T.C.
MUNZUR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SU KALİTESİNİN MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE TAHMİN EDİLMESİ

YASİN AKTEKİN

YÜKSEK LİSANS TEZİ
KİMYASAL TEKNOLOJİLER ANABİLİM DALI

DANIŞMAN
PROF. DR. MUHARREM İNCE

TUNCELİ – 2024

T.C.
MUNZUR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SU KALİTESİNİN MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE TAHMİN EDİLMESİ

YASİN AKTEKİN
(210160009)

YÜKSEK LİSANS TEZİ
KİMYASAL TEKNOLOJİLER ANABİLİM DALI

DANIŞMAN
PROF. DR. MUHARREM İNCE

TUNCELİ – 2024

T.C.
MUNZUR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SU KALİTESİNİN MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE TAHMİN EDİLMESİ

YASİN AKTEKİN
YÜKSEK LİSANS TEZİ
KİMYASAL TEKNOLOJİLER ANABİLİM DALI

Bu tez 29 / 05 / 2024 tarihinde aşağıdaki jüri üyeleri tarafından **oybirliği** ile kabul edilmiştir.

İmza	**İmza**	**İmza**
Doç. Dr. Vedat TÜMEN (Bitlis Eren Üniversitesi)	Prof. Dr. Muharrem İNCE (Munzur Üniversitesi)	Dr. Öğr. Üyesi Yusuf ÇELİK (Munzur Üniversitesi)
BAŞKAN	**DANIŞMAN**	**ÜYE**

Bu tez, Enstitümüz Kimyasal Teknolojiler Anabilim Dalı'nda hazırlanmıştır.

Prof. Dr. Altuğ KAZAR
Enstitü Müdürü

29/05/2024

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Yasin AKTEKİN

TEŞEKKÜR

Bu yüksek lisans tezinin çalışma fikrini ortaya koyan ve çalışmalarımın her aşamasında koşulsuz desteğini sunan tez danışmanım Prof. Dr. Muharrem İNCE' ye, fikirlerinden faydalandığım ve yardımlarını esirgemeyen Dr. Öğr. Üyesi Yusuf Çelik'e teşekkürü bir borç bilirim. Ayrıca, tez çalışmalarım sürecinde bana destek veren herkese teşekkür ediyorum.

Son olarak da bugüne kadar emek ve sevgilerini esirgemeyen aileme sonsuz teşekkürümü bir borç bilirim.

Yasin Aktekin
TUNCELİ-2024

İÇİNDEKİLER

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

SEMBOLLER LİSTESİ

b	: Bias terimi
$h_\theta(x)$	: Hipotez fonksiyonu veya modelin tahmin fonksiyonu
I	: Birim matris
K	: K en yakın komşu algoritmasında komşu sayısı
m	: Örnek sayısı
n	: Özellik sayısı
v	: Vektör
w	: Ağırlık vektörü
y	: Hedef değişken veya çıktı.
$\hat{y}$	: Model tarafından tahmin edilen çıktı
α	: Öğrenme oranı
λ	: Düzenleme parametresi
$\sigma(z)$	: Sigmoid fonksiyonu
$\varphi(v)$	: Fonksiyonun vektör v üzerinde işlevi

KISALTMALAR LİSTESİ

AMI	: Active Microwave Instrument
AUC	: Area Under Curve
CART	: Clasification and Regression Trees
DCF	: Deep Cascade Forest
EPA	: United States Environmental Protection Agency
EU	: European Union
GBM	: Gradient Boosted Machine
KNN	: K-Nearest Neighbors
LSTM	: Long Short-Term Memory
MSE	: Mean Square Error
NAR	: Nonlinear autoregressive network
NTU	: Nephelometric Turbidity Unit
OLS	: Ordinary Least Squares
PNN	: Probabilistic Neural Network
RF	: Random Forest
ROC	: Receiver Operating Characteristic Curve
SVM	: Support Vector Machine
TD	: Temporal Differance
UNESCO	: United Nations Educational, Scientific and Cultural Org.
UNICEF	: United Nations Children's Fund
WHO	: World Health Organization- Dünya Sağlık Örgütü
WQI	: Water Quality Index
YSA	: Yapay Sinir Ağları

ÖZET

Hayatın devamlılığı, sağlıklı yaşamın olmazsa olmazı ve kalkınmanın en önemli temel taşlarından biri olan su, hızlı sanayileşme ve çeşitli kirleticilerin ekosisteme karışması nedeniyle kullanılabilir olma özelliğini yitirmektedir. Sürdürülebilir bir yaşamın ve güçlü ekonomilerin devamı için temiz su kaynaklarına olan ihtiyaç her geçen gün artmaktadır. Yeryüzünde çok sayıda su kaynağı bulunmasına rağmen, dünyanın birçok bölgesinde temiz içme suyu kaynakları sınırlıdır ve temiz su kaynaklarının artması su kirliliğinin azaltılması ile mümkündür. Sanayileşmenin artmasına paralel olarak özellikle gelişmiş ülkelerde daha fazla olmak üzere neredeyse her ülkede su kalitesi bozulmuş durumdadır. Bazı ülkelerde ise su kıtlığı büyük bir problem olarak kendini göstermektedir. İhtiyaç duyulan kullanıma uygun su miktarının arttırılabilmesi için ekosisteme bırakılan kirleticilerin azaltılmasının yanısıra su kalitesinin takibinin sürekli yapılması gerekmektedir. Nehirler başlıca tatlı su kaynaklarıdır ve sürdürülebilir bir su yönetimi için sorunun çözümü kaynağındaki çözümde gizlidir. Çok sayıda kirletici ve kirliliğe maruz kalan yüzey suları aynı zamanda doğal deşarj alanı olarak kullanılmaktadırlar. Su kalitesi, insanların sağlığı, ekosistemlerin sürdürülebilirliği ve toplumların refahı için kritik bir öneme sahip olduğundan yüzey su kalitesinin takibinde tüm su kalite karakteristikleri yerine su kalite indekslerinin kullanılması pratiklik sağlamaktadır.

Bu çalışma, su kalitesinin makine öğrenmesi algoritmalarıyla tahmin edilmesi üzerine odaklanmıştır. Su kalitesini etkileyen faktörlerin karmaşıklığı ve çeşitliliği, geleneksel analitik yöntemlerle tespit edilmesi hem zor hem de zaman aldığından, makine öğrenmesi algoritmalarının kullanımı, su kalitesinin tahmin edilmesi ve izlenmesinde etkin bir seçenek olarak kullanıldı. Makine öğrenmesi algoritmaları ile önce otomatik sonra hiper parametre ayarlaması ile işlem yapıldı. Su kalitesinin tespiti minimum sayıda bağımsız değişken kullanımıyla gerçekleştirildi ve su kalite takibini mümkün kılan bir yaklaşım önerildi. Güney Avustralya Hükümeti veri tabanından alınan su kalite takibi için alüminyum, amonyum, demir, kalsiyum, klorür, mangan, sülfat, pH, renk, bulanıklık değişkenleri kullanılarak su kalitesi tahmin edildi. Su kalitesinin tahmin edilmesinde Gaussian Naive Bayes, K- en yakın komşu (KNN), destek vektör (Support Vector), yapay sinir ağları (Artificial Neural Network), karar ağaçları (CART), rastgele orman (Random Forests), gradyan artırmalı (Gradient Boosting Machines), kategori artırmalı (Category Boosting CatBoost), lojistik regresyon modellerinden faydalanılarak gerçekleştirildi.

Anahtar Kelimeler: Makine öğrenme yaklaşımı, su kalite sistemleri, su kalite parametreleri, yapay sinir ağları, lojistik regresyon

ABSTRACT

Water Quality Prediction Using Machine Learning Algorithms

Water, which is an essential for the continuity of life, a healthy life and one of the most important cornerstones of development, is losing its usability due to rapid industrialization and the mixing of various pollutants into the ecosystem. The need for clean water resources is increasing day by day for the continuation of a sustainable life and strong economies. Although there are many water resources on earth, clean drinking water resources are limited in many parts of the world and increasing clean water resources is possible by reducing water pollution. In parallel with the increase in industrialization, water quality has deteriorated in almost every country, especially in developed countries. In some countries, water scarcity presents itself as a major problem. In order to increase the amount of water suitable for the required use, it is necessary to constantly monitor the water quality as well as reduce the pollutants released into the ecosystem. Rivers are the main freshwater resources and the solution to the problem for sustainable water management lies in the solution at the source. Surface waters, which are exposed to many pollutants and pollution, are also used as natural discharge areas. Since water quality is of critical importance for the health of people, the sustainability of ecosystems and the welfare of societies, it is practical to use water quality indices instead of all water quality characteristics in monitoring surface water quality.

This study focused on predicting water quality with machine learning algorithms. Since the complexity and diversity of factors affecting water quality are both difficult and time-consuming to detect with traditional analytical methods, the use of machine learning algorithms has been used as an effective option in predicting and monitoring water quality. The process was performed first automatically and then with hyperparameter adjustment using machine learning algorithms. Determination of water quality was achieved with the use of a minimum number of independent variables, and an approach that enables water quality monitoring was proposed. Water quality index was estimated using aluminium, ammonium, iron, calcium, chloride, manganese, sulphate, pH, colour, turbidity variables for water quality monitoring taken from the South Australian Government database. Estimation of water quality Gaussian Naive Bayes, K-nearest neighbor (KNN), support vector (Support Vector), artificial neural networks (Artificial Neural Network), decision trees (CART), random forest (Random Forests), gradient boosting (Gradient Boosting Machines), Category Boosting CatBoost and logistic regression models were used.

Keywords: Machine learning approach, water quality systems, water quality parameters, artificial neural networks, logistic regression

1. GİRİŞ

Temiz su, dünya için çok önemli bir faktördür ve yaşamın tüm yönlerini gerçekleştirmek ve sürekli gelişmek için hayati bir rol göstermektedir. Yeryüzünde birçok su kaynağı bulunmasına rağmen, dünyanın birçok bölgesinde temiz içme suyu kaynakları sınırlıdır. Büyük, yoğun nüfus ve hızlı sanayileşme ile şekillenen büyük şehirlerde bu gerçek gözlemlenmektedir. Büyük şehirlerde, nüfusun yoğun olduğu yerlerde, hızlı sanayileşme ve teknolojik gelişmelerle birlikte çok fazla kanserojen ve yüksek derecede toksik inorganik ve organik mikro kirletici toprağı, havayı ve dolayısı ile temiz içme su kaynaklarını etkilemektedir.

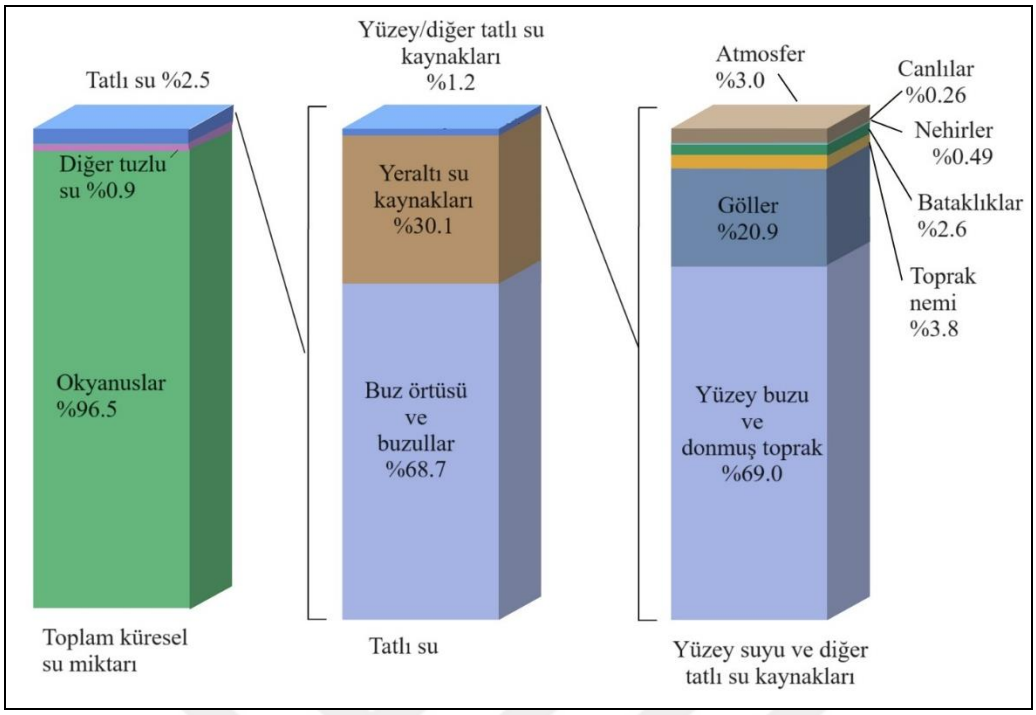
Kirli su, tarım ürünlerini ve gıdaları kontamine ettiğinden, gıda güvenliği sorunları meydana gelmektedir. Su kaynaklarının kalitesi, sucul ekosistemlerin ve biyolojik çeşitliliğin korunmasında önemli bir yer tutmaktadır. Kötü su kalitesi, suda yaşayan canlıların yaşam alanlarına zarar vermekte, çeşitliliğin ve türlerin yok olmasına ve ekosistemin zincir dengesizliğine yol açmaktadır.

Su kaynaklarının temizliği, güvenliği ve kalitesi, rekreasyonel faaliyetler ve aynı zamanda turizm için de önemlidir.

Su kalitesi, insanların sağlığı, ekosistemlerin sürdürülebilirliği ve toplumların refahı için kritik bir öneme sahiptir. Su kalitesini etkileyen faktörlerin karmaşıklığı ve çeşitliliği, geleneksel analitik yöntemlerle tespit edilmesi hem zor hem de zaman almaktadır. Bu noktada, makine öğrenmesi algoritmalarının kullanımı, su kalitesinin tahmin edilmesi ve izlenmesi açısından önemli bir yetkinliğe sahiptir. Bu çalışma, su kalitesinin makine öğrenmesi algoritmalarıyla tahmin edilmesi üzerine odaklanmıştır.

1.1. Temiz Su Kaynaklarının Önemi ve Suya Erişim Zorlukları

Yeryüzündeki su kaynaklarının toplam miktarı yaklaşık olarak 1,4 milyar km^3 olarak tahmin edilmektedir. Dünya üzerindeki su kaynaklarının sadece %2,8'i tatlı su, %97,2'si okyanus ve denizlerde bulunan tuzlu sudur (Raghunath, 2006).



Şekil 1.1. Dünya üzerindeki su kaynakları dağılımı (URL-5, 2024)

Dünyadaki tatlı suyun yaklaşık %69'u kutup buzullarında buz şeklinde kilitlenmiş durumda ve tatlı suyun diğer %30'u da yeraltı suyu şeklinde yüzeyin altında bulunmaktadır. Dünya üzerindeki su rezervi göz önüne alındığında sadece %1'nin kullanıma uygun olduğu Şekil 1.1. de görülmektedir. Bu kadar büyük hacimdeki çok küçük bir kaynak hem sınırlı hem de gün geçtikçe tükenmektedir. Dünya nüfusunun 9 milyara doğru ulaşması 2050 yılında olması tahmin edilmektedir. Bu durumda mevcut tatlı su rezervlerinin ihtiyaçları karşılaması mümkün görünmemektedir (UNESCO, 2006). Kullanılabilir ve kaliteli suya erişim gittikçe zorlaşmaktadır.

İklimlerin değişime uğraması, ülkelerin jeopolitik konumları, salgın hastalıklar, toplu göç olayları, yüksek enflasyon ve diğer krizler suya erişim zorluğunu artıran sebeplerden bazılarıdır.

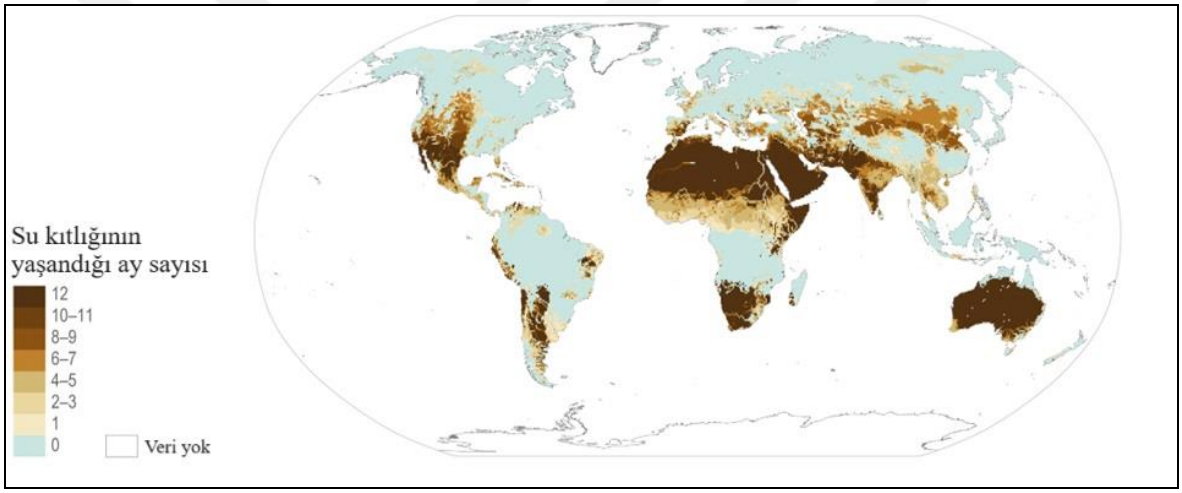
Birleşmiş milletler raporlarına göre 2030 yılına kadar temiz içme suyu, atık su ve kanalizasyonunun yeterli arıtımı ve bertarafı (sanitasyon) için yıllık yaklaşık 114 milyar ABD doları yatırım gerekecektir (UNESCO, 2024). Gelişmiş ülkelerde işlerin suya olan bağımlılığı, gelişmekte olan ülkelere göre daha düşük bir oranda kalmaktadır. Bu husus gelişmekte olan ülkeler için yüksek öncelikli bir konu durumuna gelmektedir.

Dünya nüfusunun dörtte biri, yıllık yenilenebilir tatlı su kaynaklarının %80'inden fazlasını kullanarak susuz kalma stresi ile karşı karşıya kalmaktadır (UNESCO, 2024).

Gelişmekte olan ülkelerde su kalitesi yetersiz su arıtma işlemlerinden kaynaklanırken, gelişmiş ülkelerde ise tarımsal faaliyetlerin yoğun olmasından dolayı kaynaklanan akış, suyun kalitesini olumsuz etkileyen faktörlerden biridir.

Su kalitesi için ne yazık ki dünya çapındaki veriler çok yetersiz kalmaktadır. Birleşmiş Milletler Eğitim, Bilim ve Kültür Kurumu (UNESCO) 2024 raporlarına göre, Afrika ve Asya kıtalarındaki az gelişmiş ülkelerde su kalitesini etkileyen faktörler arasında farmasötikler, endüstriyel atıklar, deterjanlar, siyanotoksinler ve nano malzemeler bulunmaktadır.

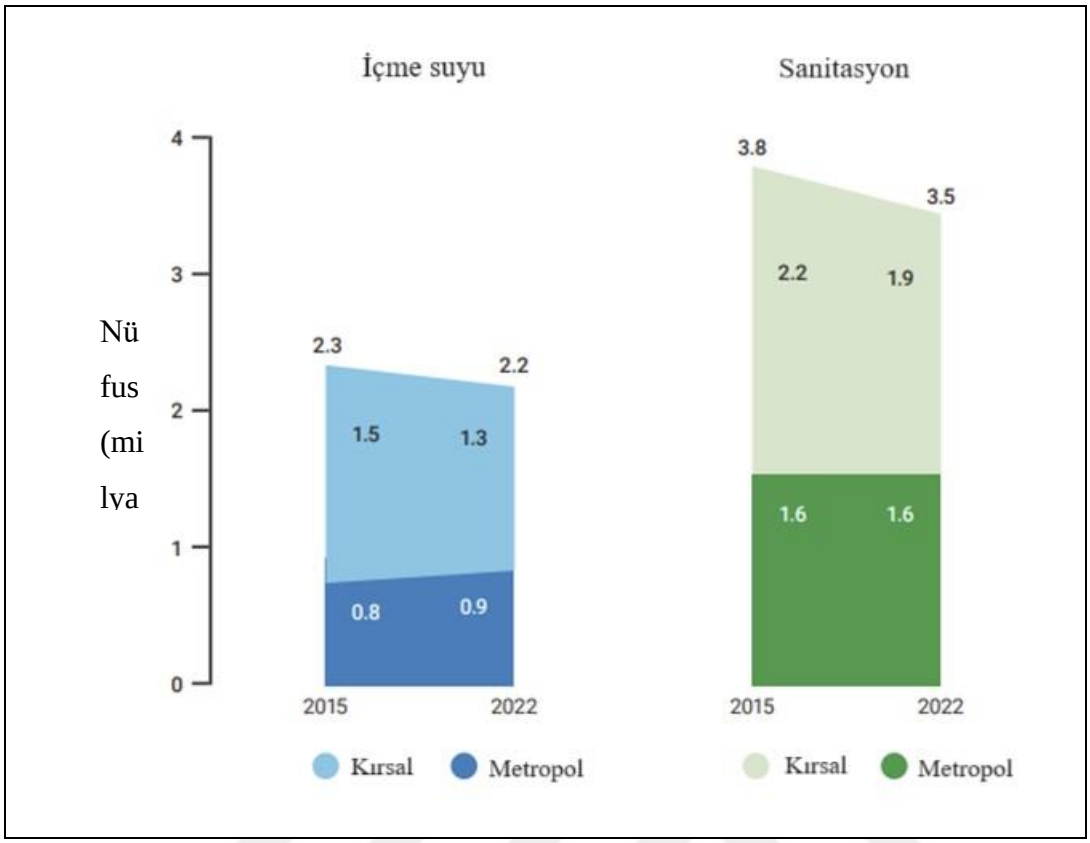
2022 yılında 2,2 milyar insan güvenli ve kaliteli suya erişim konusunda problem yaşamıştır (UNESCO, 2024).



Şekil 1.2. Yılda şiddetli su kıtlığı olan ay sayısı (Mekonnen ve Hoekstra, 2016)

Şekil 1.2'de görüleceği üzere dünya nüfusunun yarısına yakını yılın belli dönemlerinde çok ciddi su kıtlığı yaşamaktadır (Hükümetler arası iklim değişikliği paneli - The Intergovernmental Panel on Climate Change (IPCC), 2023). Bazı bölgelerde yılın tamamında su kıtlığı yaşanmaktadır.

Şekil 1.3'te, 2015 ve 2022 yılları arasında nüfus artışının yoğun olduğu metropol alanlarda temiz suya erişimin daha kolay olduğu görülmektedir. Ancak kırsal alanlarda her beş kişiden dördü temel içme suyu hizmetlerinden yoksun kalmıştır (WHO ve UNICEF, 2023).



Şekil 1.3. Kaliteli hizmet alamayan nüfus sayısı (Birleşmiş Milletler, 2023)

Temiz içme suyuna erişimde karşılaşılan bazı zorluklar:

Altyapı eksikliği: Bazı bölgelerde, suyun temizlenmesi, depolanması ve dağıtılması için gerekli altyapı eksikliği yaşanmaktadır. Bu durum temiz suya erişimi kısıtlamaktadır.

Kırsal alanlardaki uzaklık: Kırsal bölgelerde yaşayan insanlar, temiz suya ulaşmak için uzun mesafeler kat etmek zorunda kalmaktadır. Bu durum hem suya erişimi zorlaştırmakta hem de zaman kaybına neden olmaktadır.

Ekonomik kısıtlamalar: Bazı topluluklar, temiz suyun maliyetli olması nedeniyle uygun fiyatlı temiz suya erişim sağlayamamaktadır. Bu durum, gelir düzeyine bağlı olarak farklılık göstermektedir. Gelir düzeyi yüksek olan toplumlar, genellikle temiz suya daha kolay erişebilirken, düşük gelirli toplumlar bu konuda daha fazla zorlanmaktadır. Bu husus, gelir eşitsizliği ve sosyo-ekonomik adaletsizlikleri derinleştirmektedir.

