ULUBORLU MESLEKİ BİLİMLER DERGİSİ (UMBD)
Uluborlu Journal of Vocational Sciences

http://dergipark.gov.tr/umbd

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE SU KALİTESİ VE İÇİLEBİLİRLİK TAHMİNİ

Tülay TURAN[1*]

[1*] Burdur Mehmet Akif Ersoy Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Burdur, Türkiye.

[*]Sorumlu Yazar: tulayturan@mehmetakif.edu.tr

(**Geliş/Received:** 02.11.2023 **Kabul/Accepted:** 14.11.2023)

ÖZET: İçme suyu insanların yaşamlarını sürdürebilmeleri için hayati önem taşıyan temel ihtiyaçlarının başında gelmektedir. İnsan sağlığını doğrudan etkileyen bu ihtiyacın kalitesini ve içilebilirliğini anlamak önemlidir. Su kalitesi geleneksel laboratuvar ve istatistiksel analizler yoluyla tahmin edilebilir. Ancak bu çözüm genel olarak pahalı ve zaman alıcıdır. Son yıllarda hızla gelişen, hayatımızın bir çok alanına fayda sağlayan makine öğrenmesi yöntemleri ile su kullanılabilirliği hızlı ve verimli bir şekilde analiz edilebilir. Bu bağlamda gerçekleştirilen çalışmada, su kalitesinin ve içilebilirliğinin tahmini için 15 farklı makine öğrenmesi algoritması ile modeller geliştirilmiş ve elde ettikleri sonuçlar karşılaştırılmıştır. Model değerlendirmelerinde en iyi tahmin performansını LGBMClassifier ve SVC algoritmalarının sağladığı görülmüştür. En iyi tahmin performansını gösteren bu iki model için GridSearchCv nesnesi kullanılarak hiper parametre optimizasyonu gerçekleştirilmiştir. Optimizasyon işleminden sonra LGBMClassifier modeli %0,92 accuracy değeri ile en başarılı tahmin sonucunu elde etmiştir. Çalışma su kalitesi ve içilebilirliğini etkileyen faktörleri analiz etmesi, görselleştirmesi ve yüksek tahmin performansı ile gelecek çalışmalara yön verecektir.

Anahtar Kelimeler: LGBMClassifier, Makine Öğrenmesi, Su Kalitesi ve İçilebilirliği, Yapay Zeka

WATER QUALITY AND POTABILITY PREDICTION WITH MACHINE LEARNING ALGORITHMS

ABSTRACT: Drinking water is one of the basic needs of people that is vital for their survival. It is important to understand the quality and potability of this requirement, which directly affects human health. Water quality can be estimated through conventional laboratory and statistical analysis. However, this solution is generally expensive and time consuming. Water availability can be analyzed quickly and efficiently with machine learning methods, which have developed rapidly in recent years and benefit many areas of our lives. In this context, models were developed with 15 different machine learning algorithms to predict water quality and drinkability and their results were compared. In model evaluations, it was seen that LGBMClassifier and SVC algorithms provided the best prediction performance. Hyperparameter optimization was performed using the GridSearchCv object for these two models that showed the best prediction performance. After the optimization process, the LGBMClassifier model achieved the most successful prediction result with an accuracy value of 0.92%. The study will guide future studies with its analysis and visualization of the factors affecting water quality and drinkability and its high prediction performance.

Keywords: Artificial Intelligence, LGBMClassifier, Machine Learning, Water Quality and Potability

1. GİRİŞ

İnsan tüketimine yönelik su kalitesi hem insan sağlığının korunması hem de halk sağlığı politikasının bir parçasını olduğu için önemlidir. "İçme suyu" olarak bilinen, insan tüketimine uygun kaliteli suya erişim, temel bir insan hakkı olup, bireylerin ve toplumların sağlıklı yaşam ve kalkınması için bir zorunluluktur. Bu hak, Temmuz 2010'da BM'nin 64/292 sayılı kararı ile uluslararası hukukta koruma altına alınmıştır [1].

Dünyadaki çoğu içme suyu kaynağı, atık ürünlerin (kanalizasyon, plastik gibi) deşarjı, tarım ilaçları ve kentsel yerleşimlerden kaynaklanan (tortu, tuz ve mineraller gibi) etkilerden etkilenmektedir [2]. Bu durum su kullanımı ve içme suyunun güvenliğini ciddi şekilde tehdit etmektedir [3]. Ayrıca su kalitesi su kütlesinin iklimsel ve jeokimyasal konumundan, sıcaklık, yağış, sızıntı ve yer kabuğundaki elementlerin akışı yoluyla ile de etkilenmektedir [4]. Bu hayati kaynağın kalitesini etkileyen mevcut etkenleri daha iyi anlamaya ihtiyaç vardır. Su kalitesi, yerel koşullara ve ihtiyaçlara göre laboratuvarda veya evde test edilebilir [5] Su kalitesinin laboratuvar değerlendirmesi, toplanan su örneklerinin enstrümantal ve kimyasal analizlerine dayanmaktadır [6]. Laboratuvarlarda yapılan bu analizler yüksek doğrulukta sonuçlar verebilirken, gerçekleştirilen testler maliyetli ve uzun zaman gerektirmektedir. Evde yapılan testeler ise laboratuvarlarda gerçekleştirilen testlere göre daha az doğrulukla sonuçlar elde edilmesini sağladığı için çok tercih edilmemektedir [7].

Son yıllarda su kalitesini ve içilebilirliğini belirlemek için yapay zeka destekli çözümler geliştirilmeye başlanmıştır [8]. Su kalitesini etkileyen özellikleri değerlendirmek için veriye dayalı yapay zeka çözümleri geliştirmek, az maliyetle kısa sürede sonuçlar elde edilmesi yönüyle önemlidir [9]. Bu bağlamda, çalışmada su kalitesinin ve içilebilirliğinin tahmin edilmesi için denetimli makine öğrenimine dayalı bir yöntem geliştirmek amaçlanmıştır. Bu amaçla LGBM Classifier, NuSVC, SVC, XGB Classifier, Quadratic Discriminant Analysis, Extra Trees Classifier, Random Forest Classifier, Bagging Classifier, KNeighbors Classifier, Label Spreading, Decision Tree Classifier, AdaBoost Classifier, Logistic Regression, Dummy Classifier ve Ridge Classifier makine öğrenmesi algoritmaları kullanılarak modeller geliştirilmiştir. Çalışmada pandas ve numpy kütüphaneleri kullanarak verilerden veri analizi teknikleriyle değerlendirmeler elde edilmiştir. Bununla beraber veri görselleştirmeleri için matplotlib ve seaborn kütüphaneleri kullanılmıştır. Her veri görselleştirmesi ile verilerin farklı özelliklerinin vurgulanması amaçlanmıştır.

Çalışmada sırasıyla; birinci bölümünde su verimliliği ve içilebilirliği ile ilgili literatür çalışmaları incelenmiştir. İkinci bölümde çalışmada kullanılan teknolojiler ve yöntemler ayrıntılı bir şekilde açıklanmıştır. Üçüncü bölümde veri analizi ve makine öğrenmesi modellerinin geliştirilmesi, uygulama işlem basamakları açıklanmıştır. Son bölümde çalışmada elde edilen sonuçlar ve gelecekte yapılabilecek çalışmalar hakkında bilgiler verilmiştir.

