*Research Journal of Biology Sciences*
*E-ISSN: 1308-0262 14(2): 142-155, 2021*
*Research Article*

# DENETİMLİ MAKİNE ÖĞRENME YÖNTEMLERİ İLE YÜZEY SU KALİTESİNİN SINIFLANDIRILMASI

Arzu Altın Yavuz[1,a,*], Hasan Serhan Yavuz [2,b]

[1] *Eskişehir Osmangazi Üniversitesi,Fen Edebiyat Fakültesi,, İstatistik Bölümü, 26480, Eskisehir, Türkiye*
[2]*Eskişehir Osmangazi Üniversitesi,Mühendislik Mimarlık Fakültesi,, Elektrik Elektronik Mühendisliği Bölümü, 26480, Eskisehir, Türkiye*

**Corresponding Author:*
*E-mail: aaltin@ogu.edu.tr*

(Received 03th November 2021; accepted 22th November 2021)

a: ORCID 0000-0002-3277-740X, b: ORCID 0000-0002-4944-1013

**ÖZET.** Dünya yüzeyinin %70'i sularla kaplıdır ve su yaşamın devamı için hayati önem taşıyan doğal kaynaklardan biridir. Dünya nüfusunun, sanayileşmenin ve kentleşmenin hızla artması kullanılabilir tatlı su kaynaklarının hızla azalmasına ve su kirliliğinin artmasına neden olmaktadır. Su kalitesini bozan bu kirleticilerin tespiti ve kontrolü pahalı ve zaman alıcı laboratuvar çalışmaları ile sürdürülmektedir. Ancak akıllı sistemlerdeki gelişim ve küresel iklim sorunu nedeniyle daha hızlı ve ucuz tespit yöntemlerinin kullanılmasını gündeme getirmiştir. Birçok gelişmiş ülke nehir su kalitesinin izlenmesi için yapay zekaya dayalı yöntemleri kullanmaya başlamıştır. Bu çalışmada nehir su kalitesinin izlenmesinde faydalı olacak 4 farklı denetimli makine öğrenme yöntemi tanıtılmış ve Sakarya Nehri üzerinden seçilen üç istasyondaki geçmiş veriler kullanılarak su kalitesi tespitindeki performansları karşılaştırılmıştır. Destek vektör makineleri, rassal orman, lineer diskriminant analizi ve lojistik regresyon yöntemleri yardımıyla ikişerli olarak istasyonların su kalitesinin sınıflandırılması gerçekleştirilmiştir. Dengeli bir veri yapısına sahip olduğundan doğruluk kriterine göre suyun değişimini izlemede en yüksek performansa sahip yöntemin destek vektör makineleri olduğu belirlenmiştir.

**Anahtar Kelimeler:** *Su Kalitesi, Denetimli Makine Öğrenme Yöntemleri, Destek Vektör Makineleri, Lojistik Regresyon, Rassal Orman*

## GİRİŞ

Canlı yaşamının sürdürülebilirliği üzerinde atmosfer kadar suyun da önemi büyüktür. İnsan hayatının yanı sıra, tüm canlıların ve yaşam ortamının korunması için su en temel kaynaktır. Tüm ekosisteme yetecek kadar temiz su kaynaklarının bulunması her geçen gün zorlaşmaktadır. Su kaynakları, insan yaşamının ve ekonomik kalkınmanın temelidir ve sağlık ve çevre ile yakından ilişkilidir. Su kalitesinin doğru tahmini, su yönetiminin ve kirlilik kontrolünün anahtarıdır.

Dünya üzerinde 1,4 milyar $km^3$ su bulunmaktadır ve bunların sadece %2,5 i tatlı su kaynaklarından oluşmaktadır [1,2]. Bu su kaynaklarının ise sadece %0,3 ü göl ve akarsularda bulunmaktadır [3]. 2025 yılından sonra 3 milyar insanın su kıtlığı yaşayacağı öngörülmektedir. Günümüzde ise dokuz kişiden birinin güvenilir kaynaklardan içme suyuna ulaşamadığı belirlenmiştir [4]. Su kaynakları sonsuz olmadığından, su kirliliğinin azaltılması kullanılabilir su kaynaklarının artmasına neden olacaktır.

