle büyük hataların yüksek maliyetli veya kabul edilemez olduğu (aşırı yüksek kirletici

konsantrasyonunun gözden kaçırılmaması gibi) mühendislik uygulamalarında bu durum, RMSE'yi tercih edilir bir metrik haline getirir [216]. Ayrıca, karekök alma sonucu RMSE değerinin hedef değişken ile aynı fiziksel birimde olması (mg/L gibi), sonuçların daha yorumlanabilir olmasını sağlamaktadır.

3.5.2 Ortalama Mutlak Hata (MAE)

Ortalama mutlak hata (MAE), tahmin edilen değer ile gerçek değer arasındaki farkın mutlak değerlerinin ortalamasıdır. RMSE'den farkı, hataları karelemek yerine doğrudan farklarını almaktır. Matematiksel ifadesi aşağıdaki gibidir:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.14}$$

Denklem 3.14'de n örnek sayısını, y_i gerçek değeri $\hat{y}_i$ tahmin edilen değeri ifade etmektedir.

Bu hesaplama şekli (karelerin dahil olmadığı), MAE metriğini aykırı değerlere karşı daha az hassas yapar ve dolayısı ile daha dirençli bir metrik olduğunu ortaya koyar [217].

MAE, modelin ortalama veya tipik tahmin hatasının, tek bir uç değerin oluşturduğu büyük hata ile çok fazla etkilenmesine izin vermez. Bu nedenle veri setinde doğal olarak bulunabilen aykırı değer veya ölçüm hatalarından model değerlendirme sürecini korumuş olur.

3.5.3 Belirleme Katsayısı (R^2)

Belirleme katsayısı (R^2), bağımsız değişkenlerin (girdi özellikleri) bağımlı değişkendeki (hedef parametre) toplam varyansın ne kadarını açıklayabildiğini gösteren, birimsiz bir metriktir. Değer aralığı olarak 0 ile 1 arasında bir değer alır. 1'e ne kadar yakınlaşırsa modelin performansının gerçek değere o kadar yakın bir tahminde bulunduğunu gösterir. Başka bir ifade ile veri setindeki varyasyonun ne kadarının model tarafından açıklanabileceğini söyler. Literatürde, regresyon problemlerinde modelin açıklayıcılığını değerlendirmek için kullanılan en temel metriklerden biri olarak kabul edilir [218].

Metrik matematiksel olarak aşağıdaki formülle, artıkların kareleri toplamının (SS_{res}), toplam kareler toplamı (SS_{tot})'na oranının 1'den çıkarılması ile hesaplanır.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3.15)$$

Denklem 3.15'de n örnek sayısını, y_i gerçek değeri $\hat{y}_i$ tahmin edilen değeri, $\bar{y}_i$ gerçek değerlerin ortalamasını ifade etmektedir.

R^2 metriği modelin tahmin doğruluğu hakkında açıklayıcı bilgi vermesine ragmen, tek başına hata büyüklüğü hakkında doğrudan bilgi vermemesi sebebi ile diğer hata metrikleri ile birlikte kullanılıp değerlendirilmesi önerilir.

4
MATERYAL VE METOD

Tezin önceki bölümlerinde; çalışmanın konusu ve amacı, literatürdeki güncel çalışmalar ve teorik çerçeve detaylandırılmıştır. Bu bölümde ise AAT'lerin doğrusal olmayan karmaşık süreçlerini modellemek amacıyla geliştirilen veri odaklı yaklaşımın uygulama adımları sunulmaktadır. Çalışma kapsamında; Türkiye'de faaliyet gösteren, büyük ölçekli ve kentsel atık su arıtımını gerçekleştiren bir AAT'ye ait gerçek zamanlı işletme verileri kullanılmıştır. Kurumsal gizlilik ve akademik etik ilkeleri doğrultusunda tesisin tam adı ve lokasyon bilgileri paylaşılmamış; söz konusu tesis, çalışma boyunca "Tesis-A" olarak kodlanmıştır. Elde edilen veriler üzerinde veri bilimi metotları kullanıldıktan sonra, MÖ algoritmaları ile çıkış suyu kalite parametreleri tahmin edilmeye çalışılmıştır.

Bu bölümün içeriği; sırasıyla çalışma sahası ve veri setinin tanıtılması, ham verilerin analize hazır hale getirilmesi, MÖ modellerinin kurgulanması ve model değerlendirme metriklerinin yorumlanması süreçlerini kapsamakta olup; seçilen metodolojilerin gerekçeleri anlatılarak sonuçlandırılacaktır.

4.1 Çalışma Alanı ve Kullanılan Veri Seti

Veri odaklı modelleme çalışmalarının başarısı ve oluşturulan model algoritmasının genelleme yeteneği; modelde kullanılan veri setinin niteliğine, verinin elde edildiği ortamın gerçek durumu ne denli yansıttığına doğrudan bağlıdır. Bu nedenle, tez kapsamında deneysel altyapıyı doğru kurgulayabilmek için gerçek ölçekli ve karmaşık dinamiklere sahip bir AAT tercih edilmiştir. Bu bölümde; tezin her aşamasında kullanılan veri setinin temin edildiği AAT'nin teknik özellikleri, uygulanan arıtma prosesleri ve oluşturulan veri setinin yapısal detayları anlatılacaktır.

4.1.1 İncelenen Atık Su Arıtma Tesisi (Tesis-A)

Bu çalışma kapsamında incelediğimiz AAT olan Tesis-A, Türkiye'deki bir büyükşehir belediye sınırları içerisinde olup hem evsel hem de endüstriyel atık suyun birlikte arıtıldığı AÇ prosesine sahiptir. Tesis, ileri biyolojik nütrient giderimi (BNR) konfigürasyonuna sahiptir. Çalışmada bu tesisin tercih edilmesinin sebeplerinden biri; SCADA sisteminin kapsamlı ve düzenli veri teminine elverişli olması ile bünyesinde gelişmiş bir laboratuvar barındırmasıdır. Ayrıca tesise gelen atık su karakteristiği, prosesin dinamik ve kararsız yapısı ile havza sebebiyle endüstriyel atık su girişiminde mevsimsel yoğunlukların yaşanabildiği durumlar, MÖ modellerinin gerçekçi test edilmesi için uygun bir ortam sunmaktadır. Tesise ait bilgiler, kısa alt başlıklar halinde detaylandırılmıştır.

4.1.1.1 Kapasite ve Tasarım Bilgileri

Tesis; hizmet alanındaki kişi sayısı 1.600.000, ortalama arıtma debisi 400.000 m3/gün olacak şekilde projelendirilmiş olup, pik debisi 520.000 m3/gün olarak tasarlanmıştır. Tez çalışması kapsamında incelenen değerlere göre; tesiste işletme-bakım periyotlarında ortalama debinin altında, yağışlı dönemlerde ise üzerinde debiler gözlemlenmiş; ancak çalışma süresince debi ortalaması 399.274 m3/gün olarak gerçekleşmiştir. Aşırı yağışlı zamanlarda tesise gelen hidrolik yük 650.000 m3/gün'lük değerlere ulaşmış ve bypass hattı devreye girmiştir.

Tesis-A'nın projelendirdiği tarihlerde tabi olduğu deşarj limit standardı, 2006 yılı Kentsel Atık Suların Arıtılması Yönetmeliği standartlarıdır. Tesisin tasarım kriterlerinde dikkate alınan atık su karakterizasyonu ve deşarj limitleri Tablo 4.1'de verilmiştir.

Tablo 4.1 Tesis-A'ya ait tasarım kriterleri ve tabi olduğu deşarj limitleri

Parametre	Birim	Ham Atık Su	Deşarj Limiti
Toplam Kimyasal Oksijen İhtiyacı (KOİ)	mg/L	600	<125
Biyokimyasal Oksijen İhtiyacı ($BOİ_5$)	mg/L	300	<25
Askıda Katı Madde (AKM)	mg/L	500	<35
Toplam Azot (TN)	mg/L	60	<10
Toplam Fosfor (TP)	mg/L	8,0	<3

4.1.1.2 Proses Konfigürasyonu ve Temel Üniteler

Tesis-A'ya ait konfigürasyon üç kısma ayırılarak incelenebilir. Bunlar; mekanik ön arıtma, biyolojik nütrient giderimi ve çamur yönetimidir. Bu üç yapı da Şekil 4.1'de verilmiştir. Tesis-A aşağıdaki kısımlardan oluşmaktadır:

1) Mekanik Ön Arıtma: Tesis girişinde, ham atık suda bulunan kaba maddeleri tutmak üzere 50 mm çubuk açıklığına sahip kaba ızgaralar ile 6 mm çaplı perfore ve 30 mm çubuk açıklığına sahip farklı tiplerde ince ızgaralar bulunmaktadır. Ardından 4 adet her birinin kapasitesi 750 m^3 olan havalandırmalı kum ve yağ tutucu ünitesi yer almaktadır. Biyolojik kademeye gelen yükü dengelemek ve çamur çürütme ünitesine ham çamur sağlamak üzere 2 adet, 50,2 m çapında, 3,2 m derinliğinde dairesel ön çöktürme havuzu mevcuttur. Hidrolik Bekletme süresi yaklaşık 30 dakika olan ön çöktürme havuzlarının her birinin hacmi 5667 m^3'tür. Çamur arıtımından gelen üst sular da ön çöktürme havuzlarından önce sisteme verilmektedir.

2) Biyolojik Arıtma Prosesleri (A^2O): Tesis-A, Alman ATV-DVWK(ATV-131) [219] tasarım kriterleri kullanılarak tek çamurlu, karbon azot ve fosfor gideren A^2O konfigürasyonunda ve $BOİ_5$ parametresi baz alınarak tasarlanıp inşa edilmiştir. Ön arıtmadan gelen atık sular ilk önce anaerobik Bio-P havuzlarına, sonrasında ise önde denitrifikasyon yapacak şekilde anoksik ve aerobik olarak işletilen karusel tipi havuzlara alınmaktadır. Bu süreçte nitrifikasyon ve denitrifikasyon süreçleri gerçekleşmektedir. Aerobik havuz ile anoksik havuz arasında nitrat geri devri için kullanılan içsel geri devir hatları da Şekil 4.1'de gösterilmiştir.

- Bio-P Havzuları; seri veya paralel çalışabilen adet 4 adet havuzdan teşkil edilmiş ve toplam hacmi 25.000 m^3 olan bir yapıdır. Hidrolik bekletme süresi 30 dk olarak tasarlanmıştır.

- Havalandırma Havuzları; Anoksik hacim oranı %50 olarak projelendirilmiştir. Karusel tipinde eşit hacimli 16 adet oksidasyon havuzu bulunmakta olup, 2 hat şeklinde imal edilmiş ve toplam 240.000 m^3 reaktör hacmindedir. Havuzların derinlikleri 6,35 metredir.

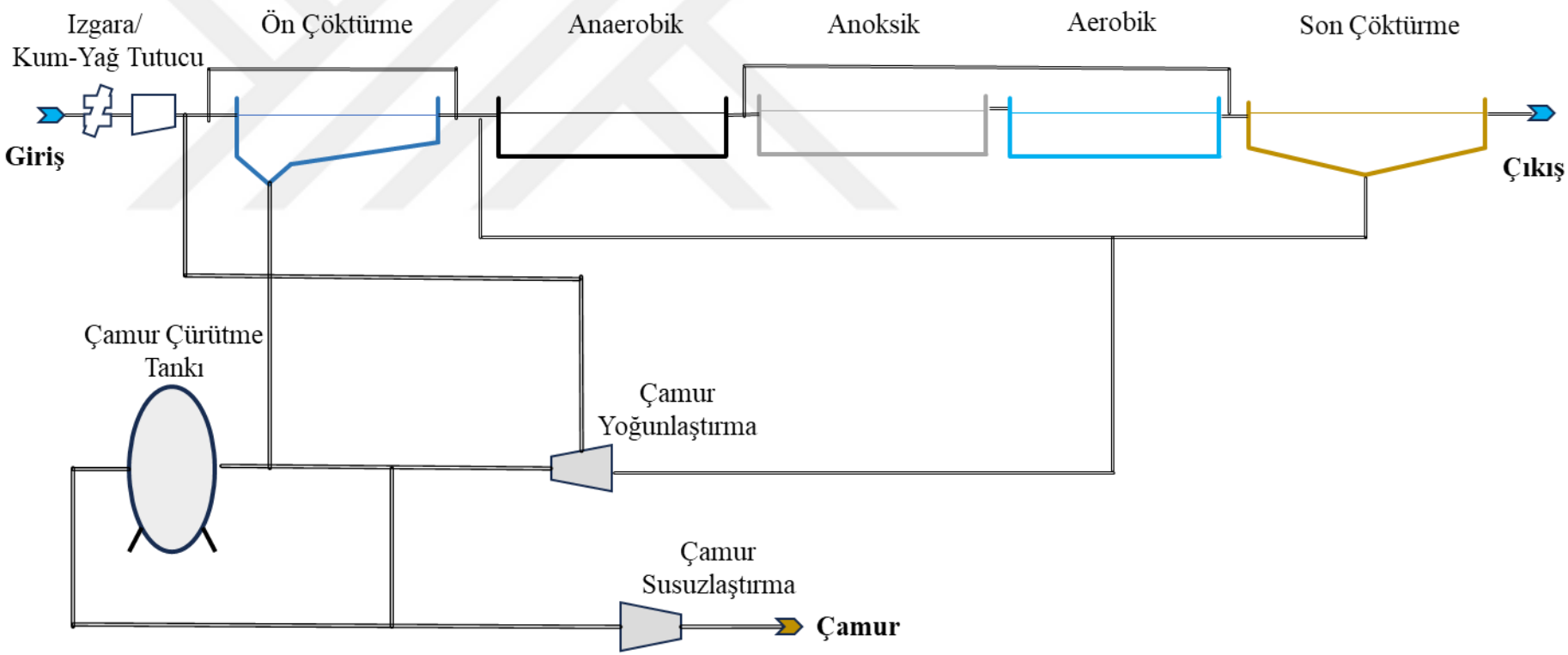
- Son çöktürme havuzları; 12 adet 2 saat bekletme süresi olan dairesel havuz şeklinde inşa edilmiştir. Havuzların çapı 51 m ve ortalama derinliği 3,8 m olup, her bir havuzun hacmi 7581 m^3'tür.

3) Çamur Yönetimi: Tesis-A'nın ön çöktürmesinden gelen çamurlar 3 adet santrifüj tipi yoğunlaştırıcıya, son çöktürmeden gelen çamurlar da 7 adet santrifüj tipi yoğunlaştırıcıya gelerek %6 katı madde oranına getirilmektedir. %6'lık çamur her biri 10.000 m^3'lük hacme sahip 6 adet çamur çürütme tankına alınarak mezofilik şartlarda stabilize edilir. Bu süreçte açığa çıkan biyogaz gaz depolama tankına gönderilir. Çürütmeden sonra %4-6 katı madde oranına sahip çamur 6 adet santrifüj tipi susuzlaştırma ünitesine alınır. %25 katı madde oranına gelen çamur; kurutma ünitelerine kurutularak sistemden uzaklaştırılır. Yoğunlaştırma ve susuzlaştırma işlemlerinden sonra oluşan süzüntü (sentrant) suları ön çöktürme havuzunun başına alınarak sisteme geri döndürülür.

4.1.1.3 Veri Toplama ve Otomasyon Altyapısı

Tesis-A'nın işletilmesi esnasında, tesisin izlenmesi ve kontrolü için kullanılan SCADA sisteminin altyapısı, Siemens firması tarafından geliştirilen WinCC 7.4 yazılımı ile sağlanmaktadır. 5 adet S7-400 PLC; tesisin tüm geneline ait verilere göre tam otomasyon görevini gerçekleştirmekte, ayrıca tesisin internet üzerinden izlenmesi için 2 adet SCADA sunucu bilgisayarı bulunmaktadır.

Tesis bünyesinde hidrolik akışın, biyolojik reaktörlerin işleyişinin ve ekipman kullanım durumlarının anlık takibini yapmak için tesis genelinde sensör ve transmiter ağı kurulmuştur. Bu doğrultuda tesiste; proses hatlarındaki debi değişimlerini izlemek için 54 adet debimetre (boru tipi ve açık kanal) kullanılmaktadır. Buna ek olarak, fiziksel ve kimyasal parametrelerin takibi için 282 adet proses ölçüm elemanı ve 23 adet de anlık ölçüm sensörü bulunmaktadır. Bu ölçüm elemanları kapsamında; iletkenlik, sıcaklık, ORP, seviye ve çözünmüş oksijen gibi cihazlar bulunmakta; AAT işletilmesi açısından kritik parametreler anlık olarak izlenmektedir.



Şekil 4.1 Tesis-A'ya ait tesis proses akış diyagramı

Yukarıda bahsedilen ölçüm ve izleme ekipmanlarından elde edilen veriler, tesis geneline yayılmış olan 867 adet transmiter aracılığıyla SCADA sistemine iletilmektedir. Bu kadar yoğun transmiter sayısının büyük kısmı çamur ünitelerinde, geri kalan kısımları ise mekanik ön arıtma, havalandırma havuzları, son çöktürme ve Bio-P havuzlarında bulunmaktadır.

Bu geniş kapsamlı ve etkin kullanılan SCADA sistemi sayesinde; zamana bağlı değişkenlik gösteren hidrolik ve organik madde yükleri (C, N, P) altında dinamik ve doğrusal olmayan (non-lineer) bir karakteristik sergileyen AAT işletme süreci, yüksek veri yoğunluğu ile kayıt altına alınmaktadır. Sürekli ve çok parametreli olarak üretilen bu büyük veri havuzu, çalışmamızın temelini oluşturarak MÖ algoritmalarının uygulanması için gerekli altyapıyı sağlamıştır.

Tez kapsamında; SCADA sisteminden elde edilen verilerle birlikte laboratuvar ortamında üretilen veriler ve hesaplanarak elde edilen diğer hibrit veriler, tesisin çıkış suyu kalitesini tahmin etmek amaçlı veri madenciliği işlemlerine tabi tutulmuştur. Diğer verilerin elde ediliş şekli, aşağıdaki bölümlerde detaylandırılmıştır.

4.1.2 Veri Seti ve Özellikleri

Çalışmanın deneysel temelini oluşturan veriler; Tesis-A'ya ait SCADA verileri, ortam koşullarının derlendiği çevresel veriler ve tesis bünyesinde işletilen laboratuvar kayıtlarından elde edilmiştir. Analizlerden üretilen veriler (KOİ, BOİ, AKM, UAKM, TN, TP vb.), Tesis-A bünyesinde bulunan laboratuvar tarafından, uluslararası kabul görmüş "Standard Methods for the Examination of Water and Wastewater" deney prosedürlerine tam uyularak oluşturulmuştur. Model eğitiminde kullanılan bu laboratuvar verilerinin analiz yöntemlerinin tamamı ve referans standartları Ek-B'de sunulmuştur.

Meydana getirilen veri seti, 1 Ocak 2022 - 8 Aralık 2024 tarihleri arasındaki yaklaşık 3 yıllık gerçek zamanlı işletme verilerinden oluşmaktadır. Bu periyot; mevsimsel geçişlerde oluşan yaz-kış sıcaklık farklarının ve değişen hidrolik yük dalgalanmalarının MÖ modeli üzerindeki etkilerini analiz edebilmek amacıyla geniş tutulmuştur. Tesis-A'ya ait tezde kullanılan tüm veriler Ek-A olarak sunulmuştur.

Oluşturulan veri seti analiz edildiğinde; toplam 1.075 adet günlük gözlemden (satır) ve 51 adet özellikten (sütun) oluştuğu görülmektedir. Veri setini oluşturan değişkenleri; girdi değişkenleri (bağımsız) ve hedef değişkenleri (bağımlı) olarak ayıran, ayrıca elde edilme şekillerini kategorik olarak sınıflayan Tablo 4.2 aşağıda sunulmuştur.

Tezdeki veri seti, Excel tabloları halinde hazırlanmış etiketli verilerden oluşmaktadır. Veri biliminde "ham veri" terimi genellikle hiç işlenmemiş ve etiketlenmemiş veriler için kullanılsa da bu çalışmada ham veriden kastımız; Excel ortamında toplanmış ve geleneksel yollarla derlenmiş, yani yarı işlenmiş verilerdir. Veri setinin istatistiksel dağılımı incelendiğinde, gerçek saha verisi olmasından kaynaklı olarak bazı parametrelerde geniş varyasyonlar görülmektedir. Veri setinin genel karakteristiğini yansıtması açısından, bazı giriş ve çıkış gözlem sonuçlarının istatistiksel özellikleri Tablo 4.3'te verilmiştir.

Tablo 4.2 Tezde kullanılan verilerin kategorizasyonu

Değişken Türü	**Kategori**	**Parametreler**
Girdi (Bağımsız) Değişken	Giriş Suyu Karakteristiği	Giriş Debisi, Giriş KOİ, Giriş $BOİ_5$, Giriş AKM, Giriş TN, Giriş TP, Giriş Çözünmüş Oksijen, Giriş UAKM, Giriş Alkalinite, pH, Giriş Suyu Sıcaklığı
	İşletme Verileri	Hav. Havuzu Çözünmüş Oksijen, AKM, UAKM, Geri Devir Oranı, İçsel Geri Devir Oranı, Çamur Yaşı, Enerji Tüketimi
	Çevresel Veriler	Hava Sıcaklığı, Yağış Durumu, Bağıl Nem,
Hedef (Bağımlı) Değişken	Çıkıs Suyu Kalitesi	Çıkış KOİ, Çıkış $BOİ_5$, Çıkış AKM, Çıkış TN, Çıkış TP

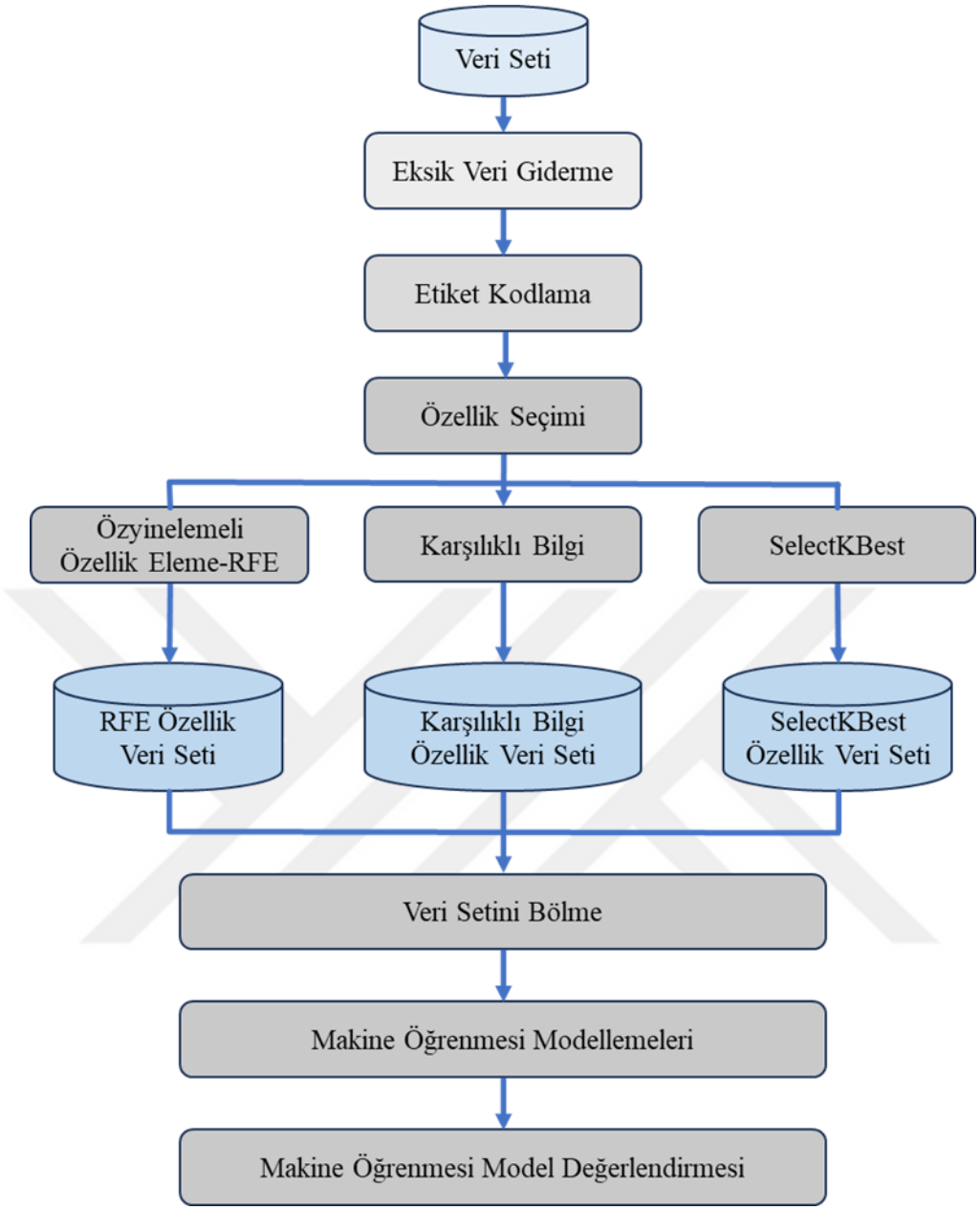
Tablo 4.3'de sunulan standart sapma değerleri tesisin stokastik yapısını göstermektedir. Burada sunulan veriler doğrudan veri analizine tabi tutulmuştur.

4.2 Önerilen Metodolojik Yaklaşımın Genel Çerçevesi

Bu tez çalışması kapsamında; bir AAT'nin karmaşık ve doğrusal olmayan süreçlerinin MÖ algoritmaları ile modellenerek çıkış suyu deşarj kalite parametrelerinin tahmini amaçlanmıştır. IWA'nın 2019 yılında başlayan, dijital su ve YZ kullanımının yaygınlaştırılmasını öngören yaklaşımından alınan motivasyon ile veri bilimi ve MÖ algoritmalarını bir araya getiren hiyerarşik bir metodoloji geliştirilmiştir. Bu kapsamda, 6 katmanlı bir sistem oluşturulmuştur. Söz konusu katmanlar Şekil 4.2'de verilmiş olup alt başlıklar halinde detaylandırılmıştır.