Kültürel ve toplumsal faktörler: Geleneksel su kaynaklarını kullanma eğiliminde olan toplumlar da vardır. Bu geleneksel kaynaklar genellikle yeraltı kuyuları, nehirler, göller veya doğal su kaynakları olmaktadır. Ancak, bu kaynaklar sıklıkla kirlenmiş olabilir ve içilebilir su sağlama konusunda güvenli değildir. Bu toplumlar, modern temiz su kaynaklarına erişimde zorlanmaktadır. Aynı zamanda bu toplumlarda su temini ve

kullanımı genellikle kadınların sorumluluğundadır. Kadınlar, suyu evlerine taşımak, temizlemek ve ailelerinin günlük ihtiyaçlarını karşılamak için uzun saatler harcamaktadır. Bu durum, kadınların eğitim ve istihdam olanaklarına erişimini kısıtlamaktadır. Tüm bu sebepler ışığında, temiz suya erişimde çözümler bulunurken kültürel ve toplumsal dinamiklerin dikkate alınması önemlidir.

İklim değişikliği: Küresel ısınma ile meydana gelen iklim değişikliği, su kaynaklarını etkileyerek suyun miktarını ve kalitesini değiştirmektedir. Kuraklık ve su seviyelerindeki düşüşler, temiz suya erişimi kısıtlamaktadır.

1.2. Makine Öğrenmesi Metotlarının Kullanılmasının Amacı ve Önemi

Makine öğrenmesi metotları kullanılarak su kalitesinin tahmin edilmesi, su kalitesini etkileyen faktörlerin incelenmesi, veri analizi ve model oluşturma süreçleri araştırılmıştır. Su kalitesini belirlemek için makine öğrenmesi algoritmaları kullanılarak doğru ve güvenilir tahminler yapılması amaçlanmıştır. Aynı zamanda, su kaynaklarının sürdürülebilirliği için veri odaklı ve bilgi tabanlı yaklaşımların ne kadar değerli olduğunu da göstermeyi amaçlamıştır. Su kaynaklarının yönetiminde ve korunmasında etkili kararlar alabilmek için makine öğrenmesi tekniklerini kullanılmasının önemi vurgulanmıştır.

Günümüzde suyun temini değil sağlıklı suyun temini önemlidir. Hastalıkların %50'sinin kirli sudan kaynaklandığı Birleşmiş Milletler kaynaklarında belirtilmiştir (Eren ve Çelebi, 2018). Her yıl 5 yaşın altındaki 297.000 çocuk kirli ve kaliteli olmayan suya bağlı ishalden ölmektedir. Kötü sanitasyon ve kirli su aynı zamanda kolera, dizanteri, hepatit A ve tifo gibi hastalıkların da kök nedenlerinden biridir. Özellikle birçok ülkenin güçlü bir nüfus artışı yaşadığı Sahra Altı Afrika'da bu durum göz ardı edilmemelidir (WHO, UNICEF, 2021).

Suyun, sanitasyonun ve hijyenin erişilebilirliği, kalitesi ve mevcudiyeti konusundaki eşitsizlik açıklarının kapatılması, hükümetlerin finansman ve planlama stratejilerinin merkezinde yer almalıdır. Bu nedenle yeryüzü su kaynaklarının kalitesinin değerlendirilmesi, bu kaynakların yönetim aksiyon planlarını geliştirmek için en önemli adımlardan biridir. Su kalitesinin doğru tahmini, su yönetimini ve kirlilik kontrolünün sağlanmasının ve iyileştirilmesinin en önemli parametresidir.

Makine öğrenme yöntemlerinin su kalitesinin değerlendirilmesinde oldukça yüksek güvenilirlikte sonuçlar verdiği, bilimsel araştırmaların sonucunda belirlenmiştir (Castrillo

ve García, 2020). Makine öğrenmesi algoritmalarının kullanımı, veri setlerinden suyun kalitesi ile ilgili önemli ve kritik ön görüler elde etmek için daha hızlı, etkili ve yenilikçi yaklaşımlar sunmaktadır. Makine öğrenmesi algoritmaları, büyük miktarda veri üzerinden hızlı ve hassas tahminler yapabilme yetkinliğine sahip olduğundan su kalitesinin tahmin edilmesi ile olumsuz durumlara zamanında müdahalelerin yapılmasına olanak sağlamaktadır.

1.3. Çalışmanın Yapısı

Bu tez çalışması beş bölümden oluşmaktadır. Birinci bölümde, çalışma kapsamında ele alınan problem, yer yüzünde kaliteli içme suyunun önemi, tatlı su kaynakların dağılımına ve tatlı su kaynaklarına erişimindeki problemlere değinilmiştir. Aynı zamanda çalışmanın amacı, önemi ve literatüre katkısı hakkında bilgiler verilmiştir.

İkinci bölümde makine öğrenmesi algoritmaları ve model değerlendirme yöntemleri detaylı bir şekilde açıklanmıştır. Uygulanan teknik ve modellerin teorik temellerine, matematiksel formüllerine değinilmiştir.

Üçüncü bölümde, kullanılan yöntem ve tekniklerle elde edilen sonuçlar analiz edilmiş ve makine öğrenmesi algoritmalarının su kalitesi tahmini üzerindeki etkisi değerlendirilmiştir.

Dördüncü bölümde elde edilen sonuçların literatürdeki diğer metotlar ile karşılaştırması ve yöntemler arasındaki farkların değerlendirilmesi yapılmıştır.

Beşinci bölümde sonuçlar değerlendirilmiş, çalışmanın önemli noktaları özetlenmiş ve gelecekteki araştırmalar için öneriler sunulmuştur.

2. MAKİNE ÖĞRENMESİ ALGORİTMALARI

Hayatımızın her alanında rol almaya başlayan makinelerin etkileri son yıllarda iyice hissedilmeye başlanmıştır. İnsan gibi düşünen bu makineler tıp, kimya, elektronik, laboratuvar çalışmaları, elektronik ticaret ve insan kaynakları yönetimi gibi birçok alanda kendine yer edinmiştir. Kimyasal uygulamalarda karşılık bulan makine öğrenimi metotları; sağlık alanında karar verme, hastalık takibi, cerrahi planlama, operasyon, tanı ve hatta tedavi gibi birçok görev için kullanılmaya başlanmıştır.

Kimyasal analiz ve laboratuvar çalışmalarında da özellikle biriken ham veriler içerisinden istenilen sonucun alınması için ihtiyaç duyulan veri okur yazarlığı, analiz ve görselleştirme işlemleri için makine öğrenmesi, derin öğrenme ve yapay zekâ gibi kavramlar literatürde sıkça yer edinmeye başlamıştır.

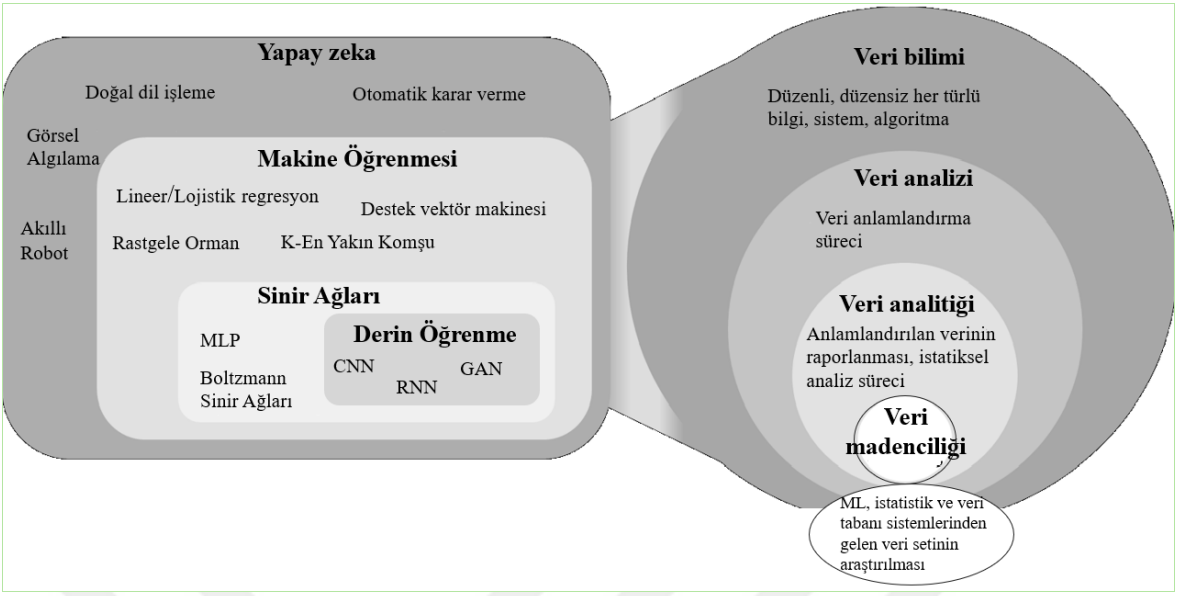
Yapay zekâ ifadesi resmi olarak ilk kez 1955 yılında Dartmouth College'da düzenlenen bir konferansta John McCarthy tarafından kullanılmıştır (McCarthy ve ark., 2006).

Zekâ, beynin bilgiyi alıp hızlı ve doğru analiz etmesi olarak tanımlanmaktadır. İdealize edilmiş bir yaklaşıma göre, yapay zekâ insan zekâsının temel özelliklerini taklit etmek üzere tasarlanmış bir işletim sistemidir. Bu sistem, algılama, öğrenme, çoğul kavramları bağlama, düşünme, fikir yürütme, sorun çözme, iletişim kurma, çıkarım yapma ve karar verme gibi yüksek bilişsel fonksiyonları veya otonom davranışları sergilemeyi amaçlar.

Geçmişten günümüze yapay zekâ ile ilgili birçok tanım yapılmıştır. Genel olarak; insanlara ait olan özelliklerin yani düşünme, anlama, faaliyete geçirmeyi sağlayacak bilgi işleme çalışması, analitik düşünme, mukayese etme gibi davranışların makine tarafından yapılması olarak tanımlanabilir (Prim, 2006).

Yapay zekâ (Artificial intelligence), Makine öğrenmesi (Machine learning), Sinir ağları (Neural networks), Derin öğrenme (Deep learning) ve Veri bilimi (Data Science) arasındaki ilişki durumu Şekil 2.1'de verilmiştir.

Veri bilimi (Data Science), düzenli veya düzensiz verilerden bilgi ve yorum elde eden bir bilim alanıdır. İstatistik, Makine öğrenmesi ve Yapay zekâ gibi çeşitli teknikleri kullanmaktadır. Verinin toplanması, temizlenmesi gibi ön işleme süreçlerini ve analiz edip yorumlama süreçlerini içermektedir.



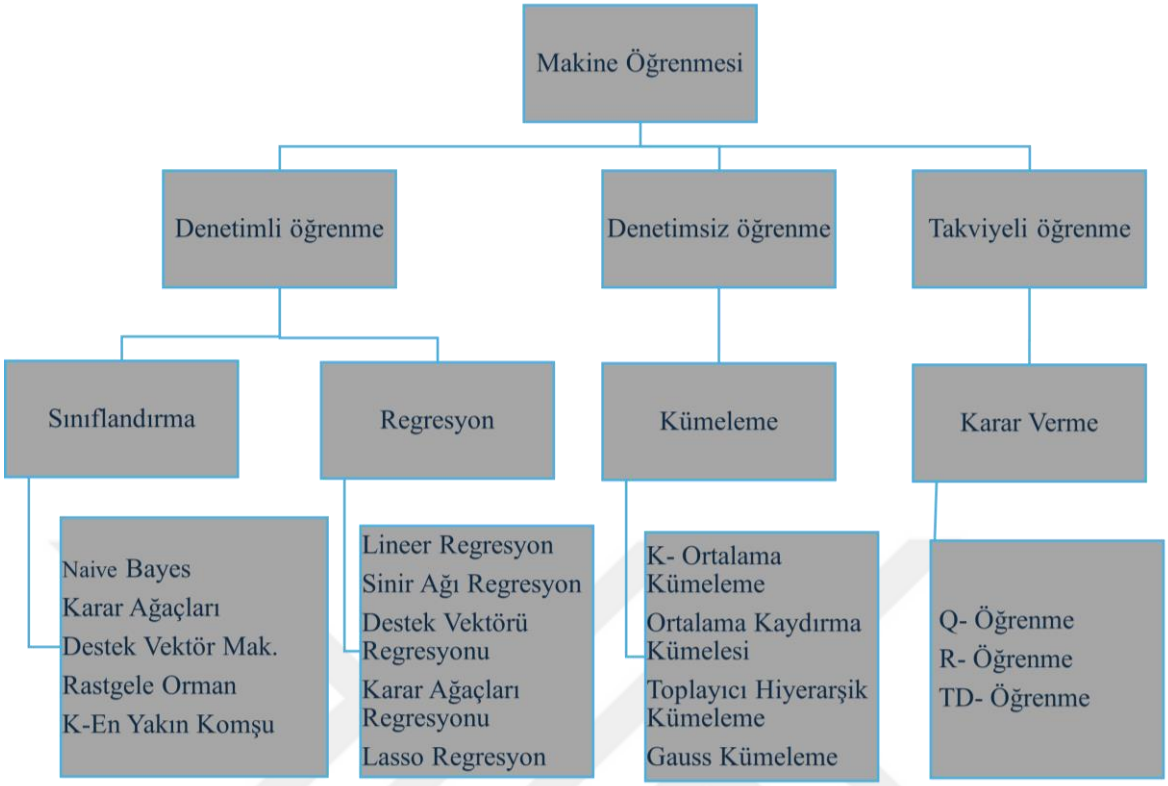
Şekil 2.1. Yapay zekâ, makine öğrenmesi, sinir ağları, derin öğrenme ve veri bilimi arasındaki ilişki

Makine öğrenmesi, sistemin geçmişteki deneyimlerinden elde edilen öğrenmelerini kullanarak bir model oluşturulmasına ve gelecekte karşılaşacağı durumlar karşısında bir tahminde bulunmasını sağlayan bir yapay zekâ alanıdır.

Makine öğrenmesi, insan zekasını taklit etmektedir. Ancak insanın yorumlayıp elle girdiği kurallara ihtiyaç duymayan algoritmalar bütünüdür.

İnsanlar, öğrenme işlevini hayatları boyunca yaşadıkları olaylar sonucunda çıkardıkları tecrübe ile gerçekleştirmektedir. Uzun süre tecrübe edinilen bir olayın sonucunu rahatlıkla kestirmek mümkün olmaktadır. Ticaret ile uğraşan insanların, tecrübeleriyle oluşturdukları müşteri profilleri sonucu hangi müşterinin hangi üründen hoşlandığını veya o ürünün alıp almayacağını kolaylıkla tahmin edebilir hale gelmesi buna örnek olarak verilebilir. Yapay zekâ ile gelinmek istenen nokta da budur. Ancak makinelerin öğrenme yöntemleri farklıdır.

Makine öğrenmesi algoritmaları, veri setlerindeki değişkenler arasındaki ilişkiyi belirlemek ve gelecekteki olayları tahmin etmek için kullanılan matematiksel modellerdir. Şekil 2.2'de makine öğrenmesi algoritmalarının genel bir şeması verilmiştir.

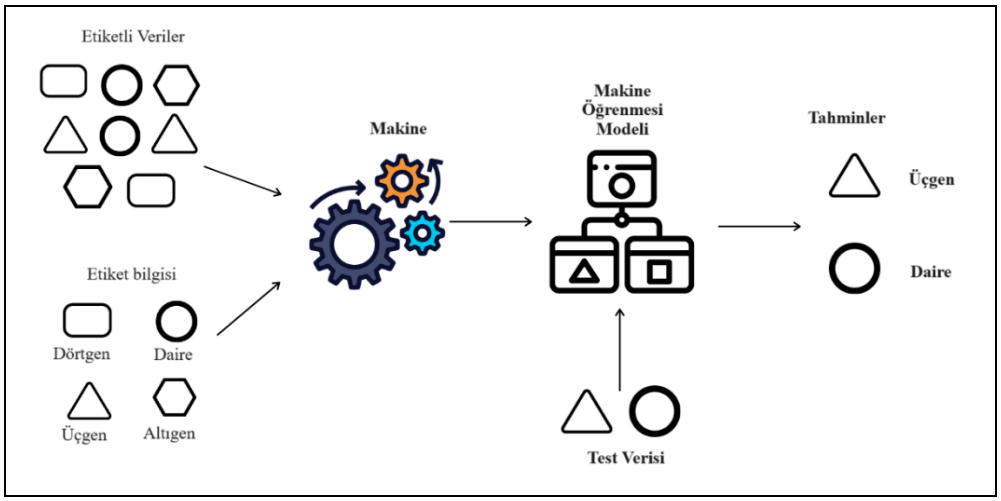


Şekil 2.2. Makine öğrenmesi hiyerarşisi

Su kalitesinin tahmininde kullanılan makine öğrenmesi algoritmaları ve bu algoritmaların etkinliğini değerlendirmek için kullanılan model değerlendirme yöntemleri detaylı bir şekilde incelenmiştir. Literatürde sıklıkla kullanılan algoritmaların teorik prensiplerine uygun olarak tekniklerin matematiksel formülleri ve çalışma koşulları aşağıda açıklanmıştır.

2.1. Denetimli Öğrenme

Belirli bir etiketli girdi verisinden belirli bir sonuç tahmin edilmek istendiğinde kullanılan bir öğrenme metodudur. Eğitim veri setini oluşturan bu girdi ve çıktı gözlemleri ile bir makine öğrenmesi modeli oluşturulmaktadır. Bu modellerin amacı daha önce hiç karşılaşmadıkları girdi verilerine göre sonucu başarılı tahmin etmektir. Eğitim seti ilk başta operatör tarafından oluşturulsa da sonraki işlemler model tarafından yapılmaktadır.



Şekil 2.3. Denetimli öğrenme modeli (URL-8, 2024)

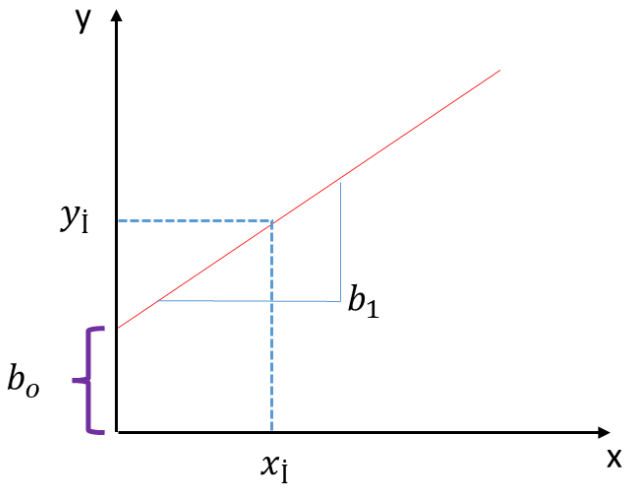
Şekil 2.3'te modeli verilen Denetimli öğrenme (Supervised learning), regresyon (regression) ve sınıflandırma (classification) problemlerine uygulanmaktadır.

2.1.1. Regresyon

Genellikle bağımlı ve bağımsız değişkenler arasındaki nedensel ilişkiyi tahmin etmek ve belirlemek için kullanılmaktadır. Sürekli değişkenlerin tahmininde avantaj sağlamaktadır. Çeşitli regresyon modelleri incelenmiştir.

2.1.1.1. Lineer regresyon

Doğrusal regresyon veya normal En Küçük Kareler (Ordinary Least Squares OLS) yöntemi, regresyon için en basit ve en klasik lineer yöntemdir. Şekil 2.4'te basit regresyon modeli verilmiştir.



Şekil 2.4. Lineer regresyon modeli

$$y_i = b_0 + b_1 x_i \tag{2.1}$$

Denklem (2.1)'de verilen tahmin modelinin formülündeki;

b_0: Doğrunun y eksenini kestiği nokta,

b_1: Doğrunun eğimi

y: Bağımlı değişken (tahmin edilen değer)

x: Bağımsız değişkendir.

2.1.1.2. Lasso regresyon

Lasso Regresyonun (Least Absolute Shrinkage and Selection Operator) amacı, regresyon katsayılarını sıfıra çekmek, küçültmek ve böylece model karmaşıklığını azaltmaktır (URL-6).

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j| \tag{2.2}$$

Lasso Regresyon modeli denklem (2.2)'deki katsayıları minimize etmektedir.

RSS, hata kareler toplamını ifade etmektedir, yani gözlemler arasındaki farkların karesinin toplamını ifade etmektedir.

p, modeldeki toplam özellik sayısını temsil etmektedir.

β_j, j'ninci özellik için regresyon katsayılarını ifade etmektedir.

λ, bir hiperparametredir ve sıfıra yakın olan katsayıları sıfıra çekerek düzenleştirmeyi kontrol etmektedir. λ ne kadar büyükse, sıfıra yakın katsayılar o kadar fazla sıfıra çekilmektedir.

2.1.2. Sınıflandırma

Veri setini farklı parametrelere ve koşullara dayalı olarak sınıflara ayırmayı hedefleyen bir algoritmadır. Literatürde sık kullanılan sınıflandırma algoritmaları aşağıda verilmiştir.

2.1.2.1. K-En yakın komşular

Bir sınıflandırma ve regresyon algoritması olan K-En Yakın Komşular (K-Nearest Neighbors) algoritması hedef noktayı tahmin veya sınıflandırma yapmak için en yakınındaki komşulara bakmaktadır. Genel çalışma prensibi aşağıdaki adımlardan oluşmaktadır.

Veri setindeki her değişkene ait gözlem birimi bir vektör olarak kabul edilmektedir. Sınıflandırma problemleri için, hedef veri etiketiyle regresyon için bir sayısal değer ile ilişkilendirilmektedir.

Sınıflandırma veya regresyon yapılmak istenen noktanın komşularını bulmak için veri noktaları arasındaki uzaklıklar hesaplanmaktadır. Bu uzaklıklar; Öklid Mesafesi, Manhattan Mesafesi ve Minkowski mesafesi olarak üç temel formülle ile bulunmaktadır. Denklem (2.3) de Öklid mesafesi hesaplama formülü verilmiştir.

$$d = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{2.3}$$

Denklem (2.3)'deki;

d mesafeyi,

k aralarındaki mesafe hesaplanacak olan toplam nokta sayısını,

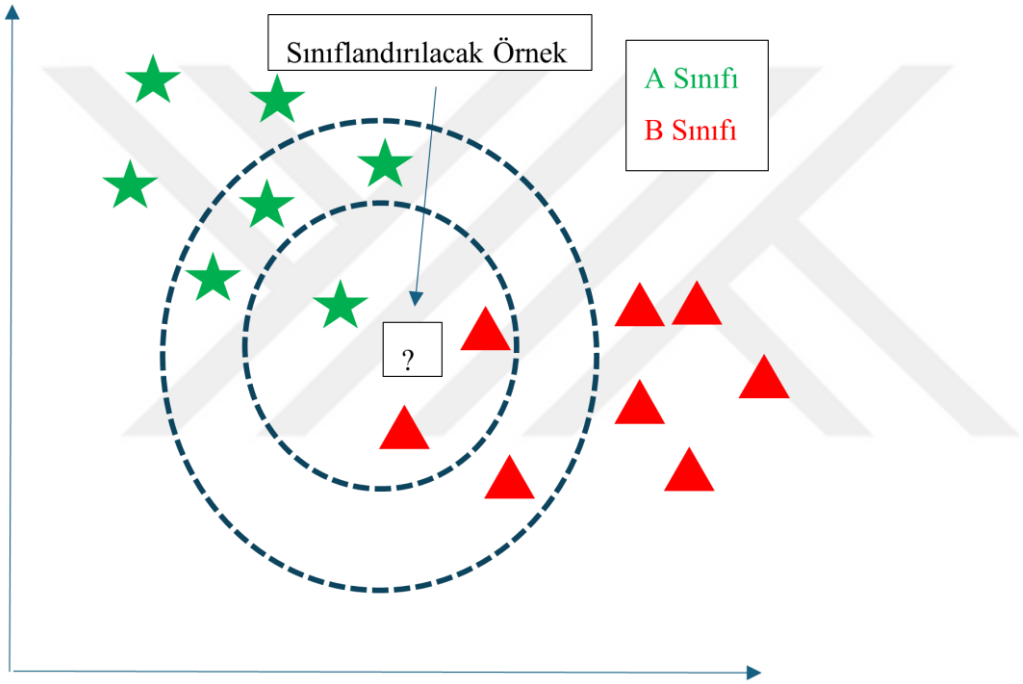
x_i ve y_i ise aralarındaki mesafe hesaplanacak noktaları ifade etmektedir.