Yüzey su kaynakları dünya çapında birincil tatlı su kaynaklarıdır. Tatlı su kaynakları içerisinde ise nehirler hayati öneme sahiptir. Yağmur, erozyon gibi doğal olaylar, kentsel ve tarımsal kirleticiler, nüfus artışı, arıtılmayan kanalizasyon ve endüstriyel faaliyetler sonucu ortaya çıkan kirleticiler gibi temel unsurlar nehir suyunun kalitesi üzerinde olumsuz etkiye sahiptir [5]. Bu su kaynaklarındaki bozulma, içme suyunun mevcudiyeti, ekonomik kalkınma ve gelecekteki stratejiler üzerinde önemli problemlere yol açabilir [5]. Nehir suyu kalitesi değerlendirmesi, su kaynakları yönetim planlarını geliştirmek için en önemli görevlerden biridir.

Su kirliliğinin takibi için su kalitesi gözlem istasyonlarından faydalanılmaktadır. Bu istasyonlardan belirli aralıklarla alınan ölçümlerle kirleticilerin tespiti yapılabilmektedir. Ancak, geleneksel yöntemlerle suyun kullanım amaçlarına uygun olarak değerlendirilmesi oldukça zaman alıcı ve pahalı bir yöntemdir. Su yönetiminin doğru ve güvenilir bir şekilde yürütülebilmesi için son yıllarda makine öğrenme yöntemlerinden yararlanılmaktadır [6, 7, 8, 9, 10, 11, 12]. Makine öğrenme yöntemlerindeki gelişme, veri analitiği, akıllı sensörlerin maliyetinin uygun hale gelmesi, elde edilen sonuçların işletme maliyetlerini azaltması yapay zekaya dayalı çözümlerin benimsenmesini arttırmıştır. Yeni nesil sürdürülebilir su yönetim sistemlerinde kısa ve uzun vadeli etkin tahminlerde bulunulabilmektedir. Bu nedenle gelişmiş ülkelerde su kalitesi izleme çalışmaları, sensörlere dayalı olarak elde edilen ölçümlerin makine öğrenme yöntemleri kullanılarak analiz edilmesiyle daha hızlı ve daha güvenilir bir şekilde yürütülmektedir [6, 7, 9, 12, 13]. ABD ve Kanada'da 2030 yılına kadar yapay zekâ ile sürdürülebilir su yönetimi için 6,3 milyar dolarlık yatırım yapmayı planlamaktadır [7].

Denetimli öğrenme yöntemleri, makine öğrenme uygulamalarında sıklıkla kullanılan yöntemlerdir. Önceden sınıfı bilinen birimlerin açıklayıcı değişkenlerle ile sınıflandırılması amacına hizmet eden bu yöntemlerde, veri seti iki kısma ayrılmaktadır. Eğitim aşamasında modelin parametrelerinin tahmini yapılırken, test aşamasında yeni elde edilecek verilerin doğru tahmin edilip edilmediğine dair performans kriterleri hesaplanmaktadır. Bu nedenle, bu yöntemlere "Denetimli Öğrenme Yöntemleri" adı verilmektedir. Denetimli öğrenme yöntemlerinden genellikle regresyon ve sınıflandırma çalışmalarında yararlanılmaktadır. Sınıflandırma amaçlı denetimli öğrenme yöntemleri arasında doğrusal diskriminant analizi (Linear Discriminant Analysis-LDA), lojistik regresyon, karar ağaçları, karar ormanı, destek vektör makineleri (DVM), Naive Bayes vb. yöntemler sıklıkla kullanılmaktadır [13, 14].

Su kalitesinin istasyonlar arasındaki değişiminin tespiti hızlı karar alma mekanizmalarına olanak sağlayarak, erken müdahale ile suyun korunmasına imkân vermektedir. Bu amaçla, çalışmada Sakarya Nehri üzerinden seçilen 3 istasyondan 2014-2018 yıllarına ait su kalite ölçüm değerleri ele alınmış ve istasyonlar arasındaki su değişiminin tespitinde denetimli makine öğrenme yöntemleri kullanılmıştır. Denetimli öğrenme yöntemleri olarak LDA, karar ormanı, DVM ve lojistik regresyon yöntemleri seçilmiştir. Bu yöntemlerin istasyondan istasyona su değişimini tespitteki performansları çeşitli kriterler ile karşılaştırılmış ve en etkin sınıflandırma yöntemi tespit edilmeye çalışılmıştır.