2. LİTERATÜR

Su kalitesinin ve içilebilirliğinin belirlenmesine yönelik çalışmalar incelendiğinde; Kaddoura (2022) çalışmasında, su kirliliği problemlerinde makine öğrenmesi (ML) algoritmalarının kullanımının etkinliği araştırmıştır. Çalışmasında, makine öğrenimi algoritmalarını karşılaştırmış ve destek vektör makineleri (DVM) modelinin %73 ile en iyi performans değerini elde ettiğini belirtmiştir. Gelecek çalışmalarda hiper parametre ayarı ile modelin geliştirilebileceğini ifade etmiştir [10].

Poudel ve arkadaşları (2022) çalışmalarında, güvenli içme suyu eksikliği günümüzde giderek artan bir sorun olduğunun altını çizmişlerdir. Çalışmalarında k-en yakın komşu (KNN), rastgele

orman (RF) ve yapay sinir ağları (ANN) algoritmaları ile suyun içilebilir olup olmadığını tahmin etmeye çalışmışlardır. Sonuç olarak RF modeli %70,42 ile en iyi tahmin sonucunu sağladığını belirtmişlerdir. İstatistiksel olarak ML algoritmalarının kullanımını başarılı bulmuşlardır [11].

Hag ve arkadaşları (2021) çalışmalarında, su kalitesini tanımlamak için karar ağacı (DT) ve naive bayes (NB) algoritmalarını kullanmışlardır. Çalışmada iki modelin performans türü karşılaştırılmıştır. Makine öğrenimi modellerini değerlendirmek için k-katlı çapraz doğrulama yöntemi kullanmışlardır. Çapraz doğrulama yönteminin modelin aşırı uyum derecesinin belirlenmesinde de önemli olduğunu vurgulamışlardır [12].

Patel ve arkadaşları (2022) çalışmalarında, ilk olarak veri setini dengelemek için sentetik azınlık aşırı örnekleme tekniği (SMOTE) kullanmışlardır. RF, XGBoost, DT, Gradient Boost ve SVM ile modeller geliştirmişlerdir. Deney sonuçlarında, Rastgele Orman ve Gradient Boost'un %81 ile en yüksek doğruluğu sağladığını belirtmişlerdir. Çalışmalarında makine öğrenimi modelinde şeffaflığın bulunmamasının, modelin sonuçlarının değerlendirilmesini imkansız kıldığına dikkat çekmişlerdir. Bu konuyu ele almak için açıklanabilir yapay zeka tekniklerinden LIME'ı çalışmalarında kullanmışlardır [13].

Kurra ve arkadaşları (2022) çalışmalarında, su kalitesi sınıfını tahmin etmek için DT ve KNN algoritmaları kullanmışlardır. Çalışmalarında "Andhra Pradesh" çevresinden elde edilen su bilgilerini içeren bir veri seti kullandıklarını belirtmişlerdir. KNN sınıflandırıcısının diğer sınıflandırıcılardan daha iyi performans gösterdiği, bulgulara göre, makine öğrenimi yaklaşımlarının doğru bir şekilde tahmin etme yeteneğine sahip olduğuna dikkat çekmişlerdir [14].

Dawood ve arkadaşları (2021) çalışmalarında, su boru hatlarına kimyasal, mikrobiyal gibi sayısız fiziksel kirletici maddelerin sızması nedeniyle içme suyunun tehlikeye girdiğine dikkat çekmişlerdir. Çalışmalarında içme suyu kalitesi etkileyen özellikleri ANN ve risk analizi tekniklerini kullanarak araştırmışlardır. ANN modellerinin %92 doğruluk oranı elde ettiğinin belirtmişlerdir. Tahmin sonuçlarının bölge sakinleri için risk düzeyini (düşük, orta, yüksek) göstereceği ve böylece herhangi bir hastalık veya hastalık salgınını önlemek için önleyici tedbirler alınabileceğinin altını çizmişlerdir [15].

Yusuf ve arkadaşları (2022), Dünya Sağlık Örgütü'ne (WHO) göre su hastalıklarına bağlı küresel ölümlerin sayısının yılda yaklaşık iki buçuk milyon kişi olduğuna dikkat çekmişlerdir. Bu nedenle güvenli su kalitesinin analizi ve sınıflandırılmasına yönelik yeni uygulamaların araştırılması gerektiğini savunmuşlardır. Araştırmalarında k-en yakın komşu, karar ağacı , rastgele orman , yapay sinir ağı , lojistik regresyon ve destek vektör makinesi sınıflandırma algoritmaları kullanılmıştır. İçme suyunun içilebilirliğini sınıflandırmada RF ve DT modelleri sırasıyla %83,78 ve %74,98 doğrulukla en yüksek performansı elde ettiği görülmüştür [16].

Nasir ve arkadaşları (2022) çalışmalarında, 2005 ile 2014 yılları arasında Hindistan'ın çeşitli yerlerinden toplanan 1600'den fazla su örneği içeren veri setini kullanmışlardır. Veri kümesinin değişkenlerinin çözünmüş oksijen, pH, iletkenlik, biyokimyasal oksijen ihtiyacı, nitrat, dışkı koliformu ve toplam koliformdan oluştuğunu belirtmişlerdir. Çalışmalarında destek vektör makinesi, rastgele orman, lojistik regresyon, CATBoost, XGBoost, karar ağacı ve çok katmanlı algılayıcı algoritmalarını kullanmışlardır. Hassasiyet, F1 puanı ve doğruluk gibi performans ölçümlerini, her sınıflandırıcının karışıklık matrisi kullanılarak tahmin edildiği görülmüştür [17].

Abulail ve arkadaşları (2023), suyun fiziksel ve kimyasal parametrelerini kullanarak su kalitesini tahmin etmek için model odaklı bir karar destek sistemi önermişlerdir. Çalışmalarında test edilen numunenin fizyokimyasal kalite parametrelerini eklemek için grafiksel bir kullanıcı arayüzü oluşturmuşlardır. Önerilen karar destek sistemi suyun kalitesini tahmin etmekte ve içmenin güvenli olup olmadığını sınıflandırmaktadır. Ayrıca suyun hangi amaçlarla kullanılabileceğini de belirlemektedir [18].

Sirikarin ve Khonthapagdee (2023) çalışmalarında, Tayland Kirlilik Kontrol Departmanından 2009'dan 2021'e kadar toplanan verileri kullanarak, Tayland'daki dört ana nehirdeki (Ping, Wang, Yom ve Nan nehirleri) su kalitesini sınıflandırmışlardır. Çalışmalarında rastgele orman, extreme gradient boosting, lojistik regresyon ve destek vektör makinesi algoritmalarını kullanmışlardır. Ayrıca, dengesiz verilerle başa çıkmaya yönelik sentetik azınlık aşırı örnekleme tekniği (SMOTE) ve Rastgele aşırı örnekleme de F1 puanını iyileştirmek için kullandıklarını belirtmişlerdir. SMOTE özellikli XGBoost'un en yüksek puanı aldığını ve su kalitesinin sınıflandırılmasında bod'un en önemli özellik olduğunu bulmuşlardır [19].

Singh ve Lilhore (2023) çalışmalarında su kalitesinin tahmin edilmesinde doğruluğu arttırmak için hibrit bir model önermişlerdir. Hibrit sınıflandırma algoritmasında PCA, K-ortalaması ve oylama sınıflandırmasının birleştirmişledir. Hibrit modelleri %96 tahmin başarısı elde etmiştir. Gelecek çalışmalarında optimizasyon yöntemleri ile tahmin başarısını arttırabileceklerini belirtmişlerdir [20].

Literatür taramalarında incelenen çalışmaların, su verimliliği ve içilebilirliği tahmininde kullandıkları algoritmalar Tablo 1'de gösterilmektedir.

Tablo 1. Literatür taramasında incelenen çalışmalar.