Hesaplanan mesafeler küçükten büyüğe doğru sıralanıp, belirlenen K değeri kadar en küçük uzaklığa sahip komşular seçilmektedir.

Sınıflandırma yapılırken, en yakın K komşunun sınıf etiketleri arasında çoğunluk sınıfı alınır ve bu sınıf tahmin olarak kullanılmaktadır.

Modelin son adımı olarak, sınıflandırma için algoritmanın doğruluğu, regresyon için tahminlerin doğruluğu değerlendirilmektedir. Değerlendirmede test veri seti veya çapraz doğrulama (cross validation) gibi parametreler kullanılmaktadır.

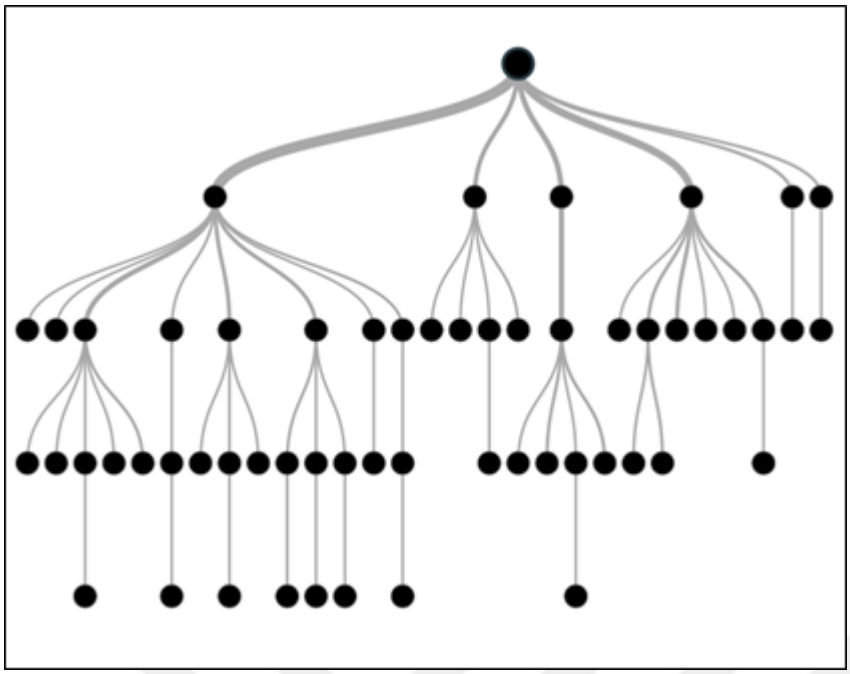
Şekil 2.5'te KNN algoritmasının sınıflandırma yapısı verilmiştir.



Şekil 2.5. KNN sınıflandırma yapısı (Patro ve ark., 2023)

2.1.2.2. Karar ağaçları

Karar ağaçları (Decision trees) algoritması, veri setinin küçük parçalara, dallara ayırıp her bir dalda tahmin işlemi yapılarak çalışmaktadır. Bu tahminler ağaç dallarındaki koşullu ifadeler kullanarak yapılmaktadır. Şekil 2.6'da karar ağaçlarının genel yapısı verilmiştir.



Şekil 2.6. Karar ağaçları yapısı (Pandey ve ark., 2022)

Genel çalışma prensibi aşağıdaki adımlardan oluşmaktadır.

Birinci aşamada bir veri seti ile ağacı eğitmektir. Veri setindeki her gözlem sınıflandırma için bir hedef değişken veya regresyon için bir hedef değer ile ilişkilendirilmektedir

Karar ağacı oluştururken her bir düğüm için hangi özelliğin kullanılacağına karar verilmesi gerekmektedir. Bu özellikler düğümlerin bölünmesi için gereklidir. En sık kullanılan bölme kriterleri, sınıflandırma işlemleri için Gini indeksi veya Entropi, regresyon işlemleri için ise ortalama kare hatasıdır (URL-1, 2022).

Gini indeksi için gerekli olan formül denklem (2.4) de verilmiştir (Hastie ve ark., 2009).

$$I_G = 1 - \sum_{i=1}^{n} (p_i)^2 \qquad (2.4)$$

Denklem (2.4) de;

n sınıf sayısı,

i sınıf,

p_i sınıf i için oranı ifade etmektedir.

Entropi için gerekli olan hesaplama formülü denklem (2.5) de verilmiştir (Hastie ve ark., 2009)

$$H = -\sum_{i=1}^{n} p_i \log_2(p_{i)} \quad (2.5)$$

Denklem (2.5) de *H* entropiyi, *i* sınıf, p_i sınıf *i* için oranı ifade etmektedir.

Ortalama kare hatası için gerekli olan hesaplama formülü denklem (2.6) da verilmiştir.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \quad (2.6)$$

Denklem (2.6) de;

y_i veri noktasının gerçek değeri,

$\overline{y}$ ise düğümdeki veri noktalarının ortalama değerini,

n ise sınıf sayısını ifade etmektedir.

Kullanılacak özellik seçildikten sonra bu özelliğe göre bir eşik değer seçilmektedir. Bu eşik değere göre bölme işlemi yapılmaktadır.

Özellik seçimi ve bölme noktası belirlendikten sonra, bu işlem, veri kümesi tamamen ayrılana kadar veya belirli bir duruma ulaşılana kadar devam etmektedir.

Ağacın aşırı öğrenme (overfit) olmasını engellemek için gerekli parametreler uygulanmaktadır.

Son adımda tahmin işlemi yapılmaktadır.

Karar ağaçların farklı türden yaklaşım ve hesaplar ile oluşturulan algoritmaları mevcuttur. Bunlardan sık kullanılanları aşağıda verilmiştir.

Sınıflandırma ve regresyon ağaçları: Bu algoritma CART (Clasification and regression trees) olarak da adlandırılmaktadır. Bu algoritma Gini impurity (kirlilik) veya Ortalama Kare Hata (MSE) gibi bölme kriterlerini kullanır. Her bir bölme noktası, veri setini iki alt kümeye bölmektedir.

Rastgele orman: Bu algoritma literatürde RF (Random Forest) olarak da bilinmektedir. Algoritma birden fazla bağımsız karar aracı oluşturup bunları bir araya getirip arasında en yüksek değerli olanın seçilmesi mantığı ile çalışmaktadır. Karar ağaçları ile arasındaki temel fark bölmenin rastgele olmasıdır.

Gradyan güçlendirilmiş ağaçlar: İngilizce Gradient Boosted Machine (GBM) olarak bilinen bu algoritma birçok zayıf tahminci ağacın bir araya gelerek güçlü bir tahminci oluşturmaya çalışmasıdır. Her bir ağaç diğerinin hatalarını düzeltmektedir.

2.1.2.3. Naive bayes

Bu algoritma Bayes teoreminin sınıflandırma işlemlerine uyarlanması mantığına dayanmaktadır. Temel çalışma prensibi aşağıda verilmiştir.

Veriler toplanıp veri seti haline getirilmektedir. Eğitim için bu veri seti kullanılmaktadır.

Her bir sınıf için, özelliklerin normal dağılımını modellemek üzere ortalama ve varyans hesaplanmaktadır. Bu, her bir özellik ve her bir sınıf için ayrı ayrı yapılmaktadır.

Bayes teoremi kullanılarak sınıf tahmini yapılmaktadır. Bayes teoremine ait formül denklem (2.7)'de verilmiştir.

$$P(A\backslash B) = \frac{P(B\backslash A)\,P(A)}{P(B)} \tag{2.7}$$

P(A\B): *B* olayı gerçekleştiği zaman *A* olayının gerçekleşme olasılığıdır.

P(A): *A* olayının gerçekleşme olasılığıdır.

P(B\A): *A* olayı gerçekleştiği zaman *B* olayının gerçekleşme olasılığıdır.

P(B): *B* olayının gerçekleşme olasılığıdır.

Bu hesaplanmış olasılıklar daha sonra test verilerinin sınıflarının tahmin edilmesinde kullanılmaktadır (URL-2, 2014).

Bayes teoremi kullanılarak, test örneği için her bir sınıfın olasılığı hesaplanmaktadır. En yüksek olasılığa sahip sınıfını tahmin olarak seçmektedir.

2.1.2.4. Destek vektör makinaları

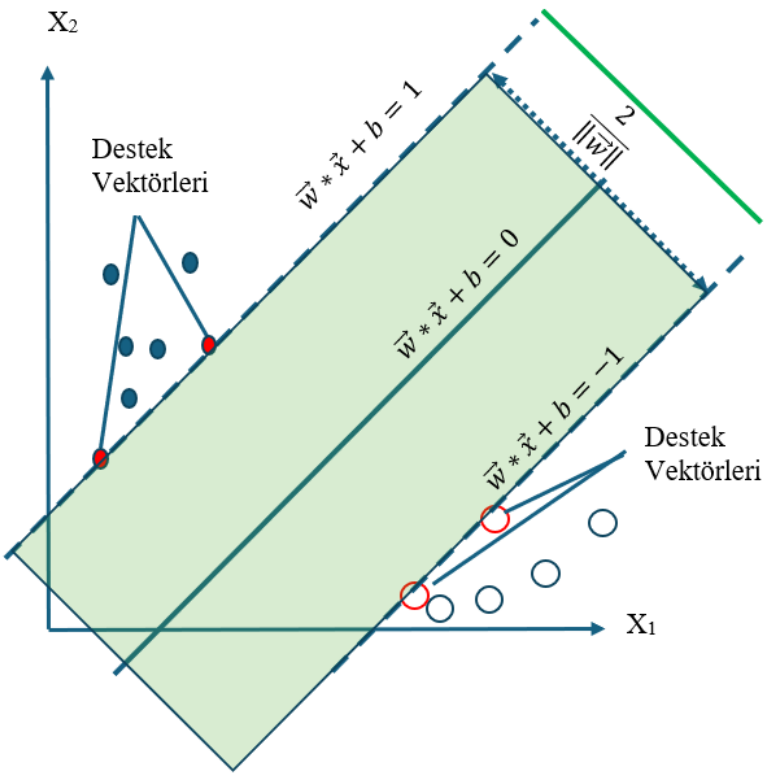
Destek Vektör Makineleri (Support Vector Machine) genellikle sınıflandırma problemlerinde kullanılan gözetimli öğrenme yöntemlerinden biridir. Bir düzlem üzerine yerleştirilmiş noktaları ayırmak için bir doğru çizilir. Bu doğrunun, iki sınıfının noktaları

için de maksimum uzaklıkta olmasını amaçlar. Karmaşık ama küçük ve orta ölçekteki veri setleri için uygundur.

Temel çalışma prensibi şu şekildedir;

Eğitim veri setini kullanarak SVM modelini eğitmek ilk adım olarak gerçekleşmektedir.

SVM, özellik uzayında veri noktalarını bölecek en iyi hiper düzlemi (margin çizgisi) bulmayı hedeflemektedir.



Şekil 2.7. Margin çizgisi

Şekil 2.7'de mavi ve beyazlar olmak üzere iki farklı sınıf vardır. Sınıflandırma problemlerindeki temel hedef, verinin doğru sınıfa atanabilmesidir. Doğru sınıflandırmayı başarmak için sınıfları ayıran bir doğru çizilir ve bu doğrunun ±1 arasında kalan yeşil bölgeye Margin denir. Margin ne kadar geniş olursa, sınıflar o kadar iyi ayrıştırılmış olur. (URL-3, 2022).

$$\hat{y} = \begin{cases} 0, eğer\ w^T * x + b < 0 \\ 1, eğer\ w^T * x + b \geq 0 \end{cases}$$

(2.8)

Denklem (2.8)'de;

w: ağırlık vektörü

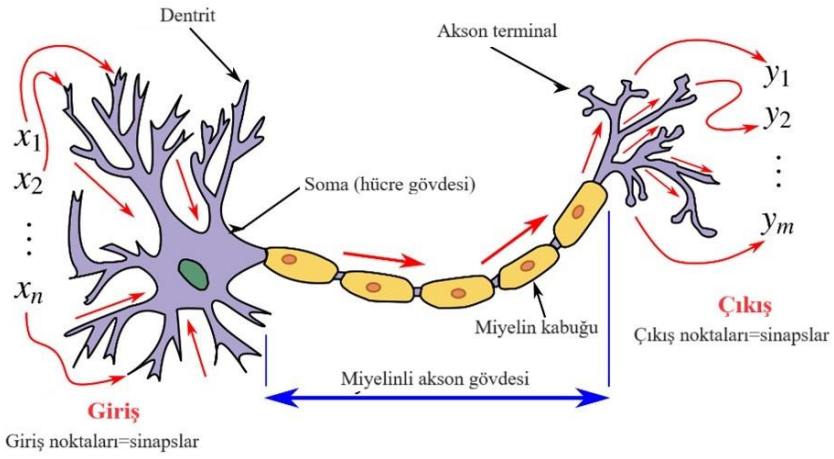
x: girdi vektörü

b: sapma

Eğer çıkan sonuç sıfırdan küçükse, yeni bir değer beyaz noktalara daha yakın olacaktır. Tersine, sonuç sıfıra eşit veya büyükse, yeni değer mavi noktalara daha yakın olacaktır.

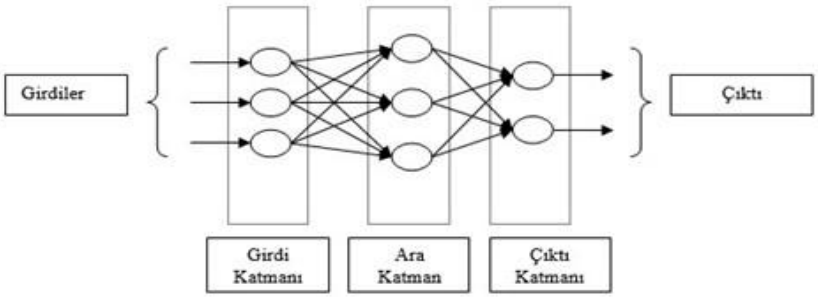
2.1.2.5. Yapay sinir ağları

Yapay Sinir Ağlarının (YSA) temeli yapısında bulunan yapay nöronlardır. Bu yapay nöronlar biyolojik bir nörona benzer prensip ile çalışırlar. Ağırlıklı girdiler yapay nöronların gövdesine iletilir. Gövde, ağırlıklı girdileri ve bias değerlerini toplar. Daha sonra bu değer tanımlanmış bir transfer fonksiyonuna göre işlenir (Hatipoğlu, 2016).

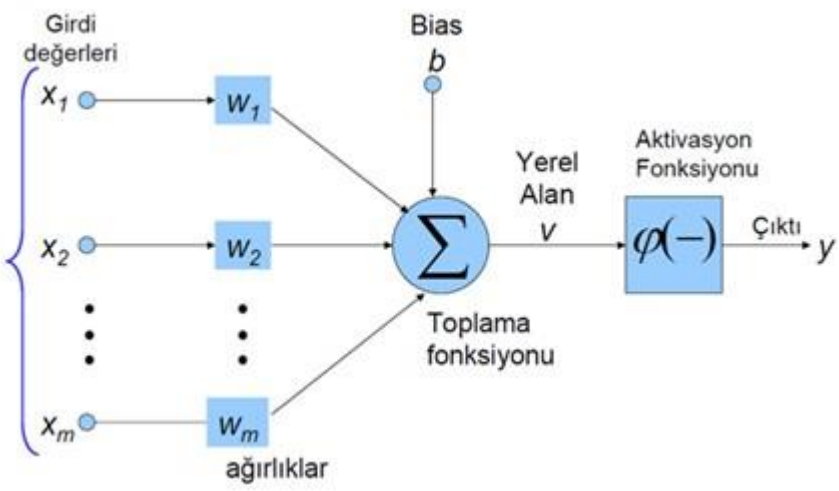


Şekil 2.8. Biyolojik nöron yapısı (URL-4, 2024)

Şekil 2.8'de verilmiş olan biyolojik nöron yapısına benzer bir yapı Şekil 2.9'da YSA ile verilmiştir.



Şekil 2.9. Yapay sinir ağı yapısı



Şekil 2.10. Yapay sinir ağı hücresine ait matematiksel modelleme (Keskenler, 2017)

$$v = \sum_{i=0}^{n} W_i x_i + b, \qquad y = \varphi(v) \tag{2.9}$$

Denklem (2.9) da;

W, hücrenin ağırlık matrisini,

x, hücrenin giriş vektörünü,

b, bias değerini,

y hücrenin çıkışını,

φ(v) aktivasyon fonksiyonunu ifade etmektedir.

Şekil 2.10'da verilen YSA'nın genel olarak şu bileşenleri vardır:

Girdi Katmanı: Verilerin giriş olarak verildiği katmandır.

Ağırlıklar: Alınan verilerin bir katsayı ile çarpıldığı katmandır. Giriş verilerinin yapay nörona bağımlılığını büyük ölçüde belirler. Başlangıç değerleri verilebileceği gibi

rastgele değerler de verilebilir. Öğrenme açısından önemli bir yer teşkil eder. Başka bir deyişle ağırlıkların en iyi olması, ağın öğrenmesi anlamına da gelmektedir.

Bias: Aktivasyon fonksiyonuna ekstra bir sinyal ekleyen bağımsız bir değişkendir. Doğrusal olmayan bir sistem oluşturulmasında yardımcı rol oynar. Aktivasyon fonksiyonunun sonucunu pozitif veya negatif tarafa kaydırmak için kullanılır.

Toplama fonksiyonu: Toplama fonksiyonu, bir yapay nörona giriş olarak gelen girdileri ağırlıklarla çarparak o hücrenin net girdisini hesaplayan bir fonksiyondur. Toplama fonksiyonu çarpım, en büyük ve en küçük gibi fonksiyonlardan da oluşabilir.

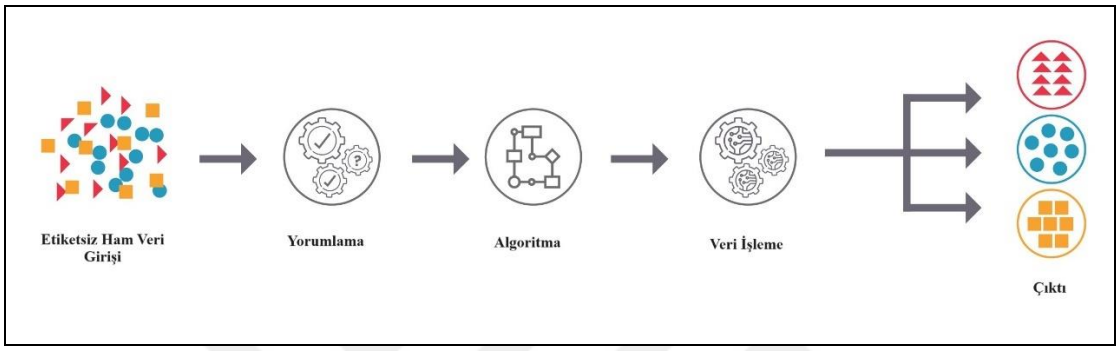
Aktivasyon fonksiyonu: Net girdiye karşı bir çıkış üreten fonksiyondur. Türevlenebilir bir fonksiyon olmasına dikkat edilmelidir. Çünkü geri beslemeli ağlarda aktivasyon fonksiyonun türevi de kullanılmaktadır. Aktivasyon fonksiyonu doğrusal olmayan bir sistem oluşturulmasına katkı sunar (Mahesh, 2020). Tablo 2.1'de yaygın kullanılan aktivasyon fonksiyonlarının formülleri ve aralıkları verilmiştir.

Tablo 2.1. Yaygın kullanılan aktivasyon fonksiyonları

Aktivasyon Fonksiyonu	Matematiksel Formülü	Aralığı
Doğrusal fonksiyon	$\varphi(v) = v$	(-∞,∞)
Basamak fonksiyonu	$\varphi(v) = \begin{cases} 0 \; için \; v < 0 \\ 1 \; için \; v \geq 0 \end{cases}$	{0,1}
Sigmoid fonksiyonu	$\varphi(v) = \sigma(v) = \frac{1}{1 + e^{-x}}$	(0,1)
Hiperbolik tanjant fonksiyonu	$\varphi(v) = \tanh(v) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$	(-1,1)
ReLu	$\varphi(v) = \begin{cases} 0 \; için \; v < 0 \\ v \; için \; v \geq 0 \end{cases}$	[0,∞)

2.2. Denetimsiz Öğrenme

Denetimsiz öğrenmenin amacı, etiketlenmemiş verilerdeki saklı örüntüleri bulmaktır. Verilerin etiketli olmadığı problemlerde denetimsiz öğrenme metotları kullanılmaktadır. Ancak veriler etiketsiz olduğundan çıktılara kesinlikle doğru denilmemektedir. Şekil 2.11'de denetimsiz öğrenme modelinin genel yapısı verilmiştir.



Şekil 2.11. Denetimsiz öğrenme modeli

Denetimsiz öğrenme problemlerinde genellikle veriler arasındaki ilişki yorumlanarak kümeleme (clustering) işlemi yapılmaktadır.

2.2.1. Kümeleme

Kümeleme (clustering), benzerlikleri içindeki veri noktalarının birbirine daha yakın, farklılıkları ise diğer gruplardaki veri noktalarından daha fazla olan birkaç gruba ayırma algoritmasıdır. Esas olarak, verilerin benzerlik ve farklılıklarına göre analiz ederek gruplara ayırır (URL-7, 2024). Literatürde sıklıkla kullanılan kümeleme algoritmaları aşağıda detaylı incelenmiştir.

2.2.1.1. K- ortalama kümeleme

K-ortalama (K-means) algoritması, en popüler kümeleme yöntemlerinden biridir. K-ortalama algoritması, bir dizi örneği, her biri kümedeki örneklerin ortalamasıyla tanımlanan ayrık kümelere böler. Tüm değişkenlerin nicel (kantitatif) tipte olduğu ve Öklid

mesafesinin karesinin alındığı durumlar için tasarlanmıştır. Denklem (2.10) ve denklem (2.11) ile hesaplama formülü verilmiştir.

Her biri n boyutlu reel vektör kümesi $\{x_1, x_2, \ldots, x_n\}$ veri kümesi olmak üzere ve K bölünecek küme sayısı olmak üzere, K- ortalama kümeleme karesel hatayı en aza indirmek için n tane veriyi K adet $S=\{S_1, S_2, \ldots, S_K\}$ kümeye bölme işlemini amaçlamaktadır.

$$\mu_i = \frac{1}{|S_j|} \sum_{x_i \epsilon S_j} x_i \tag{2.10}$$

Denklem (2.10)'da verilen μ_j, S_j'deki noktaların ortalaması olmak üzere;

$$argmin \sum_{j=1}^{K} \sum_{x_i \epsilon S_j} \|x_i - \mu_j\|^2 \tag{2.11}$$

Denklem (2.11) ile karesel hatayı en küçük yapacak *K* adet kümeyi bulmaktır. Algoritma temel olarak aşağıdaki 4 aşamadan oluşur:

1. Rastgele veya belirli bir yöntemle K adet başlangıç küme merkezi seçilmektedir.

2. Her veri noktası, en yakın küme merkezine atanmaktadır. Bu adım, her veri noktasının hangi kümede yer alacağını belirlemektedir.