## MATERYAL VE METOD

### *Çalışma Alanı ve Veri Toplama*

Sakarya Nehri, Türkiye'nin en uzun üçüncü nehri olup, Eskişehir ili Çifteler mevkiinde doğan bir nehirdir. 824 km uzunluğundaki nehir, birçok ilden geçtikten sonra Karadeniz'e

dökülmektedir. Geçtiği bölgelerden Sakarya alt havzasına gelene kadar yerleşim yerlerinden ve sanayi bölgelerinden gelen birçok kirleticiye maruz kalmaktadır. Tüm bu kirletici unsurlara rağmen geçtiği bölgelerdeki tarım arazilerinin sulanmasına katkı sağlamaktadır [29]. Bu çalışmada Sakarya Nehri üzerinden DSİ 3. Bölge alanında kalan 3 istasyon ele alınmıştır: Mudurnu Çayı Dere Mahallesi Köprü, Sarıyar Baraj Çıkışı ve Kocaçay Deresi Kayaboğazı Baraj Çıkışı.

İlgili istasyonlara ait 2014-2018 yılları arasında tutulan, BOD5 (Biyolojik Oksijen İhtiyacı-mg/L), Cl- (Klorür-mg/L), DO (Çözünmüş Oksijen- mgO2 /L), EC (Elektriksel İletkenlik- Mikromhos/cm), Ca++ (Kalsiyum-mg/L), K+ (Potasyum-mg/L), M-Al (mg/L), Mg++ (Magnezyum-mg/L), NH4N (Amonyum Azotu-mg/L), O-PO4 (Orta Fosfat-mg/L), Ph (Alkalinite), Na+(Sodyum-mg/L), Pv (Permanganat Değeri-mgO2/L), SO4 (Sülfat-mg/L), T (Sıcaklık-℃), TDS (Toplam Çözünmüş Katılar-mg/L), TH (Toplam Sertlik-mg/LcaCO3) su karakteristiklerine ilişkin veriler DSİ'den temin edilmiştir. Bu değişkenlere ait düzenli ölçümlerin yapıldığı toplam 384 adet kayıt söz konusudur. İstasyonların konumlarına ilişkin harita Şekil 1'de verilmiştir.




***Şekil 1.*** *Çalışma Alanındaki İstasyonların Konumları*

## *Metod*

Bu çalışmada Şekil 1'de konumları verilen istasyonlar arasında su kalitesinin denetimli makine öğrenme yöntemleri olan rassal orman, LDA, lojistik regresyon ve DVM yardımıyla sınıflandırılması amaçlanmıştır. Analizler R programlama dili ve MATLAB programı kullanılarak gerçekleştirilmiştir.

### *Rassal Orman*

Rassal ormanlar (Random Forest –RF) veya rassal karar ormanları, ağaç temelli bir makine öğrenme yöntemidir. Ağaç temelli yaklaşımların kapsamının genişletilerek birden fazla ağacın birlikte kullanıldığı torbalama (bagging) yöntemi ile elde edilecek tahminin

güvenilirliğinin arttırılması amaçlanmıştır [20]. Rassal ormana dayalı sınıflandırma çalışmalarında rassal ormanın tahmini, oluşturulan çok sayıda karar ağaçlarının tahminleri arasındaki en baskın sınıftır.

Eğer ormanda K adet ağaç varsa, bir m sınıfının aldığı oy sayısı Eşitlik (1)' de verilmiştir.

$$v_m = \sum_{i=1}^{K} I(\hat{y}_i == m) \qquad \textbf{Eqn. 1}$$

Burada, *I*, indikatör fonksiyondur. $\hat{y}_i$, belirli bir durumda i'inci ağacın tahminidir. Tüm bu oylar hesaplandığında, rassal ormanın tahmini en fazla oyu alan sınıftır ve

$$\hat{y} = \underset{m \in \{1,2,\ldots,K\}}{argmax} v_m \qquad \textbf{Eqn. 2}$$

şeklinde belirlenir. Regresyon çalışmalarında da kullanılan rassal orman algoritması, karar ağaçlarından daha etkin sonuçlar vermektedir [21].