Çalışma Adı	Model
Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability [10]	DVM
Comparison of machine learning algorithms in statistically imputed water potability dataset [11]	RF
Classification of Water Potability Using Machine Learning Algorithms [12]	DT
A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI [13]	RF Gradient Boost
Water Qualıty Predıctıon Usıng Machıne Learnıng [14]	KNN
Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks [15]	YSA
Classification of Water Potability Using Machine Learning Algorithms [16]	RF
Water quality classification using machine learning algorithms [17]	CATBoost
Machine Learning Techniques for Water Quality Classification of Thailand's Rivers [19]	XGBoost
Water Quality Prediction Using Hybrid Classification Model [20]	Hybrid

3. MATERYAL VE YÖNTEM

3.1. Veri Seti

Çalışmada kullanılan veri seti Kaggle veri paylaşım havuzundan alınmıştır. Veri seti, suyun insan tüketimine uygunluğu anlamına gelen içilebilirliğe ilişkin su kalitesi ölçümlerini ve değerlendirmelerini içermektedir. Veri kümesinin birincil amacı, su kalitesi parametreleri hakkında bilgi sağlamak ve suyun içilebilir olup olmadığının belirlenmesine yardımcı olmaktır.

Veri kümesindeki her satır, belirli niteliklere sahip bir su örneğini temsil eder ve "İçilebilirlik" sütunu, suyun tüketime uygun olup olmadığını göstermektedir. Veriler 10 değişken ve 3277 kayıttan oluşmaktadır [21]. Tablo 2'de veri seti içerisinde yer alan değişkenler ve açıklamaları gösterilmektedir.

Tablo 2. Veri seti alanları ve açıklamaları.

Değişkenler	Açıklaması
pH	pH değeri asit-baz dengesini belirler ve suyun asidik mi yoksa alkali mi olduğunu gösterir.
Sertlik	Suyun kalsiyum ve magnezyumun mineral içeriğinin bir ölçüsüdür.
Katılar	Suyun mineralize olup olmadığını göstermektedir.
Kloraminler	Suyun arıtılası sonucu oluşan kloramin içeriğinin ölçüsüdür.
Sülfat	Sudaki sülfat konsantrasyonu ölçüsüdür.
İletkenlik	Suyun elektriksel iletkenliği belirtmektedir.
Organik_karbon	Sudaki organik karbon içeriği ölçüsüdür.
Trihalometanlar	Sudaki trihalometan konsantrasyonu seviyesidir. Organik madde miktarına, klor miktarına ve su sıcaklığına göre değişiklik gösterir.
Bulanıklık	Bulanıklık seviyesi, suyun berraklığının bir ölçüsüdür. Atık deşarjının kalitesini göstermektedir.
Kullanılabilirlik	Hedef değişkendir; 1 (içilebilir) ve 0 (içilemez) anlamındadır.

3.2. Sınıflandırma Algoritmaları

Çalışmada kullanılan veri kümesi, su kalitesi özelliklerine göre suyun içilebilirliğini tahmin etmek için makine öğrenimi modellerinin eğitilebildiği denetimli ikili sınıflandırma görevi için uygundur. Modeller, su numunelerini içilebilir (1) veya içilemez (0) olarak sınıflandırmayı amaçlamaktadır. Gerçekleştirilen çalışmada suyun kalitesi ve içilebilirliğinin sınıflandırması için 15 farklı makine öğrenmesi algoritması kullanılmıştır.

3.2.1. LGBMClassifier

LGBMClassifier, hafif gradyan artırıcı makine sınıflandırıcı anlamına gelmektedir. Sıralama, sınıflandırma ve diğer makine öğrenimi görevleri için karar ağacı algoritmalarını kullanmaktadır [22]. LGBMClassifier, büyük ölçekli verileri doğru bir şekilde işlemek, verileri etkili bir şekilde hızlandırmak ve bellek kullanımını azaltmak için yeni bir gradyan tabanlı tek taraflı örnekleme (GOSS) ve özel özellik paketleme (EFB) tekniğini kullanmaktadır.

3.2.2. NuSVClassifier

NuSVC, çok sınıflı sınıflandırma yapabilen, scikit-learn'in sağladığı diğer bir sınıftır. Eğitim hatalarının fraksiyonunun bir üst sınırını ve destek vektörlerinin fraksiyonunun bir alt sınırını temsil etmektedir [23]. Elde edilen değer (0,1) aralığında olmalıdır. Parametrelerin ve niteliklerin geri kalanı SVC algoritması ile aynı özelliklerdedir.

3.2.3. SVClassifier

Destek vektör makineleri, sınıflandırma, regresyon ve aykırı değerlerin tespiti için kullanılan denetimli makine öğrenme yöntemlerindendir. Destek vektör makinesi algoritmasının amacı,

N boyutlu bir uzayda veri noktalarını belirgin bir şekilde sınıflandıran bir hiperdüzlem bulmaktır. İki veri noktası sınıfını ayırmak için seçilebilecek birçok olası hiperdüzlem vardır. Algoritmanın amacı maksimum kenar boşluğuna, yani her iki sınıfın veri noktaları arasındaki maksimum mesafeye sahip bir düzlem bulmaktır. Kenar boşluğu mesafesini maksimuma çıkarmak, gelecekteki veri noktalarının daha güvenle sınıflandırılabilmesi için bir miktar güçlendirme sağlamaktadır [24].

3.2.4. XGBClassifier

XGBoost, makine öğreniminde hız ve performans için tasarlanmış, degrade destekli karar ağacı algoritmasıdır. XGBoost, aşırı uyumu azaltmak için kullanılmaktadır. Başarılı tahminler üretmek için birden fazla zayıf modelin tahminlerini birleştiren topluluk öğrenme yöntemi olarak da tanımlanabilir. XGBoost'ta zayıf modeller, degrade artırma kullanılarak eğitilen karar ağaçlarını temsil etmektedir. Karar ağaçları eğitildikten sonra XGBoost, ağırlıklı ortalama kullanarak tüm ağaçların tahminlerini birleştirerek tahminler yapar. Her ağacın ağırlıkları eğitim sırasında aynı amaç fonksiyonu kullanılarak öğrenilir. Bu, algoritmanın hangi ağaçların daha önemli olduğunu ve son tahminde daha fazla ağırlık verilmesi gerektiğini otomatik olarak öğrenmesine olanak tanımaktadır [25].

3.2.5. QuadraticDiscriminantAnalysis

Quadratic Discriminant Analysis (QDA), ölçümlerin normal şekilde dağıldığının varsayıldığı doğrusal diskriminant analizi (LDA) ile yakından ilişkilidir. Ancak LDA'dan farklı olarak QDA'da her sınıfın kovaryansının aynı olduğu varsayımı yoktur [26]. İkinci dereceden ayrımda gerekli parametreleri tahmin etmek için doğrusal ayrımcılığa göre daha fazla hesaplama ve veri gerekir. Grup kovaryans matrislerinde büyük bir fark yoksa, o zaman ikincisi ikinci dereceden ayrımcılığın yanı sıra performans gösterecektir [27]. İkinci Dereceden Ayrımcılık, Bayes ayrımcılığının genel biçimidir. Diskriminant analizi, hangi değişkenlerin iki veya daha fazla doğal olarak oluşan grup arasında ayrım yaptığını belirlemek için kullanılır.

3.2.6. ExtraTreesClassifier

ExtraTrees sınıflandırıcı, varyansı ve hesaplama maliyetini rastgele orman algoritmasına göre azaltmak için kullanılan ağaç tabanlı bir makine öğrenimi yaklaşımıdır [28]. ExtraTrees sınıflandırıcı, hesaplama maliyetinin önemli olduğu ve özelliklerin dikkatle seçilip analiz edildiği senaryolarda sınıflandırma veya regresyon için kullanılabilir.