3. Her kümenin merkezi, o kümedeki veri noktalarının ortalaması alınarak güncellenmektedir. Bu yeni merkezler, kümenin ortalamasını temsil etmektedir.

4. Kararlı hale gelinene kadar 2. ve 3. adımların tekrarlanması. Veri noktalarının kümelere atanması ve küme merkezlerinin güncellenmesi adımları, küme merkezleri değişmeyene kadar veya belirli bir sayıda yineleme yapılana kadar tekrarlanmaktadır.

2.2.1.2. Ortalama kaydırma ile kümeleme

Ortalama kaydırma (mean shift) kümeleme algoritması veri kümesi üzerinde veri dağılımının en yoğun olduğu yeri bulmayı amaçlamaktadır. Veri kümesi üzerinde r yarıçaplı bir daire oluşturulmaktadır. Bu daire içerisinde yer alan noktaların ağırlık merkezleri hesaplanmaktadır. Bulunan yeni ağırlık merkezine daire kaydırılmakta bu işlem merkez değişmeyene kadar tekrar edilmektedir.

Küme sayısını otomatik olarak tespit ettiği için ilk başta küme sayısı belirlemeye gerek yoktur. Algoritma genellikle aşağıdaki adımları izler:

Başlangıçta, her veri noktası için bir pencere belirlenir ve bu pencere merkezi ile başlatılır.

Her iterasyonda, her veri noktası için pencere merkezi, o noktanın etrafındaki yoğunluğun ağırlıklı ortalaması olarak güncellenir.

Her iterasyonda, pencere merkezleri, veri noktalarının etrafındaki yoğunluğun en yüksek olduğu bölgeye doğru kaydırılır.

Algoritma, pencere merkezlerinin konverjansa ulaştığı veya belirli bir iterasyon limitine ulaşıldığında durur.

Son olarak, her veri noktası, en yakın pencere merkezine atanır ve bu merkezler küme merkezlerini oluşturur.

Temel olarak, bir veri noktasının etrafındaki yoğunluk, bir çekirdek fonksiyonu (kernel function) aracılığıyla belirlenir. Bu çekirdek fonksiyonu, veri noktasının merkezine ne kadar yakınsa, o noktanın yoğunluğunu o kadar artırır.

Matematiksel olarak, bir veri noktasının yeni konumu x_i^t (*t* zamanındaki **i.** veri noktası) denklem (2.12)'de verilmiştir (Derpanis, 2005).

$$x_i^{t+1} = \frac{\sum_{j=1}^{n} K\left(\frac{x_i^t - x_j}{h}\right) x_j}{\sum_{j=1}^{n} K\left(\frac{x_i^t - x_j}{h}\right)} \qquad (2.12)$$

Denklem (2.12) deki;

x_i^{t+1}, *(t+1)* zamanındaki *i.* veri noktasının yeni konumunu,

x_i^t, *t* zamanındaki *i.* veri noktasının mevcut konumunu,

x_j , *j.* veri noktasını,

K, çekirdek fonksiyonunu,

h, çekirdek genişliğini (bandwidth) ve

n, veri noktalarının toplam sayısını temsil eder.

Mean Shift algoritmasında kullanılan çekirdek fonksiyonu, genellikle bir yoğunluk tahmini yapmak için kullanılmaktadır. Bu çekirdek fonksiyonu, bir veri noktasının etrafındaki yoğunluğu belirlemek için kullanılan bir ağırlık fonksiyonudur. En yaygın olarak kullanılan çekirdek fonksiyonlarından biri, Gauss (normal) dağılımına dayalı bir fonksiyondur. Denklem (2.13)'te Gauss dağılımına ait çekirdek fonksiyonunun formülü verilmiştir.

$$K(x) = \frac{1}{\sqrt{2\pi}.h} e^{-\frac{x^2}{2h^2}} \tag{2.13}$$

Denklem (2.13)'te;

x, iki veri noktası arasındaki mesafeyi,

h, bant genişliğini (bandwidth) ifade etmektedir. Dağılımın genişliğini kontrol etmektedir.

2.2.1.3. Toplayıcı hiyerarşik kümeleme

Toplayıcı hiyerarşik kümelemede, kullanıcının iki gruptaki gözlemler arasındaki ikili farklılıklara dayalı olarak farklı gözlem grupları arasındaki bir farklılık ölçüsünü belirlemesini gerektirmektedir. Adından da anlaşılacağı gibi, hiyerarşinin her seviyesindeki kümelerin bir sonraki alt seviyedeki kümelerin birleştirilmesiyle oluşturulduğu hiyerarşik temsiller üretmektedirler. En düşük düzeyde, her küme tek bir gözlem içermektedir. En üst düzeyde, tüm verileri içeren tek bir küme bulunmaktadır (Hastie ve ark., 2009).

Kısaca, toplayıcı hiyerarşik kümeleme (Agglomerative hierarchical clustering), veri noktalarını bir araya getirerek bir kümeleme yapma yöntemidir. Bu yöntemde, her veri noktası kendi kümesinden başlamakta ve ardından benzerlik ölçütlerine bağlı olarak kümeleme işlemi gerçekleştirilmektedir. Tüm veri noktaları tek bir büyük kümede birleştirildiğinde işlem sonuçlanmaktadır.

Bu algoritma, veri görselleştirme, keşifçi veri analizlerinde, sınıflandırma gibi birçok alanda kullanılmaktadır. Algoritmanın temel formülasyonu veri noktalarının birbirine olan benzerliklerine dayanmaktadır. Bu benzerlikler Öklid veya Manhattan mesafesi formülleri ile hesaplanmaktadır.

2.2.1.4. Gauss kümeleme

Yöntem veri noktalarının birbirilerine olan benzerliklerine göre kümeleme işlemi yapmaktadır. Kümeleme işlemini Gauss (normal) dağılımına göre gerçekleştirmektedir. Her bir Gauss dağılımı, bir küme olarak kabul edilmektedir.

$$p(x) = \sum_{i=1}^{k} \pi_i . N(x \mid \mu_i, \sum_i) \quad (2.14)$$

Denklem (2.14)'te Gauss kümeleme modelinin olasılık yoğunluk fonksiyonunun hesaplanma formülü verilmiştir. Denklem (2.14)'te;

$p(x)$, veri noktası x için olasılık yoğunluk fonksiyonunu,

k, toplam bileşen sayısını (küme sayısını),

π_i, i. bileşenin ağırlığını (küme ağırlığını),

$N(x/\mu_i, \Sigma_i)$, çok boyutlu Gauss (normal) dağılımını,

μ_i, ortalama (merkez) vektörünü, Σ_i, kovaryans matrisini ifade etmektedir.

Çok boyutlu Gauss dağılımı aşağıdaki denklem (2.15) ile hesaplanmaktadır.

$$N(x \mid \mu_i, \sum_i) = \frac{1}{2\pi^{d/2} |\sum_i|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \sum_i^{-1} (x - \mu_i)\right) \quad (2.15)$$

d, veri noktasının boyutunu (özellik sayısını),

$| \Sigma_i |$, Σ_i kovaryans matrisinin determinantını ifade etmektedir.

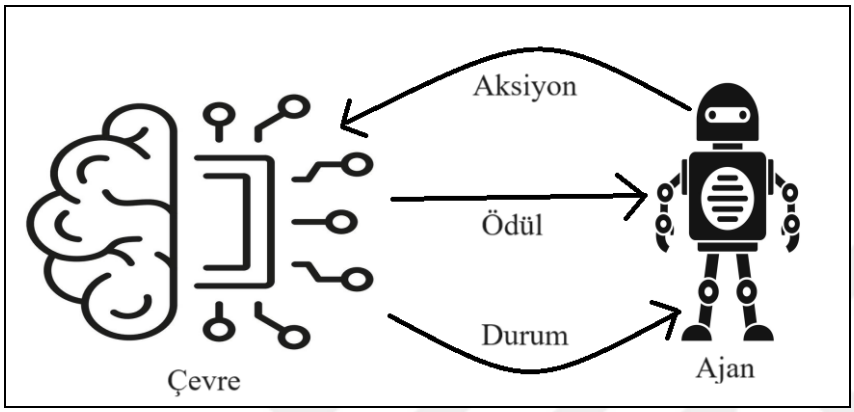
Veri setindeki dağılımı en iyi şekilde modellemek için ortalama, kovaryans ve ağırlık merkezini tahmin edilerek Gauss kümele yapılmaktadır.

2.3. Takviyeli Öğrenme

Herhangi bir ortamda, bilgisayar ile iletişimli bir robot kendi kontrol politikasına dayalı olarak bir eylem gerçekleştirmektedir. Bu eylem yürüme eylemi olarak düşünülebilir. Daha sonra durumlar güncellendikçe başarıyla gerçekleşen bu eylemin değerlendirilmesi, bir "ödül" olarak verilmektedir. Takviyeli öğrenmede bu şekilde performansı üst düzeye çıkarmak için geri bildirimler verilerek sistem eğitilmektedir (Sugiyama, 2015).

Şekil 2.12'de bu durum temsil edilmektedir. Takviyeli öğrenmede, robot (ajan) çevre ile iletişim kurarak kendi programına göre belirli bir görevi en iyi şekilde gerçekleştirmeyi öğrenmeyi amaçlamaktadır.

Karmaşık ve dinamik problemlerin çözümü için kullanılan güçlü bir makine öğrenmesi metodudur. Oyun teknolojisi, otonom araçlar, robotik alanlarda, sağlık gibi birçok alanda sıklıkla kullanılmaktadır.



Şekil 2.12. Takviyeli öğrenme (Sugiyama, 2015)

Sıklıkla kullanılan takviyeli öğrenme metotları aşağıda verilmiştir.

2.3.1. Q-Öğrenme

Watkins tarafından ilk defa 1989'da önerilmiş ve değer fonksiyonu için Q harfinin kullanılmasından dolayı yöntem bu isim ile kullanılmaktadır. Algoritmanın matematiksel modelinde kullanılan değer fonksiyonu kolayca belirlenemediğinden anlık duruma bağlı ödüller kullanılmaktadır (Watkins ve Dayan, 1992).

Q-Öğrenme algoritmasının matematiksel formülü Bellman denklemine dayanmaktadır. Bellman denklemi bir durumda için bir aksiyonun toplam ödül miktarını belirlemek için kullanılmaktadır. Tablo 2.2. de örnek bir durum tablosu verilmektedir.

Tablo 2.2. Durum tablosu

Q Tablosu	
Durum (state)-Aksiyon (action)	Ödül Değeri (Q value)

s,a	v1
s,a	v2
s,a	v3
s,a	v4

Tablo 2.2'deki durum ve ödül değerleri göz önüne alınarak aşağıdaki şekilde güncel *Q* değeri Denklem (2.16) ile hesaplanmaktadır.

$$Q(s,a) = (1-\alpha).Q(s,a) + \alpha.[r + \gamma.max_{a'}Q(s',a')] \quad (2.16)$$

Denklem (2.16)'da;

Q(s,a), ajanın durum *s* içinde aksiyon *a* için tahmin edilen *Q* değeridir. Beklenen toplam ödül miktarını ifade etmektedir.

r, anlık ödülü ifade etmektedir. Ajanın bir durumda bir eylemi gerçekleştirdikten sonra aldığı anlık ödüldü göstermektedir.

s′, bir sonraki durumu ifade etmektedir. Ajanın bir durumda bir eylemi gerçekleştirdikten sonra ulaştığı sonraki durumdur.

a′, bir sonraki aksiyonu ifade etmektedir. Ajan sonraki görevler arasında en iyisini seçmektedir.

α, öğrenme oranını ifade etmektedir. Ajanın mevcut tahminlerin güncellenme oranını kontrol etmektedir.

γ, indirim faktörü olarak ifade edilmektedir. Gelecekteki ödüllerin mevcuttaki ödüllere göre nasıl değerlendirileceğini kontrol eder. Yüksek bir *γ* değeri, uzun vadeli ödüllerin daha fazla önemli olduğunu belirtmektedir.

Q-learning algoritması, bu formülü kullanarak ajanın deneyimlerinden öğrenmesini ve en iyi eylemi seçmesini sağlamaktadır.

2.3.2. TD-Öğrenme

Temporal Difference (Zamansal Fark) kelimesinin kısaltması olan TD ile ifade edilmektedir. Q-öğrenme algoritmasında delta kuralı kullanılarak TD-öğrenme adıyla bir algoritma 1988 de Sutton tarafından literatüre sunulmuştur. (Sutton ve ark., 2008)

Delta kuralı, yeni hesaplanan Q değeri ile mevcut Q değeri arasındaki farkı belirtir ve bu farka göre Q değerlerini güncellemektedir. Denklem (2.17)'de delta kuralı ile Q değeri hesaplanmaktadır.

$$\Delta Q(s,a) = \alpha.[r + \gamma.max_{a'} Q(s',a') - Q(s,a)] \quad (2.17)$$

Denklem (2.17)'de;

$\Delta Q(s,a)$, güncellenen Q değerini ifade etmektedir.

α, öğrenme hızını ifade etmektedir.

r, anlık ödülü ifade etmektedir. Ajanın bir durumda bir eylemi gerçekleştirdikten sonra aldığı anlık ödüldü göstermektedir.

γ, indirim faktörü olarak ifade edilmektedir. Gelecekteki ödüllerin mevcuttaki ödüllere göre nasıl değerlendirileceğini kontrol eder. Yüksek bir γ değeri, uzun vadeli ödüllerin daha fazla önemli olduğunu belirtmektedir.

$max_{a'} Q(s',a')$, bir sonraki durumda ajanın gerçekleştirebileceği en iyi eylemin Q değerini ifade etmektedir.

$Q(s,a)$, mevcut durum ve eyleme karşılık gelen mevcut Q değerini göstermektedir.

TD-Öğrenme algoritmasının güncelleme formülü Denklem (2.18) ile hesaplanmaktadır.

$$V(s_t) = V(s_t) + \alpha.[r_{t+1} + \gamma.V(s_{t+1}) - V(s_t)] \quad (2.18)$$

Denklem (2.18)'de;

$V(s_t)$, durum s_t için ajanın tahmin edilen değerini ifade etmektedir.

α, öğrenme hızını ifade etmektedir.

r_{t+1}, ajanın durum s_t de bir eylem gerçekleştirdikten sonra aldığı anlık ödüldü ifade etmektedir.

γ, indirim faktörü olarak ifade edilmektedir. Gelecekteki ödüllerin mevcuttaki ödüllere göre nasıl değerlendirileceğini kontrol eder. Yüksek bir γ değeri, uzun vadeli ödüllerin daha fazla önemli olduğunu belirtmektedir.

$V(s_{t+1})$, bir sonraki durumun tahmin edilen değerini göstermektedir.

3. MATERYAL VE METOT

Su kalitesinin tahmin edilmesi için makine öğrenmesi algoritmaları kullanılarak çalışma yapıldı. Çalışmanın yöntemleri ve basamakları aşağıda açıklanmaktadır.

Su kalitesinin tahmin edilmesi çevresel etki, koruma ve sürdürülebilirlik açısından önemlidir. Su kalitesini etkileyen faktörler çok çeşitli ve karmaşık olduğundan makine öğrenmesi algoritmaları kullanılarak su kalitesinin tahmin edilmesi ve sonuçlardaki başarı oranı bu çalışma ile değerlendirildi.

3.1. Veri Setinin Kaynağı ve Özellikleri

Veri seti, Güney Avustralya Hükümeti veri tabanından alınarak çalışma yapıldı. Güney Avustralya Hükümeti adına metropol ve kırsal alanlardaki içme suyu kaynaklarını SA Water kuruluşu işletmektedir. Bir yıl içerisinde 350000'den fazla örnek analiz edilerek sonuçlar üretilmektedir. Bu sonuçlar kimyasal, biyolojik ve fiziksel olarak incelenmektedir. Yapılan bu analizler, ISO 9001 kalite sistemleri, Ulusal Test Otoriterleri Birliği şartnamelerine uygun olarak yapılmaktadır (URL-9, 2024).

Kullanılan veri seti SA Water kuruluşu tarafından toplanan 2023 yılına ait verilerden oluşmaktadır (URL-10, 2024).

Veri seti çeşitli parametreler ile ilişkilendirildi. Kullanılan kalite parametre değerleri, Dünya Sağlık Örgütü (WHO), Amerika Birleşik Devletleri Çevre Koruma Kurumu (EPA), Avrupa Birliği (EU), Türk Standartları (TS 266), Uluslararası İçme Suyu Standartları, Türkiye Cumhuriyeti Sağlık Bakanlığı'nın standartlarına ve Türk Gıda Kodeksine uygun olarak seçildi. Bu parametreler Alüminyum (Aluminium), Amonyum (Ammonia), Demir (Iron), Kalsiyum (Calcium), Klorür (Chloride), Mangan (Manganese), Sülfat (Sulphate), Hidrojen İyon konsantrasyonu (pH), Renk (Colour), Bulanıklık (Turbidity) şeklinde sıralanmaktadır.

Su kalitesi üzerindeki etkilerini değerlendirmek için, parametrelerin özellikleri aşağıda açıklanmaktadır.

Alüminyum (Aluminium): İçme suyu kaynaklarında çok fazla miktarda Alüminyum bulunması insan sağlığı üzerinde olumsuz etkilere neden olmaktadır. Böbrek hastalıkları sinir sistemi (Alzheimer) gibi sağlık problemlerine yol açmaktadır (WHO, 2003). Alüminyum, suyun tadını, kokusunu ve görünümünü değiştirmektedir. Suyun

kullanılabilirliğini etkilemekte, insanlar ve sanayi için sorunlar yaratmaktadır. Yüksek alüminyum seviyesi su ekosistemine de zarar vermektedir. Su kalitesini korumak ve sürdürülebilir su yönetimi için alüminyum seviyelerinin kontrol edilmesi gerekmektedir.

Amonyum (Ammonia): Su içerisinde bakteriler tarafından nitrata dönüştürülmektedir. Yüksek nitrat seviyesi su içerisindeki bitkilerin aşırı büyümesini sağlamaktadır. Bu durum suyun oksijen seviyesini düşürerek kalitesini etkilemektedir.

Demir (Iron): Yüksek demir yoğunluğu suyun tadını, rengini ve kokusunu etkilemektedir. Yüksek demir suya metalik tadı vermektedir. Böylece su kalitesi olumsuz etkilenmektedir.

Kalsiyum (Calcium): Suyun tat ve kokusu üzerinde bir etkisi yoktur. Suyun sertliği üzerinde etkisi bulunmaktadır. Sert su deride tahrişe, sıcak su borularında, ısıtıcılarda, kazanlarda kireç birikimine, sebzelerin katılaşmasına ve renksizleşmesine neden olmaktadır.

Klorür (Chloride): Yüksek seviyede olması suyun tuzluluğunu artırmaktadır. Bu durum içme suyunun tadını ve kalitesini olumsuz etkilemektedir. Klorür, aynı zamanda su arıtma sistemlerinin etkinliğini de etkilemekte ve borularda korozyona neden olmaktadır (DeSimone ve ark., 2014).

Mangan (Manganese): Yüksek mangan seviyeleri, suyun rengini ve tadını olumsuz etkilemektedir. İçme suyu sistemlerinde boruların korozyonuna neden olmakta ve su kalitesini daha fazla kötüleştirmektedir.

Sülfat (Sulphate): Suyun tat ve koku özelliğini çok fazla etkilememektedir. Ancak içerisinde yüksek yoğunluklu sülfat bulunan suyun tüketilmesi ishale sebep olmaktadır.

Hidrojen İyon konsantrasyonu (pH): Suyun asitlik veya alkalilik derecesini belirlemektedir. Yani suyun asit-baz dengesinin bir ölçüsüdür. Suyun kimyasal reaksiyonlarını etkilemektedir. Suda çözünmüş minerallerin çözünürlüğünü ve kimyasal dengelerini etkilemektedir. Düşük pH değeri, suyun korozyon potansiyelini artırmaktadır, bu durum borular metal ise hasara ve suyun metal içeriğinin artmasına neden olmaktadır.

Renk (Colour): Suyun kalitesinde renk önemli bir parametredir. Suyun rengi içerikteki maddelerin çeşidine ve yoğunluklarına işaret etmektedir. Yüksek demir yoğunluğu, suyun rengini olumsuz etkileyerek, kahverengi veya sarımsı bir renk almasını sağlamaktadır. Ayrıca suyun renginin farklı olması içerisinde organik maddelerin olduğunun da bir göstergesidir. Organik maddelerin (bitki kalıntıları, humus, yosun vb.) bulunduğu suyun rengi genellikle sarımsıdır. İnsanların suyu içme ve kullanma

konusundaki tercihlerini de suyun rengi belirlemektedir. Berrak su daha çekici bulunurken, bulanık veya renkli su kullanımı daha az tercih edilmektedir.

Bulanıklık (Turbidity): Çözünmemiş partiküllerin (toprak, çamur, kum, tortu, diğer organik ve inorganik maddeler) varlığı suyun bulanık olmasına sebep olmaktadır. Bu durum suyun kirliliğine işaret etmekte ve kalitesini etkilemektedir. Su analizlerinde bulanıklık birimi olarak NTU (Nephelometric Turbidity Unit) kullanılmaktadır. Ayrıca bulanık su, fotosentez yapabilen bitkilerin ışığa erişimini azaltmakta ve su altı yaşamını da olumsuz etkilemektedir.

Veri seti 10 bağımsız değişken ve bir bağımlı değişken sütunundan ve 1049 adet satırdan oluşmaktadır.

3.2. Veri Ön İşleme Süreci

Veri ön işleme süreci, makine öğrenmesi algoritmalarının değişkenleri daha iyi öğrenebilmesi ve daha doğru sonuçların çıkarılması için veri setinin dönüştürülmesi, temizlenmesi gibi bir hazırlama sürecidir.

Çalışma sırasında 2023 yılına ait veriler aylık olarak alındı. Aylık olarak alınan bu veriler, Jupyter 7.0.8 sürümlü notebook ortamında Python 3.11.7 versiyonlu programlama dili ile birleştirilerek tek dosya haline getirildi.

Tablo 3.1'de ön işleme öncesi veri setinin durumu verilmiştir. Altı farklı lokasyonda seksen adet farklı istasyondan alınmış bağımsız değişkenlere ait verilerin ilk 10 satırı verilmektedir. Makine öğrenmesi algoritmaları için veri seti çeşitli ön işleme süreçlerine tabi tutuldu.

Tablo 3.1. Ön işleme öncesi veri setinin durumu

Sıra No	Lokasyon İsmi	İstasyon İsmi	Parametre	Ortalama Değer
0	Eyre	Coffin Bay	Aluminium	0.002
1	Eyre	Coffin Bay	Calcium	51.1
2	Eyre	Coffin Bay	Chloride	130
3	Eyre	Coffin Bay	Colour	<1
4	Eyre	Coffin Bay	Iron	< 0.001
5	Eyre	Coffin Bay	Manganese	0.0001
6	Eyre	Coffin Bay	pH	7.7
7	Eyre	Coffin Bay	Sulphate	23.7
8	Eyre	Coffin Bay	Turbidity	<0.10
9	Eyre	Elliston	Aluminium	<0.001
10	Eyre	Elliston	Calcium	59.5

Öncelikli olarak Parametre sütununda yer alan bağımsız değişkenlerin indeks olarak kullanılması yani sütun bilgisi olması sağlandı.

Veri seti üzerinde yapılan bu indeks değişimi sonrası veri setinin son durumu Tablo 3.2'de verilmiştir. Parametre isimleri Türkçeye çevrildi. Ancak Python programlama dilinde problem olmaması için Türkçe karakterler kullanılmamaya dikkat edildi.