Tablo 4.3 Veri setine ait bazı temel tanımlayıcı istatistikler

Parametre	Birim	Ortalama	Std.Sapma	Medyan	Mod
G. Debi	m^3/gün	399.274,49	26.745	406.306	368,755
G. KOİ	mg/L	513,34	107,87	519,00	528,00
G. $BOİ_5$	mg/L	296,59	66,29	304,00	276,00
G. AKM	mg/L	331,15	107,53	316,00	316,00
G. UAKM	mg/L	205,78	55,23	206,00	220,00
G. TN	mg/L	53,91	9,78	54,40	57,00
G.TP	mg/L	5,16	1,29	5,20	5,00
Ç. KOİ	mg/L	47,11	25,13	41,00	37,00
Ç. $BOİ_5$	mg/L	6,56	4,82	5,00	4,00
Ç. AKM	mg/L	27,88	35,32	17,50	12,00
Ç. TN	mg/L	6,88	1,62	6,80	6,30
Ç.TP	mg/L	0,47	0,48	0,30	0,20



Şekil 4.2 Metodoloji akış şeması

Şekil 4.2'de görüldüğü gibi, tez çalışmasının uygulama kısmı 6 ayrı adımdan oluşmaktadır. Eksik veri tamamlama adımında; sensör arızası, laboratuvar çalışmaması veya programlı bakım gibi durumlarda veri üretilememiş ya da kayda alınmamış ise bu eksikliklerin yerine uygun değer atama işlemleri gerçekleştirilmektedir. Etiket kodlama (Label Encoding) aşamasında, sözel veriler sayısal ifadelere dönüştürülerek model diline çevrilmektedir. Özellik seçimi (Feature Selection) adımında ise modelin hesaplama zorluğunu gidermek ve tahmin doğruluğunu artırmak amacıyla, giriş değişkenleri arasından hedef değişkeni en fazla etkileyen parametreler belirlenmiştir. Bir sonraki adımda, modelin eğitilmesi

ve test edilmesi için veri seti “eğitim verisi” ve “test verisi” olmak üzere iki ayrı kümeye ayrılmıştır. Beşinci adımda, literatürde yer alan 5 farklı MÖ algoritması kullanılarak modeller eğitilmiştir. Son adımda ise bu modellerin tahmin başarıları değerlendirilmektedir.

Bu bölümde, söz konusu altı adımdaki işlemler alt başlıklar halinde anlatılacaktır. Tezin üçüncü bölümünde veri bilimi ve makine öğrenmesinin teorik altyapısı ile çalışma prensipleri detaylı olarak ele alınmıştır. Metodolojinin açıklandığı bu dördüncü bölümde ise hangi yöntemlerin neden tercih edildiği; Tesis-A’nın proses dinamikleri, veri setinin özellikleri (boyut, dağılım, varyans vb.) ve literatürdeki boşluk göz önüne alınarak, stratejik bir yaklaşımla gerekçeleriyle birlikte sunulmuştur.

4.3 Deneysel Ortam ve Çalışmadaki Yazılım Araçları

Tez çalışmasının tamamında kullanılan yazılım dili Python’dur. Bu dilin tercih edilme sebepleri; açık kaynaklı birçok kütüphane barındırması, kullanım kolaylığı, her adımda hatalı işlemlerin düzeltilmesine imkân tanıması ve uzmanlık kazanılmasının göreceli olarak kolay olmasıdır. Çalışmada veri ön işleme, modelleme ve model başarısının değerlendirilmesi süreçlerinde; bulut tabanlı bir çalışma altyapısı sunan Google Colab üzerindeki Jupyter Notebook arayüzü kullanılmıştır.

Tez çalışması kapsamında kullanılan yazılım kütüphaneleri ve kullanılma yönleri şunlardır:

- Veri işleme: Veri setindeki ham veriler; Excel formatında ve zaman serisi yapısında olduğu için öncelikle düzenlenmesi, tablo yapısında okunması ve eksik verilerin tespiti amacıyla Pandas kütüphanesi kullanılmıştır. Sayısal işlemler, matris tabanlı hesaplamalar ve istatistiksel yaklaşımlar için de NumPy kütüphanesi tercih edilmiştir. Bu iki kütüphane kullanılarak veri ön işleme süreci gerçekleştirilmiştir.
- Görselleştirme: Veri setinin keşifsel analizi ve model çıktılarına ait performans grafiklerinin oluşturulması için Matplotlib kütüphanesinden yararlanılmıştır.
- Modelleme: Veri setlerinin eğitim ve test verisi olarak ayrılması (veri setinin bölünmesi) ile temel makine öğrenmesi algoritmaları olan Karar Ağaçları,

Rastgele Orman ve GBM modelleri için Scikit-learn kütüphanesi kullanılmıştır. Ayrıca bu ve diğer ileri algoritmaların performansını ölçen model değerlendirme metrikleri (RMSE, MAE, R^2) için de aynı kütüphaneden yararlanılmıştır.

- İleri Algoritmalar: Daha gelişmiş MÖ algoritmaları olan XGBoost ve LightGBM için Python'da müstakil kütüphaneler mevcuttur. Bu iki topluluk öğrenme modeli için de kendilerine özgü kütüphaneler kullanılmıştır.

Bu yukarıdaki deneysel çalışmaların donanımsal altyapısında Grafik İşlem Birimi (GPU) hızlandırması kullanılmamış; yapılan tüm hesaplama işlemleri Merkezi İşlem Birimi (CPU) üzerinden gerçekleştirilmiştir. Kullanılan bulut tabanlı çalışma ortamı (Google Colab), 16 GB RAM'e sahip bir donanım üzerinde yapılandırılmıştır. Performans kıyaslamalarından önce herhangi bir hiperparametre optimizasyonu çalışması yapılmamış; kullanılan algoritmaların kütüphanelerindeki varsayılan (default) değerleri kabul edilmiştir.

4.4 Eksik Veri Giderme

Gerçek ölçekli AAT'lerden elde edilen verilerin eksik olması, zaman serisi şeklinde gerçekleştirilen veri toplama sürecinin doğasında bulunmaktadır. SCADA sistemi ile alınan verilerde; anlık iletişim kopmaları, sensörlerin kalibrasyon aralıkları veya donanımsal arızalar sebebiyle eksiklikler yaşanmaktadır. Ayrıca laboratuvar verilerinde; ekipman veya insan kaynaklı hatalar nedeniyle de eksik veriler oluşabilmektedir. Bu çalışmada veri seti incelendiğinde; belirli zaman aralıklarında ve farklı parametrelerde, toplamda %0,30'a tekabül eden 132 adet hücresel verinin eksik olduğu tespit edilmiştir.

- Tezin teorik altyapısının sunulduğu Bölüm 3'te detaylı bir şekilde anlatılan veri tamamlama yaklaşımlarından (silme, doldurma, model tabanlı tahmin vb.) "Ortalama Değer Atama (Mean Imputation)" yöntemi tercih edilmiştir. Literatürde bu yöntem; eksikliğin sistematik olmadığı ve eksik oranının bizim veri setimizde olduğu gibi düşük olduğu durumlarda önerilmektedir [176]. Ayrıca bu yöntemi tez kapsamında tercih etmemizin sebepleri şunlardır:

- Veri setindeki eksik veriler için literatürde sıklıkla başvurulan "satır veya sütun silme" yöntemi, mevcut veri setinin boyutu göz önüne alındığında çok fazla veri kaybına sebep olacaktır. Ayrıca, veri seti bir zaman serisi niteliği taşıdığı için bu

tür silme işlemleri sürekliliği bozacak; mevsimsel döngülerin ve ani yük değişimlerinin model tarafından öğrenilmesini engelleyecektir. Ortalama değer atama yöntemi sayesinde herhangi bir veri kaybı yaşanmadan zaman serisinin bütünlüğü korunmuştur.

- Veri setinin toplanma süresi geniş bir zaman aralığını kapsadığından dolayı (3 yıl), eksik verinin ortalaması, söz konusu eksik parametrenin (örneğin giriş AKM) en beklenen değerini temsil edebilmektedir.
- Yöntemin basit olması, model algoritmalarının çalıştırılması esnasında oluşabilecek gürültüye karşı sistemin daha dirençli olmasını sağlamıştır.

4.5 Kategorik Verilen Kodlanması

MÖ algoritmaları yapıları gereği matematiksel ifadelerle işlem yapabilmekte, sözel verilerle doğrudan işlem yapamamaktadırlar. Bu sebeple tez çalışmamızdaki hava durumu veya ekipmanın aktif/pasif durumu gibi sözel gözlem sonuçlarının sayısal formlara dönüştürülmesi gerekmektedir. Bu doğrultuda; literatürde yaygın kullanılan ve Bölüm 3.1.2 başlığı altında detaylı olarak anlatılan Etiket Kodlama (Label Encoding), Vektörel Kodlama (One-Hot Encoding) ve Hedef Odaklı Kodlama (Target Encoding) yöntemleri değerlendirilmiştir. Teknik değerlendirme aşağıdaki şekilde özetlenebilir:

- Vektörel Kodlama (One-Hot Encoding); her bir kategoriyi yeni bir sütun olarak veri setine ekleyerek sistemi genişlettiği ve ağaç tabanlı MÖ modellerinde işlem yükünü artırdığı için tercih edilmemiştir [187].
- Kategorik değişkenleri hedef değişkenin ortalaması ile değiştiren Hedef Odaklı Kodlama (Target Encoding); eğitim verisindeki hedef değerleri kullandığı ve çalışmadaki hedef değişkenlerin de bir deşarj limiti ile sınırlandırılmış olduğu göz önüne alındığında, bu tez çalışmasındaki veri setine uymamaktadır. Ayrıca veri setimizin boyutunun nispeten küçük olduğu, bu yöntemin daha çok yüksek sayıda kategorik değişkene sahip veri setlerinde tercih edildiği ve düşük miktarda kategorik değişken içeren veri setlerinde aşırı öğrenme (overfitting) riski oluşturduğu bilindiğinden dolayı bu yöntem seçilmemiştir [182].
- Etiket Kodlama, veri setinin yapısındaki sadeliği korumak amacıyla bu tez çalışmasında tercih edilmiştir. Bu yöntemin en büyük dezavantajı; kategorik

değişkenler sayısallaştırılırken aralarında bir büyüklük ilişkisi kurması ve bu durumun mesafe tabanlı algoritmalarda yanıltıcı sonuçlara yol açmasıdır. Ancak bu tez çalışmasında kullanılan MÖ algoritmalarından Rastgele Orman, LightGBM ve XGBoost gibi ağaç tabanlı algoritmalar, veriyi sayısal büyüklüklerine göre değerlendirmemekte; öğrenmeyi maksimize etmek için eşik değerlere göre işlemektedir [201]. Bu nedenle, ağaç tabanlı modellerin sıralı yapıdaki değerleri işleme yeteneği literatürde de doğrulandığı üzere, Etiket Kodlama (Label Encoding) yöntemi hesaplama verimliliği sağladığından tercih edilmiştir [205].

4.6 Özellik Seçimi-Boyut İndirgeme

AAT'lerden elde edilen veriler; arıtma proseslerinin çok değişkenli dinamik yapısı sebebiyle genellikle yüksek boyutlu, gürültülü ve kendi aralarında yüksek korelasyona sahip parametrelerden oluşmaktadır. Bu durum MÖ algoritmalarının tahmin becerisini azaltırken aşırı öğrenme (overfitting) riskini artırmaktadır [190].

Tezimizin literatüre katkılarından biri de; AAT tesislerinin işletilmesine yönelik karar destek mekanizması oluşturulması çalışmalarında, MÖ modellemesi yapılırken daha önce denenmemiş olan "Özellik Seçimi" yöntemlerini kullanmayı tecrübe etmektir. Bölüm 3.2'de, literatürde bilinen özellik seçimi yöntemleri, tezde verdiğimiz önem dolayısıyla detaylıca anlatılmıştır. Bu çalışmaya ait veri setinin karakteristiği ve modelleme hedeflerine uygun olan üç yöntem (SelectKBest, RFE ve Karşılıklı Bilgi) tercih edilmiştir. Bu yöntemlerin seçilmesinin mühendislik gerekçeleri ve uygulama stratejileri, alt başlıklar halinde aşağıda sunulmuştur.

4.6.1 SelectKBest Yönteminin Seçim Gerekçesi

Bu yöntemin seçilmesindeki temel neden; her özelliği bağımsız olarak değerlendirdiği için diğer yöntemlere göre hızlı olması ve ön modelleme aşamalarında hızlı iterasyon yapılmasına olanak tanımasıdır. Ayrıca bu iterasyonlarda en güçlü istatistiksel puana sahip olanları seçerek gürültülü veriyi en baştan elemektedir [189].

Veri seti ve parametreler arası ilişkiler AAT özelinde incelendiğinde, bazı parametreler arasında doğrusal ilişkiler bulunabilmektedir. Bu yöntem, doğrusal ilişkilerin tespitinde göreceli olarak daha üstündür [202]. Ancak, SelectKBest

yönteminin yalnızca tek değişkenli analiz yapması ve özellikler arasındaki etkileşimleri dikkate almaması, bu yöntemin önemli bir kısıtıdır. Bu kısıt göz önünde bulundurulduğu için çalışmada birden fazla özellik seçimi modeline başvurulmuştur.

4.6.2 Karşılıklı Bilgi Yönteminin Seçim Gerekçesi

Sadece doğrusal değil, doğrusal olmayan ilişkileri de tespit edebilen Karşılıklı Bilgi (Mutual Information) yönteminin seçilme sebeplerinden biri bu özelliğidir. AAT'de doğrusal olmayan ilişkiler, biyokimyasal süreçlerin karmaşıklığından kaynaklanmaktadır. Bu yöntem; bir giriş değişkeninin bilinmesi durumunda, çıkış değişkenindeki belirsizliğin ne kadar azaldığını ölçebilmektedir [193]. Böylelikle doğrusal ilişkisi düşük görülse bile doğrusal olmayan güçlü bir ilişkinin bulunma ihtimali değerlendirilmiş ve bu yöntem özellik seçimi sürecine dahil edilmiştir.

4.6.3 RFE-Model Tabanlı Seçimin Gerekçesi

Sarmalayıcı yöntemlerden olan RFE, filtre tabanlı yöntemlerin aksine (SelectKBest, Karşılıklı Bilgi) özelliklerin birbirleri ile olan etkileşimlerini de dikkate almaktadır. RFE ile bir özelliğin, diğer özelliklerle birlikteyken modele olan katkısı ölçülebilmektedir. Bu yaklaşım, tek başına önemli görünse de model bütününde gereksiz olan özelliklerin elenmesine yardımcı olduğundan; bu çalışma kapsamında Rastgele Orman (Random Forest) algoritması ile desteklenerek kullanılmıştır.

Bu üç ayrı metodolojinin birlikte kullanılmasının sebebi; her bir yöntemin diğerinde bulunmayan avantajlarından faydalanarak, önemli özelliklerin gözden kaçırılmasını önlemektir. Üç modelde de ortak olarak "önemli" bulunan özelliklerin seçilmesiyle, sonuçların tutarlılığı ortaya konmuş olacaktır. Ayrıca tek bir yönteme bağlı kalmamak, özellik seçim sonuçlarının güvenilirliğini artırmakta ve bilimsel olarak desteklemektedir.

Literatürde AAT çıkış suyu kalite tahminine yönelik yapılan MÖ modelleme çalışmalarının büyük bir kısmı; ham verilerle veya sınırlı ön işlemeye tabi tutulmuş verilerle gerçekleştirilmiş, algoritmaların uygulanmasından önce özellik seçimi yönetimi sistematik bir şekilde uygulanmamıştır. Tez çalışması kapsamında bu yaklaşım hayata geçirilirken, veri setine en uygun olabilecek birden fazla yöntem denenmiş ve bu yöntemlerin birbirlerinden farklı perspektiflerle bir yaklaşım

sergilemesi hedeflenmiştir. Çalışmanın birincil amacı, sadece hangi MÖ modelinin başarılı olduğu değil; hangi özellik seçim yöntemi ile entegre edilmiş MÖ algoritmasının en iyi sonucu verdiğini ortaya koymaktır.

4.7 Veri Seti Bölme

Bölüm 3.3'te veri setini bölme hakkında detaylı bilgi verilmiş; burada hangi yöntemin neden seçildiği açıklanmıştır. MÖ uygulamalarında veri setini eğitim ve test verisi olarak gruplandırmak genelde rastgele yöntemlerle gerçekleştirilmektedir. Ancak AAT'lerden elde edilen veriler; normal dönem verilerinin aniden şok yüklemelerle değiştiği, "normal olmayan" gözlem değerlerini içermektedir. Şok yüklemeler ve mevsimsel değişimler gibi sebeplerle veri setinde dengesiz bir yapı oluşmaktadır. Bu sebeple, Basit Rastgele Örnekleme yöntemi kullanıldığında pik değerlerin eğitim ve test verilerine eşit olmayan bir şekilde dağılabileceği öngörüldüğünden, bu yaklaşım benimsenmemiştir.

Bu çalışma kapsamında; örnekleme yanlılığı riskini ortadan kaldırmak için veri setini hedef değişkenin değer aralıklarına göre homojen alt gruplara ayırabilen ve her tabakadan oransal seçim yapılmasına imkân tanıyan "Tabakalı Örnekleme (Stratified Sampling)" yöntemi uygulanmıştır [184].

Bölme oranı olarak (Eğitim-Test); literatürde yaygın kullanılan, model için yeterli eğitim verisi sağlarken güvenilir bir değerlendirme için yeterli test verisi de sunduğu kabul edilen %80-%20 oranı kullanılmıştır. Çalışma kapsamında farklı bölme oranları da (%70-%30; %60-%40) denenerek model performansına etkileri araştırılmış; ancak nihai karşılaştırmalar ve sonuçların analizi için 80-20 oranı esas alınmıştır.

4.8 Makine Öğrenme Algoritmaları

AAT'lerin karmaşık ve doğrusal olmayan yapısını YZ destekli modellemek için literatürde çok sayıda yaklaşım ve her yaklaşım şekline bağlı birçok algoritma mevcuttur. Bölüm 3.4'te literatürde bulunan algoritmaların teorik yapısı detaylı bir şekilde anlatılmıştır. Ancak her algoritmanın; veri yapısına, verinin boyutuna ve ele alınan problemin türüne göre üstünlükleri veya zayıflıkları bulunmaktadır [198]. Bu tez çalışmasında kullanılacak algoritmalar seçilirken veri setinin "yapısal

özellikleri" (tablo verisi olması), gözlem sayısı ve sonuçların yorumlanabilirlik durumları dikkate alınmıştır.

Bu çalışma kapsamında; literatürde "yapısal veriler" üzerindeki başarısı kanıtlanmış ağaç tabanlı "Karar Ağacı" ile ağaç tabanlı topluluk öğrenmesi yöntemleri (Rastgele Orman, GBM, XGBoost, LightGBM) temel alınmıştır. Aşağıda alt başlıklar halinde; bu yöntemlerin neden tercih edildikleri ve Derin Öğrenme'nin (DÖ) neden kapsam dışı bırakıldığı açıklanmıştır.

4.8.1 Ağaç Tabanlı Topluluk Öğrenme Modellerinin Seçimin Gerekçesi

Doğrusal olmayan ilişkileri tahminlemede başarılı oldukları, ayrıca birden fazla zayıf öğrenicinin bir araya gelerek daha güçlü bir model oluşturabildiği (topluluk öğrenmesi prensibi) için ağaç tabanlı topluluk öğrenme modelleri tercih edilmiştir.

4.8.1.1 Karar Ağaçları (DT)

Model performanslarının kıyaslanmasında bir referans noktası oluşturulması amacıyla Karar Ağaçları seçilmiştir. Karar Ağaçları modeli tek başına kullanıldığında aşırı öğrenme (overfitting) riski taşıyan bir yapı olsa da veri yapısındaki "eğer-değilse" (if-else) mantıksal kırılma noktalarını net bir şekilde göstermesi sebebiyle, modellemeye bir başlangıç algoritması olarak dahil edilmiştir [200]. Bu model, topluluk öğrenmesi (ensemble learning) yöntemlerinin sağladığı performansa dayalı katkıyı daha net ortaya koymak adına temel bir karşılaştırma aracı olarak kullanılmıştır.

4.8.1.2 Rastgele Orman (RF)

Gürültülü sensör verileri ve ölçüm hataları içerebilen AAT verileri için; çok sayıda karar ağacının ortalamasını alarak tek bir ağacın varyansını azaltmak ve aşırı öğrenme (overfitting) ihtimalini düşürmek amacıyla, birden çok karar ağacını modelleyen Rastgele Orman (Random Forest) modelinden faydalanılması düşünülmüştür.

Bagging tekniğini kullanan bu yöntem, eğitim verisinden rastgele alt kümeler seçerek birden çok ağaç eğitir; bu süreç verideki gürültünün etkisini azaltarak daha kararlı tahminler oluşturur. Ayrıca bu modelin, literatürde birçok defa çevresel tahmin problemlerinde yüksek performans gösterdiği bilinmektedir [14, 153].

4.8.1.3 Gradyan Artırma Algoritmaları (GBM, XGBoost, LightGBM)

Bu üç algoritma da Artırma (Boosting) yöntemi ailesindendir. Rastgele Orman yönteminin aksine, ağaçları bağımsız değil sıralı olarak kurar; her yeni ağaç bir önceki ağacın hatasını düzelterek öğrenmeye devam eder [206]. Veri setimizin mevsimsel değişkenlikler gibi sebeplerle doğrusal olmadığını göz önünde bulundurarak doğrusal olmayan ilişkilere yönelmiştik. Bu yeni teknolojik yöntemler literatürde sıklıkla kullanılmaktadır. Biz de aşağıdaki sebeplerle bu algoritmalara yöneldik:

1) GBM: Kendinden sonraki diğer iki algoritmanın temelini oluşturur. Performans kıyaslamasında yeni nesil artırmalı yöntemler olan XGBoost ve LightGBM modellerinin, geleneksel GBM'e göre ne gibi iyileştirmeler getirebileceğini görmeyi amaçladık. Bu nedenle, bir referans noktası olması açısından bu algoritmayı veri setimize uyguladık.
2) XGBoost: Regresyon ve sınıflandırma problemlerinde birçok yarışmada üstün performans göstermiş; hesaplama hızı, yüksek doğruluk ve düzenlileştirme (regularization) yetenekleri ile tercih edilmiştir [185, 203]. Düzenlileştirme; modelin aşırı öğrenmesini engelleyip genelleme yeteneğini artırarak, veri setinin küçük-orta ölçekli olduğu problemlerde önemli bir avantaj sağlar.
3) LightGBM: Microsoft tarafından geliştirilen bu algoritma, büyük ölçekli ve yüksek boyutlu veri setlerinde dahi çok yüksek bir hız ve bellek verimliliği sağlamaktadır [204]. Yaprak odaklı büyüme (leaf-wise growth) stratejisi ile karmaşık problemleri kolaylıkla öğrenebilen bu algoritma, çalışmamıza hesaplama verimliliği açısından dahil edilmiştir [204].

Bu beş modelin birlikte değerlendirilmesi ile basitten karmaşıklığa doğru bir rota (DT→RF→GBM→XGBoost→LightGBM) sunulmaya çalışılmıştır. Bu seçim mantığı ile öncelikle bir referans noktası (DT) oluşturulmuş; sonrasında ise AAT verilerinin karmaşıklığı ile başa çıkabilecek en güncel ve güçlü topluluk öğrenme yöntemleri (RF, GBM, XGBoost, LightGBM) problemin çözümünde uygulanmıştır.

4.8.2 Diğer Makine Öğrenmesi Modellerinin Seçilmeme Gerekçeleri

Doğrusal modeller; basit yapıları ve yorumlanma kolaylıkları ile ilgi çekici olsalar da AAT proseslerinde etkili olan temel biyokimyasal ilişkilerin doğrusal olmaması

sebebiyle birincil model olarak seçilmemişlerdir [46]. Biyokimyasal reaksiyon hızları, mikrobiyal büyüme, içsel solunum ve diğer parametrelerin birbirleri ile olan ilişkileri doğrusal değildir. Literatür, doğrusal olmayan karmaşık ilişkileri çözümlemede doğrusal modellerin yetersiz kaldığını göstermiştir [202].

Destek Vektör Makineleri (SVM); eğitim sürelerinin uzun olması ve hiperparametre optimizasyonuna karşı aşırı hassas yapısı sebebiyle çalışmamıza dahil edilmemiştir. Bu tez çalışmasında özellik seçimine yoğunlaşıldığı için hiperparametre ayarlarında modellerin varsayılan (default) değerleri kullanılmıştır. Ayrıca ağaç tabanlı modeller, SVM'lere kıyasla daha az hiperparametre ayarına ihtiyaç duymaları ve yapılarında bulunan doğal özellik sıralama (feature importance) kabiliyetinin kolay yorumlanabilirlik sağlaması nedeniyle tercih edilmişlerdir.

4.8.3 Derin Öğrenme Yönteminin Seçilmeme Gerekçesi

Güncel literatürde Yapay Sinir Ağları ve Derin Öğrenme popüler olsa da veri seti karakteristiği dikkate alınarak bu tez çalışmasında tercih edilmemiştir. Aşağıda bu gerekçeler sıralanmıştır:

- DÖ algoritmaları genelleme yapabilmek ve kendi öğrenme stratejilerini geliştirmek için çok büyük verilere (big data) ihtiyaç duyarlar. Daha küçük veri setlerini ezberleme eğiliminde olduklarından, aşırı öğrenme (overfitting) sebebi ile hatalı sonuçlar oluşturabilektedirler [197]. Bu çalışmanın veri seti yaklaşık 1000 satırdan oluştuğundan dolayı derin öğrenme tercih edilmemiştir.
- DÖ genelde görüntü, ses ve metin gibi "yapısal olmayan" verilerde rakipsizdir. Ancak tezde kullanılan yapısal verilerde (tablolar), gradyan artırma modellerinin DÖ'ye kıyasla daha başarılı sonuçlar verdiği ve daha az hiperparametre ayarı gerektirdiği literatürde çokça kabul görmektedir [205].
- AAT işletmesinin yönetimi, neden-sonuç ilişkilerinin anlamlandırılmasını gerektirmektedir. DÖ, tam bir kara kutu (black box) şeklinde çalışır; nedensellik sorgulaması yapılmasına pek izin vermez, sadece sonuç verir. Tezde kullandığımız MÖ algoritmaları ise model sonunda özellik önem skorları (feature importance) vererek, tesis işletmecisine karar destek imkanı sunabilmektedir.

Yukarıda sıralanan model seçimi gerekçeleri doğrultusunda; hesaplama maliyeti düşük, açıklanabilirliği yüksek ve elimizdeki veri seti karakteristiğine uygun "Ağaç Tabanlı Topluluk Öğrenmesi" modelleri ile ilerlenmesine karar verilmiştir.