3.2.7. RandomForestClassifier

Rastgele ormanlar popüler bir denetimli makine öğrenme algoritmasıdır. Etiketli bir hedef değişkenin bulunduğu denetimli makine öğrenimi içindir [29]. Regresyon (sayısal hedef değişken) ve sınıflandırma (kategorik hedef değişken) problemlerini çözmek için kullanılabilir. Rastgele ormanlar bir topluluk yöntemidir, yani diğer modellerden gelen tahminleri birleştirirler [30]. Rastgele orman topluluğundaki daha küçük modellerin her biri bir karar ağacıdır.

3.2.8. BaggingClassifier

Torbalama sınıflandırıcısı, temel sınıflandırıcıların her birini orijinal veri kümesinin rastgele alt kümelerine yerleştiren ve daha sonra nihai bir tahmin oluşturmak için bireysel tahminlerini (oylama veya ortalama alma yoluyla) bir araya getiren bir topluluk meta tahmincisidir. Böyle bir meta-tahmin edici, tipik olarak, bir kara kutu tahmincisinin varyansını, yapım prosedürüne

rastgelelik katarak ve daha sonra bundan bir bütün oluşturarak azaltmanın bir yolu olarak kullanılabilir [31].

3.2.9. KNeighborsClassifier

K-en yakın komşu (KNN) algoritması hem regresyon hem de sınıflandırma için kullanılan bir tür denetimli öğrenme algoritmasıdır [32]. KNN, test verileri ile tüm eğitim noktaları arasındaki mesafeyi hesaplayarak test verileri için doğru sınıfı tahmin etmeyi amaçlamaktadır. KNN, istenen kesinlik ve doğrulukta bilinmeyen bir fonksiyonu öğrenmek için kullanılan, hedef fonksiyonun yerel minimumunu temel alan basit bir algoritmadır. Algoritma ayrıca bilinmeyen bir girişin komşuluğunu, aralığını veya uzaklığını ve diğer parametreleri de bulur. Bu, "bilgi kazancı" ilkesine dayanır; algoritma, bilinmeyen bir değeri tahmin etmek için hangisinin en uygun olduğunu bulmaktadır.

3.2.10. LabelSpreading

Algoritma Dengyong Zhou ve diğerleri tarafından. 2003 yılında “Yerel ve Küresel Tutarlılıkla Öğrenme " başlıklı makalede tanıtılmıştır [33]. Yarı denetimli öğrenmenin daha geniş yaklaşımının sezgisi, girdi uzayındaki yakındaki noktaların aynı etikete sahip olması gerektiği ve girdi uzayındaki aynı yapıdaki veya manifoldaki noktaların da aynı etikete sahip olması gerektiğidir.

3.2.11. DecisionTreeClassifier

Karar ağacı sınıflandırıcısı, bir karar ağacı oluşturarak sınıflandırma modelini oluşturmaktadır. Ağaçtaki her düğüm, bir öznitelik üzerindeki testi belirtmekte ve o düğümden inen her dal, o özniteliğin olası değerlerinden birine karşılık gelmektedir [34]. Her yaprak, örnekle ilişkili sınıf etiketlerini temsil etmektedir. Eğitim setindeki örnekler, yol boyunca yapılan testlerin sonucuna göre ağacın kökünden yaprağa kadar gidilerek sınıflandırılır. Ağacın kök düğümünden başlayarak her düğüm, örnek uzayını bir öznitelik test koşuluna göre iki veya daha fazla alt uzaya bölünmektedir. Daha sonra özniteliğin değerine karşılık gelen ağaç dalından aşağıya doğru hareket edilerek yeni bir düğüm oluşturulmaktadır. Bu işlem daha sonra yeni düğümde köklenen alt ağaç için eğitim setindeki tüm kayıtlar sınıflandırılıncaya kadar tekrarlanmaktadır [35].

3.2.12. AdaBoostClassifier

Adaptive Boosting makine öğreniminde topluluk yöntemi olarak kullanılan güçlü bir algoritma tekniğidir. Algoritmada ağırlıklar her bir örneğe yeniden atandığı ve yanlış sınıflandırılmış örneklere daha yüksek ağırlıklar atandığı için “Uyarlanabilir Arttırma” olarak da adlandırılmaktadır. Algoritmada bir model oluşturmak için ilk olarak tüm veri noktalarına eşit ağırlık verilmektedir. Daha sonra yanlış sınıflandırılan noktalara daha yüksek ağırlıklar atanmaktadır. Böylece bir sonraki modelde ağırlığı daha yüksek olan tüm noktalara daha fazla önem verilmektedir. Algoritma daha düşük bir hata alınana kadar eğitim modellerini sürdürecektir [36].

3.2.13. LogisticRegressionClassifier

Lojistik regresyon, bazı bağımlı değişkenlere dayalı olarak belirli sınıfların olasılığını tahmin etmek için kullanılan bir makine öğrenimi sınıflandırma algoritmasıdır.Algoritma lojistik regresyon modeli girdi özelliklerinin toplamını hesaplar (çoğu durumda bir önyargı terimi vardır) ve sonucun lojistiğini hesaplamaktadır. Lojistik regresyonun çıktısı her zaman (0 ile 1) arasındadır ve bu, ikili sınıflandırma görevi için uygundur [37].

3.2.14. DummyClassifier

DummyClassifier, giriş özelliklerini göz ardı eden tahminler yapmaktadır. Bu sınıflandırıcı, diğer daha karmaşık sınıflandırıcılarla karşılaştırmak için basit bir temel görevdir. Taban çizgisinin spesifik davranışı parametreyle seçilmektedir [38]. Tüm stratejiler ve Xargümanı olarak iletilen giriş özelliği değerlerini göz ardı eden tahminler yapmaktadır.

3.2.15. RidgeClassifier

Ridge sınıflandırması, maliyet fonksiyonuna karmaşıklığı önleyen bir ceza terimi ekleyerek çalışmaktadır. Ceza terimi tipik olarak modeldeki özelliklerin kare katsayılarının toplamıdır. Bu, katsayıların küçük kalmasını zorlayarak aşırı uyumu önlemektedir. Ceza süresi değiştirilerek düzenleme miktarı kontrol edilebilir. Daha büyük bir ceza, daha fazla düzenleme ve daha küçük katsayı değerleri ile sonuçlanmaktadır. Bu durum, çok az eğitim verisi mevcut olduğunda yararlı olabilir. Ancak ceza süresinin çok büyük olması, yetersiz uyumla sonuçlanabilir [39].

3.3. GridsearchCV

Makine öğreniminde hiperparametreler, öğrenme süreci başlamadan önce değerleri ayarlanan parametrelerdir. Bu parametreler verilerden öğrenilmez bu yüzden önceden tanımlanması gerekmektedir. Öğrenme sürecini kontrol etmeye yarayan bu parametreler, modelin performansını önemli ölçüde etkileyebilirler.

Makine öğrenmesi modellerinin accuracy (doğruluk) değerlerinin arttırılması için literatürde en sık kullanılan optimizasyon yöntemi GridsearchCV yöntemidir. GridsearchCV, model parametrelerinin belirlenen değerlerde kombinasyonların test edilmesi ve en iyi parametrelerin seçilmesi işlemidir. Sınıflandırma uygulamalarında GridsearchCV k-katlı çapraz doğrulama ile birlikte kullanılmaktadır. Çapraz doğrulama, modelin performansının eğitim verilerinin birden fazla alt kümesinde değerlendirilmesine yardımcı olarak aşırı uyum riskini azaltmaktadır [40] .