Tablo 3.2. Parametrelerin bağımsız değişken olarak sütun bilgisi olması

Alüminyum	Amonyum	Demir	Kalsiyum	Klorür	Mangan	Sülfat	pH	Renk	Bulanıklık
0.002	NaN	< 0.001	48.1	122	0.0001	23.5	7.8	<1	<0.10
0.002	NaN	< 0.001	59.6	211	<0.0001	34.2	7.7	<1	<0.10
0.002	NaN	0.0016	72.7	169	<0.0001	26.6	7.6	<1	<0.10
0.011	NaN	0.0037	55.7	118	0.0009	40.5	7.8	<1	<0.10
0.020	NaN	0.0438	22.0	50	0.0038	59.9	7.6	1	0.54

Tablo 3.2'de görüleceği üzere bağımsız değişkenlere ait ölçüm verilerinin standart olmadığı bazı değerlerin eksik olduğu, bazı değerlerin sayısal değerlere dönüştürülemediği bu nedenle Python tarafından NaN (Not a Number) olarak gösterildiği, bazı gözlemlerin de mantıksal değerlendirme şeklinde olduğu görünmektedir.

Bu durumda eksik verilerin, sayısal olmayan gözlemlerin belirlenmesi gerekmektedir. Keşifçi veri analizi yöntemleri ile bağımsız değişkenlere ait gözlemlerin niteliklerine ve değişken tiplerinin ne olduğuna bakıldı.

Veri setindeki değişkenlerin tipleri nesne (object) ve ondalık sayı (float) şeklinde oldukları görüldü. Makine öğrenmesi algoritmalarının öğrenme süreci esnasında nesne olarak verilen gözlemler için herhangi bir veri analizi ve öğrenme sürecine girmediğinden bu değerler sayısal, anlamlı ve standart haline getirildi.

Sayısal değer içermeyen gözlem değerlerini numerik bir değere çevirme işlemi için Python yardımı ile işlem yapıldı. Tüm bağımsız değişkenler için bu işlem yapıldı.

Sayısal değerlere dönüştürülemeyen gözlemler için NaN (Not a Number) değeri oluşturulmaktadır. NaN değerleri, veri setlerindeki eksik veya bilinmeyen değerleri temsil etmektedir. Bu değerlerin doğru bir şekilde işlenmesi ve yönetilmesi, veri analizi ve modelleme süreçlerinde önemlidir. Veri seti incelendikten sonra NaN olan gözlemler için bağımsız değişkenin ortalaması ile, sıfır ile veya bağımsız değişkenin medyanı ile doldurularak makine öğrenmesi algoritmaları için standart hale getirildi.

Veri setinin tüm ön işleme süreçlerinden sonraki durumu Tablo 3.3'te verilmiştir. analiz ve makine öğrenmesi süreçlerinde ihtiyaç duyulmayan değişkenler veri setinden çıkarıldı.

Tablo 3.3. Veri setinin ön işleme sürecinden sonraki durumu

Alüminyum	Amonyum	Demir	Kalsiyum	Klorür	Mangan	Sulfat	pH	Renk	Bulanıklık	İçilebilirlik
0.002	0.0	0.0010	48.1	122.0	0.0001	23.5	7.8	0.5	0.10	1
0.002	0.0	0.0010	59.6	211.0	0.0001	34.2	7.7	0.5	0.10	1
0.002	0.0	0.0016	72.7	169.0	0.0001	26.6	7.6	0.5	0.10	1
0.011	0.0	0.0037	55.7	118.0	0.0009	40.5	7.8	0.5	0.10	1
0.020	0.0	0.0438	22.0	50.0	0.0038	59.9	7.6	1.0	0.54	1

3.3. Keşifçi Veri Analizi Süreci

Veri setinin incelenmesi, anlaşılması, içerisindeki kalıpların belirlenmesi, değişkenler arasındaki ilişkinin tespit edilmesi ve analizi, veri setinin yapısal özelliklerinin belirlenmesi süreci keşifçi veri analizi olarak tanımlanmaktadır.

Veri setinin genel incelenmesi, istatistiksel analizler, veri görselleştirme işlemleri süresince yapılan işlemler açıklandı.

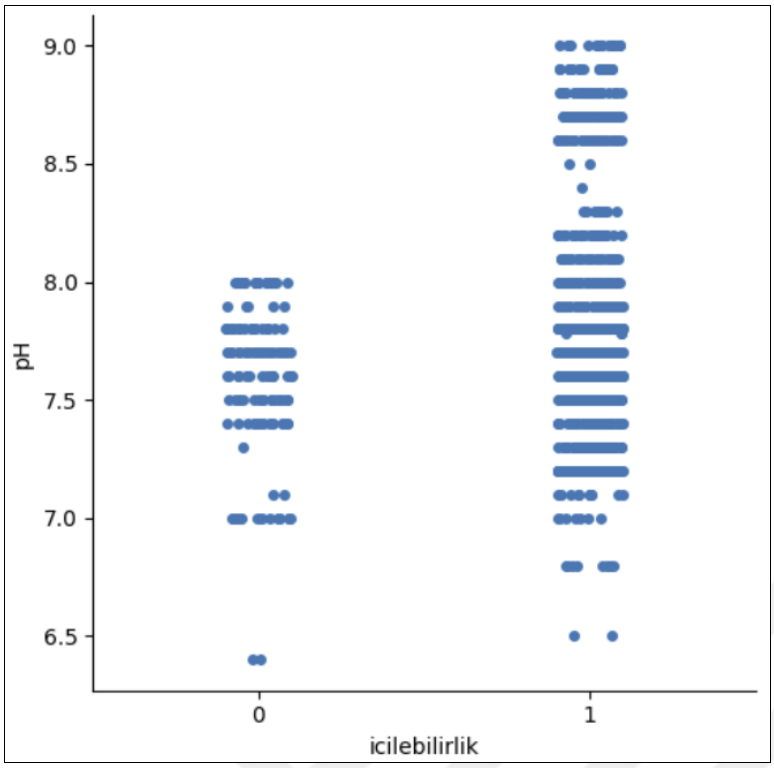
Tablo 3.4'te veri seti için temel istatistiksel analiz için özet veri verilmektedir. Tablo 3.4'te, her sütunun sayısı (count), ortalaması (mean), standart sapması (std), minimum değeri (min), 25. yüzdelik (25%), medyan (50%), 75. yüzdelik (75%) ve maksimum değeri (max) değerleri listelendi. Bu istatistik bilgileri, her bir sütunun dağılımı ve merkezi eğilimi hakkında bilgi sağlamaktadır. Veri kümesinin dağılımı, merkezi eğilimi ve yayılımı hakkında hızlı bir bilgi vermektedir. Veri setindeki ortalama değer medyan değerinden önemli ölçüde farklıysa, bu durumda aykırı değerlerin varlığından bahsedilmektedir. Tablo 3.4'te Amonyum (Ammonia) değişkeninin medyanı ile ortalamasının farklı olması bu değişkendeki anormalliğini gösterdi. Veri seti incelendiğinde Amonyum değişkeni için yeterli gözlem olmadığı anlaşıldı.

Tablo 3.4. Veri setine ait temel istatistiksel analiz özeti

	count	mean	std	min	25%	50%	75%	max
Alüminyum	1049.0	0,013	0,018	0,00	0,002	0,008	0,017	0,178
Amonyum	1049.0	0,033	0,094	0,00	0,000	0,000	0,000	0,420
Demir	1049.0	0,020	0,063	0,00	0,002	0,005	0,013	0,923
Kalsiyum	1049.0	37,431	31,611	1,42	17,200	23,600	48,200	159,000
Klorür	1049.0	125,363	124,073	0,00	51,000	84,000	166,000	682,000
Mangan	1049.0	0,002	0,004	0,00	0,000	0,001	0,002	0,025
Sülfat	1049.0	42,068	34,510	0,00	9,300	40,500	60,200	175,300
pH	1049.0	7,784	0,491	6,40	7,500	7,700	8,000	9,000
Renk	1049.0	0,834	0,702	0,50	0,500	0,500	0,834	4,000
Bulanıklık	1049.0	0,188	0,382	0,10	0,100	0,100	0,150	7,590
İçilebilirlik	1049.0	0,864	0,343	0,00	1,000	1,000	1,000	1,000

Bağımsız değişkenlerin kendi arasında ve bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi anlamak ve istatistiksel analiz yapabilmek için veri görselleştirme grafiklerinden faydalanıldı.

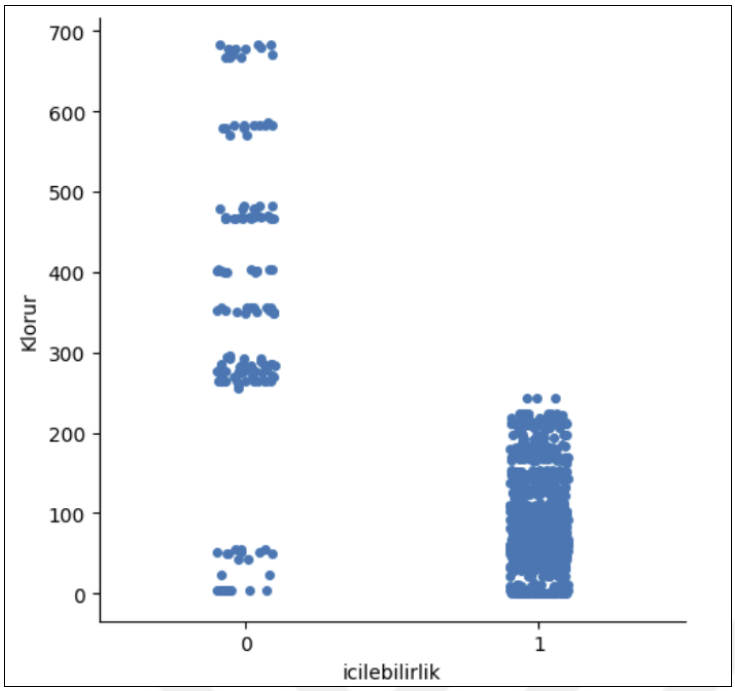
Veri setindeki değişkenlere ait gözlem değerlerine göre, suyun içilebilirliği ve pH arasındaki ilişki Şekil 3.1'de gösterildi. Suyun içilebilirlik durumuna göre pH değerinin ve içilebilirlik kategorik değerlerinin dağılımı gösterildi. Her bir içilebilirlik kategorisi için (içilebilirlik=0 ve içilebilirlik=1) pH değerlerinin genel olarak hangi aralıklarda yoğunlaştığına bakarak suyun içilebilirlik durumu ile pH arasındaki ilişki anlaşılmaktadır.



Şekil 3.1. pH ve içilebilirlik arasındaki ilişki

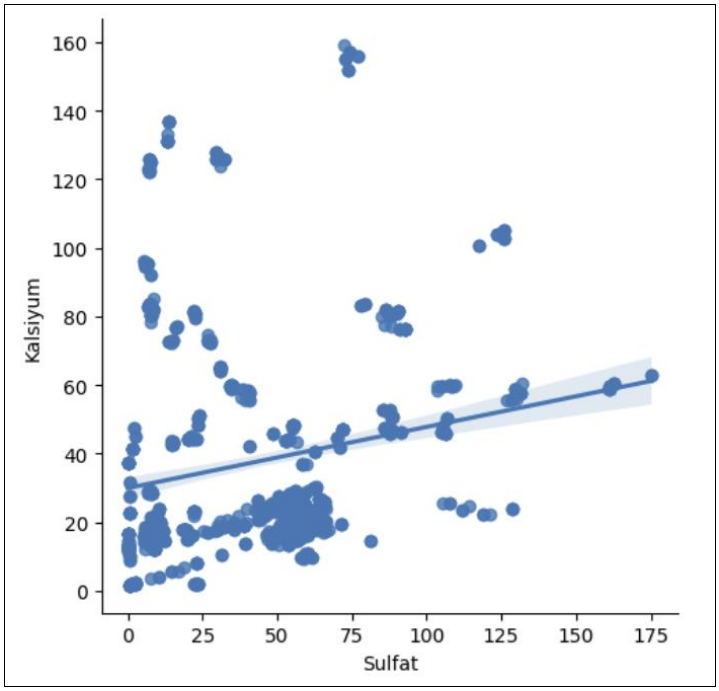
Suyun içilebilirliği üzerinde sadece bu değişken tek başına yeterli olmamaktadır. İçilebilirlik ve pH arasındaki ilişkiyi anlamak için aşağıdaki faktörler de göz önünde bulundurularak değerlendirme yapıldı.

Suyun diğer minerallerle zenginleşmiş olması pH seviyesini etkilemektedir. Yağmur suyu veya yeraltı suyu gibi farklı kaynaklardan gelen suların pH seviyeleri farklılık göstermekte, asitli topraklardan akan suyun pH seviyesi düşük olmaktadır. Ayrıca sanayi atıkları, tarımsal ilaçlar ve evsel atıklar gibi insan kaynaklı faktörler de suyun pH seviyesini etkilemektedir.



Şekil 3.2. Klorür ve içilebilirlik arasındaki ilişki

Şekil 3.2'de içilebilirlik durumunun genellikle hangi klorür düzeyiyle ilişkili olduğu anlaşılmaktadır. Klorür seviyesinin içilebilirlik üzerinde nasıl etkili olduğu görüldü.

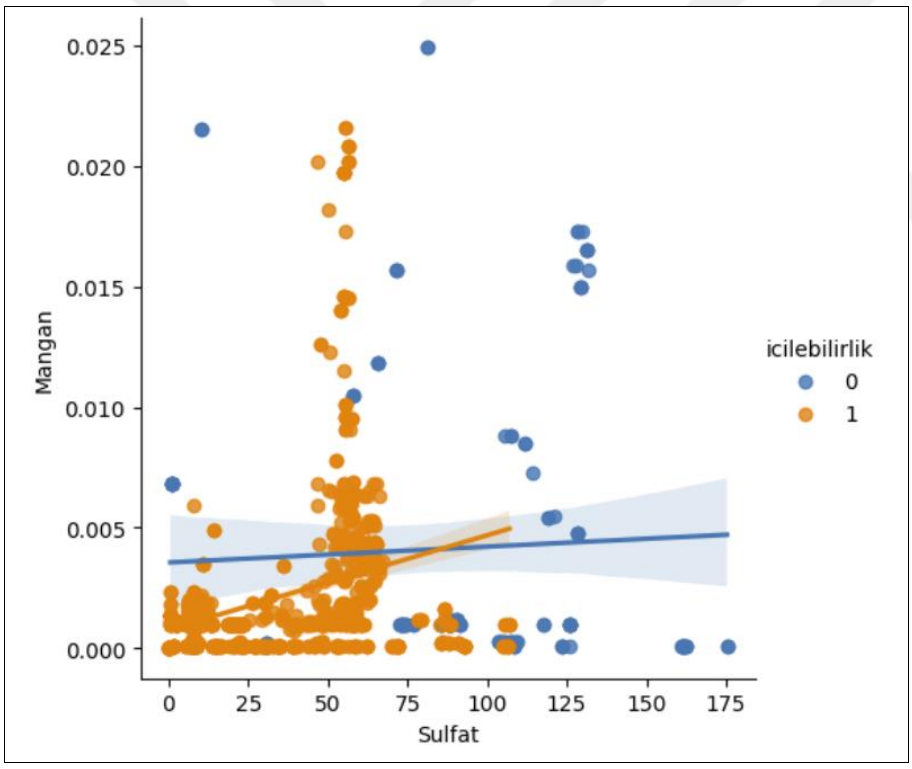


Şekil 3.3. Sülfat ile kalsiyum arasındaki doğrusal ilişki

Şekil 3.3'teki grafik incelendiğinde regresyon çizgisinin pozitif eğimli olduğu görülmektedir. Bu durum sülfat ve kalsiyum arasında doğrusal bir ilişki olduğunu göstermektedir. Gözlem değerlerinin regresyon çizgisine yakın olması doğrusal ilişkinin güçlü olduğunu göstermektedir.

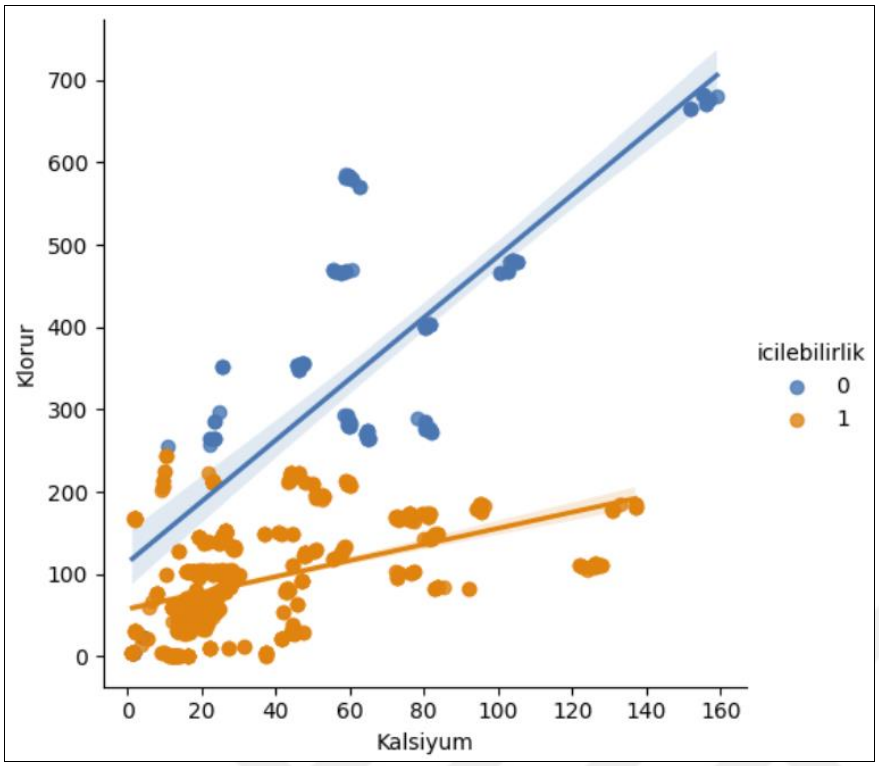
Sülfat ve mangan değişkenlerin, içilebilirlik bağımlı değişkenin sınıflarına göre dağılımı Şekil 3.4'teki grafikte gösterildi. Bağımsız değişkenlere ait gözlem değerlerinin genel dağılımına bakıldığında bu noktalar eğimli bir çizgi ile hizalandığından değişkenler arasında ilişki olduğunu belirtmektedir.

Grafikte yer alan regresyon çizgisi, sülfat ve mangan değişkenleri arasındaki ilişkiyi temsil etmektedir. Regresyon çizgisi belirgin bir eğim gösterdiğinden bu değişkenler arasındaki ilişkiyi göstermektedir.

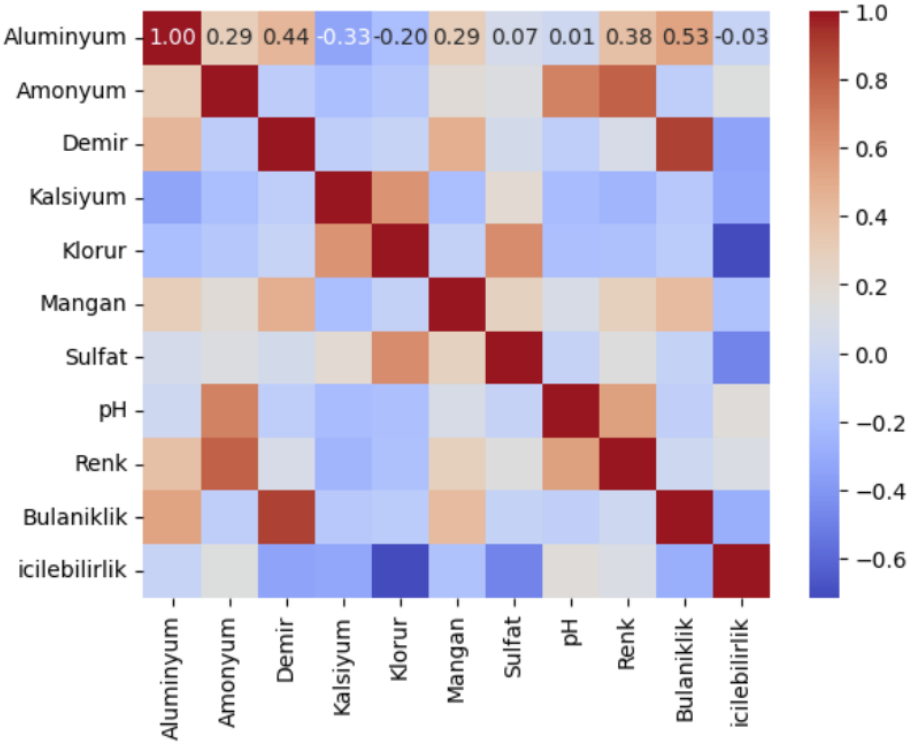


Şekil 3.4. Sülfat ve mangan arasındaki ilişkinin içilebilirlik sınıflarına göre dağılımı

Şekil 3.5'te içilebilirlik sınıf dağılımına göre kalsiyum ile klorür değişkenleri arasındaki ilişki gösterildi. Daha düşük klorür ve kalsiyum seviyelerinde içilebilirliğin daha yüksek olduğu görüldü. Sınıflar arasındaki regresyon eğrilerinin pozitif eğimli olması bu iki değişken arasındaki ilişkinin doğrusal olduğunu göstermektedir.

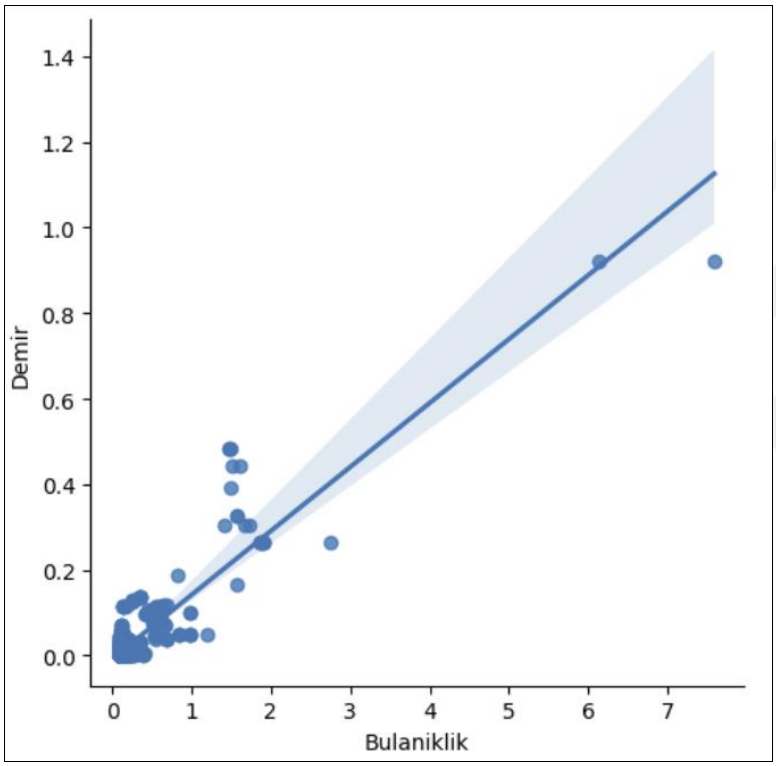


Şekil 3.5. Kalsiyum ile klorür arasındaki ilişki



Şekil 3.6. Değişkenler arasındaki korelasyon ilişkisinin ısı grafiği

Şekil 3.6'daki ısı haritası grafiği, veri setindeki her bir değişkenin diğer değişkenler ile nasıl bir korelasyon içinde olduğunu göstermektedir. Her bir hücredeki renk, ilgili iki değişken arasındaki korelasyonun gücünü göstermektedir. Pozitif bir korelasyon katsayısı sıcak renklerle (kırmızı gibi) gösterilmekte, negatif bir korelasyon katsayısı soğuk renklerle (mavi gibi) gösterilmektedir. Bulanıklık ile demir değişkenleri arasında güçlü bir korelasyonun olduğu, her iki değişkenin kesiştiği hücrenin renginden anlaşıldı. Şekil 3.7'de bu iki değişken arasındaki doğrusal ilişki bu korelasyonu teyit etmektedir.