4.9 Model Değerlendirme Metriklerinin Seçimi

Geliştirilmeye çalışılan YZ modellerinin başarısı; sadece tahmin değerlerinin gerçek değerlere yaklaşması ile değil, ayrıca hataların dağılımı ve modelin pik değerlere verdiği tepkilerle de orantılıdır. Modelin bu kritik pik değerlere gösterdiği yaklaşım, AAT'lerin modellenmesinde çıkış suyu tahmini limit değerlere uyması gerektiğinden dolayı ayrıca önemlidir. Neticede MÖ algoritmalarının tahmin becerisini ölçeklemek ve kıyaslamak için seçilen metrikler önem taşımaktadır.Bu kapsamda; teorik yapısı Bölüm 3.5'te anlatılan metrikler arasından, tezde uygulanmak üzere ilgilendiğimiz problemin yapısına ve veri setinin karakteristiğine en uygun olan 3 tanesi (RMSE, MAE, R^2) seçilmiştir.

4.9.1 Kök Ortalama Kare Hatası (RMSE) Seçilme Gerekçesi

AAT modellemesinde, yukarıda belirttiğimiz gibi pik değer hatalarının yakalanması öncelikli olmalıdır. Modelin genel ortalamayı tutturması tek başına yeterli değil; asıl risk, pik yükler geldiğinde modelin düşük tahmin yapmasıdır. Bu sebeple hataların karesi ile hesaplama yaparak büyük hataları daha şiddetli cezalandıran RMSE metriği ana gösterge olarak seçilmiştir [216].

Hataların normal dağılım gösterdiği durumlarda da RMSE metriği, literatüre göre modelin başarısını en iyi gösteren yöntemdir [217]. Bu çalışmada RMSE'nin tercih edilmesinin ana sebebi; RMSE'nin düşük çıkması durumunda modelin sadece ortalama günlerde değil, şok yüklemelerin olduğu günlerde de kararlı çalıştığını göstermesidir.

4.9.2 Ortalama Mutlak Hata (MAE) Seçilme Gerekçesi

Yukarıda açıklananlardan anlaşılacağı üzere RMSE, büyük hatalara karşı çok hassas olduğu için veri setinde bulunabilecek aykırı değerlere karşı modelin başarısını hatalı olarak düşük gösterebilir. Bu gibi durumları önlemek ve modelin gerçek hata oranını açık bir şekilde tespit etmek için MAE metriği tamamlayıcı bir yöntem olarak seçilmiştir [217].

MAE metriğinde hataların mutlak değeri alındığından, sonuçların birimi ölçülen parametrelerin birimi ile aynıdır. Bu durum operatörler açısından daha kolay yorumlanabilirlik sağlar. Örneğin; "KOİ'nin çıkış değeri 6 mg/L ortalama hata ile tahmin edilmiştir" ifadesi, daha kolay anlaşılır bir yorum oluşturur.

RMSE ile MAE arasındaki hata oranı farklarının büyümesi, yani bu ikilinin değerlerinin arasının açılması model varyansının yüksek olduğunu gösterdiğinden dolayı bu çalışmada bu iki metrik birlikte değerlendirilmiştir.

4.9.3 Belirlilik Katsayısı (R^2) Seçilme Gerekçesi

RMSE ve MAE modelin ne kadar hata ile tahmin ettiğini gösterirken, R^2 modelin ne kadar doğru çalıştığını ve veri setindeki mevcut varyansı ne derecede ölçtüğünü gösterir. Birimsizdir, farklı algoritmaların aynı hedef için performansını kıyaslamada idealdir.

Son dönem literatür çalışmalarında, R^2 metriğinin farklı ölçekteki veri setlerini ve farklı çalışmalarla kıyaslama yapabilmek için kullanılabilecek en bilgilendirici standart yöntem olarak vurgulanmaktadır [215,218]. Tez çalışmasında, R^2 giriş parametrelerindeki değişimin çıkış parametrelerini ne kadar temsil ettiğini istatiksel olarak gösterme amacıyla seçilmiştir.

R^2, modelin veri setine uyumunun genel bir göstergesidir. Ancak model değerlendirilmesinde R^2 değerinin yüksek olması modelin mükemmel olduğunu tek başına göstermez MAE ve RMSE ye de bakılmalıdır [218].

4.9.4 Diğer Metriklerin Seçilmeme Nedenleri

Ortalama Mutlak Yüzde Hata (MAPE); hatayı gerçek değerin yüzdesi olarak ifade ettiğinden dolayı, hedef değişkeninin değerinin çok düşük olduğu durumlarda (örneğin TP çıkış: 0,2 mg/L) payda sıfıra yakın olduğu için değer aşırı büyük çıkar; bu da metriği anlamsız kılar. AAT verilerinde bu gibi durumlar olabileceği için bu yöntem seçenek olarak değerlendirilmemiştir.

Literatürde Huber kaybı veya Quantile kaybı, veri mühendisliği alanında yaygın kullanılmakta olan, MAE ve RMSE'nin avantajlarını birleştiren kayıp fonksiyonlarıdır. Ancak bu yöntemlerin yorumlanması kolay değildir ve uzmanlık gerektirmektedir. Tez çalışma konumuz disiplinler arası olduğundan dolayı, çevre mühendisliği alanındaki uzmanların kolay yorumlayabileceği metrikler seçilmiştir.

Sınıflandırma metrikleri olan ve literatürde sıkça adları anılan Accuracy, Precision, Recall, F1-Skoru gibi yöntemler; tezde ele aldığımız konu bir regresyon problemi (sürekli hedef değişkenlerin tahmini) olduğu için konu dışı bırakılmıştır.

Sonuç olarak özetleyecek olursak; RMSE, MAE ve R^2 metriklerinin kombinasyonu, çalışmamızın tüm ihtiyaçlarını karşılayabilmektedir. Şöyle ki:

- RMSE, büyük hatalara ve pik durumlara olan duyarlılığı ortaya koyar.
- MAE, kolay yorumlanabilirliği ve aykırı değerlere karşı sağlamlığı gösterir.

5
BULGULAR

Bu bölümde; tez çalışması kapsamında geliştirilen MÖ modellerinin AAT çıkış suyu parametrelerinin tahmin edilmesindeki başarısı, bir önceki Bölüm 4'te verilen 6 aşamalı metodoloji çerçevesinde analiz edilmiştir. Elde edilen bulgular; veri ön işleme, istatistiksel analiz, özellik seçimi, model eğitimi ve performansların kıyaslanması alt başlıkları altında anlatılmıştır.

5.1 Veri Seti Analizi ve Ön İşleme

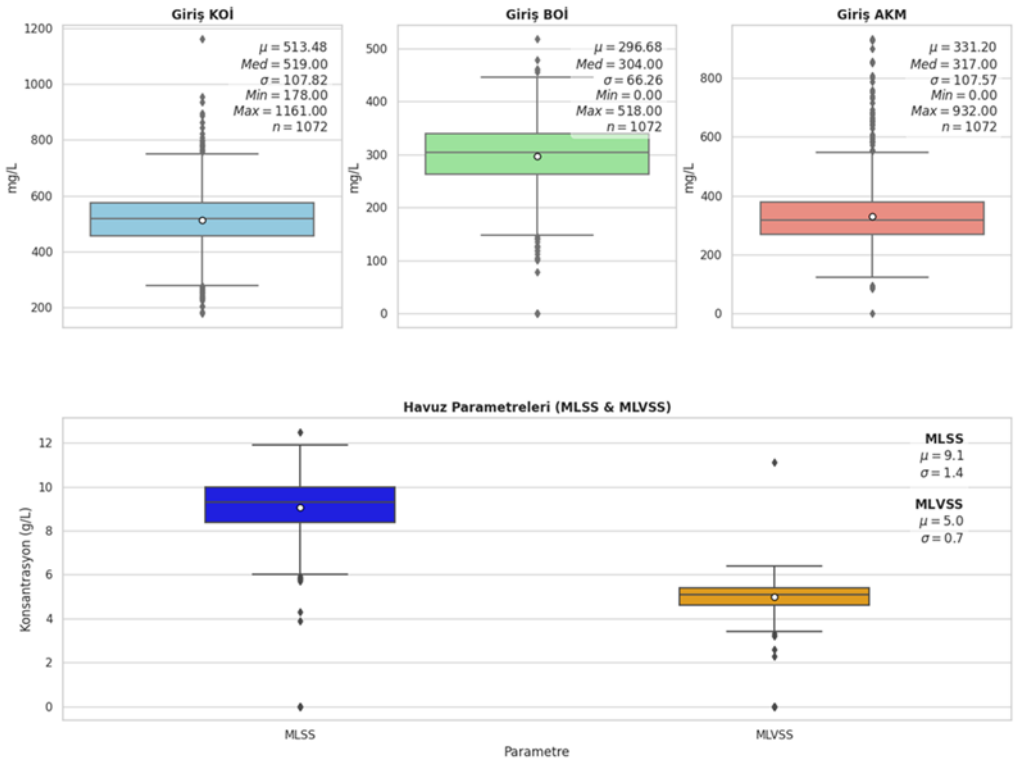
Tez çalışmasında kullanılan Tesis-A'ya ait veriler, yaklaşık 3 yıllık (1073 günlük) kesintisiz zaman serisi şeklindeki verilerdir. Veriler incelendiğinde; veri kalitesini ve model başarısını etkileyebilecek eksik verilerin ve aykırı değerlerin varlığı tespit edilmiştir.

Veri setinin bütünlüğünü sağlamak amacı ile yapılan analizlerde, veri setinin %0,3'üne tekabül eden 162 bağımsız hücrede eksik veri tespit edilmiştir. Özellikle biyolojik süreçlerde "0" olması mümkün olmayan; pH, sıcaklık, UAKM gibi değerlerdeki boşluklar "Ortalama Değer Atama" ile giderilmiştir. Böylelikle kesintisiz 1073 günlük veri seti elde edilmiştir. Eksiklikleri giderilmiş veri setinin istatistiksel analizini gösteren Tablo 5.1 ve Şekil 5.1'deki giriş parametrelerinin dağılımı incelendiğinde; Tesis-A'ya gelen kirlilik yüklerinin yüksek değişkenlik göstermesine rağmen prosesin çalıştığı gözlemlenmektedir. Tablo ve grafikte, giriş yüklerinin yüksek değişimlere uğradığı görülmektedir. Örneğin; Giriş BOİ5 ortalaması 296,68 iken standart sapması 107,57 olup yaklaşık olarak %36'lık bir değere tekabül etmektedir. Giriş AKM değerinde yaklaşık olarak %32'lik yüksek bir standart sapma vardır. Proses havuzlarının UAKM değerinin %12'lik standart

sapma oranı, dalgalı yükleri sistemin çözümleyebildiğini gösterse de bu değişimin nümerik modellerle çözümlenebilmesinin zor olacağı düşünülmektedir.

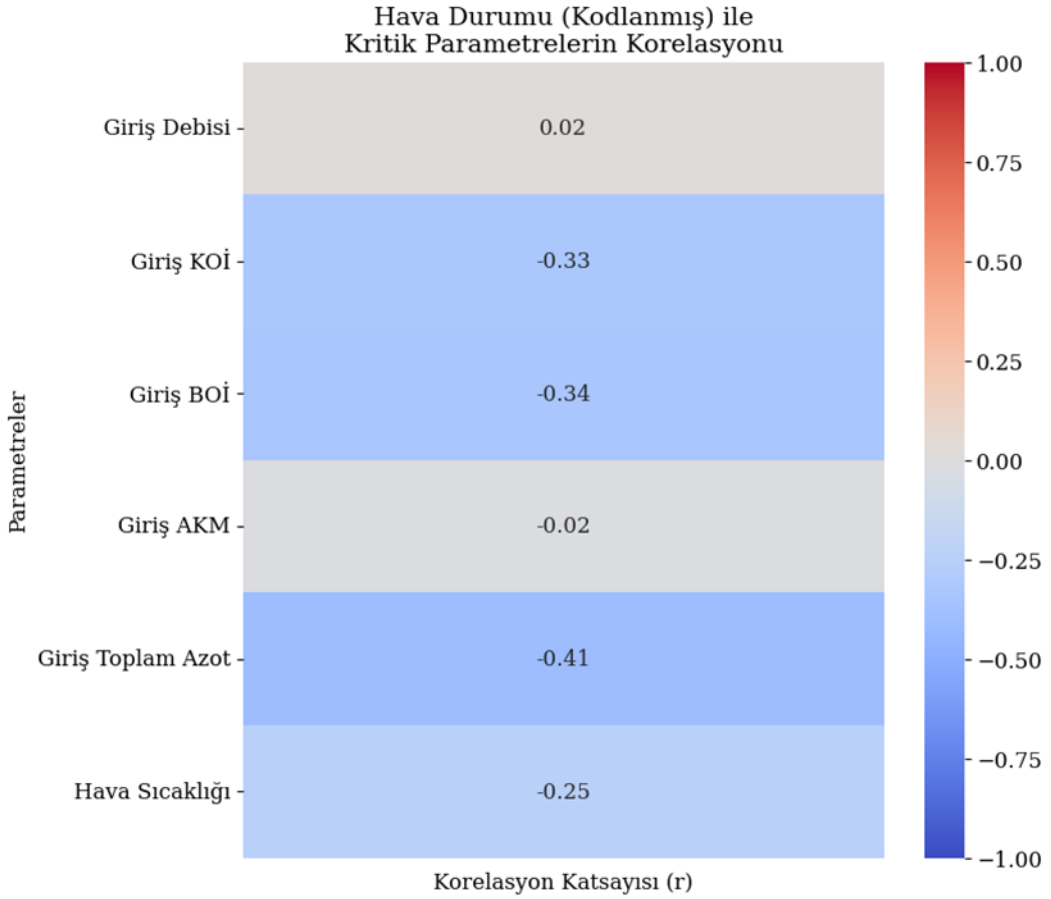
Tablo 5.1 Eksikliği giderilmiş veri setinin istatistiği

Parametre	Birim	Ortalama	Std.Sapma	Medyan	Mod
G. Debi	m3/gün	399,27	107.82	104,66	422,18
G. KOİ	mg/L	513.48	66.26	178.0	1,161.0
G. BOİ	mg/L	296.68	107.57	78.0	518.0
G. AKM	mg/L	331.20	107.06	86.0	932.0
G. TN	mg/L	53,91	9,78	18,90	84,40
G. TP	mg/L	5,17	1,27	1,50	11,30
Giriş Alkalinite	CaCO3	335.91	37.45	184.0	558.0
AKM (Havuz)	kg/m3	9,10	1,28	3,91	12,52
UAKM (Havuz)	kg/m3	5.00	0.62	2,29	11,06
pH (Havuz)	-	6,97	0,07	6,41	7,99



Şekil 5.1 Giriş parametreleri ve işletme parametrelerinin istatiksel dağılımı

Eksik verilerin giderilmesinden sonra, veri setinde yer alan kategorik (sözel) değişkenlerin kurulan MÖ algoritmaları tarafından sisteme dahil edilebilmesi için sayısallaştırma işlemi; metodoloji kısmında da anlatılan "Etiket Kodlama (Label Encoding)" yöntemi ile uygulanmıştır. Veri setinde yer alan ve Hava Durumu (Weather Condition) olarak tanımlanan değişken sözel yapıda; "D (Yağışsız)" ve "W (Yağışlı)" olarak tanımlanmıştır. Python'daki sklearn.preprocessing modülü ile bu değişkenlere 0 ve 1 değerleri atanmıştır. Böylelikle D ve W sınıfları, modelin girdi matrisine dahil edilebilir bir formata (0 ve 1) dönüştürülmüştür. Bu dönüşümle yağışlı veya kurak günlerin etkisinin, Tesis-A üzerindeki kirlilik yükü ve konsantrasyonlarını ne şekilde etkileyebileceğinin bilgisini model öğrenebilmiştir. Böylelikle debi artışının veya seyrelmenin etkisini model bir örüntü olarak yakalayabilmiştir.



Şekil 5.2 Etiket Kodlama ile işlenen hava durumu ve giriş parametrelerinin pearson korelasyon ısı haritası

Şekil 5.2'de, etiket kodlama ile sayısallaştırılmış ve modellemeye dahil edilen Hava Durumu değişkeninin kritik giriş parametreleri ile ilişkisi incelenmiştir. Hava durumu değişkeni ile giriş değişkenleri olan KOİ arasında r=-0,33; BOİ5 ile arasında r=-0,34; TN ile arasında r=-0,41 gibi anlamlı negatif bir ilişki vardır. Buradaki negatiflik, yağışla birlikte bir seyrelme olduğunun göstergesidir. Debinin korelasyon katsayısının r=0,02 olması, fazla gelen hidrolik yükün sisteme alınmadan by-pass edildiğini göstermekteyken; Giriş AKM değerinin r=-0,02 gibi düşük bir seviyede kalması, yağışlarla birlikte sisteme inert partikül maddelerin taşındığını ve bu sebeple AKM konsantrasyonunun seyrelmediğini (değişmediğini) göstermektedir.

5.2 Özellik Seçimi Senaryoları Bulguları

Veri setindeki en önemli özellikleri seçmek için Python, scikit-learn kütüphanesinden feature_selection modülleri kullanılarak gerçekleştirilen bu

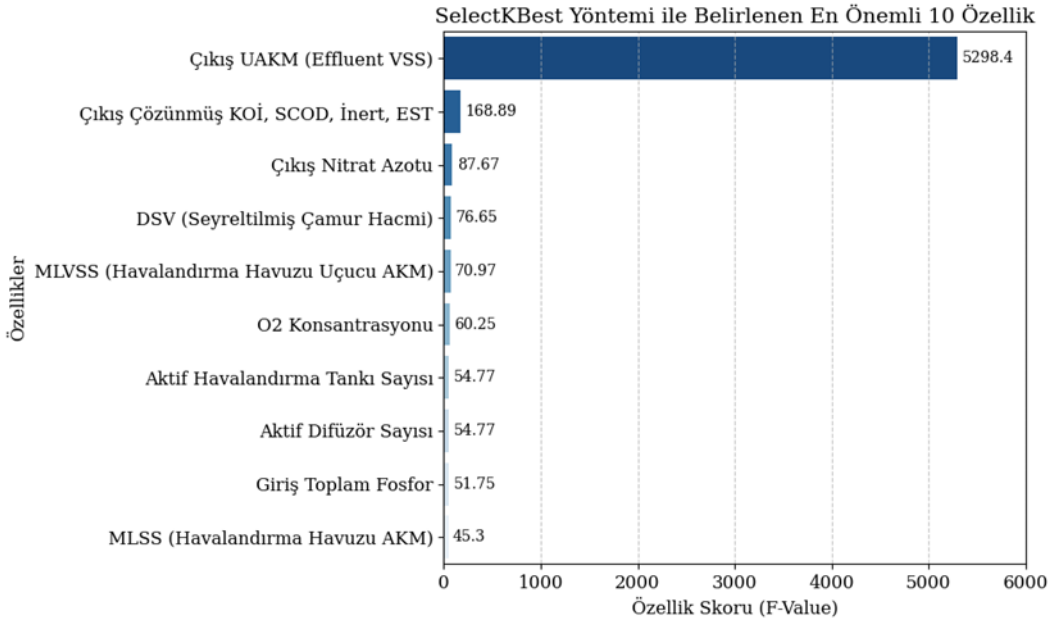
adımda; veri setinin istatistiksel özelliklerini, karmaşık ilişkilerini ve önem düzeylerini dikkate alan bir yöntem izlenmiştir. Bu kısımda ham veri setinde bulunan 51 adet girdi değişkeni, her bir senaryo için en yüksek tahmin gücüne sahip 10 temel özelliğe (k=10) indirgenmiştir. Bu üç ayrı senaryo yöntemi ile (SelectKBest, Karşılıklı Bilgi, RFE); limitlerle sınırlı olan tesisin deşarj suyu kalitesini (KOİ, BOİ, AKM, TN, TP) tahmin etmedeki başarılarına göre en uygun 10 özellik seçilmeye çalışılmıştır.

Alt başlıklar halinde her bir yöntemin bulduğu 10 önemli özellik ve kıyaslamaları aşağıda verilmiştir.

5.2.1 SelectKBest Yöntemi Sonuçları (Senaryo I)

Modellemenin ilk senaryosunda, özellik seçimi olarak tek değişkenli istatistiksel analiz yapan SelectKBest yöntemi tercih edilmiştir. Veri setinde bulunan 51 adet değişken arasından tesis performansını deşarj parametreleri yönünde temsil edebilecek 10 özellik bu yöntemle belirlenmiş ve önem skorları aşağıdaki Şekil 5.3 görsel grafiğinde verilmiştir.

Şekil 5.3'teki grafik incelendiğinde; 5298,4 gibi çok yüksek bir skorla diğer 9 özellikten büyük bir farkla ayrışan özelliğin Çıkış UAKM olduğu görülmektedir. İkinci olarak ona yaklaşan özelliğin ise yine Çıkış KOİ değeri (168,9) olduğu tespit edilmiştir. Bu kadar büyük farklarla UAKM'nin seçilmiş olması, kullanılan SelectKBest algoritmasının yapısı gereği lineer ilişkilere odaklandığını göstermekte ve çıkış suyu kalitesini en çok katı madde kaçışının açıkladığına işaret etmektedir.



Şekil 5.3 SelectKBest yöntemi ile seçilen etkili 10 adet özellik ve F-skorları

Aslında bu durum AAT işletme mantığı ile son derece örtüşmektedir. BNR proseslerinde, çıkış suyundaki KOİ ve BOİ'nin asıl kaynağı; çökelerek çamurla uzaklaştırılmayan, çıkış suyunda bulunan UAKM'dir. Dolayısı ile SelectKBest algoritmasının çıkış kalitesini tahmin etmek için Çıkış UAKM'yi en önemli parametre olarak seçmesi; aslında kirletici yükünün fiziksel olarak nasıl taşındığı mekanizmasını doğru olarak yakaladığını göstermektedir.

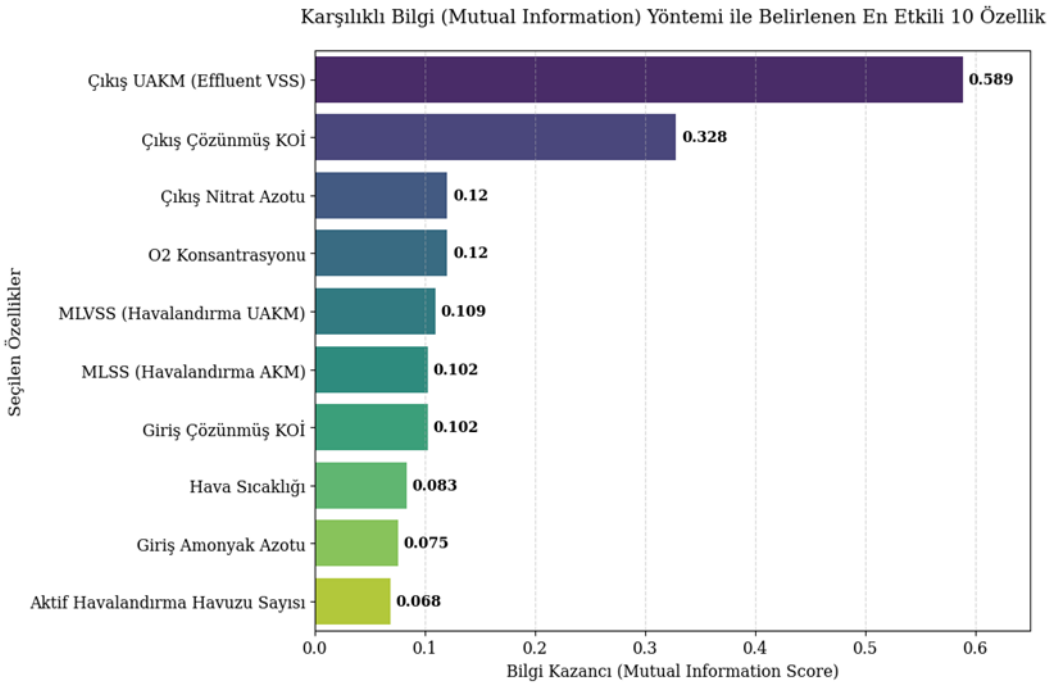
Diğer parametreler arasında; Çıkış Nitrat Azotu (87,8), Seyreltilmiş Çamur Hacmi (DSV - 76,7) ve Havalandırma Havuzu UAKM (MLVSS - 71,0) gibi parametreler, çamur kalitesinin ve biyolojik aktivitenin göstergesidir. Ayrıca Çözünmüş Oksijen Konsantrasyonu (60,3) da biyolojik aktivitenin çıkış kalitesini istatistiksel olarak etkilediğini doğrular niteliktedir.

Neticede; SelectKBest yöntemi, tesisin deşarj kalitesini etkileyen faktörleri belirlerken; çıkış suyu kalitesi (UAKM ve KOİ) ile reaktör parametreleri (MLVSS, ÇO, DSV) üzerinde yoğunlaşan bir özellik seti oluşturmuştur.

5.2.2 Karlışıklı Bilgi (MI) Yöntemi Sonuçları (Senaryo II)

AAT'lerde prosesleri temsil eden biyokimyasal reaksiyonlar, yapısı itibarıyla doğrusal ilişkilerle gerçekleşmez. Bir önceki Senaryo I'de kullanılan SelectKBest doğrusal ilişkiler üzerinden bir algoritma çalıştırarak sonuca etki eden en önemli 10 parametreyi belirlerken; "Karşılıklı Bilgi (Mutual Information - MI)" doğrusal

olmayan ilişkileri de analiz etmektedir. MI analizi sonucunda elde edilen ve sistem performansını etkileyen en önemli 10 özellik ve önem skorları Şekil 5.4'te verilmiştir.



Şekil 5.4 Karşılıklı bilgi (MI) ile seçilen en etkili 10 adet özellik ve önem skorları

Şekil 5.4 incelendiğinde; MI yönteminin tesisin biyokimyasal ve işletmesel özelliklerini hassas bir şekilde ayırt ettiği görülmektedir. Burada da yine baskın olan özellik Çıkış UAKM olup, 0,589 skorla en belirleyici özellik olmuştur. Yine aynı şekilde, çıkış çamur kaçağının; çıkış KOİ, BOİ ve TN konsantrasyonlarını etkilediği tespit edilmiştir.

Tablodan ikinci önemli bir çıkarım da biyolojik proses göstergeleri olan nitrifikasyon ve oksijen parametreleridir. Doğrusal yöntem olan SelectKBest'ten farklı olarak MI; Çıkış Nitrat Azotu ve Oksijen Konsantrasyonu parametrelerinin her ikisini de 0,12'lik önem skoru ile 3. sıraya almıştır. Bu bulgu; algoritmanın havalandırma verimliliği (O_2) ile biyolojik dönüşüm ürünü olan nitrat arasındaki doğrusal olmayan ilişkiyi yakalamış ve öğrenmiş olması açısından önemlidir.