3.4. Değerlendirme Metrikleri

Sınıflandırma, giriş verileri verilen sınıf etiketlerinin tahmin edilmesiyle ilgilidir. İkili sınıflandırmada yalnızca iki olası çıkış sınıfı vardır. Sınıflandırma performansını ölçmenin birçok yolu vardır. Çalışmada Accuracy, AUC-ROC ve F1-Score metrikleri kullanılmıştır.

3.4.1. Accuracy

Doğruluk en temel sınıflandırma ölçütüdür. İkili ve çok sınıflı sınıflandırma problemlerinin değerlendirilmesinde kullanılabilir. Doğruluk, doğru tahmin sayısının toplam girdi örnekleri sayısına oranı olarak tanımlanabilir [40]. Denklem 1'de doğruluk formülü gösterilmektedir.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Denklemde, TP, gerçek pozitif olarak adlandırılır. Modelin pozitif sınıfı doğru tahmin etmesidir. TN, gerçek negatif sayısıdır. Modelin negatif sınıfı doğru tahmin etmesi durumudur. FP yanlış pozitif anlamındadır, modelin negatif sınıfı pozitif olarak yanlış tahmin etmesidir. FN yanlış negatif anlamındadır, modelin pozitif sınıfı negatif olarak yanlış tahmin etmesi durumudur.

3.4.2. AUC-ROC

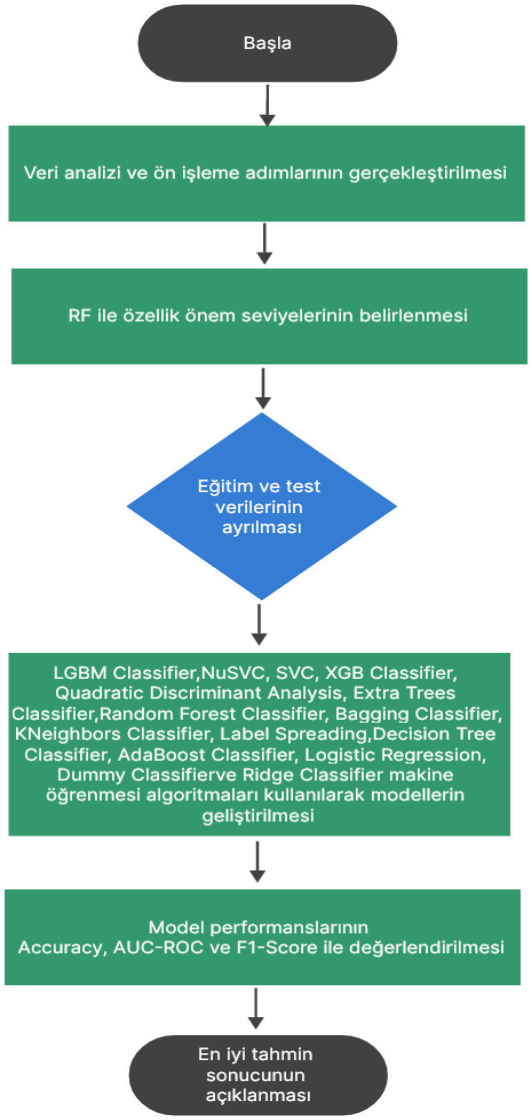
AUC-ROC, makine öğreniminde ikili sınıflandırma modellerini değerlendirmek için kullanılan bir performans ölçümüdür. AUC-ROC çalışma karakteristik eğrisi altındaki alandır. Bu ölçüm, bir modelin pozitif ve negatif sınıfları ayırt etme yeteneğinin değerlendirilmesine yardımcı olur. AUC-ROC, Gerçek Pozitif Oranını (TPR) Yanlış Pozitif Oranına (FPR) karşı farklı eşiklerde gösteren bir olasılık eğrisidir. Bu ölçüm 0 ile 1 arasında değişir; burada 0, kötü bir modeli, 1 ise mükemmel bir modeli belirtir [41].

3.4.3. F1-Score

F1 puanı, kesinlik ve geri çağırma puanlarını birleştiren bir makine öğrenimi değerlendirme ölçümüdür. F1 puanı, bir modelin tahmin becerisini, doğrulukla yapılan genel performanstan ziyade sınıf bazında performansını detaylandırarak değerlendiren alternatif bir makine öğrenimi değerlendirme ölçümüdür [42]. F1 puanı, bir modelin kesinlik ve geri çağırma puanları olmak üzere iki rakip ölçütü birleştirerek son literatürde yaygın şekilde kullanılmasına yol açar.

4. ARAŞTIRMA BULGULARI

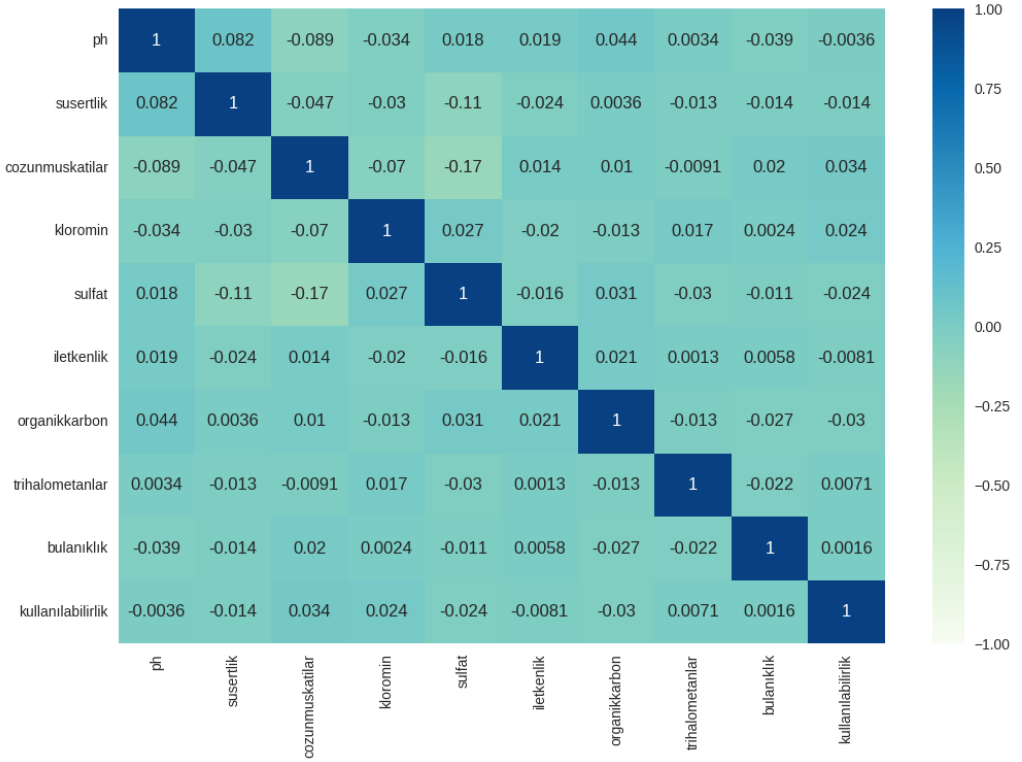
Çalışmada su verimliği ve içilebilirliği için gerçekleştirilen uygulama adımlarına ait akış şeması Şekil 1'de gösterilmektedir.



Şekil 1. Çalışma iş akış şeması.