Şekil 3.7. Bulanıklık ve demir değişkenleri arasındaki doğrusal ilişki

3.4. Model Doğrulama Yöntemleri

Makine öğrenmesi yöntemleri kullanılarak büyük veri setleri için uygun modeller elde edilmektedir. Hangi modelin daha iyi olduğunu bulmak ve gelecekte öğrenme modelinin ne kadar iyi çalışacağını anlamak için değerlendirme yöntemleri kullanılmaktadır. Makine öğrenmesi modellerine ait başarı değerlendirmesi ve model doğrulama metrikleri aşağıda sınıflandırma ve regresyon problemleri için ayrı verildi.

3.4.1. Sınıflandırma metrikleri

Sınıflandırma problemlerinde modellerin performansı aşağıdaki metrikler kullanılarak değerlendirilmektedir.

Karışıklık (confusion) matrisi: Veride var olan durum ile sınıflandırma modelinin doğru ve yanlış tahminlerinin sayısını gösterir.

Tablo 3.5. Karışıklık matrisi

		Var Olan Durum	
		Pozitif Durumlar	Negatif Durumlar
Tahmin	Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
	Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Var olan durumlar eğitim veri seti içerisinde veri setini oluşturan kişi tarafından belirtilmektedir. Bu durumlar kesinlikle doğru kabul edilmektedir. Tablo 3.5'e göre eğer tahmin işleminde normalde pozitif olarak var olan bir durum pozitif olarak tahmin ediliyorsa DP bir tahmin yapılmaktadır. Eğer var olan durum negatif ve tahmin negatif ise DN bir tahmin yapılmaktadır. Başka bir deyişle yanlış bir durum yanlış olarak doğru bir şekilde tahmin edilmiş demektir. Eğer var olan durum negatif ise ancak tahmin sistemi pozitif olarak tahmin ederse birinci tip hata YP durumu oluşur. Eğer var olan durum pozitif ise ve tahminci negatif olarak tahmin ederse YN ikinci tip hata oluşur.

Veri seti içerisindeki değişkenlerin makine tarafından algılanması, öğrenmesi ve bunun sonucunda kullanıcıya en doğru sonucu vermesi için önemli performans metrikleri bulunmaktadır. Bu metrikler aşağıda verilmiştir.

Doğruluk (Accuracy): Doğru tahmin edilen gözlemlerin tüm gözlemlere oranıdır. Denklem (3.1)'deki formül ile hesaplanmaktadır.

$$Accuracy\ (\%) = \frac{(DP+DN)}{(DP+YP+DN+YN)} x100 \qquad (3.1)$$

Kesinlik (Precision): Doğru tahmin edilen pozitif gözlemlerin toplam tahmin edilen pozitif gözlemlere oranıdır. Denklem (3.2)'deki formül ile hesaplanmaktadır.

$$Precision\ (\%) = \frac{DP}{(DP+YP)} x100 \quad (3.2)$$

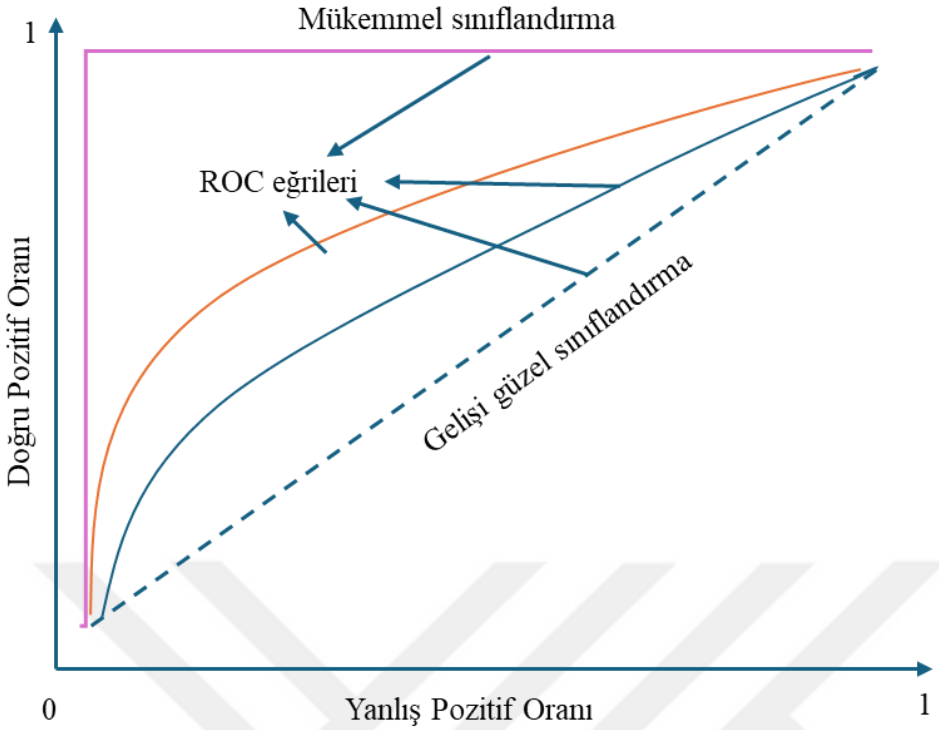
Duyarlılık (Recall): Doğru sınıflandırılan gözlemlerin o sınıftaki tüm gözlemlere olan oranıdır. Denklem (3.3) ile hesaplanmaktadır.

$$Recall\ (\%) = \frac{DP}{(DP+YN)} x100 \quad (3.3)$$

F1 Skor (F1 score): Precision ve Recall değerlerinin ağırlıklandırılmış bir ortalamasıdır. Denklem (3.4) ile hesaplanmaktadır.

$$F1\ Score = \frac{Precision\ x\ Recall\ x\ 2}{Precision+Recall} x100 \quad (3.4)$$

İşlem karakteristik eğrisi (ROC-Receiver Operating Characteristic Curve): Bu eğri eşik değeri değiştirilerek Doğru Pozitif Oranı-Yanlış Pozitif Oranı grafiği veya doğruluk oranının grafik olarak işlenmesidir.



Şekil 3.8. ROC eğrisi

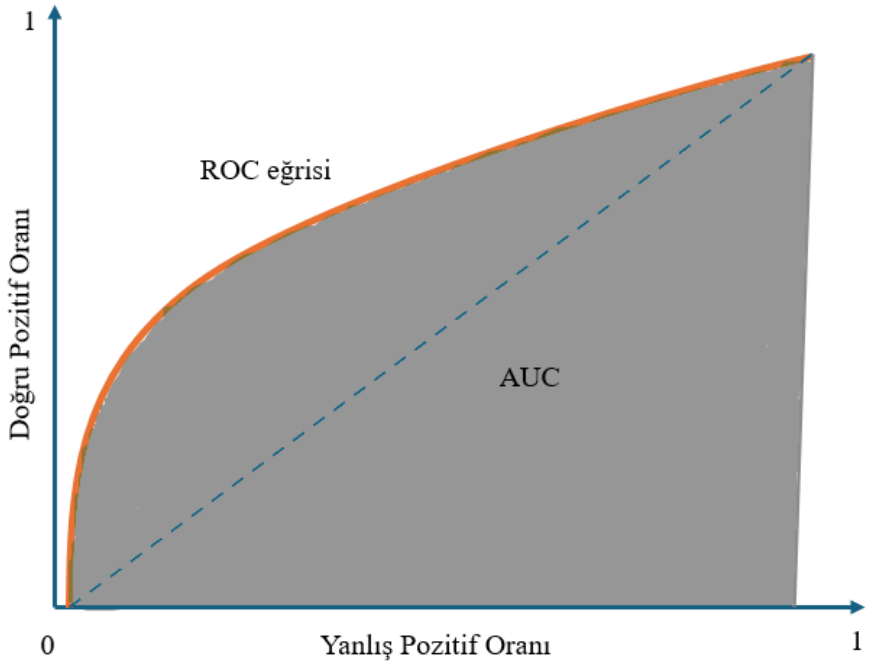
Şekil 3.8'de verilen ROC eğrisi incelendiğinde yatay eksende Yanlış Pozitif Oranı, Dikey eksende ise Doğru Pozitif Oranı yer almaktadır. Yanlış Pozitif oranı, karışıklık matrisinde yer alan Yanlış Pozitif (YP) gözlemlerin tüm yanlış gözlemlere oranını göstermektedir. Doğru Pozitif oranı ise karışıklık matrisinde yer alan Doğru Pozitif (DP) gözlemlerin tüm doğru gözlemlere oranını göstermektedir. Denklem (3.5) ve Denklem (3.6) ile oranların hesaplanması verilmiştir.

$$Doğru\ Pozitif\ Oranı = \frac{DP}{DP+YN} \tag{3.5}$$

$$Yanlış\ Pozitif\ Oranı = \frac{YP}{DN+YP} \tag{3.6}$$

Bu iki oran yatay ve dikey eksende 0 ile 1 arasındaki değerler üzerinde bir eğri oluşturmaktadır. Oluşan bu eğriye ROC eğrisi (Receiver Operating Characteristic Curve- İşlem karakteristik eğrisi) denilmektedir. Bu eğrinin altında kalan alana ise AUC (Area Under Curve- Eğri altındaki alan) denilmektedir. Şekil 3.9'da bir modele ait ROC eğrisi

altındaki alana ait AUC gösterilmiştir. AUC'nin (eğri altında kalan alan) büyük olması modelin başarılı olduğu, küçük olması ise modelin başarısız olduğu anlamına gelmektedir.



Şekil 3.9. AUC alanı

3.4.2. Regresyon Metrikleri

Regresyon problemlerinde modellerin performansı aşağıdaki metrikler kullanılarak değerlendirilmektedir.

Ortalama Kare Hatası (MSE-Mean Square Error): Gerçek değerlerle tahmin edilen değerler arasındaki kare farkların ortalamasıdır. MSE değeri sıfıra ne kadar yakın ise model o kadar başarılı demektir. Denklem (3.7) ile hesaplanmaktadır.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (3.7)$$

Denklem (3.7)'deki;

n: gözlem sayısını

y_i: gerçek değerleri

$\hat{y}_i$: tahmin edilen değerleri ifade etmektedir.

$y_i - \hat{y}_i$: hatayı vermektedir.

Hata Kareler Ortalamasının Karekökü (RMSE-Root Mean Square Error): MSE'nin karekökü alınarak hesaplanan bir metriktir. MSE de negatiflik durumunu ortadan kaldırmak için alınan kare işlemi bu metrikte karekök işlemi ile geri alınmaktadır. Denklem (3.8)'de formülü verilmiştir.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.8)$$

Ortalama Mutlak Hata (MAE-Mean Absolute Error): Gerçek ve tahmin edilen değerler arasındaki mutlak farkın ortalamasıdır. Daha küçük bir Ortalama Mutlak Hata (MAE) daha başarılı bir model olduğunu açıklamaktadır. Denklem (3.9) ile hesaplanmaktadır.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (3.9)$$

Model başarısı hatalar ile değerlendirildiğinden hataların önemi büyüktür.

3.5. Model Seçimi ve Değerlendirilmesi

Makine öğrenmesi modellerinin daha iyi öğrenmesi ve daha isabetli tahminler yapabilmesi için veri seti üzerinde gerekli ön işlem süreçleri gerçekleştirildi. Sonraki aşamada veri seti, eğitim ve test olmak üzere iki kısma ayrıldı. Veri setinin %80'i eğitim veri seti, %20'si test veri seti olarak ayrıldı.

Rastgele Orman (Random Forest-RF) algoritması ile tahmin işlemi gerçekleştirildi. Model doğruluğu (accuracy): 0.9904 olarak çıktı.

Çapraz doğrulama (cross validation), makine öğrenme algoritması yani modelin başarı performansını değerlendirmek için kullanılan bir metriktir. Bir modelin değerlendirilmesi yapılırken gerçek dünya verilerine ne kadar iyi uyarlandığı çapraz doğrulama metriği ile belirlenmektedir. Test ve eğitim veri setleri ile yapılan doğrulama, bölme işleminin nasıl yapıldığı ile ilgilidir. Çapraz doğrulamada ise veri setini belirli parçalara (katmanlara) ayırmaktadır. Her bir katman sırayla test seti olarak kullanılmakta,

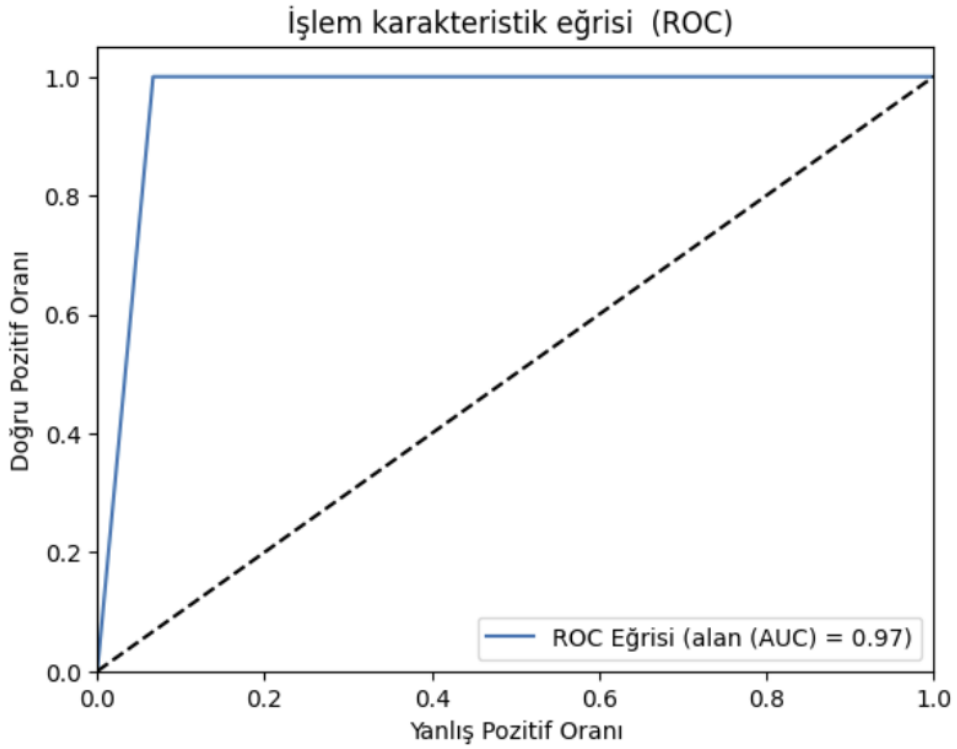
diğer katmanlar birleştirilerek eğitim setini oluşturmaktadır. Bu işlem, her katmanın sırayla test seti olarak kullanılmasına kadar tekrarlanmaktadır. Sonuç olarak, her veri noktası hem eğitimde hem de testte kullanılmakta ve modelin performansı bu farklı veri bölümleri üzerinden yapılan doğrulama işlemlerinin ortalaması alınarak değerlendirilmektedir.

Çapraz doğrulama işleminin en yaygın türü k katmanlı çapraz doğrulama (k-fold cross-validation) işlemidir.

Modelin çapraz doğrulaması yapıldı. Burada önemli bir hiper parametre değeri olan cv değerine 5 verilerek, 5 katmanlı doğrulama işlemi yapıldı. Bir modelin performansını güvenilir bir şekilde ölçmek ve genelleştirilebilirliğini belirlemek için kullanılan bu yöntemin sonucu şu şekilde çıktı:

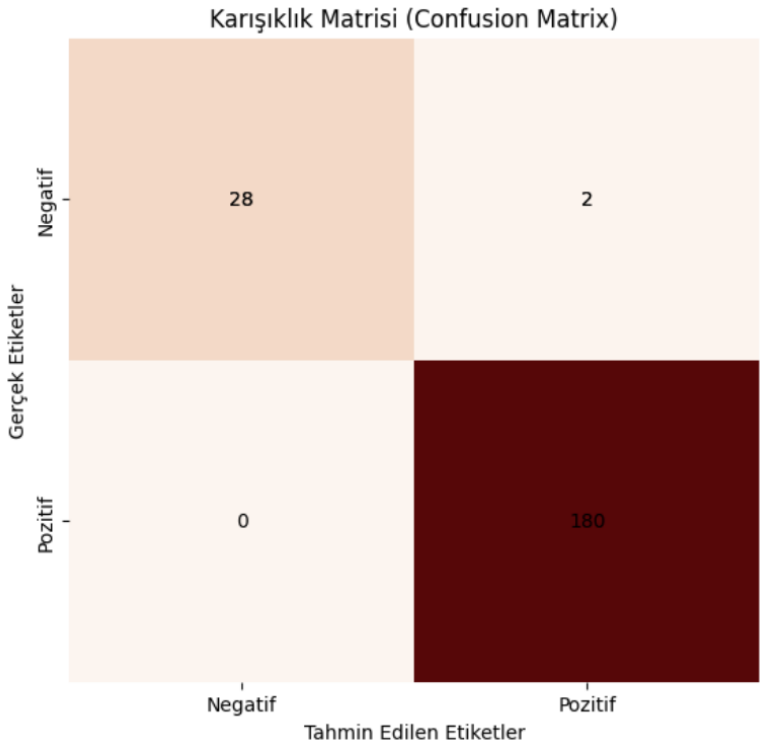
Modele ait Çapraz doğrulama skorları (Cross validation scores): 0.995, 1, 1, 0.990, 1 modelin Ortalama Çapraz doğrulama skoru (Mean Cross validation score): 0.9971 olarak çıktı.

Makine öğrenmesi modellerinin başarı değerlendirme metriklerinden olan ROC eğrisi (Receiver Operating Characteristic Curve- İşlem karakteristik eğrisi) Şekil 3.10'da verilmiştir.



Şekil 3.10. İşlem karakteristik eğrisi

Modele ait karışıklık matrisi Şekil 3.11'de verilmiştir.



Şekil 3.11. Karışıklık matrisi

Rastgele Orman (Random Forest) algoritması ile oluşturulan model için sınıflandırma başarı metrikleri aşağıdaki şekilde hesaplandı. Model ait başarı metriklerinin sonucu Tablo 3.6'da verildiği gibi çıktı.

Tablo 3.6. Modele ait performans metrikleri

Metrik	Değer
Kesinlik (Precision)	0.989
Duyarlılık (Recall)	1.0
F1-Skor (Score)	0.994

Literatürde sıkça kullanılan diğer makine öğrenmesi algoritmaları ile modeller oluşturulmuştur.

Öncelikli olarak modellere ait hiper parametre değerlerine müdahale edilmeden Python programlama dilindeki ön tanımlı (default) değerler ile veri keşfi ve model seçimi yapıldı.

İşlem sonucunda oluşturulan modellere ait performans metrikleri Tablo 3.7'de verilmiştir. Hiper parametre değerlerine müdahale edilmeden oluşturulan modeller içerisinde en başarılı model Karar ağaçları (Decision tree) modeli oldu.

Tablo 3.7. Dabl kütüphanesi ile oluşturulan modellerin başarı sonuçları

Model Adı	Metrik	Değer
Dummy Classifier	Doğruluk (accuracy)	0.865
Gaussian NB	Doğruluk (accuracy)	0.542
Multinomial NB	Doğruluk (accuracy)	0.868
DecisionTree Classifier	Doğruluk (accuracy)	0.999
Logistic Regression	Doğruluk (accuracy)	0.982

Sınıflandırıcı modeller oluşturulduktan sonra, test verisi üzerindeki tahmin işleminin sonucu test doğruluğu olarak tanımlanır. Bu durumda, test doğruluğu 0.966 olarak hesaplandı.

Makine öğrenmesi algoritmalarına ait modeller üzerinde hiper parametre ayarlaması (tunning) işlemi yapıldı. Modellere ait doğruluk (accuracy) başarı metriği sonuçları yüzde olarak Tablo 3.8'de verilmiştir.

Tablo 3.8. Hiper parametre ayarlaması yapılan modellere ait başarı performansları

Model Adı	Doğruluk (Accuracy)
Gaussian Naive Bayes	91.429
K- En Yakın Komşu (KNN)	99.048
Destek Vektör (Support Vector)	99.048
Yapay Sinir Ağları (Artificial Neural Network)	99.048
Karar Ağaçları (CART)	99.048
Rastgele Orman (Random Forests)	99.048
Gradyan Artırmalı (Gradient Boosting Machines)	99.048
Extreme Gradient Boosting (XGBoost)	99.048
Kategori Artırmalı (Category Boosting CatBoost)	99.048
Logistic Regression	99.048

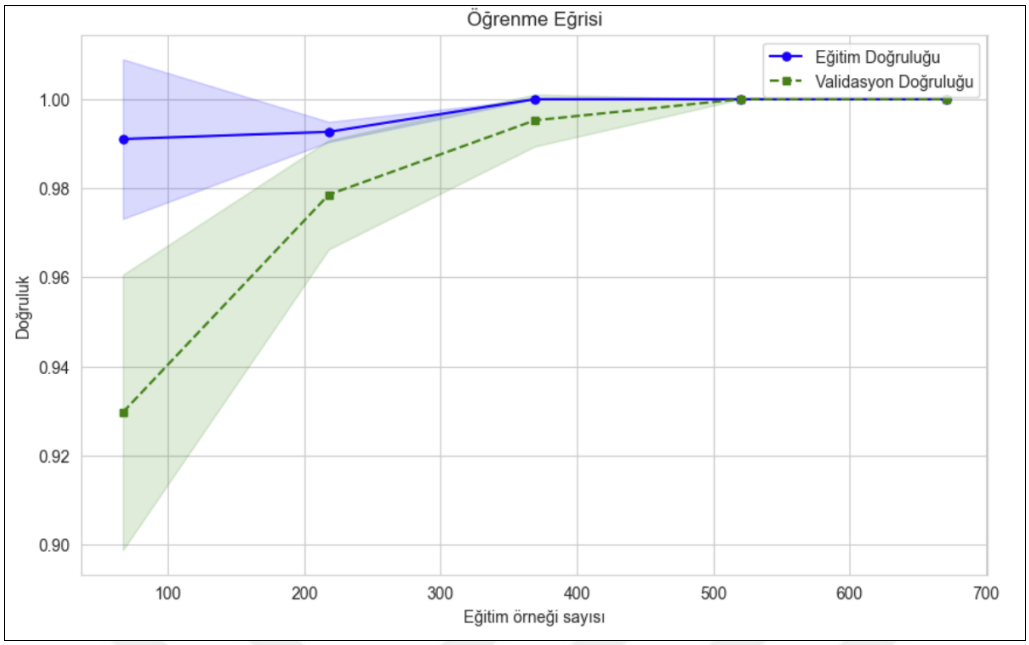
Gaussian Naive Bayes modeli diğer modellere göre daha düşük bir doğruluk oranı verdi. Gaussian Naive Bayes algoritması diğer modellere göre karmaşık veri setlerinde daha az performans göstermektedir.

K-En Yakın Komşu (KNN), Destek Vektör (Support Vector), Yapay Sinir Ağları, Karar Ağaçları, Rastgele Orman, Gradyan Artırmalı, Extreme Gradient Boosting ve Kategori Artırmalı modeller, oldukça yüksek doğruluk oranını verdi. Bu modeller karmaşık veri setlerinde daha güvenilir ve daha genellenebilir sonuçlar vermektedir. Yüksek performans sonucu veren bu modeller genellikle süre açısından daha uzun ve maliyeti daha yüksektir.

Lojistik Regresyon (Regression) isminde regresyon kelimesi olsa da bir sınıflandırma modelidir. Bu model diğer modellere benzer bir performans sergiledi. Diğer modellere göre yorumlanması ve anlaşılması daha basittir. Maliyet ve süre açısından daha uygun olduğu görüldü.