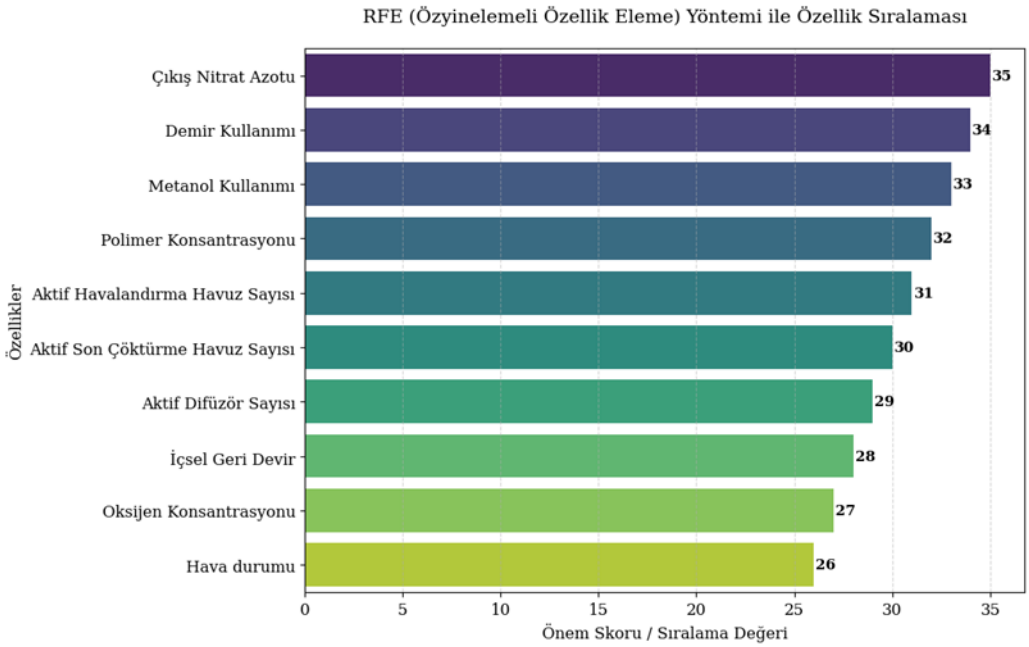
İkinci sıradan önem önceliği verdiği ve 0,328'lik bir skora sahip olan Çıkış Çözünmüş KOİ ile seçilen diğer özelliklerden olan, havalandırma havuzundaki aktif biyokütleyi temsil eden MLVSS (0,19) ve MLSS (0,102) parametreleri de;

organik yük ve biyokütle dengesinin algoritma tarafından önceliklendirildiğini göstermektedir.

Ayrıca burada düşük de olsa (0,083) hava sıcaklığını anlamlı bir değişken olarak sisteme dahil etmiştir.

5.2.3 RFE (Özyinelemeli Özellik Eleme) Yöntemi Sonuçları (Senaryo III)

RFE, diğer algoritmaların aksine en başarılı olanları değil, en az katkısı olanları eleyerek sonuca ulaşmakta; yani "geriye doğru eleme" mantığı ile çalışmaktadır. Şekil 5.5'te görüldüğü gibi RFE, sisteme en az etkisi olanları sistemden çıkartarak son kalan 10 özelliği seçmiştir.



Şekil 5.5 RFE ile seçilen en etkili 10 adet özellik ve önem skorları

Şekil 5.5 dikkatle incelendiğinde; yukarıda bahsettiğimiz RFE'nin çalışma mantığından kaynaklı bazı ilginç seçilimler oluşmuştur. Algoritma sürekli değişen giriş yükü verilerini elerken, çok sık rastlanılmayan ve çok fazla değişim göstermeyen operasyonel verileri sona bırakarak seçmiştir. Bu durum mühendislik açısından şu şekilde yorumlanabilir:

- Çıkış Azotundan (35) sonra en yüksek skor oranına sahip olan Demir (34), Metanol (33) kullanımı ve Polimer Konsantrasyonu (32); veri setinde sayıca en az gözlemlenen değişkenlerdir. RFE katkısı en az olanı elerken, bu kimyasalların kullanımı çıkış suyu kalitesini en hızlı artıran faktör olduğundan, en önemli

faktörler olarak sona kalmışlardır. Bu değişkenler yılın büyük bir kısmında kullanılmasalar da sisteme dahil olduklarında en yüksek etkiye sahip oldukları için, sistemin en önemli ögeleriymiş gibi üst sıralara yerleşmişlerdir.

- Ayrıca genelde operasyonel olarak sabit olan Aktif Havalandırma Havuz Sayısı (31), Aktif Son Çöktürme Sayısı (30) ve Aktif Difüzör Sayısı (29), yukarıda anlatılan benzer sebeplerle RFE'de seçilmişlerdir. Bu operasyonel parametreler genelde sabit kalmakta; sadece bakım zamanlarında, ani yük değişimlerinde veya öngörülemeyen büyük arızalarda değişmektedir. Bu da söz konusu ani değişiklikler nedeniyle, bu parametrelerin son 10 özellik arasında kalmasına yol açmıştır.

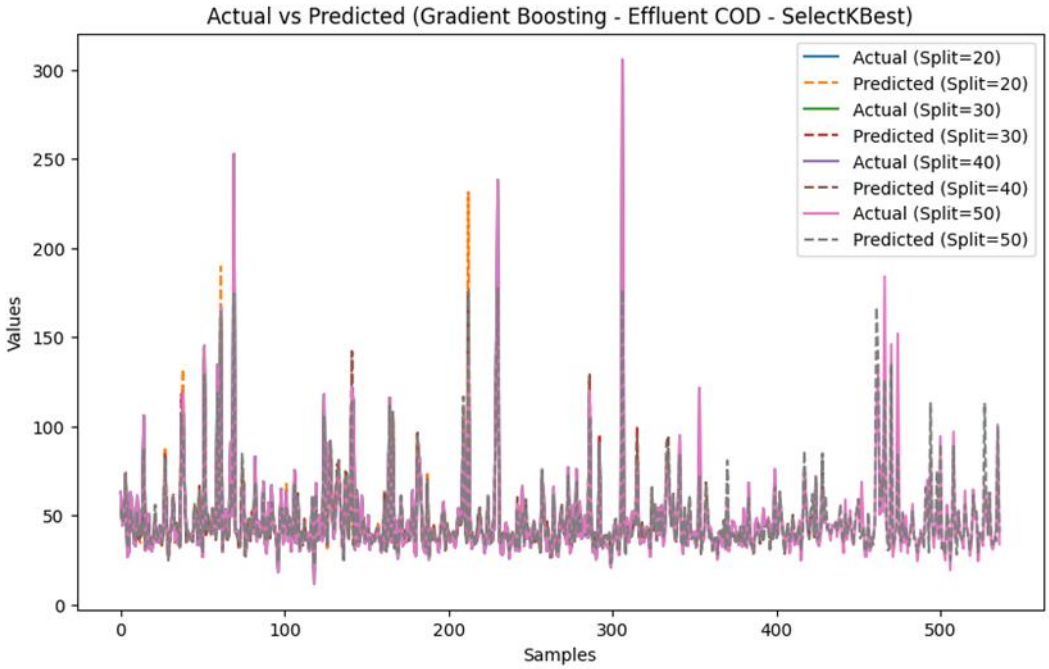
Bu tez kapsamında, yukarıdaki bulgular sebebiyle RFE yöntemini eleştirmiş olsak da seçilmiş özellikler arasındaki önem skorlarının birbirine yakın olması; Çıkış Nitrat Azotu (35), İçsel Geri Devir Oranı (28), O_2 Konsantrasyonu (27) ve Hava Durumu (26) gibi AÇ sistemindeki nümerik modellerde ve proses kinetiklerinde kritik öneme sahip parametreleri yakalamıştır. Bu sebeple, söz konusu yöntemle de çalışılmaya devam edilmiştir.

5.3 Veri Setini Bölme İşlemi

Modelin veri bölme işlemi; Python'un scikit-learn kütüphanesinde yer alan train_test_split fonksiyonundan yararlanılarak "Tabakalı Bölme (Stratified Splitting)" yöntemi ile yapılmıştır. Burada rastgelelik etkilerini engelleyebilmek ve sonuçların tekrarlanabilirliğini sağlamak için sabit bir rastgelelik parametresi (random_state) tanımlanmıştır.

Bu tez kapsamında veri seti; %80-%20, %70-%30, %60-%40 ve %50-%50 oranlarında eğitim ve test verisi olmak üzere 4 farklı şekilde bölümlenmiştir. Literatürde genel kabul gören veri bölümleme oranı %80-%20 olsa da bu çalışmada diğer oranlar da test edilmiştir. Bu testlerin amacı, veri setimizin özelliklerinden kaynaklanabilecek ve aşırı öğrenmeye (overfitting) sebep olabilecek bir durumun varlığını kontrol etmektir. Tüm modelleme çalışmaları, bu 4 ayrı senaryo için de yürütülmüştür. Çalışmadaki bu veri bölme stratejisini gösteren Şekil 5.6 aşağıda verilmiştir.

Yapılan incelemeler, veri bölme oranlarının model performansı üzerinde belirleyici bir etkiye sahip olduğunu göstermiştir.



Şekil 5.6 Dört ayrı veri bölümlemesi ile çalışıldığını gösterir grafik.

Test oranının %20'den %50'ye kademeli olarak çıkarıldığı durumlarda, eğitim verisi azaldığı için modelin öğrenme kapasitesinin düştüğü gözlemlenmiştir. Şok yüklemelerin yaşanabildiği dinamik süreçlerden dolayı AAT verileri yüksek değişkenlik gösterir. Eğitim verisinin %50'ye düşürülmesi, bu tür uç durumların (outliers) veri setinde daha az temsil edilmesine ve dolayısıyla öğrenme veriminin azalmasına neden olmuştur. Eğitim verisindeki bu kısıtlılık, modelin şok yükleme senaryolarını yeterince kavrayamamasına ve Tablo 5.2'de görüldüğü gibi hata oranının (MAE) göreceli olarak daha yüksek (6,13) çıkmasına sebep olmuştur.

Denemeler sonucunda, Tablo 5.2'de görüldüğü üzere, en yüksek R^2 (82,05) değeri %80-%20 oranında gerçekleşmiştir. Tesis-A'nın 1073 günlük verisi ile mevsimsel döngülerin yakalanması ve AAT proses kinetiğine adaptasyonun sağlanması için; eğitim verisinin en fazla olduğu %80-%20 oranı ile modelleme sonuçları yorumlanmış ve model değerlendirme metrikleri bu şekilde uygulanmıştır.

Tablo 5.2 GBM-çıkış KOİ tahmin başarısının veri bölme oranlarına göre durumu

Veri Bölme Oranı	Eğitim Verisi (Gün)	Test Verisi (Gün)	MAE (Hata)	R^2 (Başarı)	Durum
%80-%20	858	214	6,03	%82,05	Yüksek Başarı
%70-%30	750	322	5,90	%83,07	Dengeli
%60-%40	643	429	5,64	%81,22	R^2 Düşüyor
%50-%50	536	536	6,13	%79,29	Düşük Performans

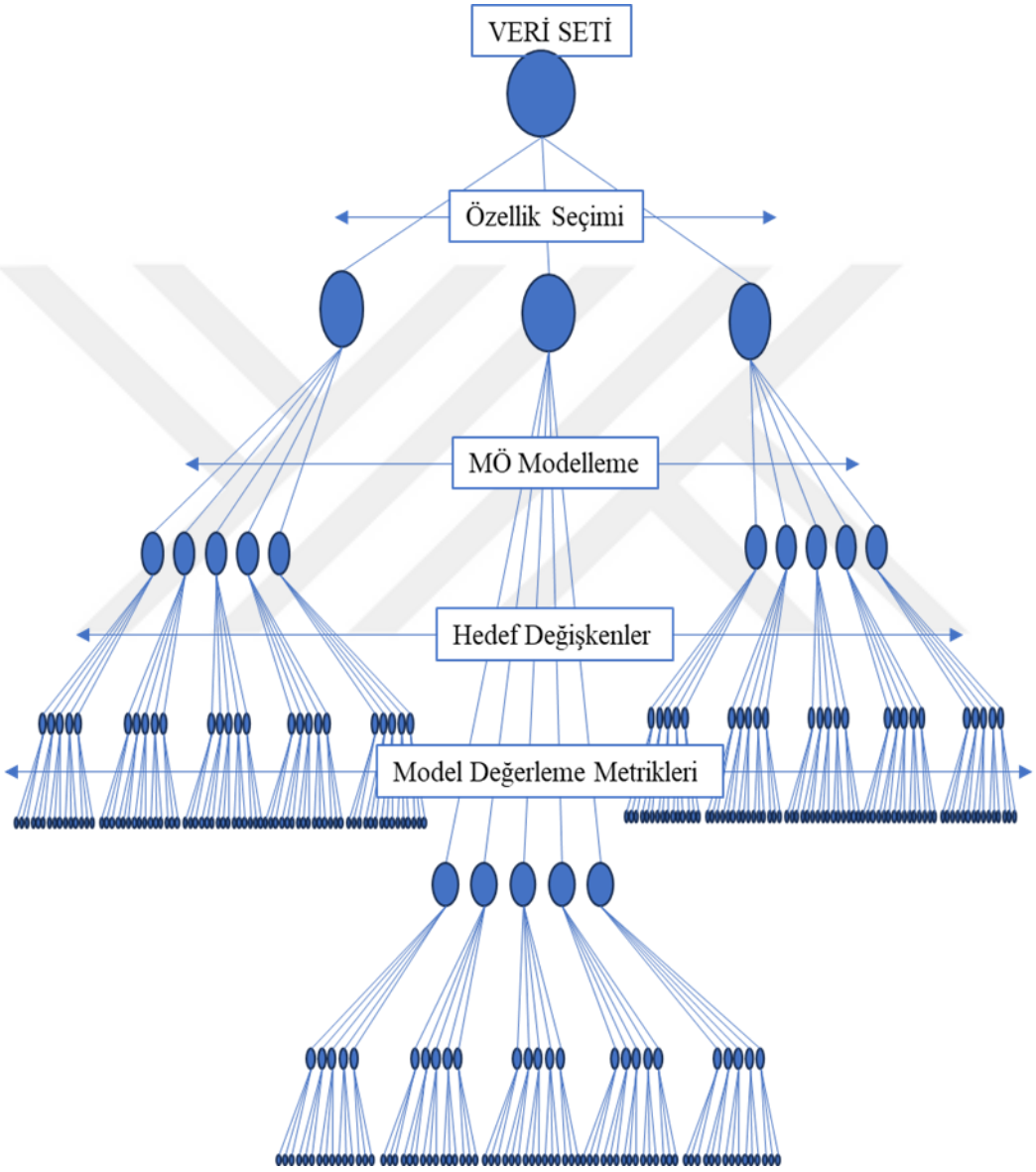
Sonuç olarak bu bulgu, bu tez kapsamında çalışılan modeller içerisinde aşırı öğrenme (overfitting) durumunun olmayacağını göstermesi açısından önemlidir. Ayrıca AAT modellerinde, verinin büyüklüğünden ziyade dönemsellik (seasonality) temsilinin de önemli olduğunu göstermektedir. Eğitim verisi olarak daha fazla pay ayrılmasının (%80); modelin, tesisin hidrolik ve biyolojik hafızasını daha iyi öğrendiğini kanıtlarken; %20'lik test verisinin de modelin genelleme yeteneği için yeterli olduğunu ortaya koymuştur.

5.4 MÖ Model Uygulamaları ve Değerleme Metrikleri Sonucu

Bu çalışmanın nihai amacı; Tesis-A verilerini kullanarak MÖ algoritmalarının, literatürdeki nümerik modelleme yaklaşımlarına bir alternatif olup olmayacağının değerlendirilmesidir. AAT'ler, çok fazla veri üretebilen tesislerdir. Veri gürültüsünden korunmak ve literatürde daha önce denenmemiş olan model algoritmalarını uygulamadan önce, özellik seçimi (feature selection) yöntemlerinin uygulanması hedeflenmiştir.

5.3 başlığı altında SelectKBest, MI ve RFE yöntemleri ile seçilen 10'ar adet özellik; bu bölümde MÖ algoritmaları kullanılarak tesis deşarj standartları için kritik öneme sahip olan Çıkış KOİ, Çıkış BOİ, Çıkış AKM, Çıkış TN ve Çıkış TP tahminlemesi için geliştirilecek modellere altlık olarak kullanılacaktır. Burada kullanılan MÖ algoritmaları (DT, RF, GBM, XGBoost ve LightGBM) için aynı veri bölme oranları (%80-%20) kullanılmıştır. Ayrıca her MÖ modeli; üç ayrı değerlendirme metriği (MAE, RMSE ve R^2) ile değerlendirilmiştir.

Çalışmada yapılan işlemler; veri setinden 3 ayrı özellik seçimine, özellik seçiminden 5 ayrı model algoritmasına, model algoritmalarından 5 ayrı hedef değişkenine ve hedef değişkenlerinden 3 ayrı değerlendirme metriğine doğru ilerlemektedir. Bu da Şekil 5.6'da görüldüğü gibi, analiz edilmesi gereken 225 adet sonuç demektir. Bu sebeple bulguların daha anlaşılır bir şekilde değerlendirilmesi için sistem 3 ayrı senaryoya ayrılmıştır.



Şekil 5.7 Tez kapsamında yapılan çalışmalarda ulaşılan sonuçlar

Yukarıda belirtilen karmaşıklığın çözümlenmesi için 3 ayrı senaryo, aşağıda Tablo 5.3'te gösterildiği şekilde kurgulanmıştır.

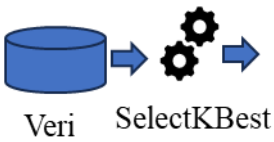
Tablo 5.3 Senaryoların kurgu yapısı

Senaryo	Özellik Seçimi Yöntemi	Odaklandığı İlişki Türü	Uygulanan Modeller	Hedeflenen Çıktılar	Değerleme Metriği
Senaryo-1	SelectKBest	Doğrusal İlişkiler	DT RF GBM XGBoost LightGBM	Çıkış KOİ Çıkış BOİ Çıkış AKM Çıkış TN Çıkış TP	MAE RMSE R^2
Senaryo-2	MI	Doğrusal Olmayan İlişkiler	DT RF GBM XGBoost LightGBM	Çıkış KOİ Çıkış BOİ Çıkış AKM Çıkış TN Çıkış TP	MAE RMSE R^2
Senaryo-3	RFE	Model Performansı ve Etkileşimi	DT RF GBM XGBoost LightGBM	Çıkış KOİ Çıkış BOİ Çıkış AKM Çıkış TN Çıkış TP	MAE RMSE R^2

Tezde alt başlıklar halinde tüm yukarıdaki kurgulanan senaryo yapısına göre, model sonuçları ve değerleme metrikleri bazında değerlendirilecektir.

5.4.1 Senaryo-1 Analizi: SelectKBest Özellik Seçimi Sonuçları

Bu senaryo kapsamında, aşağıda Şekil 5.7'de sunulan 10 adet özellik SelectKBest algoritması ile seçilmiştir. Bu yöntem; girdi değişkenleri ile hedef değişkenler olan KOİ, BOİ, AKM, TN ve TP arasındaki doğrusal ilişkileri ölçerek, korelasyonu en yüksek özellikleri modele dahil etmektedir.



Seçilen Özellikler	
Çıkış UAKM	Ç.O
Çıkış Çöz. KOİ	Aktif Tank Sayısı
Çıkış Nitrat-N	Aktif Dif. Say.
DSV	Giriş TP
MLVSS	MLSS

Şekil 5.8 Senaryo-1'de SeelctKBest ile seçilmiş özellikler

5.4.1.1 Makine Öğrenme Algoritmalarının Performans Karşılaştırmaları

Bu başlık altında, SelectKBest kullanılarak indirgenmiş veri setleri üzerinde beş farklı algoritmanın (RF, GBM, XGBoost, LightGBM ve DT) tüm değerlendirme

metrikleri (MAE, RMSE, R^2) kıyaslanarak sunulacaktır. Bu bağlamda hazırlanmış olan Tablo 5.4 aşağıda verilmiştir.

Tablo 5.4 Senaryo-I'e göre MÖ model ve değerleme sonuçları

Özellik Seçim Metriği: SelectKBest				
Hedef Değişken	Model	MAE	RMSE	R^2 (%)
KOİ (Çıkış)	Random Forest	6.019023	11.307	82.21%
KOİ (Çıkış)	Gradient Boosting	6.063454	11.649	81.12%
KOİ (Çıkış)	XGBoost	6.251666	10.919	83.41%
KOİ (Çıkış)	LightGBM	6.75571	12.567	78.02%
KOİ (Çıkış)	Decision Tree	9.039535	18.171	54.05%
BOİ (Çıkış)	Random Forest	1.623498	2.466	74.28%
BOİ (Çıkış)	Gradient Boosting	1.645596	2.465	74.29%
BOİ (Çıkış)	XGBoost	1.686399	2.516	73.21%
BOİ (Çıkış)	LightGBM	1.768605	2.748	68.04%
BOİ (Çıkış)	Decision Tree	2.295814	3.860	36.97%
AKM (Çıkış)	Random Forest	4.187265	9.378	95.39%
AKM (Çıkış)	Gradient Boosting	3.732114	7.514	97.04%
AKM (Çıkış)	XGBoost	4.629108	11.485	93.09%
AKM (Çıkış)	LightGBM	6.028417	17.678	83.63%
AKM (Çıkış)	Decision Tree	5.166512	11.278	36.97%
TN (Çıkış)	Random Forest	0,621885	0,942	64.15%
TN (Çıkış)	Gradient Boosting	0.673567	1.057	37.23%
TN (Çıkış)	XGBoost	0.746523	1.398	42.50%
TN (Çıkış)	LightGBM	0.772575	1.338	-84.08%
TN (Çıkış)	Decision Tree	1.083744	2.394	-57,7%
TP (Çıkış)	Random Forest	0.227669	0.491	18.23%
TP (Çıkış)	Gradient Boosting	0.223368	0.489	18.85%
TP (Çıkış)	XGBoost	0.224297	0.477	22.75%
TP (Çıkış)	LightGBM	0.230975	0.459	28.68%
TP (Çıkış)	Decision Tree	0.330349	0.682	-57.70%

Tablo 5.4'teki bulgular; tekil ve topluluk öğrenme mimarilerinin tahmin başarıları arasındaki belirgin performans farkını ortaya koymaktadır.

- Topluluk Modelleri: Topluluk modellerinden hem Bagging (RF) hem de Boosting (GBM, XGBoost, LightGBM) algoritmaları, tekil bir model olan DT'ye kıyasla ezici bir üstünlük sağlamıştır. Özellikle AKM parametresinde GBM %97,04'lük bir R^2 başarı oranına sahipken, XGBoost algoritması da %83,41'lik R^2 değeri ile yüksek tahmin doğruluğuna erişmiştir. SelectKBest yönteminin seçtiği doğrusal ilişkilere sahip özellikler; gürültülü olabilmelerine rağmen, topluluk öğrenmelerinin içsel yapısındaki hata düzeltme mekanizmaları sayesinde başarılı sonuçlar vermiştir.
- Tekil Modeller: Çalışmada kullanılan tekil model, bir referans noktası olarak ele alınmış ve beklentilerimizi karşılamıştır. Bu algoritma; KOİ ve BOİ gibi nispeten kararlı ve AAT proses kinetiklerinde nümerik olarak daha kolay hesaplanabilen parametrelerde dahi oldukça düşük (%36-54 R^2) performans sergilemiştir. Öte yandan TN ve TP gibi daha karmaşık mekanizmalara sahip parametrelerde ise negatif başarı oranları (-%57,7 ve -%57,7) vererek modelleme yeteneğini tamamen kaybettiğini göstermiştir.

Bu durumun mühendislik yorumu şu olabilir: AAT verileri doğası gereği; yük ve işletme koşullarındaki zamansal değişimler nedeniyle dalgalı, ölçüm kaynaklı belirsizlikler nedeniyle de gürültülü olabilmektedir. Tek bir karar ağacı, bu gürültüyü öğrenmeye çalışırken veriyi ezberlemiş ve aşırı öğrenme (overfitting) sorunu ile karşı karşıya kalmıştır. Tesis verileri modellenirken, tekil modellemelerden ziyade verideki varyansı sönümleyebilen ve genelleme yeteneği yüksek olan topluluk modelleri (ensemble models) tercih edilmelidir.

5.4.1.2 Parametre Bazlı Model Analizi

MÖ Algoritmaları arasındaki genel karşılaştırmalar sonrasında, her bir parametre için en düşük hata oranı (MAE, RMSE) ve en yüksek başarı yüzdesi (R^2) değerlerini sağlayan "En İyi Modeller" belirlenmiş ve Tablo 5.5'te sunulmuştur.

Tablo 5.5 Senaryo-1' e ait en iyiler tablosu

Hedef Değişken	Seçilen En İyi Model	MAE	RMSE	R^2 (%)
KOİ	XGBoost	6.251666	10.919	83.41%
BOİ	Gradient Boosting	1.645596	2.465	74.29%
AKM	Gradient Boosting	3.732114	7.514	97.04%
TN	Random Forest	0.621885	0.942	64.15%
TP	LightGBM	0.230975	0.459	28.68%

Tablo 5.5'teki makine öğrenmesi algoritmalarının tahmin başarıları; AAT'nin fiziksel, biyokimyasal ve nütrient (besin maddesi) giderim süreçleri temelinde, aşağıda maddeler halinde irdelenmiştir.

1) Fiziksel Artıma Süreci-AKM: Bu parametreyi en iyi tahmin eden model Gradient Boosting (GBM) olmuştur R^2: %97,04; RMSE: 7,515). Bu sonuç, Senaryo-1'in elde ettiği en büyük başarıdır. SelectKBest ile seçilen özellikler arasında bulunan UAKM, MLSS, MLVSS ve DSV gibi parametreler ile AKM arasındaki lineer ilişkilerin %97 oranında açıklanabilmesi oldukça önemli bir bulgudur. GBM, bu deşarj kalite parametresini bir "sanal sensör" hassasiyeti ile tahmin edebilmiştir.
2) Organik Madde Giderimi-KOİ ve BOİ: KOİ'yi XGBoost R^2: %83,41 başarı oranı ile tahmin edebilirken, BOİ'yi parametresini ise Gradient Boosting R^2: %74,29'luk bir oranla tahmin edebilmiştir. Karbon giderimi, fizikokimyasal çökelme ile birlikte biyolojik oksidasyona dayanır. SelectKbest yöntemi, karbon giderimi ile ilgili olan parametreleri seçmede başarılı olduğu söylenebilir.