Veri ön işleme, yapay zeka modelleri geliştirilmeden önce, veri seti üzerinde gerçekleştirilen işlemler olarak tanımlanabilir. Veri ön işleme ile veriler üzerinde veri dönüşümü, tekrarlanan verilerin silinmesi, gürültülü, eksik ya da aykırı verilerin düzenlenmesi işlemleri gerçekleştirilir. Bu işlem sonucunda model performansı ciddi şekilde artmaktadır. Veri seti içerinde ph değişkeni için 491, sülfat degişkeni için 781 ve trihalomentanlar için 162 kaydın "NULL" değer içerdiği görülmüştür. Boş olan bu kayıt değerleri için her değişkenin ortalaması (mean) alınıp ilgili alanlara eklenmiştir.

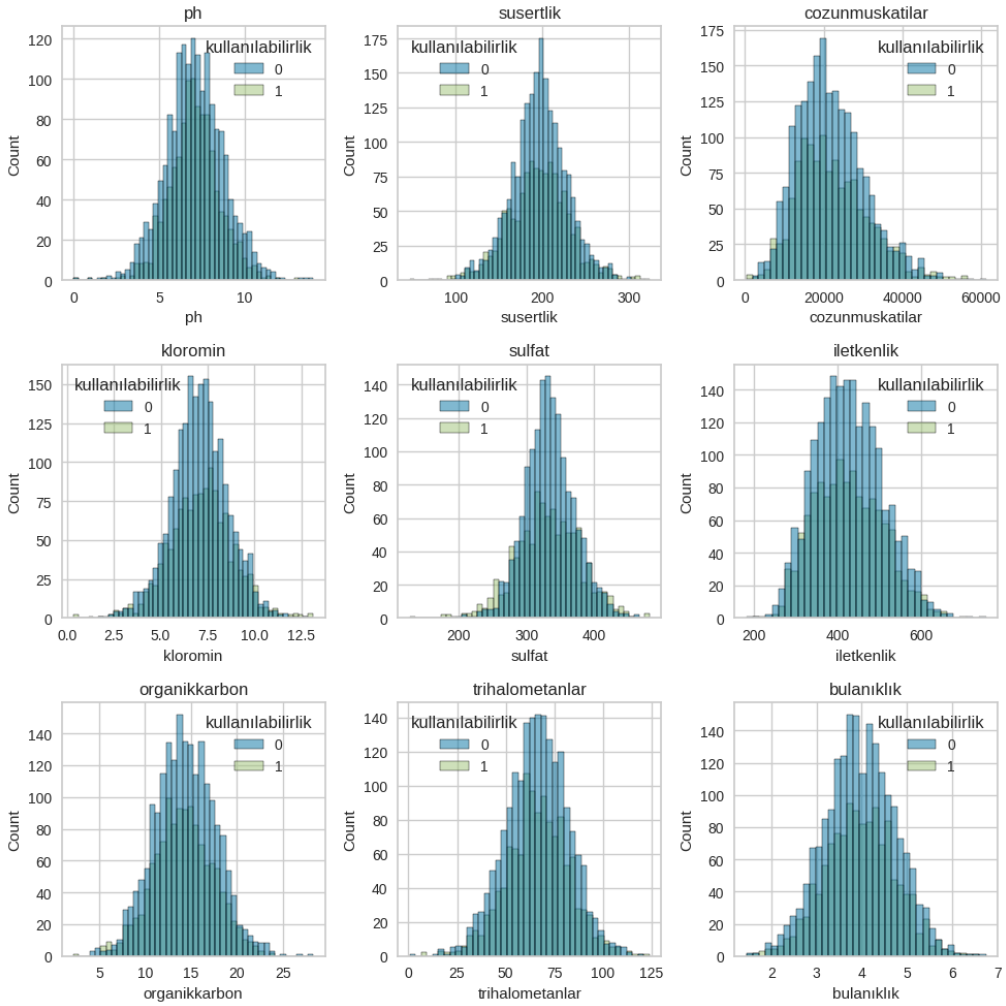
Korelasyon Analizi, iki değişken arasında bir ilişki olup olmadığını ve bu ilişkinin ne kadar güçlü olabileceğini keşfetmek için kullanılan istatistiksel yöntemdir. Korelasyon ısı haritası, birden fazla değişken arasındaki korelasyonu renk kodlu bir matris olarak görüntüleyen grafiksel bir araçtır. Farklı değişkenlerin ne kadar yakından ilişkili olduğunu bize gösteren bir renk şeması gibidir Korelasyon ısı haritasında her değişken bir satır ve bir sütunla temsil edilir ve hücreler bunlar arasındaki korelasyonu gösterir. Her hücrenin rengi korelasyonun gücünü ve yönünü temsil eder; koyu renkler ise daha güçlü korelasyonları gösterir. Şekil 2'de çalışmada yer alan değişkenler arasındaki korelasyon ısı haritası görülmektedir. Grafiğe göre ph ve susertlik nicel değişkenlerinin 0.082 korelasyon değeri ile pozitif yönde en yüksek ilişki seviyesine sahip olan iki değişken oldukları görülmektedir. Bu sonuca göre ph değeri arttıkça susertlik değerinin de arttığı söylenebilir. Grafikte negatif yönde en yüksek ilişkinin sülfat ve cozunmuskatilar değişkenleri arasında olduğu görülmektedir. Buna göre cozunmuskatılar değişken değerindeki artışın sulfat değerinin azalmasına neden olduğu yorumlanabilir.



Şekil 2. Değişkenler arasındaki korelasyon analiz grafiği

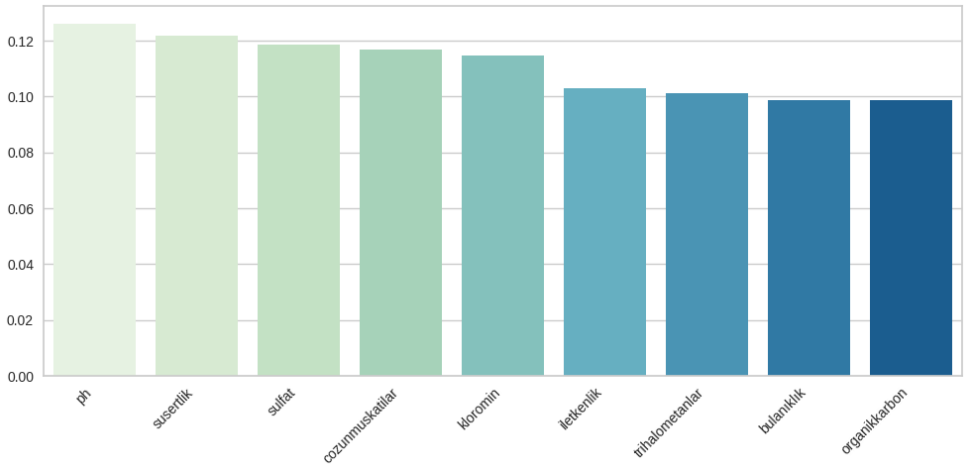
Çalışmada kullanılan veri seti 10 değişken alanı sahiptir. Bu alanlardan pH, su sertliği, çözünmüş katılar, kloromin, sülfat, iletkenlik, organik karbon, trihalometan ve bulanıklık alanları bağımsız değişken, kullanılabilirlik alanı bağımlı değişken olarak çalışmada

kullanılmıştır. Bağımsız değişkenlere ait su kullanılabilirlik (içilebilirlik) sınıf sayıları Şekil 2'de gösterilmektedir.