Makine öğrenmesi algoritmasına ait modelin başarısını Şekil 3.12'deki öğrenme eğrisi ile de değerlendirildi. Öğrenme eğrisi modelin eğitim ve doğrulama (validation) setlerindeki performansını eğitim örneği sayısına göre göstermektedir. Eğitim veri setine ait doğruluk oranı eğitim örneği sayısı artıkça artmaktadır. Bu durum modelin eğitim verisine uyum sağladığını göstermektedir.

Validasyon doğruluğu ise, eğitim örneği sayısı artıkça artmakta belli bir örnek sayısına ulaşınca sabit kalmaktadır. Ayrıca eğitim ve validasyon doğruluğu arasındaki alanın miktarının az olması modelin aşırı öğrenme (overfit) yapmadığını göstermektedir. Model iyi bir genelleme yeteneğine sahip olduğu görüldü.



Şekil 3.12. Öğrenme eğrisi

Model seçimi ve değerlendirilmesi hesaplama süresi, maliyeti ve performans metrikleri göz önüne alınarak yapılmalıdır.

Derin öğrenme modeli olan Uzun Kısa Süreli Bellek (Long Short-Term Memory-LSTM) algoritması ile aynı veri seti ile çalışıldı.

LSTM modelinin genel çalışma prensibinde, hücre durumu (cell state) ve üç ana kapı (giriş, unutma, çıkış) kullanarak bilgiyi işlemektedir. Bu kapılar, hangi bilginin saklanacağı, hangi bilginin unutulacağı ve hangi bilginin çıkışa verileceği konusunda karar vermektedir.

LSTM'nin temel bileşenlerinden olan hücre durumu (cell state) modelin bellek kapasitesini temsil etmektedir. Bilgiyi uzun süre saklamak için kullanılmaktadır. Giriş kapısı (Input gate) hücre durumuna ne kadar yeni bilgi ekleneceğine karar vermektedir. Unutma kapısı (Forget gate) hücre durumundan ne kadar bilginin unutulacağına karar vermektedir. Çıkış kapısı (Output gate) ise bir sonraki gizli duruma ne kadar bilginin aktarılacağını belirlemektedir.

Veri setinin LSTM modeli ile tahmin yapabilmesi için gerekli normalleştirme, değişken hazırlama, veriyi eğitim ve test olarak ayırma işlemi gerçekleştirildi.

Normalize etme işlemi modelin daha kararlı ve hızlı öğrenmesini sağladı, ayrıca farklı ölçeklerdeki verilerin eğitim sürecinde model üzerinde olumsuz etkiler yapmasını önledi.

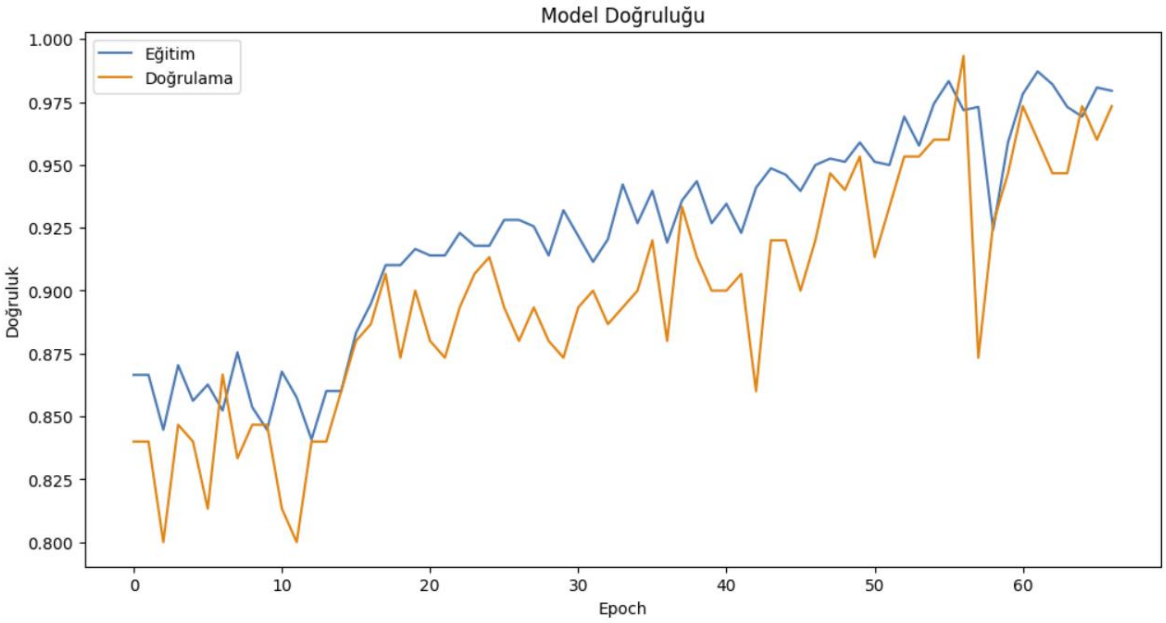
Veri çerçevesindeki (data frame) her bir örnek için belirli bir zaman adımı kadar veri alınarak döngü işlemi yapıldı. Eğitim ve test verileri LSTM modeline uygun olarak örnek sayısı, zaman adımı ve özellik sayısı şeklinde 3 boyutlu formata getirildi.

LSTM modeline ait performans sonuçlar Tablo 3.9'da verilmiştir. LSTM modelinin eğitim ve test aşamalarında oldukça iyi performans verdiği görüldü.

Tablo 3.9. LSTM modelinin performans metrikleri

Doğruluk (Accuracy)	Kayıp (Loss)	Test Doğruluğu
98,29	0,13	97,33

Modele ait doğruluk grafiği Şekil 3.13'te verilmiştir.



Şekil 3.13. Model doğruluğu

Modelin test verileri üzerindeki doğruluğu %97.33 olarak hesaplandı. Bu değer, modelin daha önce görülmemiş veriler üzerinde ne kadar iyi performans gösterdiğini ifade etmektedir. Yüksek bir test doğruluğu, modelin genelleme yeteneğinin iyi olduğunu ve gerçek dünya verileri üzerinde benzer performans gösterebileceğini belirtmektedir. Ayrıca hem eğitim (%98.29) hem de test (%97.33) doğruluk değerleri oldukça yüksek olarak hesaplandı. Bu değerler, modelin hem eğitim verileri üzerinde iyi öğrendiğini hem de test verileri üzerinde iyi genelleme yaptığını göstermektedir.

Derin öğrenme modeli LSTM yüksek bir performans gösterse de Makine öğrenmesi algoritmaları daha yüksek ve başarılı sonuçlar gösterdi.

4. BULGULAR VE TARTIŞMA

Yapılan bu çalışmada literatürdeki diğer çalışmalardan farklı olarak suyun içilebilirliği ile ilgili parametrelerin veri seti lokasyonuna uygun olarak uluslararası standartlardaki eşik değerlerine göre sınıflandırma yapıldı. Makine öğrenmesi algoritmaları ile önce otomatik ön tanımlı (default) değerler ile sonrasında hiper parametre ayarlaması ile işlem yapıldı. Ayarlama işlemi yapılan hiper parametreler EK 1'de verildi. Kullanılan makine öğrenmesi algoritmaları, Gaussian Naive Bayes, K- En Yakın Komşu (KNN), Destek Vektör (Support Vector), Yapay Sinir Ağları (Artificial Neural Network), Karar Ağaçları (CART), Rastgele Orman (Random Forests), Gradyan Artırmalı (Gradient Boosting Machines), Kategori Artırmalı (Category Boosting CatBoost), Logistic Regression modelleriydi.

Veri ön işleme sürecinin parametrelerin kimyasal özelliklerine ve etki değerlerine bağlı olmak üzere yapılmıştır. Parametrelerin kimyasal özelliklerindeki değişiklik farklı etkiler yaratacağından buna uygun olarak eksik verilerin doldurulması işlemi yapıldı. Parametrelerin eşik değerlerine göre yapılan sınıflandırma ile sınıf ayrımı %90'a %10 gibi bir oran verdiğinden tüm modeller oldukça başarılı sonuçlar verdi.

Maliyet ve zaman kriterleri göz önüne alındığı zaman kullanılan tüm modeller içerisinde Rastgele Orman (Random Forests) ve Logistic Regression en başarılı modeller olarak gözlemlendi.

Ayrıca, Derin Öğrenme algoritmalarından biri olan Uzun Kısa Süreli Bellek (LSTM) ile aynı veri seti üzerinde çalışıldı. Ancak, yapılan tespitler, Makine Öğrenmesi modellerinin (örneğin Rastgele Orman ve Lojistik Regresyon gibi) daha yüksek performans sergilediğini gösterdi.

Yapılan çalışma ve sonuçların literatür ile kıyaslanması, mevcut literatür ile örtüştüğü yönleri ve elde edilen başarılı sonuçların değerlendirilmesi için daha önceden yapılan çalışmalar incelendi. Bu kapsamda incelenen verilerin bilimsel, pratik ve teorik açıdan ne anlama geldiği tartışıldı.

Shams ve arkadaşları su kalitesinin sınıflandırılması için makine öğrenmesi algoritmalarından Rastgele Orman (Random Forest), Extreme Gradyan artırmalı (XGBoost Machine), Uyarlamalı Artırma (AdaBoost Machine), ve Gradyan Artırmalı modelleri kullandı. Ayrıca regresyon modeli olarak K-En Yakın Komşu (KNN), Destek Vektör Regresyonu, Karar Ağaçları ve Çok Katmanlı Alıcılar (MLP-Multi Layer Perceptron)

algoritmalarını kullandılar. Bu çalışmada su kalitesinin sınıflandırılması için kullandıkları Gradyan Artırmalı (GB) algoritması %99,5 doğrulukla ile buldular. Regresyon modeli olarak da en iyi R^2 değerini %99,8 ile MLP modeli ile buldular (Shams ve ark., 2023). İlgili çalışmada Hindistan'ın çeşitli bölgelerindeki sulardan alınan veriler ile çalışıldı. Çözünmüş oksijen, pH, iletkenlik, biyolojik oksijen, nitrat, dışkı koliform, toplam koliform olmak üzere 7 değişken ile çalışıldı.

Juna ve arkadaşları suyun kalitesini daha hızlı ve ekonomik bir şekilde tespit etmek için K-En Yakın Komşu (KNN) ile Çok Katmanlı Alıcılar (MLP-Multi Layer Perceptron) algoritmalarını kullanarak çalışma yaptılar (Juna ve ark., 2022). Bu çalışmada kullanılan veri setini Kaggle ortamından aldılar. Veri setinde, pH, sertlik (hardness), çözünmüş katılar (solids), klorominler (chloromines), sülfat (sulfate), iletkenlik (conductivity), organik karbon (organic carbon), trihalometanlar (trihalomethanes) ve bulanıklık (turbidity) olmak üzere 9 değişken bulunmaktaydı. Bu çalışma sonucunda %99,9 doğruluk oranı elde ettiler. Eksik veriler için KNN algoritmasının önemine vurgu yaptılar.

Nasir ve arkadaşları, makine öğrenimi algoritmalarını kullanarak su kalitesi sınıflandırması yapmak için topladıkları veriler ile su kalitesi indeksinden faydalandı. Destek Vektör Makinesi (SVM), Rastgele Orman (RF), Lojistik Regresyon (LR), Karar Ağacı (DT), CATBoost, XGBoost ve Çok Katmanlı Algılayıcı (MLP) algoritmaları kullanarak çalışma yapmışlardır (Nasir ve ark., 2022). CATBoost algoritması ile %94,51 doğrulukla sınıflandırmayı gerçekleştirdiler. Veri setlerinde çözünmüş oksijen, pH, iletkenlik, nitrat, dışkı koliform ve toplam koliform değişkenlerine yer verdiler. Bu değişkenler ile su kalite indeksi hesaplanıp tahmin algoritmalarında kullandılar.

Khan ve See makine öğrenimini kullanarak su kalitesini tahmin etme ve analiz etme isimli çalışmalarında Yapay Sinir ağları ve zaman serisi analizi kullanarak su kalitesi tahmin modeli geliştirdiler (Khan ve See, 2016). Bu çalışmada su kalitesini etkileyen 4 değişken ele alındı. Bu parametreler klorofil, özgül iletkenlik, çözünmüş oksijen ve bulanıklıktır. Doğrusal Olmayan Otoregresif (Nonlinear autoregressive network-NAR) modeline sahip ileri beslemeli bir sinir ağı kullanılarak regresyon başarı metrikleri hesaplanarak su kalitesi tahmin ettiler.

Kaddoura yaptığı çalışma ile su kalitesinin tahmini için makine öğrenmesi algoritmalarının verimliliğini değerlendirdi (Kaddoura, S., 2022). Kullandığı veri setinde pH, sertlik, çözünmüş katılar, klorominler, sülfat, iletkenlik, organik karbon, trihalometanlar, bulanıklık özelliklerine değindi. Denetimli makine öğrenmesi metotları

kullandı ve performanslarını değerlendirdi. İşlem karakteristik eğrisi altında kalan alan (ROC-AUC) değerlendirdi ve en başarılı modelin Destek Vektör Makineleri (SVM) olduğunu tespit etti.

Ahmed ve arkadaşları, su kalite indeksi ve su kalite sınıflarını baz alarak temelinde makine öğrenmesi algoritmaları olan bir su tahmin modeli üzerinde çalıştılar (Ahmed ve ark., 2019). Sıcaklık, bulanıklık, pH ve toplam çözünmüş katılar olmak üzere dört bağımsız değişken parametresi kullanarak su kalite indeksini hesaplayıp hem regresyon hem de sınıflandırma metriklerini hesapladılar. Çok katmanlı algılayıcı (MLP) modeli en başarılı kalite sınıflandırmasını gerçekleştirdiğini gözlemlediler. Üç parametre ile değerlendirme yaptıklarında Gradyan artırmalı (Gradient Boosting) modeli, dört parametre ile değerlendirme yaptıklarında Polinom Regresyon en iyi regresyon sonucunu verdiğini değerlendirdiler.

Su Kalitesi Sınıflandırma Modellerinin Makine Öğrenmesi ile Karşılaştırılması adlı çalışmada pH, biyokimyasal oksijen ihtiyacı, çözünmüş oksijen, elektriksel iletkenlik özellikleri yardımı ile su kalitesi tahin çalışması yapan Radhakrishnan ve Pillai**,** Destek Vektör Makineleri (SVM), Karar Ağaçları (Decision Tree) ve Naive Bayes gibi makine öğrenimi algoritmalarını kullanan su kalitesi sınıflandırma modelleri geliştirdiler (Radhakrishnan ve Pillai, 2020). Karar Ağaçları algoritması ile %98,50 doğrulukla sınıflandırma yaptılar.

Chen ve arkadaşları, karşılaştırmalı yüzey suyu kalitesinin tahmin edilmesi ve makine öğrenmeleri algoritmaları ile temel su parametrelerinin belirlenmesi üzerine çalışma yaptılar (Chen ve ark., 2020). Çin deki nehir ve göllerden alınan 2012 ve 2018 yılları arasında 33614 gözlem verisiyle çalışma yaptılar. Temel su parametresi olarak kimyasal oksijen ihtiyacı, pH, çözünmüş oksijen ve amonyak nitrojeni seçtiler. Çalışmada büyük verilerin makine öğrenmesi metotları ile su kalitesi tahminindeki etkisini incelediler. Çalışmada en başarılı modeller Karar Ağaçları (Decision Tree), Rastgele Orman (Random Forest) ve Derin kaskat orman (DCF -Deep Cascade Forest) olarak genelleştirildi.

Modaresi ve Araghinejad, su kalitesi sınıflandırması için, Olasılıksal Sinir Ağı (Probabilistic Neural Network- PNN) ve K-En Kakın Komşu (KNN), Destek Vektör Makinesi (SVM) üç denetimli sınıflandırma yönteminin performansını araştırdı (Modaresi ve Araghinejad, 2014). Çalışmada Tahran Ovası'nın su kalitesi sınıflandırması incelendi. Destek Vektör Makineleri (SVM) algoritmasının daha iyi sonuçlar verdiği tespit edildi.

Malek ve arkadaşları, 2005 ile 2020 yılları arasında Kelantan Nehri'ne ait verileri ile makine öğrenmesi metotları kullanarak su kalitesinin sınıflandırılması konusunda çalıştılar (Malek ve ark., 2022). Biyokimyasal oksijen ihtiyacı, kimyasal oksijen ihtiyacı, amonyak azotu, çözünmüş oksijen, pH, toplam askıda katılar, sıcaklık, elektriksel iletkenlik, tuzluluk, bulanıklık, azot, fosfor ve escherichia coli olmak üzere toplamda 13 bağımsız değişken ile çalışma yaptılar. Karar Ağacı, Yapay Sinir Ağları, K-En Yakın Komşular, Naif Bayes, Destek Vektör Makinesi, Rastgele Orman ve Gradyan Artırma olmak üzere yedi makine öğrenmesi modelini kullandılar. Bu çalışmada en başarılı model %94,90 doğruluk oranı ile Gradyan artırmalı (Gradient Boosting Machines) algoritma ile oluşturulan model oldu.

Deng ve arkadaşları, Hong Kong'daki Tolo Limanı'nda yaşanan alg patlamasıyla ilgili makine öğrenmesi metotları ile bir çalışma yaptılar (Deng ve ark. 2021). Su kalitesi sorunlarının modellenmesi ve tahmin edilmesinde makine öğrenimi yöntemleri ile çalışan bu ekip Yapay sinir ağları (ANN) ve Destek vektör makinesi (SVM) yöntemlerini kullandılar. Toplam inorganik nitrojen, fosfor, klorofil, çözünmüş oksijen, su sıcaklığı, seki disk derinliği değişkenlerini kullanarak model oluşturdular. Bu çalışmada Yapay sinir ağlarının Destek vektör makinelerine göre daha verimli sonuçlar ortaya koydu. Destek vektör makinelerinin performansı, su kalitesi tahmin sonuçları açısından tüm Yapay sinir ağları modellerinden daha iyidir, ancak değişkenler ve çıktılar arasındaki doğrusal olmayan ilişkilerin dahil edilmesi nedeniyle daha düşük hesaplama verimliliğine sahip olduğu tespit edildi. Çalışma hem çevre yönetimi hem de su kalitesi tahmini için verimli bir çalışma olmuştur.

Vietnam'daki La Buong Nehri'nin yüzey suyu kalitesini tahmin etme çalışmasında Khoi ve arkadaşları Adaptif artırmalı (AdaBoost), Gradyan artırmalı (Gradient Boosting), Histogram tabanlı gradyan artırmalı (Histogram Based Gradient Boosting), Işık gradyan artırma (Light Gradient Boosting), Ekstra (aşırı) gradyan artırma (Extreme Gradient Boosting), Karar ağaçları, Rastgele orman, Ekstra ağaçlar, Yapay sinir ağları makine öğrenmesi yöntemlerini kullandılar (Khoi ve ark., 2022). Toplamda 10 adet değişken ile çalışma yaptılar. Makine öğrenimi modellerinin tahmin performansı, iki regresyon verimlilik istatistiği olan R^2 ve RMSE kullanılarak değerlendirildi. Aşırı gradyan artırma (XGBoost) modelinin R^2 = 0.989 ve RMSE = 0.107 olarak en iyi performansa sahip olduğu görüldü. Makine öğrenimi modellerinin yüksek derecede doğrulukla su kalitesini

tahmin etmek için kullanılabileceği ve bunun da su kalitesi yönetimini daha da iyileştireceği argümanını güçlendirdi.

Slatnia ve arkadaşları özellik seçimi analizi ile hibrit makine öğrenmesi metotlarını kullanarak su kalite indeksinin tahmini ve sınıflandırmasını iyileştirme çalışması ile ilgili araştırma yaptılar (Slatnia ve ark., 2022). Su kalite indeksi belirli bir konumdaki ve belirli bir zaman aralığındaki su kalitesi değişkenleri ile hesaplanmaktadır. Cezayir'deki Tilesdit Barajı'nda su kalite indeksi ile ilgili araştırma yapmak için, Yapay sinir ağları ve Destek vektör makineleri algoritmalarını kullandılar. Kullanılan her ki modelde başarılı sonuçlar verdi.

Su kalitesi tahmini ve sınıflandırması için Temel bileşen regresyonu ve Gradyan artırmalı modelleri kullanarak çalışma yapan Khan ve arkadaşları her iki modelden de başarılı sonuçlar elde etti (Khan ve ark., 2022). Temel bileşen regresyonu (Principal component regression) %95 tahmin doğruluğu, Gradyan artırmalı model ise %100 sınıflandırma doğruluğu sonucunu verdi.

Dogo ve arkadaşları, makine öğrenmesi tekniklerini kullanarak su kalitesi verilerindeki anormalleri tespit etmek için çalışma yaptılar (Dogo ve ark., 2019). Aynı zamanda Derin öğrenme (Deep learning) metotlarını da içeren bu çalışma makine öğrenmesi metotları ile karşılaştırma yapılmıştır. Ancak net bir sonuca varılmamıştır.

Su kalitesi esnekliği tahmin modeli üzerine çalışma yapan Imani ve arkadaşları, Yapay sinir ağlarını kullandılar (Imani ve ark., 2021). Su kalitesinin esnekliğini hesaplamak maliyetli, zorlu ve zaman alıcı olduğundan makine öğrenmesi yöntemlerinden faydalandılar. Brezilya'nın São Paulo Eyaletindeki 22 su havzasından elde edilen 17 yıllık bir su kalitesi veri seti, ANN modelini eğitmek ve test etmek için kullandılar. Su kalitesi esnekliği bu çalışmada yüzey suyunun kirlenmeyle başa çıkma ve kirlenmeden kurtulma kapasitesi olarak tanımlanmaktadır. Ölçülen ve simüle edilen Su kalitesi indeks (WQI - Water Quality Index) esneklik değerleri oldukça başarılı çıktı. Modelin daha dinamik bir esneklik tahmin sistemi için gerçek zamanlı veri izleme sistemlerinin entegrasyonu ile desteklenebileceğini savundular.

Haghiabi ve arkadaşları, makine öğrenmesi yöntemlerini kullanarak su kalitesi tahmini çalışması yaptılar (Haghiabi ve ark., 2018). Çalışmada İran'ın Tireh Nehri'nin su kalitesi parametrelerini incelediler. Çalışmada Yapay sinir ağları (ANN), Destek vektör makineleri (SVM) modellerini kullandılar. Su kalitesi bileşenlerinden arasında çözünmüş oksijen, kimyasal oksijen ihtiyacı, biyokimyasal oksijen ihtiyacı, elektriksel iletkenlik, pH,

sıcaklık, potasyum, sodyum, magnezyum, kalsiyum değişkenleri vardı. Sonuçlar, uygulanan modellerin su kalitesi bileşenlerini tahmin etmek için uygun performansa sahip olduğunu, ancak en iyi performansın SVM ile ilgili olduğunu gösterdi. Ayrıca model performansları karşılaştırıldığında SVM modellerinin sonuçlarının YSA'ya kıyasla daha güvenilir olduğu görüldü.

Muhammad ve arkadaşları, makine öğrenmesi metotlarını kullanarak su kalitesinin sınıflandırılmasının tahmin edilmesi üzerine çalıştılar (Muhammad ve ark. 2015). Malezya'nın Perak kentindeki Kinta Nehri'nin su kalitesinin sınıflandırılması için bir model geliştirmek için çalışma yapıldı. Çalışmada; çözünmüş oksijen, biyokimyasal oksijen çözünmesi, kimyasal oksijen çözünmesi, askıda katı madde, pH ve amonyak azotu (NH3-N) olmak üzere 6 değişken kullandılar. Weka ile yapılan bu çalışmada K-star algoritması %86.67 ile en üstün doğruluğa sahip olduğu görüldü. Örnek tabanlı bir yaklaşım olan K-star algoritması, test örneklerini bir benzerlik fonksiyonu kullanarak eğitim örnekleriyle karşılaştırarak sınıflandırmaktadır. Benzer örneklerin ortak bir sınıf etiketine sahip olduğu varsayılmaktadır. Algoritma, örnekler arasındaki benzerliği belirlemek için entropik mesafeyi kullanmaktadır.