   KOİ, biyolojik olarak ayrışabilen ve ayrışamayan inert kısımlardan oluşur. XGBoost'un düzenleme (regularization) yeteneği, bu çok faktörlü yapıyı aşırı öğrenmeye (overfitting) kaçmadan kavrayabilmiştir. BOİ tahminindeki %74'lük nispeten düşük başarı oranı ise, bu parametrenin Monod kinetiği gibi doğrusal olmayan ilişkilere sıkı sıkıya bağlı olmasından kaynaklandığı şeklinde yorumlanabilir. SelectKBest yöntemi öncelikle doğrusal ilişkilere odaklandığı için BOİ'nin bu karmaşık dinamiklerini bir miktar kaçırmış olabilir.

3) Nutrient Giderimi-TN ve TP: Bu iki parametre AAT modellemesinin en zor kısmını oluşturmaktadır. SelectKBest yöntemi gölgesinde eğitilen modeller değişken ama literatürle uyumlu sonuçlar sergilemişlerdir [9,46]. Ancak bu parametreler değerlendirilirken sadece R^2 hata metriğine bakmak doğru bir yaklaşım olmayacaktır. TN ve TP'nin yasal deşarj limitleri sırası il e10 mg/L ve 1 mg/L dir. Başarı oranı metrikleri düşük çıktığında, veri seti bu iki parametre açısından incelenmiş ve her birinin kendi içinde değer değişiminin çok az olduğu gözlemlenmiştir. Bu sebeple Metrikler değerlendirilirken MAE ve RMSE ye de bakmak gerekmektedir R^2

Toplam Azot (TN) tahmininde en yüksek başarıyı (%64,15 R^2) Rastgele Orman (RF) algoritması sergilemiştir. Azot giderimi (nitrifikasyon / denitrifikasyon); birçok ara kademeye ve hassas proses şartlarına (Ç.O, pH, Alkalinite) bağlıdır. Ancak bu prosesin temelinde stokiyometrik denklemler vardır (örneğin; 1 g amonyak azotu giderimi için 4,57 g oksijen gerekmesi gibi) ve SelectKBest yöntemi, gerekli özelliklerden bazılarını (ÇO, Çıkış-N, Aktif Difüzör Sayısı) seçmeyi başarmıştır. Bu değerlendirmeye şu açıdan yaklaşmak mühendislik gereğidir: TN deşarj standardı 10 mg/L'dir. MAE metrik değeri ise en iyi model olan RF'de 0,622 mg/L olarak ölçülmüştür. Bu değer, yasal limitin sadece %6,22'sine tekabül etmektedir. Kıyaslama yapmak gerekirse; en başarılı model olarak kabul ettiğimiz AKM'nin göreli hata oranı (Deşarj limiti: 35 mg/L; MAE: 3,73) %10,6'ya denk gelmektedir ki bu, TN'nin göreceli hata oranından daha yüksektir. Bu da modelin aslında TN tahmini açısından da operatör kararlarına güçlü bir destek sağlayabileceğini kanıtlamaktadır.

Toplam Fosfor (TP) için en iyi model tahminlemesi LightGBM ile elde edilmiş olsa da, %28,68 gibi oldukça düşük bir R^2 oranı gözlemlenmiştir. Bu durumun temel nedeni; veri setinde TP giderim mekanizmasını doğrudan açıklayabilecek analiz sonuçlarının (örneğin ORP - Oksidasyon İndirgenme Potansiyeli) bulunmaması olabilir. Ayrıca bu tür veriler mevcut olsa dahi, fosfor giderimi ile aralarında tam bir doğrusal ilişki bulunmamaktadır. SelectKBest yöntemiyle seçilen özellikler incelendiğinde; TP giderimini doğrudan temsil edecek spesifik bir parametrenin seçilemediği görülmektedir.Buna rağmen, modele sadece R^2 metriği üzerinden yaklaşmamak gerekir. MAE metriğindeki 0,231 mg/L'lik hata

değeri; 1 mg/L olan deşarj limiti göz önüne alındığında, modelin tamamen kullanışsız olmadığını ortaya koymaktadır. Model, fosforun anlık pik salınımlarını yakalamakta zorlansa da 0,23 mg/L'lik bir hassasiyetle operatöre bilgi sunabileceği görülmektedir.

5.4.1.3 Senaryo-1'in Genel Değerlendirilmesi

Senaryo-1 ile elde edilen bulgular; istatistiksel tabanlı bir yöntem olan SelectKBest özellik seçiminin, özellikle AKM ve KOİ gibi fiziksel ve stokiyometrik ilişkilerin baskın olduğu parametrelerde, MÖ algoritmalarının yüksek doğruluklu tahmin sonuçlarına ulaşabilmesini sağladığını göstermiştir.

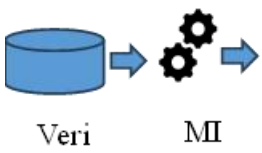
Senaryo-1'in güçlü yönü; AKM (%97) ve KOİ (%83) parametrelerinde sergilediği yüksek tahmin performansı ile işletme kontrolü için güçlü bir zemin oluşturmasıdır.

Senaryo-1'in zayıf yönü; biyokimyasal kinetiğin karmaşıklığı nedeniyle BOİ, TN ve TP parametrelerinde sadece doğrusal ilişkilere odaklanılmasının, kurulan MÖ algoritmalarının nispeten düşük bir başarı sergilemesine yol açmasıdır. Ancak MAE ve RMSE hata değerleri yönünden incelendiğinde; sonuçların yine de operatör uyarıcı olarak kullanılabilecek seviyede olduğu görülmektedir.

Neticede; Senaryo-1, tesisin genel kirlilik yükünü ve genel çerçevede neler olabileceğini izlemek için hızlı ve hesaplama maliyeti düşük bir yol göstermiş olsa da, doğrusal olmayan yöntemleri de çözümlemek için geliştirilmiş yöntemlere ihtiyaç vardır.

5.4.2 Senaryo-2 Analizi MI Özellik Seçimi Sonuçları

Bu senaryoda, Karşılıklı Bilgi (MI) yöntemi, lineer olmayan karmaşık ilişkileri de hesaba katarak özellik seçimi yaptığı için tercih edilmiş ve Şekil 5.9'da sunulan 10 adet özelliği belirlemiştir.



Seçilen Özellikler	
Çıkış UAKM	MLSS
Çıkış Çöz. KOİ	Giriş Çöz.KOİ
Çıkış Nitrat Azotu	Hava Sıcaklığı
Ç. O	Giriş Amonyak-N
MLVSS	Aktif Hav.Hvz.Sayısı

Şekil 5.9 Senaryo-2'de MI ile seçilmiş özellikler

AAT proseslerindeki mikrobiyal büyüme ve substrat tüketimi gibi doğrusal olmayan süreçleri daha iyi temsil etmek amacıyla MI (Karşılıklı Bilgi) yöntemi tercih edilmiş; bu yöntemle seçilen özelliklerin MÖ algoritmaları ile modellenmesi ve tahmin başarıları aşağıda alt başlıklar halinde sunulmuştur.

5.4.2.1 Makine Öğrenme Algoritmalarının Performans Karşıaltırmaları

Burada MI ile veri setinden seçilen ve özellikler ile kurulan 5 adet farklı algoritmanın (RF, GBM, XGBoost, LightGBM ve DT), tüm değerleme metrikleri (MAE, RMSE, R^2) ile kıyaslandığı Tablo 5.6 aşağıda irdelenmiştir.

Tablo 5.6 incelendiğinde, özellik seçimi yönteminin değişmesi MÖ algoritmalarının hiyerarşik yapısını korurken, parametre bazında modellerin davranışını değiştirdiği görülmektedir. Şöyle ki:

- Senaryo-1'de olduğu gibi burada da GBM ve XGBoost ile yapılan modellemeler en etkili yöntemler olmuştur. Bu durum özellik seçiminden bağımsız olarak topluluk öğrenme modellerinin AAT erisindeki örüntüleri kavramada başarılı olduğunu göstermektedir..
- MI ile yapılan özellik seçimi, biyokimyasal bir parametre olan BOİ'nin tahmin başarısını yükseltirken, bunun aksine stokiyometrik parametre olan TN'daki tahmin başarısını düşürmüştür. Buradan şu sonuç çıkartılabilir: AAT karakterizasyonu tek bir matematiksel yaklaşımla (lineer ya da non-lineer) açıklanamayacak bir düzeydedir.

Tablo 5.6 Senaryo-I'e göre MÖ model ve değerleme sonuçları

Özellik Seçim Metriği: MI				
Hedef Değişken	Model	MAE	RMSE	R^2 (%)
KOİ (Çıkış)	Random Forest	6.248535	12.504	78.24%
KOİ (Çıkış)	Gradient Boosting	6.275448	12.882	76.91%
KOİ (Çıkış)	XGBoost	6.474302	12.222	79.21%
KOİ (Çıkış)	LightGBM	6.838371	12.989	76.52%
KOİ (Çıkış)	Decision Tree	10.41163	21.083	38.14%
BOİ (Çıkış)	Random Forest	1.610291	2.389	75.86%
BOİ (Çıkış)	Gradient Boosting	1.663382	2.505	73.45%
BOİ (Çıkış)	XGBoost	1.589808	2.194	79.64%
BOİ (Çıkış)	LightGBM	1.771592	2.713	68.86%
BOİ (Çıkış)	Decision Tree	2.377209	3.750	40.49%
AKM (Çıkış)	Random Forest	3.972572	7.674	96.92%
AKM (Çıkış)	Gradient Boosting	3.667118	6.873	97.53%
AKM (Çıkış)	XGBoost	4.624556	11.683	92.85%
AKM (Çıkış)	LightGBM	6.148256	17.523	83.92%
AKM (Çıkış)	Decision Tree	4.970698	9.506	95.27%
TN (Çıkış)	Random Forest	0.938949	1.596	18.13%
TN (Çıkış)	Gradient Boosting	0.915302	1.598	17.99%
TN (Çıkış)	XGBoost	0.925648	1.615	16.22%
TN (Çıkış)	LightGBM	0.902001	1.563	21.57%
TN (Çıkış)	Decision Tree	1.230209	2.394	6.19%
TP (Çıkış)	Random Forest	0.236964	0.526	-84.14%
TP (Çıkış)	Gradient Boosting	0.249694	0.589	-17.71%
TP (Çıkış)	XGBoost	0.235134	0.502	-1.07%
TP (Çıkış)	LightGBM	0.244911	0.489	14.58%
TP (Çıkış)	Decision Tree	0.334651	0.735	18.76%

5.4.2.2 Parametre Bazlı Model Analizi

Her bir hedef değişken parametresi için en düşük hatayı ve en yüksek başarı oranını sağlayan "En İyi Modeller" aşağıdaki Tablo 5.7'de verilmiştir.

Tablo 5.7 Senaryo-2'ye ait en iyiler tablosu

Hedef Değişken	Seçilen En İyi Model	MAE	RMSE	R^2 (%)
KOİ	XGBoost	6.474302	12.222	79.21%
BOİ	XGBoost	1.589808	2.194	79.64%
AKM	Gradient Boosting	3.667118	6.873	97.53%
TN	LightGBM	0.902001	1.563	21.57%
TP	Decision Tree	0.334651	0.735	18.76%

Tablo 5.7'deki makine öğrenmesi algoritmalarının tahmin başarıları, parametre bazlı olarak kirletici türlerine göre aşağıda ayrıntılı olarak tartışılmıştır.

1) Fiziksel Arıtma Süreci-AKM: GBM ile yine çok yüksek oran ile (R^2: %97,53; RMSE: 6,873) tahmin edilmiştir. Senaryo-1'e göre kısmi bir artış göstermiştir. AKM giderimi Stokes Yasası'na dayanarak fiziksel bir özellik gösterse de çamur partiküllerinin flokülasyonu ve biyopolimer oluşumu sebebiyle çamur kabarması veya battaniyesi gibi biyolojik süreçler doğrusal olmayan ilişkilerdir. MI yöntemi, bu gizli ilişkileri de model çerçevesinde öğrenilmesine imkân tanımış, az da olsa tahmin sonuçlarının yükselmesini sağlamıştır. Her iki senaryoda da GBM en iyi performansı göstermeye devam etmektedir.

2) Karbon Bazlı Organik Besin Giderimi (KOİ ve BOİ): KOİ tahmin doğruluğu az da olsa düşmüş, hata payı artmıştır (R^2: %79,21; RMSE: 12,222). Sebebi MI ile seçilmiş olan KOİ özelliklerinin çözünmüş KOİ'ye odaklanmış, inert kısmı ile ilgilenmemiş olması olabilir. Ancak BOİ'nin yükselmiş olması (%74'ten yaklaşık %80'e) bu MI ile özellik seçiminin başarısı olarak kabul edilebilir.

   BOİ, ortamdaki mikroorganizmaların biyolojik aktivitesini temsil eder. Biyokütlenin büyümesi Monod Kinetiği ile eğrisel bir ilişki ile ifade edilirken, içsel solunumda doğrusal değildir. MI özellik seçimi ile BOİ'deki bu %6'lık

artışın sebebi, doğrusal olmayan ilişkileri yakalamasına bağlıdır. MI ile indirgenen veri setinde uygulanan XGBoost algoritması, literatürdeki kayda değer başarı değerlerinin (R^2: %70-75) üzerinde (~%80) gerçekleşmiştir [9,46].

3) Nutrient Giderimi-TN ve TP: Azot ve fosfor giderimi, karbon giderimine göre daha karmaşık ilişkilere sahipken, MI ile özellik seçimi daha yüksek tahmin doğruluğu beklenirken, Senaryo-1'in gerisinde kalmıştır.

   TN parametresi Senaryo-2'de LightGBM algoritması ile çok düşük bir doğruluk oranı (R^2: %21,57) ile çıkış tahmin değerine ulaşmışken, Senaryo-1'deki (R^2: %64) çok gerisinde kalmıştır. Ancak çok da yüksek olmayan bir hata payı vardır (MAE: 0,92 mg/L). Deşarj limiti yönetmeliklerine göre TN : 10 mg/L'dir. Modelin yaptığı 0,902 mg/L'lik hata, yasal limitlerin altında yaklaşık %10'luk bir derecededir. Çalışmanın en başarılı değeri olan AKM ile kıyaslandığında (yasal limit: 35 mg/L; MAE: 3,66 mg/L) hata oranı yaklaşık %10,5 seviyesindedir. Dolayısıyla TN için bulunan R^2 düşük olsa da, limit değerler göz önüne alındığında göreceli hata bakımından en az AKM kadar başarılı ve hassastır. Düşük R^2 değeri, modelin başarısız tahmin etmesinden değil, tesis çıkışı TN değerinin çok dar bir aralıkta seyretmesinden (düşük varyansa sahip olmasından) kaynaklanmaktadır.

   TP beklendiği üzere Senaryo-1'den yüksek bir başarı oranına (R^2: %18,76) ve daha az hata payına (MAE: 0,245 mg/L) sahip olmuştur. TP'nin deşarj değeri 1 mg/L'dir. Havuz bazlı ORP verisinin eksik olması, TP giderimini tanımlayıcı veri eksikliğine sebep olmaktadır. Buna rağmen hata miktarının (MAE) düşük olması, deşarj TP değerinin tahminlemesi yine de yasal limitin aşılıp aşılamayacağı konusunda bir fikir verecektir.

5.4.2.3 Senaryo-2'in Genel Değerlendirilmesi

Senaryo-2 bulguları değerlendirilirken Senaryo-1'in bulguları ile kıyaslı düşünüldüğünde, AAT tesislerinin modellenmesinde tek bir doğru özellik seçimi olmadığı görülmektedir.

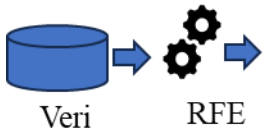
AKM parametresi diğer parametrelere göre doğrusal ilişkilere daha bağımlı olmasına rağmen MI yöntemi ile yapılan özellik seçiminde tahmin doğruluk oranını arttırmıştır. Deşarj standardına göre göreli hata oranı %10 civarında seyretmektedir.

BOİ doğrusal olmayan ilişkileri daha iyi tahmin ederek hem hata payında düşüş sağlamış hem de başarı oranı %6 gibi bir değerle yükselmiştir. Bu, tesisin işletme esnasında karşılaşılan şok yüklenmeleri daha iyi öğrendiğini göstermektedir.

TN ve TP yönünden incelendiğinde, her iki parametrenin tahmin başarısı MI ile özellik seçimi yapılmasına rağmen beklenenin aksine düşüş göstermiştir. Bu durum, veri setinde bu parametrelerin karmaşık yapısını temsil edecek yeterli parametrelerin bulunmaması ile açıklanabilir. MI seçimi ile ortaya konulan MÖ algoritmaları incelendiğinde, R^2 değeri düşük olmasına rağmen (R^2: %21 ve %14), MAE hata büyüklüğü açısından incelendiğinde (MAE: 0,902 ve 0,334), aslında her iki parametrede de deşarj limit değerlerinin %10'un altında kaldığı görülmektedir. Modeller, her iki parametre için de veri setindeki düşük varyans nedeniyle anlık trend takibinin göstergesi olan R^2 metrik değeri düşük olsa da, çıkış suyunun yasal sınırlara uygunluğunu denetlemede hâlâ kullanılabilir tahmin sunabilmektedir.

5.4.3 Senaryo-3 Analizi: RFE Özellik Seçimi Sonuçları

Tez çalışmasının bu son senaryosunda özellik seçimi olarak kullanılan, sarmalayıcı yöntem olarak nitelendirilen RFE (Özyinelemeli Yaklaşım)'ın seçmiş olduğu 10 özellik aşağıdaki Şekil 5.9'da verilmiştir. Burada bulgular değerlendirilmeden önce, bu yöntemin diğer iki seçim yaklaşımından farklı olarak en çok temsil edeni seçmek değil, en az temsil eden özellikleri eleyerek son kalan 10 özelliğe göre seçimi belirlediği hatırlanmalıdır.



Seçilen Özellikler	
Çıkış Nitrat Azotu	Akf. Son Çök.Hvz.Sayı
Demir Kullanımı	Aktif Dif. Sayısı
Metanol Kullanımı	İçsel Geri Devir
Polimer Kons.	Ç.O
Aktif Hav.Hvz.Sayısı	Hava Durumu

Şekil 5.10 Senaryo-3'de RFE ile seçilmiş özellikler

5.4.3.1 Makine Öğrenme Algoritmalarının Performans Karşılaştırmaları

Bu bölümde RFE kullanılarak indirgenmiş veri setleri üzerinde beş farklı algoritmanın (RF, GBM, XGBoost, LightGBM ve DT) ve tüm değerleme metrikleri (MAE, RMSE, R^2) kıyaslanarak anlatılacaktır. Bu bağlamda hazırlanmış olan Tablo

5.8 incelendiğinde, önceki özellik seçim yöntemlerinin aksine RFE uygulanan tüm MÖ algoritmalarının tüm parametrelerin tahminindeki başarıları dramatik bir düşüş göstermiştir.

Hiçbir MÖ algoritması R^2 başarı oranını %11'in üzerine çıkaramadığı gibi, özellikle bir önceki senaryolarda çok büyük başarı gösteren AKM tahminlerinde modelin hata metriği olan RMSE değeri tüm modellerde parametrenin deşarj limit standardı olan 35 mg/L'nin de üstünde yaklaşık 42 mg/L değerlerini vermiştir..

Ayrıca TN tahmininde tüm algoritmalar negatif değerler vermiştir. Bu durum MAE ve RMSE'nin yorumlanmasının da önüne geçmiştir. Negatif doğruluk oranı, hiç tahmin etmeden doğrudan parametrenin ortalamasını tahmin etmek yerine değerlendirilse bile daha doğru değer vereceği anlamına gelmektedir.

Tablo 5.8 Senaryo-3'e göre MÖ model ve değerleme sonuçları

Özellik Seçim Metriği: RFE				
Hedef Değişken	Model	MAE	RMSE	R^2 (%)
KOİ (Çıkış)	Random Forest	14.98364	25.460	9.79%
KOİ (Çıkış)	Gradient Boosting	14.95776	25.471	9.71%
KOİ (Çıkış)	XGBoost	14.95892	25.470	9.73%
KOİ (Çıkış)	LightGBM	14.97676	25.485	9.62%
KOİ (Çıkış)	Decision Tree	14.95892	25.470	9.73%
BOİ (Çıkış)	Random Forest	2.795836	4.597	10.59%
BOİ (Çıkış)	Gradient Boosting	2.798389	4.599	10.52%
BOİ (Çıkış)	XGBoost	2.799759	4.599	10.53%
BOİ (Çıkış)	LightGBM	2.791269	4.599	10.52%
BOİ (Çıkış)	Decision Tree	2.799772	4.599	10.53%
AKM (Çıkış)	Random Forest	20.84885	41.886	8.13%
AKM (Çıkış)	Gradient Boosting	20.82707	41.888	8.12%
AKM (Çıkış)	XGBoost	20.82904	41.888	8.12%
AKM (Çıkış)	LightGBM	20.73614	41.918	7.99%
AKM (Çıkış)	Decision Tree	20.82904	41.888	8.12%
TN (Çıkış)	Random Forest	1.069232	1.773	-1.01%
TN (Çıkış)	Gradient Boosting	1.073397	1.776	-1.30%
TN (Çıkış)	XGBoost	1.073701	1.776	-1.32%
TN (Çıkış)	LightGBM	1.077928	1.780	-1.75%
TN (Çıkış)	Decision Tree	1.073708	1.776	-1.32%
TP (Çıkış)	Random Forest	0.265439	0.522	7.47%
TP (Çıkış)	Gradient Boosting	0.26712	0.523	7.21%
TP (Çıkış)	XGBoost	0.267285	0.523	7.17%
TP (Çıkış)	LightGBM	0.262978	0.518	9.15%
TP (Çıkış)	Decision Tree	0.267298	0.523	7.17%

5.4.3.2 Parametre Bazlı Model Analizi

Senaryo-3'teki başarısızlık sadece başarı oranı olan R^2 değerindeki düşüklük değil, deşarj limitlerine göre artan hata oranları ile (MAE/Limit) da kendini

göstermektedir. Başarısızlığın nedenleri parametre bazında Tablo 5.9'a göre irdelenmiştir.

Tablo 5.9 Senaryo-3'ye ait en iyiler tablosu

Hedef Değişken	Seçilen En İyi Model	MAE	RMSE	R^2 (%)
KOİ	Random Forest	14.95776 (GB)	25.470 (RF, XGB, DT)	9.79%
BOİ	Random Forest	2.791269 (LGBM)	4.597 (RF)	10.59%
AKM	Random Forest	20.73614 (LGBM)	41.886 (RF)	8.13%
TN	Random Forest	1.069232 (RF)	1.773 (RF)	-1.01%
TP	LightGBM	0.262978 (LGBM)	0.518 (LGBM)	9.15%

1) Fiziksel Arıtma Süreci-AKM: Önceki senaryolarda R^2: %97 olan parametre tahmin doğruluk değerleri bu senaryoda %8,13 olarak gerçekleşmiştir. RFE bir eleme mantığı ile çalıştığı için iterasyon sürecinin başlarında AKM ile yüksek korelasyonu olan gereksiz olarak değerlendirdiği özellikleri elemiştir. Elenen değişkenlerin yerine kalan değişkenler çok az ama ani değişiklik göstererek sistemi büyük etkileyen (arıza, kimyasal dozlama gibi) özellikler olmuştur. Ortaya çıkan tahmin başarısı kullanılabilir bir bilgi desteği değildir.

2) Karbon Bazlı Organik Besin Giderimi (KOİ ve BOİ): Tüm modellerde R^2 metriğinin değeri %10 civarında kalmıştır. MAE ve RMSE hata değerleri de herhangi bir değerlemeye uygun boyutta değildir.

3) Nutrient Giderimi-TN ve TP: Azotun metrikleri sırasıyla R^2: -%1,01 ve MAE: 1,07 mg/L olarak gerçekleşmiştir. Başarı oranının negatif çıkması, hiç tahmin yapılmamasının, sadece ortalama ile işlem yapılmasının daha doğru olacağının göstergesidir. Model TN giderim mekanizmasını hiç öğrenememiştir. RFE yönteminin seçtiği özelliklere bakıldığında (Çıkış TN ve Ç:O), doğrudan azot giderimini ilgilendiren bir parametre seçilmemiş olmasına da bağlanabilir.

   TP'nin değerleme metriklerinin en iyileri R^2: %9,15 ve MAE: 0,26 mg/L'dir. Veri setindeki modelin hedef değerlerinin varyansına bakıldığında düşük olduğu

görülmektedir. RFE yöntemi ile özellik seçimi yaparak indirgenmiş MÖ algoritmalarının, modeli öğrenmediği, sadece bu düşük varyanstan faydalandığı görülmektedir.

5.4.3.3 Senaryo-3 Genel Değerlendirilmesi

Senaryo-3'ün değerlendirilmesinden elde edilen tanı, bu tez çalışmasının elde edilen en önemli "negatif bulgu"sudur. Bu, AAT verilerine RFE özellik seçiminin uygulanmaması gerektiğinin iki ana sonucunu ortaya koymuştur.

1- RFE yöntemi, hata oranlarını ve hata değerlerini diğer yöntemlere kıyasla dramatik ölçüde olumsuz ortaya koymuştur. Hatta bazı parametrelerde tahmin hatası, o parametrenin yasal sınırlarının dahi üstündedir.

2- RFE'nin AAT veri setine uygulanması bir yapısal uyumsuzluktur. Şöyle ki AAT verileri birbirleriyle yüksek derecede bağımlıdır. Küçük değişkenlikleri dahi birbirleri üzerinde yansıtır. RFE'nin agresif yapısı küçük değişimleri tespit edip gereksiz görerek asıl öğrenmesi gereken korelasyonları elemiş, nadir gerçekleşen kimyasal dozlama gibi pik karakter değişimlerine sebep olan özellikleri sona bırakmıştır. Bu da hatalı öğrenmeye sebep olmuştur.

Neticede AAT'lerde sarmalayıcı özellik seçim yöntemleri yerine SelectKBest ve MI gibi filtre yöntemlerinin kullanılması daha doğru bir yaklaşım olacaktır.

5.5 Senaryoların Çapraz Karılaştırılmalı Analizi

Bu bölümde uygulanan üç ayrı "Özellik Seçim" yöntemi olan SelectKBest, MI ve RFE ile seçilen özelliklere ait veri setleri üzerinden uygulanan MÖ model algoritmalarının hata ve başarı durumları, hedef değişkenler (KOİ, BOİ, AKM, TN ve TP) özelinde karşılaştırılmıştır. Aşağıdaki Tablo 5.10'da her bir hedef değişken için en iyi MÖ modeli baz alınarak tablo hazırlanmıştır.