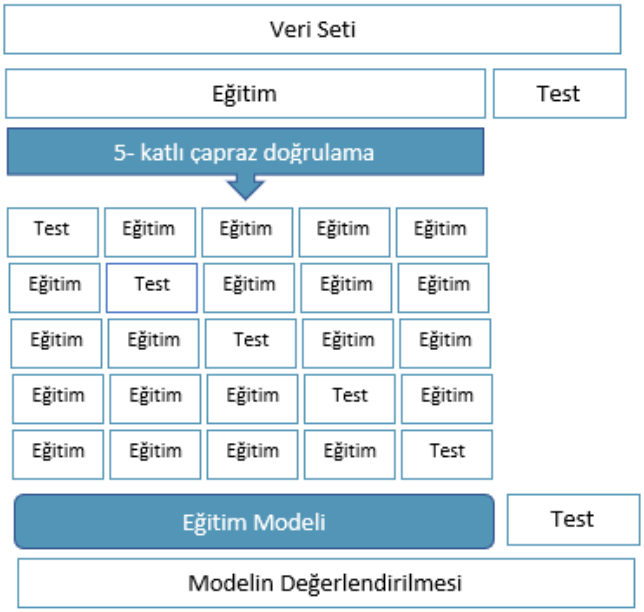
Şekil 2. Bağımsız değişkenlere ait kullanılabilirlik (içilebilirlik) sınıf sayıları

Değişken özellik önemi, bağımlı değişkeni tahmin etmede bağımsız değişkenlerin ne kadar yararlı olduklarını gösteren, bir puan atama yöntemidir. Çalışmada bağımsız değişkenlerin önem dereceleri Rastgele Ormanlar Önem Hesaplaması ile ortaya koyulmuştur. Şekil 3'de sınıflandırma görevi için bağımsız değişkenlerin önemini seviyeleri gösterilmektedir. Şekle göre ph, susertlik, sulfat, cozunmuskatilar ve klomin değişkenleri en önemli değişkenler olarak görülmektedir.



Şekil 3. RF ile değişken özellik önem grafiği

Çalışmada veri analizi ve veri ön işleme adımlarından sonra makine öğrenmesi modellerinin geliştirilmesi için veri seti %80 eğitim ve %20 test verilerine ayrılmıştır. Daha sonra test setinde aşırı uyum sorunu ile karşılaşmamak için eğitim seti k-katlı çapraz doğrulama yöntemi ile kendi içerisinde bölünmüştür. Bu yöntem ile her bölüm kendi içerisinde bir alt kümeyi test, diğer alt kümeleri eğitim işlemi için kullanmıştır. Bu işlem tüm bölümler için tekrarlanmıştır. Şekil 4'te çalışmada gerçekleştirilen 5-katlı çapraz doğrulama yapısı görülmektedir.



Şekil 4. 5–katlı çapraz doğrulama adımları

Çalışmada makine öğrenmesi yöntemlerinden LGBM Classifier, NuSVC, SVC, XGB Classifier, Quadratic Discriminant Analysis, Extra Trees Classifier, Random Forest Classifier, Bagging Classifier, KNeighbors Classifier, Label Spreading, Decision Tree Classifier, AdaBoost Classifier, Logistic Regression, Dummy Classifier ve Ridge Classifier algritmaları ile modeller geliştirilmiştir. Model sonuçlarını değerlendirmek için accuracy, ROC-AUC ve F1 Score metrikleri kullanılmıştır. Modellere ait performans değerlendirme sonuçları Tablo 3'de gösterilmektedir.

Tablo 3. Makine öğrenmesi modellerinin değerlendirme sonuçları

Model	Accuracy	ROC AUC	F1- Score
LGBMClassifier	0.87	0.83	0.85
NuSVC	0.84	0.80	0.83
SVC	0.87	0.80	0.82
XGBClassifier	0.83	0.80	0.82
QuadraticDiscriminantAnalysis	0.86	0.80	0.82
ExtraTreesClassifier	0.85	0.79	0.81
RandomForestClassifier	0.85	0.78	0.81
BaggingClassifier	0.83	0.78	0.81
KNeighborsClassifier	0.81	0.77	0.80
LabelSpreading	0.78	0.76	0.78
DecisionTreeClassifier	0.77	0.76	0.78
AdaBoostClassifier	0.81	0.74	0.76
LogisticRegression	0.80	0.70	0.65
DummyClassifier	0.80	0.70	0.65
RidgeClassifier	0.80	0.70	0.65

Tabloya göre %0,87 doğruluk değerleri ile LGBMClassifier ve SVC modelleri en iyi tahmin değerlerini elde etmişlerdir. Çalışmada LGBMClassifier ve SVC modellernin tahmin performanslarının artırılması için scikit–learn kütüphanesindeki GridSearchCv nesnesi kullanılarak hiper parametre optimizasyonu gerçekleştirilmiştir. Optimizasyon işleminden sonra LGBMClassifier modeli %0,92, SVC modeli %0,91 doğruluk değeri ile başarılı bir tahmin sonucu elde edilmiştir.

5. SONUÇ

Çalışmada su kalitesi ve içilebilirliğini etkili bir şekilde sınıflandırmak ve belirlemek için on beş farklı makine öğrenmesi modeli geliştirilmiştir. Modeller arasında ile LGBMClassifier ve SVC modelleri en iyi tahmin değerlerini elde etmiştir. Bu modellerin tahmin performanslarının artırmak amacıyla GridSearchCv nesnesi kullanılmıştır. Optimize edilmiş LGBMClassifier modeli %92 doğruluk oranı ile su kalitesi ve içilebilirlik durumunu sınıflandırmayı başarmıştır. Çalışmada kullanılan veri seti ile literatürde gerçekleştirilen çalışmaların tahmin başarı oranları ve kullandıkları algoritmalar Tablo 4'te verilmiştir.

Tablo 4. Gerçekleştirilen çalışma ve literatürde yer alan çalışmaların tahmin başarı oranları

Çalışma Adı	Model	Accuracy
Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability (Kaddoura , 2022)	DVM	%73
Comparison of machine learning algorithms in statistically imputed water potability dataset (Poudel vd., 2022)	RF	%70,42
A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI (Patel vd., 2022)	RF Gradient Boost	%81
Water Qualıty Predıctıon Usıng Machıne Learnıng (Kurra vd., 2022)	KNN	%61,7
Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks (Dawood vd., 2021)	YSA	%92
Classification of Water Potability Using Machine Learning Algorithms (Yusuf vd., 2022)	RF	%83,78
Water quality classification using machine learning algorithms (Nasir vd., 2022)	CATBoost	%94,51
Machine Learning Techniques for Water Quality Classification of Thailand's Rivers (Sirikarin ve Khonthapagdee, 2023)	XGBoost	%89,27
Gerçekleştirilen Çalışma	LGBMClassifier	%92

Su kalitesini tahmin etmek, su kaynağı tahsisini optimize etmek, su kaynağı kıtlığını yönetmek gibi çalışmalarda makine öğreniminin tam olarak uygulanmasında çeşitli zorluklar devam etmektedir. Su arıtma ve yönetim sistemlerinde yeterli verinin yüksek doğrulukla elde edilmesi, maliyet veya teknoloji sınırlamaları nedeniyle çoğu zaman zordur. Gerçek su arıtma ve yönetim sistemlerindeki koşullar son derece karmaşık olabileceğinden, mevcut algoritmalar yalnızca belirli sistemlere uygulanabilmektedir. Bu durum makine öğrenimi yaklaşımlarının geniş çapta uygulanmasını engellemektedir. Gelecek çalışmalarda su kalitesini daha iyi değerlendirilebilmesi için büyük miktarda gerçek ölçüm verilerinin paylaşılmasının önemli ve gerekli olduğu görülmüştür.

KAYNAKLAR

[1] Yalçın, L., & Musa, G. Ö. K. (2013). Su Hakkının Bir Temel İnsan Hakkı Olarak Tanınma Süreci ve Türkiye'de Uygulanabilirliği. Memleket Siyaset Yönetim, 8(19-20), 25-62.