Rassal ormana dayalı sınıflandırma çalışmalarında iki düzenleme parametresi bulunmaktadır. Ormandaki ağaç sayısı (K) ve rassal olarak her düğümde seçilen tahmin edici sayısı m'dir. Sınıflandırma çalışmalarında genellikle $m = \sqrt{p}$ olarak kullanılmakta ve p, değişken sayısını göstermektedir. Rassal orman algoritmasında, veri seti çok fazla sınıftan oluşuyorsa düğüm sayısı artmaktadır. Bunun sonucunda sınıflama performansı düşmektedir [22].

*Doğrusal Diskriminant Analizi*

Doğrusal diskriminant analizi, birimlere ait p tane özelliği bilinen k adet sınıftan birine, üzerinde yeni ölçüm yapılmış bir birimin atanmasını sağlayan denetimli bir makine öğrenme yöntemidir. Birimlerin sınıflandırılmasında diskriminant fonksiyonu adı verilen matematiksel eşitliklerden faydalanılır Bu atama yapılırken birimlerin en az hata ile ait oldukları sınıfa atanması amaçlanmaktadır. LDA aracılığı ile elde edilen ayırıcı fonksiyonların etkin sonuçlar vermesi, LDA'nin dayandığı varsayımların sağlanmasına bağlıdır [23]. Bunlardan en önemlisi her değişkenin normal dağılıma sahip olması gerektiğidir. $A_1$ *ve* $A_2$ iki farklı kitle olsun. $X$, *k*-boyutlu bir gözlem vektörünü göstermek üzere iki grup için diskriminant fonksiyonu,

$$D_{xi} = d_{0i} + d_{1i}X_1 + \cdots + d_{ki}X_k \qquad \textbf{Eqn. 3}$$

şeklindedir. Burada $D_{xi}$ diskriminant fonksiyonunun değerini, $d_{ki}$ ise k'ıncı değişkenin katsayısını göstermektedir [24].

*Lojistik Regresyon Analizi*

Lojistik regresyon, sınıflandırma problemlerinde veri setinin normal dağılıma sahip olma varsayımının sağlanmaması durumunda diskriminant analizinin alternatifi olarak sıklıkla kullanılan bir makine öğrenmesi yöntemidir. Sigmoid (lojistik) fonksiyonu yardımıyla doğrusal olarak iyi bir şekilde sınıflandırılamayan birimlerin sınıflandırılmasında kullanılır.

İkili lojistik regresyon analizinde, bağımlı değişkenin düzeylerini için 1 ve 0 değerleri kullanılabilir. Bu nedenle bağımlı değişken Bernoulli dağılımına sahiptir. $\beta' = [\beta_0, \beta_1, \ldots, \beta_p]$ bilinmeyen parametreler vektörünü, $x_j$, $p \times 1$ boyutlu bağımsız değişkenler vektörünü göstermek üzere, $E(Y|X = x_i) = \pi(x_i)$ koşullu olasılığı olmak üzere, lojistik regresyon modelinin koşullu olasılığı Eşitlik (4)' de verilmiştir [21, 24].

$$\pi(x_i) = P(Y = 1|x_i) = \frac{e^{(\beta_0+\beta_1 x_1)}}{1+e^{(\beta_0+\beta_1 x_1)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1)}} \qquad \textbf{Eqn. 4}$$

*Destek Vektör Makineleri*

Destek vektör makineleri ilk kez Vapnik vd. tarafından doğrusal ve doğrusal olmayan sınıflandırma problemlerinin çözümü için geliştirilmiş popüler bir sınıflandırma yöntemidir [25]. Veri yapısına bağlı olarak farklı çekirdek fonksiyonları kullanılabilmektedir. Bu sayede doğrusal ve doğrusal olmayan sınıflandırma işlemlerini başarılı bir şekilde gerçekleştirdiğinden, gerçek dünya problemlerinde yaygın bir kullanıma sahiptir.

DVM'nde amaç iki boyutlu uzayda bir sınıflandırma yaparken, yüksek boyutlu öznitelikler uzayında hiper düzlemlerin maksimal marjlı olarak ayrılmasını sağlamaktır. Bu amaçla, DVM doğrusal aşitsizlik kısıtları altında kuadratik fonksiyonun çözümü için konveks optimizasyonu kullanır. Sınıflandırma işleminde tam olarak ayrıştırılabilir verilerle çalışıldığında tüm veriler bir hiper düzlem ile ayrıştırılabilmektedir.