Tablo 5.10 Hedef değişkenler ve özellik seçimleri karşılaştırılması

Hedef Değişken	SelectKBest (En İyi R^2)	MI (En İyi R^2)	RFE (En İyi R^2)	Yorum
KOİ	%83,41 (XGBoost)	%79,21 (XGBoost)	%9,79 (RF	SelectKBest: KOİ'nin doğrusal bileşenlerini iyi yakalamıştır. RFE başarı gösterememiştir.
BOİ	%74,29 (GBM)	%79,64 (XGBoost)	%10,59 (RF)	MI: BOİ'nin doğrusal olmayan yapısını öğrenebilmiştir.
AKM	%97,04 (GBM)	%97,53 (GBM)	%8,13 (RF)	MI/SelectKBest: Her iki yöntemde iyi sonuç vermiştir. RFE ise burada hiç tahmin yapamamıştır.
TN	%64,15 (RF)	%21,57 (LGBM)	%-1,01 (RF)	SelectKBest: TN parametresindeki doğrusal ilişkiler MÖ tarafından öğrenebilse de karmaşık ilişkiler için daha fazla veriye ihtiyaç vardır.
TP	%28,26 (LGBM)	%18,76 (DT)	%9,15 (LGBM)	SelectKBest: Tüm Yöntemler düşük performans gösterse de, bu yöntem göreceli iyidir.

Literatürde, sarmalayıcı modeller (RFE gibi), kurulan modelle daha etkileşimli çalıştıkları için filtre yöntemlerine (SelectKBest ve MI gibi) göre daha yüksek başarı puanı gösterdiği ifade edilmiştir [219]. Ancak bu tez çalışmasının bulguları, AAT verisi için bu durumun tam tersini ortaya koymuştur.

Filtre yöntemlerinden olan SelectKBest, özellikle KOİ ve TN gibi doğrusal ilişkileri bulunan parametrelere ait deşarj tahminlerini en yüksek doğrulukla modele yansıtabilmiştir. Yine bir diğer filtre yöntemi olan MI özellik seçimi ile indirgenen veri setinde uygulanan MÖ algoritmaları, BOİ ve KOİ için en iyi tahmin başarısını göstermiştir.

Ancak yukarıdaki tabloda da görüleceği üzere RFE sarmalayıcı yöntemi, 5 hedef parametrenin tamamında da kayda değer bir model çıktısı üretememiştir.

Her özellik seçimine ait eğitim veri setleri ile seçilen 5 adet MÖ algoritması eğitilmiştir. Makine öğrenmesi algoritmalarının parametre bazlı başarılarının kıyaslanması Tablo 5.11'de verilmiştir.

Tablo dikkatle incelendiğinde, başarılı yöntemler arasında boosting algoritmaları tartışmasız üstün olmuştur. Tablo incelendiğinde en önemli çıkarım şudur: özellik seçimi, modelin başarısını en çok etkileyen unsurdur.

Tablo 5.11 Hedef değişkenler ve MÖ algoritmalarının kıyaslanması

Hedef Değişken	MÖ Modeli (En İyi R^2)	MÖ Modeli (En Kötü R^2)	Yorum ve Değerlendirme
KOİ	XGBoost (SKT %83,41)	RF (RFE %9,79)	XGBoost, regülasyon özellikleri ile doğrusal ilişkileri yakalayarak en iyi sonuçları vermiştir. RFE stabil bir veri seti oluşturduğundan dolayı ağaç yapısı düzgün oluşmamıştır.
BOİ	XGBoost (MI %79,64)	RF (RFE %1,59)	XGBoost, doğrusal olmayan ilişkileri destekleyen MI ile regülasyonda başarılı olmuştur. RFE stabil bir veri seti oluşturduğundan dolayı ağaç yapısı düzgün oluşmamıştır.
AKM	GBM (MI %97,04)	RF (RFE %8,13)	GBM, hem doğrusal hem de doğrusal olmayan ilişkileri en iyi öğrenmiştir. RFE stabil bir veri seti oluşturduğundan dolayı ağaç yapısı düzgün oluşmamıştır.
TN	RF (SKT%64,15)	LGBM (RFE %-1,75)	Aslında TN'nin çıkış değerleri incelendiğinde çok dar bir varyansta bulunduğu görülmektedir. Bu sebeple model başarı oranı ile değil deşarj limitlerine göre hata payı ile değerlendirilmiştir.
TP	LGBM (SKT %28,26)	RF (SKT %-84,14)	Aslında TP'nin çıkış değerleri incelendiğinde çok dar bir varyansta bulunduğu görülmektedir. Bu sebeple model başarı oranı ile değil, deşarj limitlerine göre hata payı ile değerlendirilmiştir.

6
TARTIŞMA VE DEĞERLENDİRME

6.1 Giriş

Bu çalışmada, AAT'lerin içerdiği süreçlerin geleneksel deterministik yaklaşımlarla modellenmesine (ASM'ler), dijital simülasyon araçlarının karar destek mekanizması olarak kullanılmasına; IWA'nın "Dijital Su" vizyonu çerçevesinde Endüstri 4.0 bileşenlerinden biri olan YZ'nin alt dalı MÖ algoritmaları ile geliştirilen modelleme yöntemlerinin bir alternatif olarak değerlendirilebilirliğinin araştırılması amaçlanmıştır. Özellikle Tesisi-A gibi doğrusal olmayan dinamiklere sahip, havzasındaki değişkenliğin ve endüstriyel atık su girişimlerinin yoğun olduğu tesislerde, tesisin operasyonel becerisini kısıtlayan pik yüklemelerin MÖ ile modellenmesinde sadece algoritma başarısına odaklanılmamış, verinin temsil kalitesini arttırmak amacıyla "Özellik Seçimi" yöntemleri de uygulanmıştır.

Bu bölümde, ortaya koyduğumuz veri odaklı yaklaşımın performansı, elde edilen bulguların literatür karşısındaki anlamı ve metodolojimizin güçlü ve zayıf yönleri değerlendirilecektir. Ayrıca alana katkısı ve gelecek perspektifi de tartışılacaktır.

6.2 Özellik Seçimi Senaryolarının Değerlendirilmesi

Bu çalışmada literatüre temel katkı olarak, AAT verilerine dayalı yapay zeka uygulamalarından önce "Özellik Seçimi" yöntemlerinin uygulanmasıdır. Burada literatüre katkı sağlayan en önemli bulgu, farklı özellik seçimi yöntemlerinin AAT verisi üzerindeki farklı performanslarla tahmin başarısı göstermesidir. Literatürdeki yaygın kanı, hesaplama maliyeti daha yüksek olan sarmalayıcı yöntemlerin daha başarılı olduğu yönündeyken, bu çalışmada AAT verileri için filtre yöntemleri çok daha başarılı olmuştur.

SelectKBest yöntemi, veri setinde istatistiksel üstünlüğüne dayanarak etkili ve AAT karakteristiğine uygun özellikleri filtreleyebilmiştir. Özellikle AKM ve KOİ parametrelerinde gösterdiği başarı, hem fiziksel hem de stokiyometrik ilişkileri ayırabildiğini kanıtlamıştır. SelectKBest, TN tahmininde başarılı diyebilecek bir performans sergilemese de çalışmada kullanılan diğer iki yönteme göre göreceli üstünlük sağlamıştır. SelectKBest'in AAT'nin birincil parametreleri sayılabilen giriş debi, giriş konsantrasyonu, çökelme ile giderim mekanizmalarının (çökelme ve adsorpsiyon) arasındaki doğrusal ilişkileri tespit etmesinden başarılı olduğu söylenebilir. Ancak TN ve TP gibi daha karmaşık ilişkileri temsil edebilecek özellikleri seçememiştir. Bu bulgu, literatürde filtreleme yöntemlerinin karmaşık ilişkileri bulmada yetersiz olduğu savını doğrular niteliktedir. Özetleyecek olursak, SelectKBest, AKM için yaptığı özellik seçiminde modellerin göstermiş olduğu başarı, parametrenin doğrusal ilişkiler dolayısıyla çok yüksek olurken, TP hariç diğer parametrelerin (KOİ, AKM ve TN) tahminlemesinde parametrelerin doğrusal ilişkilerini yakalarken doğrusal olmayan ilişkilerini tam kavrayamamıştır. Ancak yine de göreceli olarak başarılıdır.

Karşılıklı Bilgi (MI), ondan bekleneni sergileyerek hem doğrusal hem de doğrusal olmayan özellikleri yakalayabilmiştir. Bu yöntem çok belirgin olmasa da AKM tahmininde SelectKBest'ten daha yüksek doğruluk değerleri elde etmiştir. Bunun sebebi, AKM giderimindeki doğrusal olmayan ilişkiyi de başarılı bir şekilde seçtiği özellikler kümesine yansıtabilmesidir. MI yöntemi; yüzey yükü, çamur çökelme mekanizması (SVI) ve hidrolik bekleme gibi parametreler arasındaki karmaşık ve doğrusal olmayan ilişkileri MÖ modellerine yansıtabilmiştir. Biyolojik parametre olan BOİ tahmininde çalışma kapsamındaki en yüksek başarıyı sağlamıştır ki bu beklenen bir sonuçtur. MI'nın bu başarılı yaklaşımı, biyolojik süreçteki doğrusal olmayan ilişkileri dikkate alabilme becerisidir. Ancak bu yöntem TN ve TP'de düşük başarı oranları sergilemiştir. Bunun sebebi bu iki parametreyi kontrol eden mekanizmanın daha karmaşık olması ve bu mekanizmaları temsil edecek bazı verilerin veri setinde bulunmaması olabilir (örneğin tank bazlı ORP seviyeleri). Ayrıca bu iki parametre, çıkış değerleri incelendiğinde varyansının çok düşük olduğu tespit edilmiştir. Bu çalışma çıkış değerlerinin limit değerlerinin altında kalmasında odaklanmıştır. Tesis A'nın operasyonel güvenlik faktörü de büyük olduğundan tesis çok büyük oranla deşarj limitlerini sağlamıştır. Bu da neredeyse

sabit bir değeri etkileyen özelliklerin tespitine odaklanmak gibi bir durum oluşturduğundan genelde hem özellik seçimlerinin hem de MÖ algoritmalarının tahmin doğrulukları (R^2) düşük gerçekleşmiştir.

Senaryo-3'te incelenen Özyinelemeli Özellik Seçimi (RFE), güncel literatürde popüler bir yöntem olmasına rağmen bizim çalışmamızda kendisinden bekleneni verememiş, neredeyse tüm hedef değişkenlerin tahmininde uygulanan modelleri çökertmiştir. Sarmalayıcı yöntemlerin her veri setinde üstün olamayabileceğinin bir kanıtı olarak bu çalışmanın literatür çıktısı değerlendirilebilir. Bu neticenin şu sebeplerle ortaya çıktığı düşünülebilir:

- AAT'lerdeki önemli parametrelerin birbirleriyle yüksek korelasyonla ilişki içinde olması, sensör kaynaklı pik yüklerin zamansal belirsizliği ve mevsimsel döngülerin tam olarak öğrenilebilmesi için daha uzun dönemli verilerin bulunması gerekliliği gibi sebeplerle veri kalitesinin bu yönteme uygun olmayışı başarısızlığı getirmiştir. Yani AAT verisinin yüksek bağımlılığı ve gürültüsü, RFE'nin agresif eleme stratejisi ile karşılaşınca sonuç başarısızlığa doğru evrilmiştir. Bu yöntemle seçilen özellikler incelendiğinde, sistemde çok seyrek olarak beliren ama büyük değişiklik yapan özelliklerin seçildiği gözlemlenmiştir (kimyasal dozlama, arıza sebebiyle devre dışı kalan ve azalan aktif havalandırma havuz sayısı gibi). Bu da modelin genel davranışı öğrenmek yerine aykırı durumları öğrenmesine yol açtığından test verilerinde mevcut olabilen aykırılık sayısı kadar başarı sergilemiştir. Bu tez kapsamında ulaşılan bu sonuca istinaden, eğer bu yöntem kullanılacaksa veri setinin çok daha dar ancak derin olması; tesisi temsil eden tüm parametrelerden ziyade doğrudan incelenmesi düşünülen spesifik parametrelere odaklanarak tek bir hedef değişken üzerinden çalışılması durumlarında tercih edilmelidir.
- RFE, özellik seçimi işini yaparken temel aldığı modelin performansına dayalı olarak gerçekleştiren bir yöntemdir. Bu çalışmada RFE'nin Rastgele Orman modeli temel alınarak uygulanmıştır. Çalışmanın geneline baktığımızda özellik seçiminden bağımsız olarak MÖ modeli olarak uygulanan RF algoritmasında başarı düşük olmuştur. Bu da Rastgele Orman metodolojisine uymayan bir veri seti içerisinden özellik seçimi yaptığımızı göstermektedir. Muhtemelen özellik seçimi olarak RF destekli RFE kullandığımız için RFE, veri setinde seçilim yaparken yerel optimumlarda sıkışmıştır. Başka bir şekilde ifade etmek

gerekirse, RFE yöntemi tarafından belirlenen özellik kümesi veri setinin tüm yapısal çeşitliliğini temsil edemeyip, rastgele ormanın yaklaşımı ile oluşturulmuş olan konfigürasyona aşırı uyum sağlamış olabilir.

Genel bir değerlendirme yapmak gerekirse, hiçbir özellik seçimi yöntemi tek başına tüm hedef değişkenler için genel bir başarı sağlayamamıştır. Bu durumda tezden literatüre katkı olarak şu perspektif ortaya çıkmıştır: "Her bir hedef değişken için kendi veri seti ile kendine ait özellik seçimi ve model uygulaması gerçekleştirilmelidir." Bir AAT için işletmesel bir karar destek uygulaması yapılacaksa, deşarj kalite parametreleri için farklı özellik seçimi yöntemleri belirlenmelidir.

6.3 Makine Öğrenmesi Algoritmalarının Tartışılması

Çalışmada MÖ algoritmalarını eğitmek üzere veri kümesi 4 ayrı oranda (%80-20, %70-30, %60-40 ve %50-50) eğitim ve test verisi olarak ayrılmıştır. Tüm modeller bu oranların tamamıyla denenmiştir. Burada hangi oranın daha optimum olduğunu belirlemek ve AAT verisinin herhangi bir MÖ algoritmasında aşırı öğrenme sorunu ile karşılaşıp karşılaşmayacağı öğrenilmeye çalışılmıştır. Tüm algoritmalarda en yüksek tahmin doğruluğu, eğitim verisinin %80, test verisinin de %20 olduğu oranda gerçekleşmiştir. Herhangi bir aşırı öğrenme olmamış, aksine %50-50 eğitim-test oranında eksik öğrenme gerçekleşmiş ve model başarısı dramatik bir düşüş yaşamıştır. Burada da bulduğumuz sonuç literatürle uyumludur.

Çalışmada Karar Ağacı (DT) algoritması, bir referans olması açısından seçilmiştir. DT modeli, tüm özellik seçimleri ile oluşturulmuş indirgenmiş veri setleri ile kurulan tüm algoritmalarda en başarısız model olarak hem literatüre uygun bir davranış sergilemiş hem de bu çalışmada kıyaslama açısından bir referans noktası olmuştur. TN ve TP gibi karmaşık özellik gösteren parametrelerde negatif R^2 değerleri üreterek tamamen başarısız olmuş ve AAT veri setinin tek bir karar ağacı ile modellenemeyecek kadar karmaşık ve gürültülü olduğunu doğrulamıştır.

Topluluk öğrenme modelleri ise başarılı olmuştur. Bu algoritmaların varyans düşürme (bagging – RF) ve yanlılığı azaltma (boosting – GBM, XGBoost ve LightGBM) özellikleri, AAT verilerindeki belirsizliklerle başa çıkma konusunda başarılı olduğunu göstermiştir.

XGBoost ve GBM, temel kirletici parametrelerinde başarılı olmuştur. LightGBM, büyük verilerde daha hızlı ve kolay işlem yapabilen bir algoritmadır. XGBoost ve GBM birbirleriyle göreceli üstünlük sağlarken, LightGBM bu ikisinin biraz gerisinde kalmıştır. Bu durumun veri setimizin küçük-orta ölçekli olmasından ve LightGBM'in hızlı olmaya çalışmasından dolayı öğrenme eksikliği yaşanmasından kaynaklandığı düşünülmektedir. Tartışmasız en başarılı modelleme XGBoost olmuştur. Rastgele Orman (RF) sadece TN'de göreceli olarak daha iyi model başarısı göstermiştir; ancak yine de başarı oranı R^2 = %64'te kalmıştır.

Bu başlık altında parametre bazlı farklılıkların sebeplerine de değinecek olursak;

- AKM ve KOİ'nin yüksek başarı ile modellenmesi, bu parametrelerin giderim mekanizmalarının nispeten daha kolay olması ve seçilen özelliklerde daha yüksek temsile sahip olmalarından kaynaklanmaktadır. Özellikle AKM'nin tahmin doğruluğunun %98 mertebelerine yaklaşması, operasyonel karar destek sistemi olarak kullanılabilecek bir erken uyarı aracı olabileceğini göstermektedir. Sadece AKM korelasyonu bile tesisin olası diğer kalite parametre bozulmalarının da önüne geçecektir.
- BOİ'nin çıkış tahmin doğruluk oranı KOİ tahmininden yaklaşık %3,5 daha azdır. Bu da literatürde bilinen bir durumdur: biyolojik süreçlerin karmaşıklığı KOİ'nin giderimine nispetle daha etkilidir (KOİ gideriminde biyokimyasal etki haricinde adsorpsiyon ve fiziksel tutunma mekanizmaları da işlemektedir).
- Nutrient giderimi (TN, TP) konusunda modeller tamamen başarısız olarak görünmektedir. Bu iki parametrede de literatür incelendiğinde model başarısı genelde R^2 metriği ile değerlendirildiği ve başarılı sonuç almanın zor olduğu belirtilmiştir. Bizim çalışma veri setimizde, özellikle TP'ye etki eden faktörleri yansıtabilecek ORP değerleri tank bazında mevcut değildir. Bu sebeple veriden modelin eğitilmesi eksik kalmıştır. Ayrıca hem TN'nin hem de TP'nin tesis giriş-çıkış değerleri oldukça lineerdir. Bu iki parametrenin deşarj standartları göz önüne alındığında (sırasıyla 10 mg/L ve 1 mg/L) olup, Tesis-A'nın çıkış değerlerinin 1073 günün neredeyse tamamında bu limitleri sağlarken çok az bir standart sapma gösterdiği görülmektedir. Tesis-A inşa edilirken ATV 131-E standardı ile tasarlanmıştır. TN ve TP mekanizmaları da tasarım standardının yapısından dolayı yüksek güvenlik katsayılarına sahiptir ve başarılı bir giderim

prosesi oluşmuştur. Bu sebeple çıkış değerleri birbirine çok yakın gerçekleşmiştir. Bu parametrelerin çıkış değerlerinin çok düşük bir varyansla seyretmesi, istatistiksel bir metriği olan R^2'nin düşük değerlere sahip olmasına sebep olmuştur. Ancak Mutlak Hata Değerleme metriği (MAE) değerlerinin yasal deşarj limitlerinin yaklaşık %10 ve altında gerçekleşmiş olması, bu modellerin pratikte kullanılabileceği anlamını taşımaktadır. Model trendi tam yakalayamasa da hata değerlerini nümerik olarak doğru tahmin edip limit aşımı durumlarını operatöre bildirebilir.

6.4 Metodolojik Yaklaşımın Güçlü Yönleri

Tesis-A'dan elde edilen 3 yıllık ve 53 parametreli kapsamlı ve gerçek veri setinin kullanılması, laboratuvar veya pilot ölçekli çalışma sonuçları değil gerçek dünya koşullarının test edilmesini sağlamıştır.

Literatürde AAT kaynaklı veriler ile MÖ algoritmaları çalışılırken genellikle özellik seçimi yöntemi ihmal edilmiştir. Bu çalışmada özellik seçimi yöntemi üç ayrı metodoloji ile uygulanmış ve çalışmanın merkezine alınmıştır. Tezin en özgün katkılarından biri bu yaklaşım olmuştur.

Beş farklı MÖ algoritmasının ve üç farklı değerleme metriğinin kullanılması ve bunların her ikisinin de tez kapsamında çapraz karşılaştırmaya tabi tutulması, çalışmanın performans değerlendirmesini nesnel ve oldukça kapsamlı kılmıştır.

Tek bir hedef özelliği değil, bir AAT'nin kritik çıkış parametrelerinin tamamının aynı anda ve yukarıda anlatıldığı şekilde her yönüyle modellenmeye çalışılması, bu çalışmayı literatürdeki birçok çalışmadan ayırmaktadır. Ayrıca bu yaklaşım, tesis genelini kapsayacak bir karar destek sistemine altlık yapmanın bir adımı niteliğinde olmuştur.

IWA'nın Dijital Su vizyonundan ilham alarak yapılan bu çalışma, deterministik modellerden veri odaklı, tahmine dayalı karar destek sistemlerine geçiş vizyonunu desteklemektedir. Çalışma, yüksek model karmaşıklığı ve anlamlandırılması zor nümerik işlemlerden uzak; operatör dostu ve kolay yorumlanabilir (özellik önem skorları, hata ve başarı oranı) çıktılar sunabilen, mevcut SCADA sistemlerine kolay entegre edilebilecek bir karar destek sistemi geliştirilmesine uygun bir yaklaşım gerçekleştirmiştir. Önerilen bu yaklaşım, özellikle uzman personel sıkıntısı olan

küçük ve orta ölçekli tesisler için operasyonel kolaylık sağlama potansiyeli taşımaktadır. Bu yönüyle çalışma salt akademik bir yöne odaklanmamış; pratikte uygulanabilirlik yönünden faydayı önceliklendirmiştir.

6.5 Çalışmanın Kısıtları, Zorluklar ve Gelecek Perspektifi

Bir modelin performansı o modele ait veri setinin kalitesine ve genelleme yeteneğine çok büyük ölçüde bağımlıdır. Bu tez, bir tesise (Tesis-A) odaklanmıştır. Farklı iklim, farklı proses konfigürasyonuna ve farklı ölçekteki tesislerin verileri ile de bu yaklaşım denenmelidir. Ayrıca nutrient giderimi için eksik olan verilerin (ORP, mikrobiyal tür verileri, gerçek zamanlı spektroskopik zamansal veriler vb.) sisteme dahil edilmesi gerekmektedir.

Tez çalışmasında veri setindeki gözlemler bağımsız gözlemler olarak ele alınmıştır. Ancak AAT'ler sürekli veri üreten, zaman serisi kapsamında değerlendirilebilecek kaynaklardır. Bu bağlamda gelecek çalışmalarda veri seti zamansal olarak daha uzun süreli incelenerek modellemeye "Zaman Serisi Yöntemleri (LSTM, zaman serisi uyarlı GBM vb.)"nin de eklenmesi denenmelidir. Böylelikle hidrolik bekleme süresi, çamur yaşı, mevsimsel etkiler ve geçmiş olayların bağımlı etkileri de modele dahil edilebilecektir.

Bu çalışma salt kara kutu modeli olmuştur. Bir sonraki hedef, giderek literatürde de yaygınlaşan veri odaklı yaklaşımın tahmin gücü ile deterministik yöntemlerin kesinliğini birleştiren hibrit modeller olmalıdır. Örneğin, deterministik simülasyon modellerinin AAT'ye adaptasyonunu kalibre etme ve bunu sürekli güncelleyerek sistemde canlı tutmaya yönelik MÖ algoritmaları geliştirilebilir.

Tezin odaklandığı çerçeve çıkış suyu kalitesi tahmini olmuştur. Bu odak, tahmin edilen parametreye göre kullanılan enerjinin optimizasyonu gibi operasyonel verimliliğe yönlendirilmelidir. Bunun için pekiştirmeli öğrenme modelleri ile otonom kontrol sistemleri geliştirilmelidir.

Bu çalışmaların ulaşacağı en nihai yer, açıklanabilir yapay zeka modelleri olabilir. Bu modeller tahminlerin perde arkasını da irdeleyerek operatörlere ne olacağını değil, neyin neden öyle olacağını da anlatan yöntemlerdir.

6.6 Tezin Literatüre Katkısı ve Çıkarımlar

Tez çalışması, çevre mühendisliğinin atık su arıtma çerçevesi özelinde veri bilimi ile kesişimine önemli katkı sunmaktadır.

- Metodolojik bir rehber özelliği barındırmaktadır. Tezin başında tarihsel modelleme yaklaşımından başlayarak gelinen noktaya kadarki serüven detaylı bir şekilde, nedensellik kavramları çerçevesinde anlatılmış, MÖ modellemesi yapılması için, veri seçiminden özellik seçimine her bir adım, tercih sebepleri ile verilmiş ve model değerleme metrikleri ile yapılan her adımın çapraz kıyaslamasının yöntemleri ortaya konulmuştur.
- AAT veri setlerinde özellik seçimi yöntemlerinden RFE yaklaşımının uygun olmadığı deneysel olarak kanıtlanmıştır. Ayrıca Filtre Yöntemlerinin neden tercih edilmesi gerektiği noktasında yön gösterici olmuştur. Karşılıklı Bilgi Yöntemi (MI)'nin AAT verilerindeki doğrusal olmayan ilişkileri tespit etmede diğer yöntemlere kıyasla daha başarılı olduğu ispat edilmiştir.
- Özellikle AKM, KOİ ve BOİ parametrelerinde ulaşılan yüksek tahmin doğrulukları ve TN ile TP parametrelerinde kabul edilebilir mutlak hata seviyeleri ile çıkış suyu kalite tahminlerini gösteren bir arayüz programı ile karar destek sistemine entegre edilebilirlik için düşük maliyetli bir araç olabileceği gösterilmiştir.
- Veri Setinin bölünme oranlarının en kararlı hali olan %80 - %20 oranı deneysel yöntemle ortaya konulmuş, bu oranla aşırı öğrenme gerçekleşmediği doğrulanmıştır.
- Topluluk öğrenmesi algoritmalarının tercih edilmesi gerektiği, tekil modellerin AAT verilerindeki parametrelerin birbirleriyle güçlü korelasyon ilişkilerini öğrenemedikleri vurgulanmıştır.
- Çalışma, IWA Dijital Dönüşüm perspektifine sadece akademik bir önermede bulunmamış, ayrıca kullandığı gerçek verilerle uygulanabilirliği de göstermiştir.

7
SONUÇ

Bu tez, atık su arıtma tesislerinin doğrusal olmayan, karmaşık ve belirsizlik içeren dinamiklerinin modellenmesinde, geleneksel deterministik yaklaşımların (ASM'ler) kanıtlanmış başarılarına kıyasla, yapay zekanın veri odaklı bir alt dalı olan makine öğrenmesi modelleme yöntemlerinin ne ölçüde işlevsel bir alternatif oluşturabileceği, kapsamlı bir şekilde incelenmeye çalışılmıştır. Çalışma, IWA'nın 2019 yılında deklare ettiği "Dijital Su" vizyonu çerçevesinde, özellikle Endüstri 4.0 yöntemleri ile veriden makine öğrenimine doğru bir yaklaşımla, AAT işletilmesinde kullanılabilecek bir karar destek sisteminin geliştirilebileceğine dair metodolojik bir perspektif çizmeye odaklanmıştır.