Literatürde bilim insanları, akademik çalışanlar ve öğrenciler tarafından su kalitesinin tahmini üzerine farklı model çeşitleri ile çalışılmıştır. Düzenli veri setlerinin toplanması ve periyodik ölçümlerin ve aynı parametrelerin sistematik toplanmasındaki problemler nedeniyle bu konuda daha fazla zaman ve emek harcanması gerekmektedir.

5. SONUÇ VE ÖNERİLER

Bu çalışma, Güney Avustralya Hükümeti veri tabanından alınan veriler su kalite takibi ve tahmini için on bağımsız (alüminyum, amonyum, demir, kalsiyum, klorür, mangan, sülfat, pH, renk, bulanıklık) ve on makine öğrenmesi modeli (Gaussian Naive Bayes, K- en yakın komşu, destek vektör, yapay sinir ağları, karar ağaçları, rastgele orman, gradyan artırmalı, kategori artırmalı, ekstra gradyan artırmalı, lojistik regresyon model) kullanılarak değerlendirildi. Amacımız sadece su kalite takibi ve tahmini için değil aynı zamanda su kalitesine etkisi bakımından en önemli parametreleri kullanarak ölçüm istasyonlarından elde edilen verileri işleyerek su biliminin diğer yönleri için de yeni algoritmalar geliştirmek ve önermekti.

Modelleme süreci klorür, sülfat, demir, kalsiyum parametrelerinin konsantrasyonunun su kalitesinin en önemli belirleyicisi olduğunu ortaya çıkardı. Bunu önem sırasına göre bulanıklık, alüminyum, mangan, pH, renk ve amonyum takip etti. Elde edilen sonuçlar, karşılaştırıldığında, tahmin doğruluklarının gelişmiş olduğunu ancak her durumda o kadar başarılı olmayabileceğini gösterdi. Rastgele orman ve Lojistik regresyon modellerinin tahmin düzeyi diğer tüm modellerden daha iyiydi. Gaussian Naive Bayes modeli yüksek performansa sahip olmasına rağmen su kalite sınıflandırmasını diğer modellere göre yüksek doğrulukla tahmin edemedi.

Bu algoritmaların kısa zaman dilimini kapsayan bir veri seti kullanarak güvenilir sonuçlar ürettiğinde, daha uzun zaman dilimlerini kapsayan veri setleri için çok daha sağlam sonuçlar vereceğini belirtelim. Bu nedenle, bu algoritmalar, daha sınırlı ölçüm ağlarına sahip olan ve ölçüm ağlarının daha yakın zamanda kurulduğu gelişmekte olan bölgeler için özellikle yararlı olabilir. Sonuçlar, önerilen Rastgele orman ve lojistik regresyon algoritmalarının, veri setinin alındığı Güney Avustralya bölgesinin yüzey suyu kalite yönetimini iyileştirmek için güvenilir ve uygun maliyetli bir algoritma olabileceğini göstermektedir. Kullanılan bu modellerin, bazı su kalitesi parametrelerini ölçme maliyetlerinin yüksek olduğu ve genel olarak engelleyici olabileceği gelişmekte olan ülkelerde çok daha yararlı olması muhtemeldir. Ancak bu sonuçlar genelleştirilemez ve diğer çalışma alanlarına uygulanamaz veya diğer hidrolojik verilerle eşitlenemez.

Dünyadaki temiz su kaynaklarının hızlı bir şekilde kirletilmesiyle birlikte kullanılabilir su miktarındaki azalma gezegenimizin geleceği için en önemli problemlerinden bir tanesi olarak etkisini göstermektedir. Ekosisteme etkisinin yanısıra sağlık ve gelişmişlik

üzerindeki etkileri değerlendirildiğinde enerji probleminden bile çok ciddi bir kuvvet çarpanına sahiptir. Birincil enerji kaynakları olan petrol, kömür, doğal gazın alternatifi (Güneş enerjisi, rüzgâr enerjisi) varken maalesef temiz su kaynaklarının alternatifi mevcut değildir. Temiz ve kullanılabilir su kaynaklarının arttırılması ancak kirli kaynakların temizlenmesi mevcut kaynakların etkin kullanımı ve kirleticilere anında müdahale edilmesi ile mümkün olacaktır. Bu hayati durum birçok gelişmiş ülkenin gelecek planlarında yer almakla birlikte gelecek nesiller için ciddi bir kaygı kaynağı olmuştur. Dünyanın birçok bölgesinde olduğu gibi ülkemizde de su kalitesi tespiti ve takibi geleneksel olarak numune alımı ve laboratuvar analizleriyle gerçekleştirilmektedir. Bu tespit ve takip tekniği birçok kirliliğin ve kirlilik kaynağının geç tespitine ve giderek temiz su kaynaklarının kaybına yol açmakta ve temiz ve kullanılabilir su fakiri bir ülke olmamıza neden olmaktadır. Bu hızlı ve kötü gidişatı kısa vadede yavaşlatmak ve orta vadede durdurmak için kirlilik riskine maruz kalan su kaynaklarına yerleştirilecek sensörler yardımıyla su kalitesinin takibinin yanısıra kirlilik kontrolü için de zaman kazanılmış olacaktır. Unutmamak gerekir ki bütün su kalite parametrelerinin sensörlerle takibi maliyetli ve zahmetli bir uygulamadır. Bu nedenle bu çalışmada su kalitesini belirlemek ve takibini sağlamak için az sayıda parametreye dayalı sürdürülebilir yöntemlerin kullanımı tercih edildi. Kullanılan veri setinde ölçülen her parametre su kalitesi tespiti için kullanmak yerine en etkili olan on bağımsız değişkenin değerleri değeri kullanılarak su kalitesi tahmin edilmeye çalışıldı. Az sayıda değişkenle daha etkin sonuçlar elde edilebilmesi amacıyla birçok makine öğrenmesi yönteminin kullanılmasının yararlı olacağı fikri oluştu. Tam da bu perspektifle, bu değişkenlerin belirlenmesi sonucunda yeterli sayıda sensör yardımıyla sürdürülebilir bir su yönetiminin mümkün olacağı öngörülmektedir. Sonuç olarak, ileri teknoloji ürünü sensörlerin kullanılarak verilerin uydu üzerinden kontrol merkezlerine iletilmesiyle su kalitesinin sürdürülebilir olmasına, analizlerin periyodik ve anında yapılmasına, zamanında müdahale ve etkili bir su kirliliği kontrolünün yapılmasına olanak sağlayacaktır.

6. KAYNAKLAR

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., García-Nieto, J., 2019. Efficient water quality prediction using supervised machine learning. Water, 11(11), 2210.

Castrillo, M., García, Á. L., 2020. Estimation of high frequency nutrient concentrations from 62erce quality surrogates using machine learning methods. Water research, 172, 115490.

Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Ren, H., 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water research*, *171*, 115454.

Deng, T., Chau, K. W., ve Duan, H. F., 2021. Machine learning based marine water quality prediction for coastal hydro-environment management. Journal of Environmental Management, 284, 112051.

Derpanis, K. G., 2005. Mean shift clustering. Lecture Notes, 32(1), 4.

DeSimone, L. A., McMahon, P. B., Rosen, M. R., 2014. Water quality in principal aquifers of the United States, 1991–2010, circular 1360. *US Geological Survey. https://doi. org/10.3133/cir1360*.

Dogo, E. M., Nwulu, N. I., Twala, B., Aigbavboa, C., 2019. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. *Urban Water Journal*, *16*(3), 235-248.

Eren, B., Çelebi, A., 2018. Sakarya Nehri Su Kalitesinin İstatistiksel Veri Değerleme Yöntemleri Kullanılarak Değerlendirilmesi. Academic Perspective Procedia, 1(1), 1347-1356.

Haghiabi, A. H., Nasrolahi, A. H., Parsaie, A., 2018. Water quality prediction using machine learning methods. *Water Quality Research Journal*, *53*(1), 3-13.

Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H., 2009. The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 309). New York: springer.

Hatipoğlu, P. U., 2016. Time series classification using deep learning Time series classification using deep learning [M.S. – Master of Science]. Middle East Technical University.

Imani, M., Hasan, M. M., Bittencourt, L. F., McClymont, K., Kapelan, Z., 2021. A novel machine learning application: Water quality resilience prediction Model. *Science of the Total Environment*, *768*, 144459.

IPCC (Intergovernmental Panel on Climate Change)., 2023. Summary for Policymakers. H. Lee and J. Romero (eds.), Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, IPCC, pp. 1–34. www.ipcc.ch/report/ar6/syr/downloads/report/ IPCC_AR6_SYR_SPM.pdf.

Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., Ashraf, I., 2022. Water quality prediction using KNN imputer and multilayer perceptron. Water, 14(17), 2592.

Kaddoura, S., 2022. Evaluation of machine learning algorithm on drinking water quality for better sustainability. Sustainability, 14(18), 11478.

Keskenler, M. F., Keskenler, E. F., 2017. Geçmişten günümüze yapay sinir ağları ve tarihçesi. Takvim-i Vekayi, 5(2), 8-18.

Khan, M. S. I., Islam, N., Uddin, J., Islam, S., Nasir, M. K., 2022. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University-Computer and Information Sciences*, *34*(8), 4773-4781.

Khan, Y., See, C. S., 2016. Predicting and analyzing water quality using machine learning: a comprehensive model. In 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT) (pp. 1-6). IEEE.

Khoi, D. N., Quan, N. T., Linh, D. Q., Nhi, P. T. T., Thuy, N. T. D., 2022. Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. Water, 14(10), 1552.

Mahesh, B., 2020. Machine learning algorithms-a review. International Journal of Science and Research (IJSR). [Internet], 9(1), 381-386.

Malek, N. H. A., Wan Yaacob, W. F., Md Nasir, S. A., Shaadan, N., 2022. Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques. Water, 14(7), 1067.

McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E., 2006. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine, 27(4), 12-12.

Mekonnen, M. M., Hoekstra, A. Y., 2016. Four billion people facing severe 63erce scarcity. Science advances, 2(2), e1500323.

Modaresi, F., ve Araghinejad, S., 2014. A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. *Water resources management*, *28*, 4095-4111.

Muhammad, S. Y., Makhtar, M., Rozaimee, A., Aziz, A. A., Jamal, A. A., 2015. Classification model for water quality using machine learning techniques. International Journal of software engineering and its applications, 9(6), 45-52.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., Al-Shamma'a, A., 2022. Water quality classification using machine learning algorithms. Journal of Water Process Engineering, 48, 102920.

Pandey, M. K., Singh, M. K., Pal, S., Tiwari, B. B., 2022. Analysis of Phishing Base Problems Using Random Forest Features Selection Techniques and Machine Learning Classifiers. In Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2022 (pp. 53-64). Singapore: Springer Nature Singapore.

Patro, K. S. K., Yadav, V. K., Bharti, V. S., Sharma, A., Sharma, A., Senthilkumar, T., 2023. IoT and ML approach for ornamental fish behaviour analysis. Scientific Reports, 13(1), 21415.

Pirim, H., 2006. Yapay zekâ. Yaşar Üniversitesi E-Dergisi, 1(1), 81-93.

Radhakrishnan, N., Pillai, A. S., 2020. Comparison of water quality classification models using machine learning. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1183-1188). IEEE.

Raghunath, H. M., 2006. Hydrology: principles, analysis and design. New Age International.

Schwartz, H. M., 2014. Multi-agent machine learning: A reinforcement approach. John Wiley & Sons.

Shams, M. Y., Elshewey, A. M., El-kenawy, E. S. M., Ibrahim, A., Talaat, F. M., Tarek, Z., 2023. Water quality prediction using machine learning models based on grid search method. Multimedia Tools and Applications, 1-28.

Slatnia, A., Ladjal, M., Ouali, M. A., Imed, M., 2022. Improving Prediction and classification of Water Quality Indices using Hybrid Machine learning Algorithms with features selection analysis. In Online International Symposium on Applied Mathematics and Engineering (ISAME22) January 21-23, 2022, Istanbul-Turkey (p. 16).

Sugiyama, M., 2015. Statistical reinforcement learning: modern machine learning approaches. CRC Press.

Sutton, R. S., Maei, H., Szepesvári, C., 2008. A convergent on temporal-difference algorithm for off-policy learning with linear function approximation. Advances in neural information processing systems, 21.

T.C. Resmi Gazete, İçme Suyu Temin Edilen Suların Kalitesi ve Arıtılması Hakkında Yönetmelik (30823), 06.07.2019.

T.C. Resmi Gazete, İnsani Tüketim Amaçlı Sular Hakkında Yönetmelik (28580), 07.03.2013.

UNESCO, 2006. Water Shared Responsibility, The United Nations World Water Report II, Paris, s.121.

UNESCO, 2024. The United Nations World Water Development Report 2024

United Nations, 2023. The Sustainable Development Goals Report – Special Edition. New York, United Nations. https://unstats.un.org/sdgs/report/2023/.

URL-1, 2022. https://ceng.cu.edu.tr/uorhan/DersNotu/Ders03.pdf 02 Aralık 2022

URL-2, 2014. https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/. 02 Aralık 2022

URL-3, 2022. https://www.researchgate.net/figure/An-example-of-SVM-classification-An-example-of-SVM-classification_fig1_336085357 02 Aralık 2022

URL-4, 2024. https://en.wikipedia.org/wiki/Artificial_neuron

URL-5, 2024. https://water.usgs.gov/edu/gallery/watercyclekids/earth-water-distribution.html

URL-6, 2024. https://medium.com/@batubilgili1907.bb/lasso-regression-lasso-regresyon-46dcea98868a

URL-7, 2024. https://medium.com/machine-learning-türkiye/adım-adım-makine-öğrenmesi-bölüm-3-denetimsiz-örenme-nedir-f890ada49a40

URL-8, 2024. https://www.enjoyalgorithms.com/blog/classification-of-machine-learning-models

URL-9, 2024. https://data.sa.gov.au/data/dataset/water-quality

URL-10, 2024. South Avustralian Government Data Directory, Sourced on 01 January 2024, https://data.sa.gov.au/data/dataset/996ec2ae-d52c-4d7e-be9c-d4dab1c1aa45

Velev, D., Zlateva, P., 2023. Issues of Artificial Intelligence Application in Digital Marketing. In Digitalization and Management Innovation II (pp. 52-59). IOS Press.

Watkins, C. J., Dayan, P., 1992. Q-learning. Machine learning, 8, 279-292.

WHO, UNICEF., 2021. Progress on household drinking water, sanitation and hygiene 2000-2020: five years into the SDGs. World Health Organization.

World Health Organization, 2003. Aluminium in drinking-water: background document for development of WHO Guidelines for drinking-water quality (No. WHO/SDE/WSH/03.04/53). World Health Organization.

ÖZGEÇMİŞ

EKLER

EK 1: Hiper Parametreler

Hiper parametre ayarlaması veya hiper parametre optimizasyonu, makine öğrenmesi modellerinin performansını artırmak için modelin hiper parametrelerini en iyi değerlerle ayarlama sürecidir. Hiper parametreler, modelin öğrenme sürecini ve mimarisini doğrudan etkileyen parametrelerdir.

Doğru hiper parametre seçimi, modelin performansını optimize etmek ve aşırı uyumu azaltmak için önemlidir. Bu nedenle, Grid Search veya Randomized Search gibi tekniklerle bu parametrelerin en iyi değerlerini bulmak yaygın yöntemlerdir.

Makine öğrenmesi algoritmalarına ait hiper parametreler ve açıklamaları aşağıda verildi.

Gaussian Naive Bayes Hiper Parametreleri

var_smoothing: Hesaplama sırasında sayısal kararlılığı artırmak için varyansa eklenen küçük bir pozitif değerdir. Bu, modelin aşırı küçük varyans değerlerinden dolayı kararsız hale gelmesini önler.

K- En Yakın Komşular Hiper Parametreleri

n_neighbors (k): Kullanılacak komşu sayısıdır. Bu, en önemli hiper parametredir ve modelin esnekliğini belirler. Küçük bir k değeri modelin varyansını artırırken, büyük bir k değeri modelin yanlılığını artırabilir.

weights: Komşuların ağırlıklandırma stratejisidir. Değerler aşağıdaki gibidir:

'uniform': Tüm komşular eşit ağırlığa sahiptir.

'distance': Komşuların mesafelerine göre ağırlıklandırma yapılır (yakın komşular daha yüksek ağırlığa sahiptir).

Özel bir fonksiyon: Kullanıcı tanımlı ağırlıklandırma fonksiyonu.

algorithm: Komşuları bulmak için kullanılan algoritmadır. Argümanları aşağıdaki gibidir:

'auto': En iyi algoritmayı otomatik olarak seçer.

'ball_tree': BallTree algoritmasını kullanır.

'kd_tree': KDTree algoritmasını kullanır.

'brute': Brute-force arama yapar.

leaf_size: BallTree veya KDTree algoritmaları kullanılırken yapının yaprak boyutudur. Bu parametre, yapılandırma süresini ve sorgu hızını etkiler.

p: Minkowski mesafe metriğinde kullanılan güç parametresidir. p=1 Manhattan mesafesini (L1), p=2 ise Öklid mesafesini (L2) belirtir. Minkowski mesafesi, genelleştirilmiş bir metrik olup, p=1 ve p=2'nin ötesindeki değerler için de kullanılabilir.

metric: Kullanılan mesafe metriğidir. Varsayılan olarak 'Minkowski' metriği kullanılır ve bu, p parametresine bağlıdır. Diğer yaygın metrikler aşağıdaki gibidir:

'euclidean': Öklid mesafesi.

'manhattan': Manhattan mesafesi.

'minkowski': Minkowski mesafesi.

Destek Vektör Makineleri (Support Vector Machine) Hiper Parametreleri

kernel: Veriyi daha yüksek boyutlu bir uzaya dönüştürerek doğrusal olarak ayrılamayan veri setlerinin ayrılabilmesini sağlar. Yaygın çekirdek fonksiyonları aşağıdaki gibidir:

'linear': Doğrusal çekirdek.

'poly': Polinom çekirdek.

'rbf': Radial Basis Function (RBF) veya Gaussian çekirdek.

'sigmoid': Sigmoid çekirdek.

C: SVM modelinin ceza parametresidir ve modelin eğitim verilerine uymasını kontrol eder.

n_jobs: Özellikle büyük veri setleri veya karmaşık modellerle çalışırken, işlemleri paralelleştirerek hesaplama süresini azaltmak için kullanılır.

n_jobs=-1: Mevcut tüm işlemci çekirdeklerini kullanır ve işlemleri paralel olarak yürütür. n_jobs=1: Paralel işlemleri devre dışı bırakır ve yalnızca tek bir iş parçacığı kullanır.

verbose: Makine öğrenimi algoritmalarının eğitim süreci sırasında çıktıların ne kadar ayrıntılı olacağını kontrol etmek için kullanılır. Bu parametre, algoritmanın ilerlemesini izlemek veya hata ayıklama yapmak için yararlı olabilir. Ancak, gereksiz

ayrıntılar, çıktıların okunmasını zorlaştırabilir ve konsol ekranını gereksiz yere kalabalıklaştırabilir. Bu nedenle, verbose parametresini doğru bir şekilde ayarlamak önemlidir.

Yapay Sinir Ağları Hiper Parametreleri

alpha: L2 düzenleme (regularization) parametresidir. Ağırlıkların büyüklüğünün cezalandırılmasını kontrol eder ve aşırı uymayı azaltır. Alpha'nın değeri ne kadar büyükse, düzenleme miktarı o kadar yüksek olur ve modelin karmaşıklığını azaltır. Genellikle, alpha'nın belirli bir aralıkta seçilmesi ve cross-validation kullanılarak en iyi değerin bulunması önerilmektedir.

hidden_layer_sizes: Gizli katmanların boyutunu belirlemektedir.

solver: Ağırlık optimize edici algoritmayı belirtmektedir. "lbfgs", "adam" ve "sgd" gibi farklı optimizasyon algoritmaları kullanılabilir.

activation: Aktivasyon fonksiyonunu belirlemektedir. "relu" (Rectified Linear Activation), "logistic" (Logistic Sigmoid Activation) gibi farklı aktivasyon fonksiyonları seçilebilir. Aktivasyon fonksiyonları, modelin temsili gücünü artırabilir ve modelin öğrenme sürecini etkilemektedir.

Karar Ağaçları Hiper Parametreleri

max_depth: Karar ağacının maksimum derinliğini belirtmektedir. Karar ağacı ne kadar derin olursa, model o kadar karmaşık hale gelmektedir. Ağacın aşırı uyumunu önlemek için maksimum derinlik sınırlanmaktadır.

min_samples_split: Bir düğümün bölünmeden önce minimum örnek sayısını belirtmektedir. Bir düğümün bölünmeden önce kaç tane en az örneğe sahip olması gerektiğini kontrol etmektedir.

max_features: Her bir bölünme için dikkate alınacak maksimum özellik (değişken) sayısını belirtmektedir.

n_estimators: Ormandaki karar ağacı sayısını belirlemektedir. Ağaç sayısının artması modelin performansını iyileştirebilir, ancak hesaplama maliyetini de artırmaktadır.

learning_rate: Her bir ağacın katkısının azaltılması için kullanılan orandır. Düşük öğrenme oranları daha fazla ağaç gerektirmektedir, ancak modelin genel performansını artırmaktadır.

subsample: Her bir ağacı eğitmek için kullanılacak eğitim veri setinin oranıdır.

Feature Importance: Özellik (değişken) önemini belirlemek için kullanılmaktadır. Hangi özelliklerin model için daha önemli olduğunu göstermektedir.

iterations: Bu parametre, modelin kaç tur (iterasyon) çalıştırılacağını belirlemektedir. Daha yüksek bir değer, modelin daha fazla öğrenmesini sağlar ancak eğitim süresini de artırmaktadır.

Uzun Kısa Süreli Bellek (LSTM) Derin Öğrenme Modelinin Hiper Parametreleri

Hidden Units (Units): Her LSTM katmanındaki gizli birimlerin sayısıdır. Bu, LSTM'nin kapasitesini ve öğrenme gücünü belirlemektedir.

Number of Layers: LSTM katmanlarının sayısıdır. Daha fazla katman, modelin daha karmaşık veri ilişkilerini öğrenmesine olanak tanır, ancak aynı zamanda daha fazla hesaplama gerektirir ve overfitting (aşırı öğrenme) riskini artırmaktadır.

Dropout Rate: Ağırlıkların rastgele sıfırlanması için kullanılan dropout oranıdır. Aşırı öğrenmeyi önlemek için kullanılmaktadır.

Recurrent Dropout Rate: LSTM hücrelerindeki tekrarlanan bağlantılarda kullanılan dropout oranıdır.

Learning Rate: Optimizasyon algoritmasının öğrenme hızıdır. Modelin ağırlıklarını güncelleme hızını belirlemektedir.

Batch Size: Modeli eğitmek için kullanılan örneklerin sayısıdır. Modelin her adımda kaç örnek üzerinde güncelleneceğini belirlemektedir.

Epochs: Tüm eğitim veri seti üzerinden kaç kez geçileceğini belirtmektedir. Modelin eğitileceği yineleme sayısını belirlemektedir.

Optimizer: Modelin öğrenme sürecini optimize etmek için kullanılan algoritmayı belirtmektedir. Yaygın olarak kullanılan optimizasyon algoritmaları arasında Adam, RMSprop ve SGD bulunmaktadır.

Activation Function: LSTM katmanlarındaki aktivasyon fonksiyonlarıdır. Genellikle tanh ve sigmoid fonksiyonları kullanılmaktadır.

Sequence Length (Time Steps): LSTM'ye girilen zaman adımlarının sayısıdır. Modelin aynı anda kaç zaman adımını işleyeceğini belirlemektedir.