[2] Yaseen, Z. M. (2021). An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. *Chemosphere*, *277*, 130126.

[3] Akhtar, N., Syakir Ishak, M. I., Bhawani, S. A., & Umar, K. (2021). Various natural and anthropogenic factors responsible for water quality degradation: A review. *Water*, *13*(19), 2660.

[4] Peng, H., Yang, W., Ferrer, A. S. N., Xiong, S., Li, X., Niu, G., & Lu, T. (2022). Hydrochemical characteristics and health risk assessment of groundwater in karst areas of southwest China: A case study of Bama, Guangxi. *Journal of Cleaner Production*, *341*, 130872.

[5] Zainurin, S. N., Wan Ismail, W. Z., Mahamud, S. N. I., Ismail, I., Jamaludin, J., Ariffin, K. N. Z., & Wan Ahmad Kamil, W. M. (2022). Advancements in monitoring water quality based on various sensing methods: a systematic review. *International Journal of Environmental Research and Public Health*, *19*(21), 14080.

[6] Panigrahi, N., Patro, S. G. K., Kumar, R., Omar, M., Ngan, T. T., Giang, N. L., ... & Thang, N. T. (2023). Groundwater Quality Analysis and Drinkability Prediction using Artificial Intelligence. *Earth Science Informatics*, *16*(2), 1701-1725.

[7] Pandey, J., & Verma, S. (2022). Water Quality Analysis and Prediction Techniques Using Artificial Intelligence. In *ICT with Intelligent Applications: Proceedings of ICTIS 2021, Volume 1* (pp. 279-290). Springer Singapore.

[8] Yurtsever, M., & Murat, E. M. E. Ç. (2023). Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability. *Ege Academic Review*, *23*(2), 265-278.

[9] Khot, I. M., & Surve, A. R. (2020). IoT Assisted Drinkable Water Quality Analysis System using Machine Learning Techniques. *International Journal for Research in Applied Science and Engineering Technology*, *8*, 228-236.

[10] Kaddoura, S. (2022). Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability. *Sustainability*, *14*(18), 11478.
[11] Poudel, D., Shrestha, D., Bhattarai, S., & Ghimire, A. (2022). Comparison of machine learning algorithms in statistically imputed water potability dataset. *Preprint, February*.
[12] Haq, M. I. T. K., Ramadhan, F. D., Az-Zahra, F., Kurniawati, L., & Helen, A. (2021, October). Classification of water potability using machine learning algorithms. In *2021 International Conference on Artificial Intelligence and Big Data Analytics* (pp. 1-5). IEEE.
[13] Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., ... & Ratna, R. (2022). A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI. *Computational Intelligence and Neuroscience: CIN*, *2022*.
[14] Kurra, S. S., Naidu, S. G., Chowdala, S., Yellanki, S. C., & Sunanda, D. B. E. (2022). Water quality prediction using machine learning. *International Research Journal of Modernization in Engineering Technology and Science*, *India*.
[15] Dawood, T., Elwakil, E., Novoa, H. M., & Delgado, J. F. G. (2021). Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks. *Journal of Cleaner Production*, *291*, 125266.
[16] Yusuf, H., Alhaddad, S., Yusuf, S., & Hewahi, N. (2022, October). Classification of Water Potability Using Machine Learning Algorithms. In *2022 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 454-458). IEEE.
[17] 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.
[18] Abulail, N., Owda, A. Y., & Owda, M. (2023, August). Water Quality Classification Decision Support System. In *2023 International Conference on Information Technology (ICIT)* (pp. 73-78). IEEE.
[19] Sirikarin, K., & Khonthapagdee, S. (2023, June). Machine Learning Techniques for Water Quality Classification of Thailand's Rivers. In *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 470-475). IEEE.
[20] Singh, R. I., & Lilhore, U. K. (2023, July). Water Quality Prediction Using Hybrid Classification Model. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-5). IEEE.
[21] Kaggle, Dataset, Erişim Linki: https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability Erişim Tarihi: 26.06.2023
[22] Chang, Y. Y., Lin, L., Pan, H. A., Lin, C. A., Hsieh, B. C., Bottrell, C., & Wang, P. W. (2022). SDSS-IV MaNGA: Unveiling Galaxy Interaction by Merger Stages with Machine Learning. The Astrophysical Journal, 937(2), 97.
[23] Zhu, H., Zhou, M., Liu, G., Xie, Y., Liu, S., & Guo, C. (2023). NUS: Noisy-Sample-Removed Undersampling Scheme for Imbalanced Classification and Application to Credit Card Fraud Detection. IEEE Transactions on Computational Social Systems.
[24] Raudys, Š. (2000). How good are support vector machines?. Neural Networks, 13(1), 17-19.
[25] Chang, C. C., Li, Y. Z., Wu, H. C., & Tseng, M. H. (2022). Melanoma detection using XGB classifier combined with feature extraction and K-means SMOTE techniques. Diagnostics, 12(7), 1747.
[26] Ghojogh, B., & Crowley, M. (2019). Linear and quadratic discriminant analysis: Tutorial. arXiv preprint arXiv:1906.02590.
[27] Srivastava, S., Gupta, M. R., & Frigyik, B. A. (2007). Bayesian quadratic discriminant analysis. Journal of Machine Learning Research, 8(6).
[28] Abhishek, L. (2020, June). Optical character recognition using ensemble of SVM, MLP and extra trees classifier. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-4). IEEE.
[29] Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014.
[30] Bari Antor, M., Jamil, A. H. M., Mamtaz, M., Monirujjaman Khan, M., Aljahdali, S., Kaur, M., ... & Masud, M. (2021). A comparative analysis of machine learning algorithms to predict alzheimer's disease. Journal of Healthcare Engineering, 2021.
[31] Rayaroth, R. (2019). Random bagging classifier and shuffled frog leaping based optimal sensor placement for leakage detection in WDS. Water Resources Management, 33, 3111-3125.
[32] Pandya, V. J. (2016, December). Comparing handwritten character recognition by AdaBoostClassifier and KNeighborsClassifier. In 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 271-274). IEEE.
[33] Tudisco, F., Benson, A. R., & Prokopchik, K. (2021, April). Nonlinear higher-order label spreading. In Proceedings of the Web Conference 2021 (pp. 2402-2413).

[34] Swain, P. H., & Hauska, H. (1977). The decision tree classifier: Design and potential. IEEE Transactions on Geoscience Electronics, 15(3), 142-147.
[35] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.
[36] Liao, X., Xue, Y., & Carin, L. (2005, August). Logistic regression with an auxiliary data source. In Proceedings of the 22nd international conference on Machine learning (pp. 505-512).
[37] Liu, Z., Chen, G., Li, Z., Kang, Y., Qu, S., & Jiang, C. (2022). Psdc: A prototype-based shared-dummy classifier model for open-set domain adaptation. IEEE Transactions on Cybernetics.
[38] Singh, A., Prakash, B. S., & Chandrasekaran, K. (2016, April). A comparison of linear discriminant analysis and ridge classifier on Twitter data. In 2016 International Conference on Computing, Communication and Automation (ICCCA) (pp. 133-138). IEEE.
[39] Kabir, F., Siddique, S., Kotwal, M. R. A., & Huda, M. N. (2015, March). Bangla text document categorization using stochastic gradient descent (sgd) classifier. In 2015 International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 1-4). IEEE.
[40] Foody, G. M. (2002). Status of land cover classification accuracy assessment. Remote sensing of environment, 80(1), 185-201.
[41] Narkhede, S. (2018). Understanding auc-roc curve. Towards Data Science, 26(1), 220-227.
[42] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the first workshop on evaluation and comparison of NLP systems (pp. 79-91).