Tesis-A'ya ait yaklaşık 3 yıllık bir zaman serisine sahip, 53 parametre barındıran ve gerçek işletme koşullarını oldukça iyi yansıtan veri seti ile yapılan çalışmalarda, yalnızca tesis çıkış değerlerinin tahmin başarılarına odaklanmamış, verinin temsil gücünü arttırmak için "Özellik Seçimi" yaklaşımlarını merkeze almıştır. Literatürde verinin bu yönü genellikle ihmal edilirken, burada sistematik şekilde, kıyaslamalı olarak ele alınmış, AAT'lerinden elde edilen veri setleri için "modelden önce veri" ilkesi önceliklenmiştir. Çalışmanın en önemli bulgusu olan; "özellik seçimin stratejisinin belirlenmesi, model başarısı için makine öğrenmesi algoritmasının kendisinden daha belirleyicidir" ifadesi yukarıdaki yaklaşımla ortaya çıkmıştır.

Elde edilen bulgular değerlendirildiğinde, AAT hedef değişkenlerinin tahmin modelleme başarısı için, tek bir özellik seçimi yöntemi ve tek bir makine öğrenmesi algoritmasının genel bir başarı sağlayamayacağı net bir şekilde ortaya konmuştur. Literatür genelde, Sarmalayıcı (Wrapper) türdeki özellik seçim yöntemlerini (RFE) hesaplama güçlerinin yüksek olması ve karmaşık yöntemleri daha kolay ortaya

çıkarıcı yönlerinden dolayı üstün bulurken, tezde Filtre tabanlı yöntemlerden olan SelectKBest ve Karşılıklı Bilgi yaklaşımlarının AAT karakteristiğine daha iyi uyum sağlayarak daha yüksek doğrulukta tahminler oluşturduğu vurgulanmıştır. Bu durum, özellik seçiminde genel doğrular üzerinden değil veri setinin yapısal özellikleri dikkate alınarak hareket edilmesi gerektiğini açıkça göstermiştir.

Model algoritmaları üzerinden bir değerlendirme yapıldığında ise Karar Ağaçları gibi tekil algoritmaların, Topluluk Öğrenme modellerine kıyasla bariz geride kaldığı görülmüştür. Topluluk öğrenmelerinde özellikle gradyan artırmalı modellerin AAT verilerinde daha başarılı oldukları görülmüştür. Özellikle XGBoost algoritması hem tahmin doğruluğu hem de kararlılığı bakımından öne çıkmıştır. AKM, KOİ ve BOİ parametrelerinin tahmin edilmesinde görülen yüksek başarı oranları, karar destek sistemlerinde erken uyarıcı mekanizması olarak geliştirilen modellerin kullanılabileceğini göstermektedir. Buna karşılık TN ve TP parametrelerinde düşük başarı oranları (R^2) elde edilmiştir. Ancak bu durumun modelin başarısızlığından ziyade, bu kirletici parametrelerin veri setindeki değerlerinin yapısal özelliklerinden ve tesisin dizayn karakteristiğinden kaynaklandığını öngörmek gerekmektedir. Şöyle ki: bu parametrelerin çıkış değerleri çok stabil hareket etmiş, varyansı çok dar bir aralıkta kalmıştır. Bu sebeple istatistiksel hesaplamaya dayalı R^2 metriği verideki trendi yakalayamamıştır. Çalışmada elde edilen mutlak hata değerleri (MAE) incelendiğinde bu metriğe göre değerlendirmenin daha doğru olacağı düşünülmüştür. Modeller bu iki parametre için trend tahmininde başarılı görünmeseler de işletmesel açıdan önemli olan, deşarj limit aşımı risklerini makul bir hata ile (MAE) öngörebilmişlerdir.

Bu tezin en önemli çıktılarından biri; AAT'ler için geliştirilecek veri odaklı bir çıkış suyu kalite tahmin sisteminin, parametre bazlı, esnek bir mimariye sahip olması gerektiğidir. Yani kurulacak dijital yapının tek tip (one-size-fit-all) olamayacağıdır. Her kirletici parametrenin farklı bir matematiksel karakteristiğinin olması sebebiyle, ayrı ayrı özellik seçimleri, ayrı ayrı model algoritmaları ve belki de ayrı ayrı değerleme metriklerine sahip olmaları gerekmektedir.

Sonuç olarak bu tez, nümerik temellere dayanan deterministik modelleme geleneğini dışlamadan, veri odaklı tahmine dayalı ve işletmesel profesyonelliğe destek olacak sistemlere yönelmenin gerekli ve gelecek olduğunu ortaya koymaktadır. Özellikle uzman personel eksikliği yaşayan küçük-orta büyüklükteki

tesislerin operasyonel becerilerini arttırmada bir fırsat taşıdığı değerlendirilmelidir. Ayrıca bu çalışma, Açıklanabilir Yapay Zeka uygulamalarının gelecek çalışmalarda kullanılması ile işletme profesyonellerinin kara kutu algısını kırarak, nedenselliği mühendislik dili ile açıklayabilen bir karar destek mekanizmasına geçişin yapı taşlarından biri olabilir.

KAYNAKÇA

[1] N. Choudhary, V. K. Yadav, K. K. Yadav, A. I. Almohana, S. F. Almojil, G. Gnanamoorthy, D.-H. Kim, S. Islam, P. Kumar, and B.-H. Jeon, "Application of green synthesized MMT/Ag nanocomposite for removal of methylene blue from aqueous solution," Water, vol. 13, no. 22, p. 3206, 2021.

[2] S. Rajendran, G. K. Inwati, V. K. Yadav, N. Choudhary, M. B. Solanki, M. H. Abdellattif, K. K. Yadav, N. Gupta, S. Islam, and B.-H. Jeon, "Enriched catalytic activity of TiO_2 nanoparticles supported by activated carbon for noxious pollutant elimination," Nanomaterials, vol. 11, no. 11, p. 2808, 2021.

[3] O. Modin, N. Fuad, M. Abadikhah, D. I'ons, E. Ossiansson, D. J. Gustavsson, E. Edefell, C. Suarez, F. Persson, and B.-M. Wilén, "A relationship between phages and organic carbon in wastewater treatment plant effluents," Water Research X, vol. 16, p. 100146, 2022

[4] M. G. Cantwell, D. R. Katz, J. C. Sullivan, D. Shapley, J. Lipscomb, J. Epstein, A. R. Juhl, C. Knudson, and G. D. O'Mullan, "Spatial patterns of pharmaceuticals and wastewater tracers in the Hudson River estuary," Water Research, vol. 137, pp. 335–343, 2018.

[5] S. Long, L. Zhao, H. Liu, J. Li, X. Zhou, Y. Liu, Z. Qiao, Y. Zhao, and Y. Yang, "A Monte Carlo-based integrated model to optimize the cost and pollution reduction in wastewater treatment processes in a typical comprehensive industrial park in China," Science of the Total Environment, vol. 647, pp. 1–10, 2019.

[6] S. Bijekar, H. D. Padariya, V. K. Yadav, A. Gacem, M. A. Hasan, N. S. Awwad, K. K. Yadav, S. Islam, S. Park, and B.-H. Jeon, "The state of the art and emerging trends in the wastewater treatment in developing nations," Water, vol. 14, no. 16, p. 2537, 2022.

[7] R. Batool, S. A. Ahad, Q. Maqsood, S. W. Ali, S. M. Abbas, et al., "Redefining sustainability: Next-gen wastewater treatment breakthroughs," Cleaner Water, vol. 1, p. 100018, 2024.

[8] F. Bellamoli, Machine Learning Methods for Wastewater Treatment Plants, Ph.D. dissertation, Dept. of Industrial Engineering, Univ. of Trento, Trento, Italy, 2025, p. 1.

[9] M. S. Zaghloul, Predictive Modelling of Advanced Wastewater Treatment Technologies Using Artificial Intelligence, Ph.D. dissertation, University of Calgary, Calgary, Canada, 2020, p. 20.

[10] F. B. Nasir, Application of Explainable Artificial Intelligence in Wastewater Treatment Plants, Ph.D. dissertation, University of Wisconsin-Milwaukee, Milwaukee, WI, USA, 2025, p. 2.

[11] M. Henze, W. Gujer, T. Mino, and M. C. Van Loosdrecht, Activated Sludge Models ASM1, ASM2, ASM2d and ASM3, IWA Publishing, 2000.

[12] S. Sözen, et al., İSKİ Rehabilitation of Wastewater Treatment Plants Project, İstanbul Teknik Üniversitesi, İnşaat Fakültesi, Çevre Mühendisliği Bölümü, İstanbul, Türkiye, 2008.

[13] J. J. W. Hulsbeek, J. Kruit, P. J. Roeleveld, and M. C. M. van Loosdrecht, "A practical protocol for dynamic modelling of activated sludge systems," Water Science and Technology, vol. 45, no. 6, pp. 127–136, 2002.

[14] A. Trew, "Spatial takeoff in the first industrial revolution," Rev. Econ. Dyn., vol. 17, no. 4, pp. 707–725, Oct. 2014.

[15] V. Vijayaraghavan and J. Rian Leevinson, "Internet of things applications and use cases in the era of industry 4.0," in The Internet of Things in the Industrial Sector: Security and Device Connectivity, Smart Environments, and Industry 4.0, 279–298, 2019

[16] L. Benedetti, Probabilistic Design and Upgrade of Wastewater Treatment Plants in the EU Water Framework Directive Context, Ph.D. dissertation, Ghent University, Belgium, 2006, p. 304.

[17] M. E. Erşahin, R. K. Dereli, H. Özgün, Z. A. Akmirza, and İ. Öztürk, "Atıksu arıtma tesislerinde enerji verimliliğinin incelenmesi," Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, c. 21, no. 2, pp. 380–387, 2017.

[18] G. Insel, B. Güder, G. Güneş, and E. Ubay Cokgor, "Are standard wastewater treatment plant design methods suitable for any municipal wastewater?" Water Science & Technology, 2012.

[19] A. N. Angelakis, E. Kavoulaki, and M. G. Dialynas, "Sanitation and stormwater and wastewater technologies in Minoan Era," in Evolution of Sanitation and Wastewater Management through the Centuries, A. N. Angelakis and J. B. Rose, Eds. London, UK: IWA Publishing, 2014.

[20] G. De Feo, G. Antoniou, H. F. Fardin, F. El-Gohary, X. Y. Zheng, I. Reklaityte, D. Butler, S. Yannopoulos, and A. N. Angelakis, "The historical development of sewers worldwide," Sustainability, vol. 6, no. 6, pp. 3936–3974, 2014.

[21] S. Halliday, The Great Stink of London: Sir Joseph Bazalgette and the Cleansing of the Victorian Capital, Stroud, U.K.: The History Press, 2001.

[22] J. Hatcher, The Black Death: A Personal History, Cambridge, MA, USA: Da Capo Press, 2008

[23] J. Navarro Palazón and P. Jiménez Castillo, "Water in Andalusian housing: supply, storage and evacuation," Verdolay, vol. 7, pp. 401–412, 1995.

[24] P. Lens, G. Zeeman, and G. Lettinga, Eds., Decentralised Sanitation and Reuse: Concepts, Systems and Implementation (Integrated Environmental Technology Series), London, U.K.: IWA Publishing, 2001, p. 14.

[25] G. Tchobanoglous and F. L. Burton, Wastewater Engineering: Treatment, Disposal, and Reuse, 3rd ed., New York, NY, USA: McGraw-Hill, 1991.

[26] E. Ardern and W. T. Lockett, “Experiments on the oxidation of sewage without the aid of filters,” J. Soc. Chem. Ind., vol. 33, no. 10, pp. 523–539, 1914.

[27] "The activated sludge process - invented by Ardern and Lockett in 1913," International Water Association, 2014. [Online]. Available: http://www.iwa100as.org/history.php.

[28] G. Lofrano and J. Brown, Wastewater Management Through the Ages: A History of Mankind. Salerno, Italy: Univ. of Salerno, Dept. Civil Eng., 2009, p. 11.

[29] E. Debik, N. Manav, and T. Coşkun, Biyolojik Temel İşlemler Ders Notları. İstanbul, Turkey: Yıldız Teknik Univ., 2008

[30] G. Tchobanoglous, F. L. Burton, and H. D. Stensel, Wastewater Engineering: Treatment and Reuse, 4th ed. New York, NY, USA: McGraw-Hill, 2003.

[31] H. Seeger, “The history of German wastewater treatment,” European Water Management, vol. 2, pp. 51–56, 1999.

[32] T. Egli, "Growth Kinetics, Bacterial," in “Encyclopedia of Microbiology” (3rd ed.), M. Schaechter, Ed. Oxford, U.K.: Academic Press, 2009, pp. 320-335.

[33] D. Butler and J. W. Davies, “Urban Drainage”, 3rd ed. London, U.K.: Spon Press, 2011.

[34] S. Judd, The MBR Book: Principles and Applications of Membrane Bioreactors for Water and Wastewater Treatment, 2nd ed. Oxford, U.K.: Butterworth-Heinemann, 2011.

[35] P. Reiter, The Founding and Evolution of IWSA and IAWQ 1947–1999 and the Creation of IWA in 2000 and its Evolution through 2015, Final Draft. London, UK: International Water Association (IWA), Sep. 1, 2022.

[36] P. A. Wilderer and H. H. Hahn, “Global water management: Challenges and responses,” Water Supply, vol. 1, no. 1, pp. 1–7, 2001.

[37] M. K. de Kreuk, N. C. G. Tan, and M. C. M. van Loosdrecht, "Aerobic granular sludge: From idea to pilot plant," Water Research, vol. 44, no. 3, pp. 761–768, Jan. 2010.

[38] J. Altmann, A. S. Ruhl, F. Zietzschmann, and M. Jekel, “Direct comparison of ozonation and adsorption onto powdered activated carbon for micropollutant removal in advanced wastewater treatment,” Water Research, vol. 55, pp. 185–193, 2014.

[39] J. Margot, L. Rossi, D.-A. Barry, and C. Holliger, "Treatment of micropollutants in municipal wastewater: Ozone or powdered activated carbon?," Science of The Total Environment, vol. 461–462, pp. 480–498, 2013

[40] O. F. Schoumans, F. Bouraoui, C. Kabbe, O. Oenema, and K. C. van Dijk, "Phosphorus management in Europe in a changing world," Ambio, vol. 44, pp. 180–192, 2015.

[41] C. Tarayre, L. De Clercq, R. Charlier, E. Michels, E. Meers, M. Camargo-Valero, and F. Delvigne, "The design of sustainable bioprocesses for nutrient recovery from waste," Bioresource Technology, vol. 206, pp. 264–274, 2016.

[42] J. E. Drewes and S. J. Khan, "Water reuse for drinking water augmentation," in Water Reuse: An International Survey of Current Practice, Issues and Needs, B. Jimenez et al., Eds. London, U.K.: IWA Publishing, 2010, pp. 1–22

[43] B. D. Shoener et al., "Energy positive wastewater treatment: the state of the art and future directions," Environmental Science: Water Research & Technology, vol. 2, no. 3, pp. 512–521, 2016.

[44] W. P. F. Barber, "Thermal hydrolysis for sewage treatment: A critical review," Water Research, vol. 104, pp. 53–71, 2016.

[45] A.-J. Wang, H. Li, Z. He, Y. Tao, H. Wang, M. Yang, D. Savic, G. T. Daigger, and N. Ren, "Digital twins for wastewater treatment: A technical review," Engineering, vol. 36, pp. 21–35, 2024.

[46] M. Alvi, D. Batstone, C. K. Mbamba, P. Keymer, T. French, A. Ward, J. Dwyer, and R. Cardell-Oliver, "Deep learning in wastewater treatment: a critical review," Water Research, vol. 245, p. 120518, 2023.

[47] C. Martin and P. A. Vanrolleghem, "Analysing, completing, and generating influent data for WWTP modelling: A critical review," Environmental Modelling & Software, vol. 60, pp. 188–201, 2014.

[48] Y. Gu, Y. Li, F. Yuan, and Q. Yang, "Optimization and control strategies of aeration in WWTPs: A review," Journal of Cleaner Production, vol. 418, p. 138008, 2023.

[49] G. Mannina, T. F. Rebougas, A. Cosenza, M. Sanchez-Marre, and K. Gibert, "Decision support systems (DSS) for wastewater treatment plants - A review of the state of the art," Bioresource Technology, vol. 290, p. 121814, 2019.

[50] U. Jeppsson, Modelling Aspects of Wastewater Treatment Processes. Ph.D. dissertation, Dept. Indust. Elect. Eng. and Automat., Lund Univ., Lund, Sweden, 1996.

[51] M.S. Duarte, G. Martins, P. Oliveira, B. Fernandes, E.C. Ferreira, M.M. Alves, F. Lopes, M.A. Pereira, P. Novais, “A review of computational modeling in wastewater treatment processes”, ACS EST Water 2023.

[52] Y. Yang, K.-R. Kim, R. Kou, Y. Li, J. Fu, L. Zhao, and H. Liu, “Prediction of effluent quality in a wastewater treatment plant by dynamic neural network modeling,” Process Safety and Environmental Protection, vol. 158, pp. 515–524, 2022.

[53] Y. Wang, Y. Cheng, H. Liu, Q. Guo, C. Dai, M. Zhao, and D. Liu, “A review on applications of artificial intelligence in wastewater treatment,” Sustainability, vol. 15, no. 17, p. 13557, 2023.

[54] A. Osmane, K. Zidan, R. Benaddi, S. Sbahi, N. Ouazzani, M. Belmouden, and L. Mandi, “Assessment of the effectiveness of a full-scale trickling filter for the treatment of municipal sewage in an arid environment: Multiple linear regression model prediction of fecal coliform removal,” Journal of Water Process Engineering, vol. 64, 2024.

[55] Y. Tharrault, G. Mourot, and J. Ragot, “WWTP diagnosis based on robust principal component analysis,” in Proc. 7th IFAC Symp. Fault Detection, Supervision and Safety of Technical Processes (SAFEPROCESS), Barcelona, Spain, June 30–July 3, 2009.

[56] S. Yang, K. Behzadian, C. Coleman, T. G. Holloway, and L. C. Campos, “Application of AI-based techniques for anomaly management in wastewater treatment plants: A review,” Journal of Environmental Management, vol. 392, p. 126886, 2025.

[57] S. Estrada-Flores, I. Merts, B. De Ketelaere, and J. Lammertyn, “Development and validation of ‘grey-box’ models for refrigeration applications: a review of key concepts,” International Journal of Refrigeration, vol. 29, no. 6, pp. 931–946, 2006.

[58] B. Chen, Y. Wang, R. Wang, Z. Zhu, L. Ma, X. Qiu, and W. Dai, “The gray-box based modeling approach integrating both mechanism-model and data-model: The case of atmospheric contaminant dispersion,” Symmetry, vol. 12, no. 2, art. no. 254, 2020.

[59] J. Monod, “The growth of bacterial cultures,” Annual Review of Microbiology, vol. 3, no. 1, pp. 371–394, 1949.

[60] M. S. Duarte et al., "A Review of Computational Modeling in Wastewater Treatment Processes," ACS ES&T Water, vol. 4, no. 4, pp. 784–804, 2024.

[61] W. W. Eckenfelder and D. J. O'Connor, Biological Treatment of Sewage and Industrial Wastes. New York, NY, USA: Reinhold Publishing, 1961.

[62] A. W. Lawrence and P. L. McCarty, “Unified basis for biological treatment design and operation,” Journal of the Sanitary Engineering Division, vol. 96, no. 3, pp. 757–778, 1970.

[63] Metcalf & Eddy, Wastewater Engineering: Collection, Treatment, Disposal, New York, NY, USA: McGraw-Hill, 1972.

[64] J. F. Andrews, "Dynamic models and control strategies for wastewater treatment processes," Water Res., vol. 8, no. 5, pp. 261–289, May 1974.

[65] P. L. Dold, G. A. Ekama, and G. v R. Marais, "A general model for the activated sludge process," Prog. Water Technol., vol. 12, pp. 47–77, 1980.

[66] M. Henze, "Nitrification in the activated sludge process," Prog. Water Technol., vol. 8, no. 4/5, pp. 323–336, 1976.

[67] M. K. Stenstrom and R. A. Poduska, "The effect of dissolved oxygen concentration on nitrification," Water Res., vol. 14, no. 6, pp. 643–649, 1980.

[68] U. Jeppsson, Modelling Aspects of Wastewater Treatment Processes. Lund, Sweden: Dept. of Industrial Electrical Engineering and Automation, Lund Institute of Technology, Lund University, 1996.

[69] M. Henze, C. P. L. Grady, W. Gujer, G. V. R. Marais, and T. Matsuo, "A general model for single-sludge wastewater treatment systems," Water Research, vol. 21, pp. 505–515, 1987.

[70] M. Mulas, "Modelling and Control of Activated Sludge Processes," Ph.D. dissertation, Univ. of Cagliari, Cagliari, Italy, 2006.

[71] M. C. M. Van Loosdrecht and M. Henze, "Maintenance, endogenous respiration, lysis, decay and predation," Water Science and Technology, vol. 39, pp. 107–117, 1999.

[72] M. Henze, W. Gujer, T. Mino, and M. C. M. van Loosdrecht, Activated Sludge Models ASM1, ASM2, ASM2d and ASM3. London, United Kingdom: IWA Publishing, 2002

[73] W. Gujer, M. Henze, T. Mino, and M. C. M. van Loosdrecht, "Activated Sludge Model No. 3," Water Science and Technology, vol. 39, pp. 183–193, 1999.

[74] S. W. H. Van Hulle, G. Sin, and P. A. Vanrolleghem, "Limitations of ASM1 and ASM3: A comparison based on batch oxygen uptake rate profiles from different full-scale wastewater treatment plants," Water Science and Technology, vol. 52, no. 8, pp. 13–20, 2005

[75] L. D. Alejo Alvarez, "On a deeper understanding of data-driven approaches in the current framework of wastewater treatment: looking inside the black-box," Ph.D. dissertation, Department of Civil and Environmental Engineering, Technische Universität Darmstadt, Darmstadt, Germany, 2020.

[76] K. V. Gernaey, M. C. M. van Loosdrecht, M. Henze, M. Lind, and S. B. Jørgensen, "Activated sludge wastewater treatment plant modelling and simulation: State-of-the-art," Environmental Modelling & Software, vol. 19, no. 9, pp. 763–783, 2004.

[77] D. Maruejouls, H. Chai, M. Sperandio, J. P. Steyer, N. Bernet, and N. G. F. Sanz, "Modeling the partial nitrification in sequencing batch reactor for biomass adapted to high ammonia concentrations," Science of The Total Environment, vol. 599-600, pp. 1935-1944, Dec. 2017.

[78] Q. Su, J. Zhang, G. Pan, and Y. Li, "Establishment of Nitrous Oxide (N_2O) dynamics model based on ASM3 model during Biological Nitrogen Removal via Nitrification," Bioprocess and Biosystems Engineering, vol. 43, pp. 1–23, 2020.

[79] L. Rieger, G. Koch, M. Kühni, W. Gujer, and H. Siegrist, "The EAWAG Bio-P module for activated sludge model No. 3," Water Research, vol. 35, no. 16, pp. 3887–3903, 2001.

[80] G. Levin and J. Shapiro, "Metabolic uptake of phosphorus by wastewater organisms," J. Water Pollut. Control Fed., vol. 37, no. 6, pp. 800-821, 1965.

[81] G. W. Fuhs and M. Chen, "Microbiological basis of phosphate removal in the activated sludge process for the treatment of wastewater," Microb. Ecol., vol. 2, no. 2, pp. 119-138, 1975.

[82] . L. Barnard, "A review of biological phosphorus removal in the activated sludge process," Water SA, vol. 2, no. 3, pp. 136-144, 1976.

[83] G. v. R. Marais, R. E. Loewenthal, and I. P. Siebritz, "Observations supporting phosphate removal by biological excess uptake," Water Sci. Technol., vol. 15, no. 3-4, pp. 15-41, 1983.

[84] M. C. Wentzel, G. A. Ekama, R. E. Loewenthal, P. L. Dold, and G. v. R. Marais, "Enhanced polyphosphate organism cultures in activated sludge

systems. Part II: Experimental behaviour," Water SA, vol. 15, no. 2, pp. 71-88, 1989.

[85] Henze, W. Gujer, T. Mino, T. Matsuo, M. C. Wentzel, and G. v. R. Marais, Activated Sludge Model No. 2, IAWQ Scientific and Technical Report No. 3, London, UK, 1995

[86] M.C.Wentzel, M. B. Mbewe, and G. A. Ekama, "Batch test for characterisation of the carbonaceous materials in municipal wastewaters," Water SA, vol. 21, no. 2, pp. 117–124, 1995.

[87] H. Siegrist, A. Brunner, G. Koch, L. C. Phan, and V. C. Le, "Reduction of biomass decay rate under anoxic and anaerobic conditions," Water Sci. Technol., vol. 39, no. 1, pp. 129–137, 1999.

[88] Kuba, G. Smolders, M. C. M. van Loosdrecht, and J. J. Heijnen, "Biological phosphorus removal from wastewater by anaerobic-anoxic sequencing batch reactor," Water Sci. Technol., vol. 27, no. 5–6, pp. 241–252, 1993.

[89] J. P. Kerrn-Jespersen and M. Henze, "Biological phosphorus uptake under anoxic and aerobic conditions," Water Res., vol. 27, no. 4, pp. 617–624, 1993.

[90] M. Henze, W. Gujer, T. Mino, T. Matsuo, M. C. Wentzel, G. v. R. Marais, and M. C. M. van Loosdrecht, "Activated Sludge Model No. 2d," Water Sci. Technol., vol. 39, no. 1, pp. 165-182, 1999.

[91] K. V. Gernaey, M. C. M. van Loosdrecht, M. Henze, M. Lind, and S. B. Jørgensen, "Activated sludge wastewater treatment plant modelling and simulation: state-of-the-art," Environ. Model. Softw., vol. 19, no. 9, pp. 763–783, 2004.

[92] A. Oehmen, M. T. Vives, H. Lu, Z. Yuan, and J. Keller, "The effect of pH on the competition between polyphosphate-accumulating organisms and glycogen-accumulating organisms," Water Res., vol. 39, no. 16, pp. 3727-3737, Oct. 2005.

[93] S. C. F. Meijer, M. C. M. van Loosdrecht, and J. J. Heijnen, "Metabolic modelling of full-scale biological nitrogen and phosphorus removing WWTPs," Water Res., vol. 35, no. 11, pp. 2711-2723, Aug. 2001.

[94] T. Panswad, A. Doungchai, and J. Anotai, "Temperature effect on microbial community of enhanced biological phosphorus removal system," Water Res., vol. 37, no. 2, pp. 409-415, Jan. 2003

[95] ATV-DVWK, Dimensioning of Single-Stage Activated Sludge Plants, ATV-DVWK-A 131E, Hennef, Germany, 2000.

[96] M. Henze, M. C. M. van Loosdrecht, G. A. Ekama, and D. Brdjanovic, Biological Wastewater Treatment: Principles, Modelling and Design. London, UK: IWA Publishing, 2008.

[97] Standard ATV-A 131: Dimensioning of Single-Stage Activated Sludge Plants, Abwassertechnische Vereinigung e.V. (ATV), St. Augustin, Germany, 1991.

[98] Standard ATV-DVWK-A 131E: Dimensioning of Single-Stage Activated Sludge Plants, German Association for Water, Wastewater and Waste (ATV-DVWK), Hennef, Germany, 2000.

[99] Standard DWA-A 131: Dimensioning of Single-Stage Activated Sludge Plants, German Association for Water, Wastewater and Waste (DWA), Hennef, Germany, 2016.

[100] D. S. Parker, E. J. Wahlberg, and H. Z. Gerges, "Improving secondary clarifier performance and capacity using a structured diagnostic approach," Water Sci. Technol., vol. 41, no. 9, pp. 201–208, May 2000.

[101] G. Olsson et al., "Instrumentation, control and automation in wastewater systems," Water Sci. Technol., vol. 52, no. 12, pp. 477–492, 2005.

[102 J. Makinia, Mathematical Modelling and Computer Simulation of Activated Sludge Systems. London, U.K.: IWA Publishing, 2010.

[103] K. V. Gernaey, U. Jeppsson, P. A. Vanrolleghem, and J. B. Copp, Benchmarking of Control Strategies for Wastewater Treatment Plants. London, U.K.: IWA Publishing, 2014.

[104 H. G. İnsel et al., "Atıksu Yönetimi Standardizasyon Projesi-1: İleri Biyolojik Atıksu Arıtma Tesisi Deneysel Karakterizasyon ve Dinamik Koşullarda Modellenmesi – Final Raporu". İstanbul, Türkiye: İstanbul Teknik Üniversitesi, Nov. 2020.

[105] L. Corominas, G. Sin, M. Puig, N. Balaguer, and M. D. Colprim, "Model calibration and validation," in Guidelines for Using Activated Sludge Models, L. Rieger et al., Eds. London, U.K.: IWA Publishing, 2012.

[106] . Rosso, L. E. Larson, and M. K. Stenstrom, "Aeration of large-scale municipal wastewater treatment plants: state of the art," Water Sci. Technol., vol. 57, no. 7, pp. 973–978, 2008.

[107] L. Amand and B. Carlsson, "Aeration control with gain scheduling in a full-scale wastewater treatment plant," in Proc. 19th World Congr. IFAC, Cape Town, South Africa, Aug. 2014.

[108 K. B. Newhart, R. W. Holloway, A. S. Hering, and T. Y. Cath, "Data-driven performance analyses of wastewater treatment plants: A review," Water Res., vol. 157, pp. 498–513, 2019.

[109] L. Rieger, S. Gillot, G. Langergraber, T. Ohtsuki, A. Shaw, I. Takács, and S. Winkler, Guidelines for Using Activated Sludge Models, Scientific and Technical Report No. 21. London, U.K.: IWA Publishing, 2012.

[110] T. Duarte, A. F. E. Ramos, A. R. L. Ribeiro, P. M. L. Castro, and A. M. T. Silva, "A review of computational modeling in wastewater treatment processes," ACS ES&T Water, vol. 4, no. 3, pp. 784–804, 2024.

[111] "GPS-X," Hatch. Alındığı tarih: 15 Ara. 2025. [Çevrimiçi]. Erişim adresi: https://www.hatch.com/hydromantis/gps-x/

[112] N. D. Mu'azu, O. Alagha, and I. Anil, "Systematic modeling of municipal wastewater activated sludge process and treatment plant capacity analysis using GPS-X," Sustainability, vol. 12, no. 19, Oct. 2020.

[113] "BioWin," EnviroSim. Alındığı Tarih: 15 Ara. 2025. [Çevrimiçi]. Erişim adresi: https://envirosim.com/products#bio

[114 J. L. Callahan, A. R. Pfluger, L. A. Figueroa ve J. Munakata-Marr, "BioWin modeling of anaerobic sludge blanket treatment of domestic wastewater," Bioresource Technology Reports, Cilt 20, s. 101231, 2022.

[115] "WEST" DHI. Alındığı Tarih: 15 Ara. 2025. [Çevrimiçi]. Erişim adresi: https://www.dhigroup.com/technologies/mikepoweredbydhi/west

[116] Gabrielli, M. et al., "Modeling and prediction of the effect of operational parameters on the fate of contaminants of emerging concern in WWTPs," Sci. Total Environ., vol. 856, p. 159200, 2023.

[117] "STOAT" WRc. Alındığı Tarih: 15 Ara. 2025. [Çevrimiçi]. Erişim adresi: https://www.wrcgroup.com/services/plan-it-stoat-specialised-wastewater-plant-design-tool/

[118] S. Yang and K. Wu, "The simulation of A2/O process in municipal wastewater treatment plant by STOAT," in Proc. 4th Int. Conf. Mech. Mater. Manuf. Eng. (ICMMME 2016), vol. 61, Guilin, China, 2016.

[119] "ASIM" EAWAG. Alındığı Tarih: 15 Ara. 2025. [Çevrimiçi]. Erişim adresi: https://www.eawag.ch/en/department/eng/software/

[120] Gujer, W. ve Larsen, T. A., "The implementation of biokinetics and conservation principles in ASIM," Water Sci. Technol., c. 31, sayı 2, ss. 257-266, 1995.

[121] "SIMBA" IFAK. Alındığı Tarih: 15 Ara. 2025. [Çevrimiçi]. Erişim adresi: https://www.ifak.eu/en/produkte/simba

[122] M. Schütze, A. Campisano, H. Colas, W. Schilling, and P. A. Vanrolleghem, "Real time control of urban wastewater systems—where do we stand today?" Journal of Hydrology, vol. 299, no. 1-4, pp. 335-348, 2004

[123] "SUMO" Dynamita. Alındığı Tarih: 15 Ara. 2025. [Çevrimiçi]. Erişim adresi: https://dynamita.com/products.html

[124] G. Insel, G. Ozyıldız, D. Okutman-Tas, D. Guven, G. E. Zengin, I. Pala-Ozkok, E. Sagir Kurt, E. Atli, N. Artan, I. Takács, and E. Cokgor, "A comprehensive evaluation of process kinetics: A plant-wide approach for nutrient removal and biogas production," Water Research, vol. 217, Art. no. 118410, 2022.

[125] H.Hauduc, L. Rieger, A. Oehmen, M. C. M. Van Loosdrecht, Y. Comeau, A. Héduit, and S. Gillot, "Critical review of activated sludge modeling: State of process knowledge, modeling concepts, and limitations," Biotechnology and Bioengineering, vol. 110, no. 1, pp. 24–46, 2013.

[126] International Water Association (IWA), IWA Annual Report 2017-2018, London, UK, 2018.

[127] International Water Association (IWA), IWA Annual Report 2018–2019, London, UK, 2019.

[128] International Water Association (IWA), Digital Water: Industry Leaders Chart the Transformation Journey, London, UK, 2019.

[129 International Water Association (IWA), Digital Water: Artificial Intelligence Solutions for the Water Sector, London, UK, 2020.

[130] International Water Association (IWA), "Digital Water White Papers Series," Alındığı Tarih: 16 Ara. 2025. [Çevrim İçi]. Erişim Adresi: https://www.iwa-network.org/our-work/digital-water-white-papers-series

[131] X. Chen, Z. Lei, J.-S. Chang, ve D.-J. Lee, "Navigating future wastewater treatment plants with artificial intelligence: Applications, challenges, and innovations," Journal of Cleaner Production, cilt 504, madde no. 145467, 2025

[132 J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955," AI Mag., vol. 27, no. 4, p. 12, 2006.

[133] A. M. Turing, "Computing machinery and intelligence," Mind, vol. LIX, no. 236, pp. 433–460, Oct. 1950.

[134] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2020.

[135] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagation errors," Nature, vol. 323, pp. 533-536, 1986.

[136] V. N. Vapnik, Statistical Learning Theory. New York, NY, USA: Wiley, 1998.

[137] T. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill,

[138 . Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[139] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[140] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[141 J. Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

[142] A. Rajaraman and J. D. Ullman, Mining of Massive Datasets. Cambridge, UK: Cambridge University Press, 2012.

[143] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[144 A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105, 2012.

[145] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[146] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[147] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Advances in

Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017.

[148] Y. Tay, M. Dehghani, S. Abnar, Y. Chung, W. Fedus, J. Rao, S. Narang, V. Vasudevan, A. Roberts, and D. Metzler, "Scaling Laws and Interpretability of Learning from Heterogeneous Data," ACM Comput. Surv., vol. 55, no. 2, pp. 1-35, 2022.

[149] X. Li, J. Su, H. Wang, G. Boczkaj, J. Mahlknecht, S. Singh, and C. Wang, "Bibliometric analysis of artificial intelligence in wastewater treatment: Current status, research progress, and future prospects," J. Environ. Chem. Eng., vol. 12, no. 4, p. 113152, Aug. 2024.

[150] X. Yu et al., "Research progress of the artificial intelligence application in wastewater treatment during 2012–2022: A bibliometric analysis," Water Sci. Technol., vol. 88, no. 7, pp. 1750–1766, Oct. 2023.

[151] . Zhou, B. Wu, and X. Zhang, "Worldwide research progress and trends in application of machine learning to wastewater treatment: A bibliometric analysis," Water, vol. 17, no. 9, p. 1314, May 2025.

[152] A. Elmotawakkil, N. Enneya, S. Bhagat, and M. Ouda, "Advanced machine learning models for robust prediction of water quality index and classification," J. Hydroinf., vol. 27, no. 2, pp. 299–319, Jan. 2025.

[153] D. Wang et al., "A machine learning framework to improve effluent quality control in wastewater treatment plants," Sci. Total Environ., vol. 784, p. 147138, Jul. 2021.

[154] M. Bagheri, K. Bagheri, N. Farshforoush, A. Velazquez, and Y. Liu, "A novel hybrid deep learning model for real-time monitoring of water pollution using sensor data," J. Water Process Eng., vol. 68, p. 106595, 2024.

[155] J. Pyo, Y. Pachepsky, S. Kim, and A. Abbas, "Long short-term memory models of water quality in inland water environments," Water Res. X, vol. 21, p. 100207, 2023.

[156] S. B. Patel, K. Shah, S. Vaghela, and M. Aglodiya, "Water potability prediction using machine learning," Sci. Rep., vol. 13, p. 12421, 2023.

[157] Y. Alali, F. Harrou, and Y. Sun, "Predicting energy consumption in wastewater treatment plants through light gradient boosting machine: A comparative study," in Proc. 10th Int. Conf. Syst. Control (ICSC), 2022.

[158] A. S. Qambar and M. M. Al Khalidy, "Optimizing dissolved oxygen requirement and energy consumption in wastewater treatment plant aeration tanks using machine learning," J. Water Process Eng., vol. 50, p. 103237, 2022.

[159 M. Colwell and M. Abolghasemi, "Digital twins for forecasting and decision optimisation with machine learning: Applications in wastewater treatment," arXiv, Preprint, 2024.

[160] E. Mohammadi et al., "Application of soft actor-critic algorithms in optimizing wastewater treatment," arXiv, Preprint, 2024.

[161] D. S. M. S. Al-Kathiri et al., "Membrane fouling prediction and control using AI and machine learning: A comprehensive review," J. Environ. Earth Sci., vol. 7, no. 6, pp. 315–350, 2025.

[162] H. Kazim, M. Sabri, A. Al-Othman, and M. Tawalbeh, "Artificial intelligence application in membrane processes and prediction of fouling for better resource recovery," J. Resour. Recovery, vol. 1, p. 1008, 2023.

[163] M. I. Ejimofor et al., "Experimental and artificial intelligence optimization of paint wastewater (PWW) coagulation using novel phaseolus vulgaris seed extract (PVSE)," Heliyon, vol. 10, no. 14, p. e34229, 2024.

[164] G. C. ChazUZA, F. O. Afolabi, and Musonge P., "Response surface methodology and deep learning modelling for the removal of Pb2+ ions from wastewater using apple pomace," J. Environ. Earth Sci., vol. 7, no. 10, pp. 100–125, 2025.

[165] A. S. Mahmoud, N. Y. Mohamed, M. Mostafa, and M. S. Mahmoud, "Effective chromium adsorption from aqueous solutions and tannery wastewater using bimetallic Fe/Cu nanoparticles: Response surface methodology and artificial neural network," Air, Soil Water Res., vol. 14, pp. 1–15, 2021.

[166] A. Ajao, "Application of machine learning to understand PFAS occurrence, distribution, transport and removal in water," Preprints, p. 2024030627, 2024.

[167] M. A. Dada et al., "Review of smart water management: IoT and AI in water and wastewater treatment," World J. Adv. Res. Rev., vol. 21, no. 1, pp. 1373–1382, 2024.

[168] M. Mousazadehgavan et al., "The critical role of artificial intelligence in optimizing electrochemical processes for water and wastewater remediation," ACS ES&T Water, vol. 5, no. 1, pp. 238–255, 2025.

[169 W. Sun, Y. Gao, J. Zhou, K. J. Shah, and Y. Sun, "An overview of the latest developments and potential paths for artificial intelligence in wastewater treatment systems," *Water*, vol. 17, no. 16, p. 2432, 2025.

[170] S. M. Hocaoğlu et al., "Artificial intelligence in wastewater treatment plants: A review of current trends and adaptation strategies in the face of climate-driven rainfall fluctuations," J. Water Clim. Change, vol. 16, no. 9, pp. 2742–2759, 2025.

[171] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.

[172] S. Garcia, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining. Cham, Switzerland: Springer, 2015.

[173] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2017.

[174] Y. Pandya, "How do you handle missing data in a dataset?", Medium, Dec. 6, 2023. Alındığı Tarih: 20 Ara. 2025 [Çevirimiçi]. Available: https://medium.com/@yagnesh.pandya/how-do-you-handle-missing-data-in-a-dataset-337a89c1b4a1

[175] . L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," Psychological Methods, vol. 7, no. 2, pp. 147–177, 2002.

[176] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.

[177] D. A. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," International Statistical Review, vol. 78, no. 1, pp. 40–64, 2010.

[178 D. A. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," International Statistical Review, vol. 78, no. 1, pp. 40–64, 2010.

[179] S. van Buuren, Flexible Imputation of Missing Data, 2nd ed. Boca Raton, FL, USA: Chapman & Hall/CRC, 2018.

[180] . A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, no. 1, pp. 1–22, 1977.

[181] J. L. Schafer, Analysis of Incomplete Multivariate Data. London, UK: Chapman & Hall, 1997.

[182 F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized Target Encoding Outperforms One-Hot Encoding," arXiv preprint arXiv:2104.00629, 2022.

[183] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2nd ed. New York, NY, USA: Springer, 2021.

[184] A. Géron, burkov-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.

[185] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[186] F. Chollet, Deep Learning with Python, 2nd ed. Shelter Island, NY, USA: Manning Publications, 2021.

[187] E. Poslavskaya and A. Korolev, "Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding?" IEEE Access, vol. 11, pp. 123456-123467, 2023.

[188] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," Frontiers in Bioinformatics, vol. 2, p. 927312, Jun. 2022.

[189] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," Journal of Machine Learning Research, vol. 5, pp. 1205–1224, Oct. 2004.

[190] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, Mar. 2003.

[191] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, "Feature selection: A review and comparative study," E3S Web of Conferences, vol. 351, p. 01046, 2022.

[192] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[193] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.

[194] . Liu and H. Motoda, Eds., Computational Methods of Feature Selection. Boca Raton, FL: Chapman & Hall/CRC, 2008.

[195] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," "Machine Learning", vol. 46, no. 1-3, pp. 389-422, Jan. 2002.

[196] A. L. Samuel, "Some studies in machine learning using the game of checkers," IBM Journal of research and development, vol. 3, no. 3, pp. 210-229, 1959.

[197] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[198] E. Alpaydın, Introduction to Machine Learning, 2nd ed. Cambridge, MA, USA: MIT Press, 2010.

[199] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[200] A. Burkov, The Hundred-Page Machine Learning Book. Quebec City, QC, Canada: Andriy Burkov, 2019.

[201] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[202] F. Dikmen, A. Demir, B. Özkaya, M. O. Raza, J. Rasheed, T. Asuroglu, and S. Alsubai, "AI-driven wastewater management through comparative analysis of feature selection techniques and predictive models," Scientific Reports, vol. 15, no. 1, Art. no. 25347, 2025.

[203] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, pp. 785–794 2016.

[204] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NIPS), vol. 30, Long Beach, CA, USA, 2017, pp. 3146–3154.

[205] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2018.

[206] J. H. Friedman, "Stochastic gradient boosting," Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367–378, 2002.

[207] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.

[208 I. Jolliffe, Principal Component Analysis, 2nd ed. New York: Springer, 2002.

[209] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.

[210] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), 1994..

[211] Y R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[212] C. J. C. H. Watkins and P. Dayan, "Q-learning," Machine Learning, vol. 8, no. 3, pp. 279–292, 1992.

[213] V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, 2015.

[214] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, Jan. 2015.

[215] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Computer Science, vol. 7, p. e623, 2021.

[216] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," Geosci. Model Dev., vol. 7, no. 3, pp. 1247–1250, 2014.

[217] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," Geoscientific Model Development, vol. 15, no. 14, pp. 5481–5487, 2022.

[218] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” PeerJ Computer Science, vol. 7, p. e623, 2021.

[219] J. M. Cadenas, M. C. Garrido ve R. Martínez, "Feature subset selection Filter-Wrapper based on low quality data," Expert Systems with Applications, c. 40, no. 16, ss. 6241–6252, Kas. 201

A
PHYTHON KODLARI

Bu tez çalışmasında uygulanan yöntemlere ait python kodları (veri işleme modelleme) fikri mülkiyet hakları nedeniyle tam metin olarak paylaşılmamıştır. Çalışmanın şeffaflığını ve tekrarlanabilirliğini sağlamak amacıyla, geliştirilen modelin temel algoritma mantığı ve işlem adımları aşağıda "Sözde Kod (Pseudo-Code)" şeklinde verilmiştir.

A.1 Veri Ön İşleme ve Regresyon Modelleri

```
BAŞLAT Veri_Ön_İsleme
GİRDİ: Ham Veri Dosyası (CSV)
ÇIKTI: Ölçeklendirilmiş Özellik Matrisi (X_scaled), Hedef Vektörler (y)

ADIM 1: Veri Yükleme ve Temizlik
  Veri setini oku (pandas.read_csv)
  Analiz dışı sütunları (Örn: 'Tarih') veri setinden çıkar.

ADIM 2: Kategorik Dönüşüm (Encoding)
  "Hava Durumu" (Weather Condition) sütununu seç.
  Etiket Kodlama (Label Encoding) uygula:
   - Yağışsız (Dry) -> 0
   - Yağışlı (Wet) -> 1

ADIM 3: Hedef ve Özellik Ayrımı
  Hedef Değişkenler (Targets) = [Çıkış KOİ, BOİ, AKM, TN, TP]
  Girdi Özellikleri (X) = Tüm Sütunlar - Hedef Değişkenler

ADIM 4: Standardizasyon (Scaling)
  Standart Ölçekleyici (StandardScaler) başlat.
  X matresindeki tüm verileri (x - u) / s formülü ile dönüştür.
  (u: ortalama, s: standart sapma)
```

BİTİR ve DÖNDÜR (X_scaled, y)

A.2 Öznitelik Seçimi ve Optimize Edilmiş Modeller

BAŞLAT Ozellik_Secimi

GİRDİ: Tüm Özellikler (X), Hedef Değişken (y), Seçilecek Sayı (k=10)
ÇIKTI: En İyi 10 Özellik Listesi (Selected_Features)

TANIMLA Yöntemler Listesi:
1. İstatistiksel Filtre: SelectKBest (f_regression skorlama)
2. Entropi Tabanlı: Mutual Information (mutual_info_regression)
3. Sarmalayıcı (Doğrusal): RFE (LinearRegression tabanlı)
4. Sarmalayıcı (Ağaç): RFE (RandomForest tabanlı)

DÖNGÜ (Her Yöntem İçin):
EĞER Yöntem == "Filtre Tabanlı" İSE:
Her özelliğin hedef değişkenle olan skorunu (F-skoru veya MI) hesapla.
Skorları büyükten küçüğe sırala.

EĞER Yöntem == "Sarmalayıcı (RFE)" İSE:
Modeli tüm özelliklerle eğit.
En az öneme sahip özelliği ele.
İstenen sayıya (k=10) ulaşana kadar tekrarla.

SONUÇ:
Her yöntem için en yüksek skora sahip 10 özelliği belirle.
Veri setini sadece bu 10 özelliği içerecek şekilde indirge.

BİTİR ve DÖNDÜR (X_reduced)

BAŞLAT Model_Egitimi

Model Eğitimi ve Performans Analizİ

GİRDİ: İndirgenmiş Veri Seti (X_reduced), Hedefler (Targets)

PARAMETRELER: Eğitim/Test Oranları = [%80-%20, %70-%30, %60-%40, %50-%50]

MODELLER: [Linear, Ridge, Lasso, Random Forest, GBM, XGBoost, LightGBM, DT, KNN]

DÖNGÜ 1: Her Bir Hedef Değişken (KOİ, BOİ, AKM...) İçin:

DÖNGÜ 2: Her Bir Makine Öğrenmesi Modeli İçin:

DÖNGÜ 3: Her Bir Veri Bölme Oranı (Split) İçin:

ADIM 3.1: Veri Bölme

Veriyi rastgele (random_state=42) Eğitim ve Test setlerine ayır.

ADIM 3.2: Model Eğitimi (Fitting)

Modeli Eğitim seti (X_train, y_train) üzerinde eğit.

ADIM 3.3: Tahmin (Prediction)

Test seti (X_test) üzerinde tahmin üret -> y_pred.

ADIM 3.4: Hata Hesaplama (Metrics)

MAE = Mutlak Ortalama Hata (y_test, y_pred)

MSE = Ortalama Kare Hata (y_test, y_pred)

R^2 = Belirlilik Katsayısı (y_test, y_pred)

ADIM 3.5: Kayıt

Sonuçları "Results" veri tabanına/listesine ekle.

Gerçek vs. Tahmin grafiklerini çizdir.

ADIM 4: Raporlama

Tüm sonuçları (Model Adı, Hedef, Metrikler) birleştir.

Sonuçları CSV dosyası olarak dışa aktar (export).

BİTİR

B
DENEY YÖNTEMİ VE REFERANSLARI

DENEY ADI	DENEY YÖNTEMİ	REFERANS METOD
Askıda Katı Madde (AKM) ve Uçucu Askıda Katı Madde (UAKM)	Gravimetrik Analiz Yöntemi	Standard Metot 2540D (AKM) Standard Metot 2540B (UAKM)
Toplam Katı Madde (TKM) ve Toplam Uçucu Katı Madde (TUKM)	Gravimetrik Analiz Yöntemi	Standart Metot 2540B
Kuru Madde Tayini (%TKM)	Enstrümental Analiz Yöntemi (Nem Tayin Cihazı)	TS ISO 11465:1997
Çökebilir Katı Madde Tayini (ÇKM)	Volumetrik Analiz Yöntemi	Standart Metot 2540F
Kimyasal Oksijen İhtiyacı (KOİ)	Volumetrik Analiz Yöntemi (Open Reflux Metodu)	Standart Metot 5220B Open Reflux
Kimyasal Oksijen İhtiyacı (KOİ)	Volumetrik Analiz Yöntemi (Closed Reflux Metodu)	Standart Metot 5220C Closed Reflux
Biyokimyasal Oksijen İhtiyacı ($BOİ_5$)	Enstrümental Analiz Yöntemi (Oxitop Başlığı)	Standart Metot 5210D Respirometric Metod
Yağ-Gres	Volumetrik Analiz Yöntemi (Sokslet Ekstraksiyon Metodu)	Standart Metot 5520D
Amonyum Azotu (NH_4^+-N)	Enstrümental Analiz Yöntemi (Seçici Elektrot Metodu)	Standart Metot 4500 NH_3 D
Çözünmüş Oksijen	Enstrümental Analiz Yöntemi (Elektrokimyasal Prob Metodu)	EN 25824: 1990-04 Elektrokimyasal Prob Metodu
pH	Enstrümental Analiz Yöntemi (Elektrometrik Metod)	Standart Metot 4500 H+B - Elektrometrik Metod
İletkenlik	Enstrümental Analiz Yöntemi	Standart Metot 2510B - Laboratuar Metodu
Sıcaklık	Enstrümental Analiz Yöntemi	Standart Method 2550B - Laboratuvar ve Saha Metodu
Alkalinite	Enstrümental Analiz Yöntemi (Elektrometrik Metod)	Standart Metot 4500 H+B - Elektrometrik Metod
Uçucu Yağ Asiti (UYA)	Volümetrik Analiz Yöntemi (Titrimetrik Metot)	Anderson G. K., Yang G. 1992, ”Determination Of Bicarbonate And Total Volatile Acid Concentration Anaerobic Digesters Using A Simple Titration”, Water Environmental Research, Cilt 64, Sayfa 53-59.
F.coli Sayımı	Membran Filtre Yöntemi	Standart Metot 9222-D Fecal coliform

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. F. Dikmen, "Intelligent Effluent Monitoring: Machine Learning-Based Threat Classification and Quality Prediction for Wastewater Treatment Plants," presented at the Third International Conference on Artificial Intelligence of Things (AIoT), Istanbul, Turkiye, Dec. 26-28, 2024.

Makaleler

1 F. Dikmen, A. Demir, B. Özkaya, M. O. Raza, J. Rasheed, T. Asuroglu, ve S. Alsubai, "AI-driven wastewater management through comparative analysis of feature selection techniques and predictive models," *Sci. Rep.*, c. 15, no. 1, Makale no. 25347, 2025. doi: 10.1038/s41598-025-07124-0.

2 B. Özkaya, F. Dikmen, A. Demir, M. O. Raza, S. Alsubai, O. Osman, ve J. Rasheed, "Harnessing machine learning to mitigate water pollution in support of climate action," *Discov. Artif. Intell.*, 2025. doi: 10.1007/s44163-025-00728-5.