İkili sınıflandırma probleminde, eğitim verilerine ait bağımsız değişkenler vektörü, $x_i = \{x_1, x_2, \ldots x_n\}$, $x_i \in R^d$ ve bunların karşılık gelen bağımlı değişken $y_i \in \{-1,1\}$ olarak gösterilsin SVM'nin temel amacı, +1 etiketli vektöre ait bir düzlem ve -1 etiketli vektöre ait düzlem yaratarak, gözlemleri bu iki düzlemden birisine atamaktır. Düzleme en yakın eğitim verileri, destek vektör olarak adlandırılır [19, 26].

Eğitim veri setindeki gözlemleri kullanarak DVM için amaç fonksiyonu Eşitlik (5)' deki gibi yazılabilir.

$$D(x,w,b) = \sum_{i=1}^{n} a_i^* y_i H(x_i,x). \qquad \textbf{Eqn. 5}$$

Sınıflandırma için en iyi ayırıcı hiper düzlem, D(x) karar fonksiyonunun sıfıra eşit olduğu x değerleri ile tanımlanır. Burada $a_i^{*\prime}ler$ Eşitlik (6)' da verilen kısıtlı quadratik problemin maksimal çözümüdür ve H, inner product kernel'dir [26, 27].

$$Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j H(x_i,x). \qquad \textbf{Eqn. 6}$$

*Hata Matrisi*

Hata matrisi (Confusion Matrix) bir sınıflandırıcının performansını ölçmek için kullanılan bir ölçüttür. Sınıflandırma yönteminin farklı sınıftaki bireyleri hangi ölçüde sınıflandırdığı gösteren bir analiz aracıdır. İki düzeyli bir sınıflandırma probleminde 2x2'lik bir matris şeklindedir.

Doğru Pozitif (DP): Sınıflandırıcı tarafından doğru olarak tahmin edilmiş pozitif sınıf etiketidir.

Doğru Negatif (DN): Sınıflandırıcı tarafından doğru olarak tahmin edilmiş negatif sınıf etiketidir.

Yanlış Negatif (YN): Sınıflandırıcı tarafından yanlış olarak tahmin edilmiş negatif sınıf etiketidir.

Yanlış Pozitif (YP): Sınıflandırıcı tarafından yanlış olarak tahmin edilmiş pozitif sınıf etiketidir.

Hata matrisi yardımıyla oluşturulacak performans göstergeleri doğruluk (accuracy), kesinlik (Precision), duyarlılık (recall, sensivity) ve F1 skoru şeklindedir. Performans ölçütleri kullanılırken veri setindeki etiketli durumların dengeli olup olmadığı gözönünde bulundurulmalıdır. İki sınıf sınıflandırma için bu durum her iki sınıftaki gözlemlerin yaklaşık eşit olması anlamına gelmektedir [28].

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN} \quad \textbf{Eqn. 7}$$

$$Kesinlik = \frac{DP}{DP + YP} \quad \textbf{Eqn. 8}$$

$$Duyarlılık = \frac{DP}{DP + YN} \quad \textbf{Eqn. 9}$$

$$F1\ Skoru = 2 * \frac{Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik} \quad \textbf{Eqn. 10}$$

Doğruluk ölçütünden elde edilecek sonucun güvenilirliği verinin dengeli dağılımına bağlıdır. Dengeli olmayan veri setleri için kesinlik ve duyarlılık ölçütleri daha güvenilirdir. Veri setinin dengeli olmaması durumunda F1 skoru, kesinlik ve duyarlılığı bir arada kullandığından daha güvenilir sonuçlar vermektedir [28].

## TARTIŞMA VE SONUÇ

İstasyonlar arasında su kalitesi değişiminin takibinde makine öğrenme yöntemlerinin performansı araştırılacağından, üç istasyon kendi aralarında ikişerli olarak ele alınmış ve aşağıdaki gibi kodlanmıştır.

I. karşılaştırma için bağımlı değişken aşağıdaki gibi tanımlanmıştır.

$$y_i = \begin{cases} 1 & , \text{Mudurnu Çayı Dere Mahallesi Köprü} \\ 0 & , \text{Sarıyar Baraj Çıkışı} \end{cases}$$

II. karşılaştırma için bağımlı değişken aşağıdaki gibidir:

$$y_i = \begin{cases} 1 & , \text{Mudurnu Çayı Dere Mahallesi Köprü} \\ 0 & , \text{Kocaçay Deresi Kayaboğazı Baraj Çıkışı} \end{cases}$$

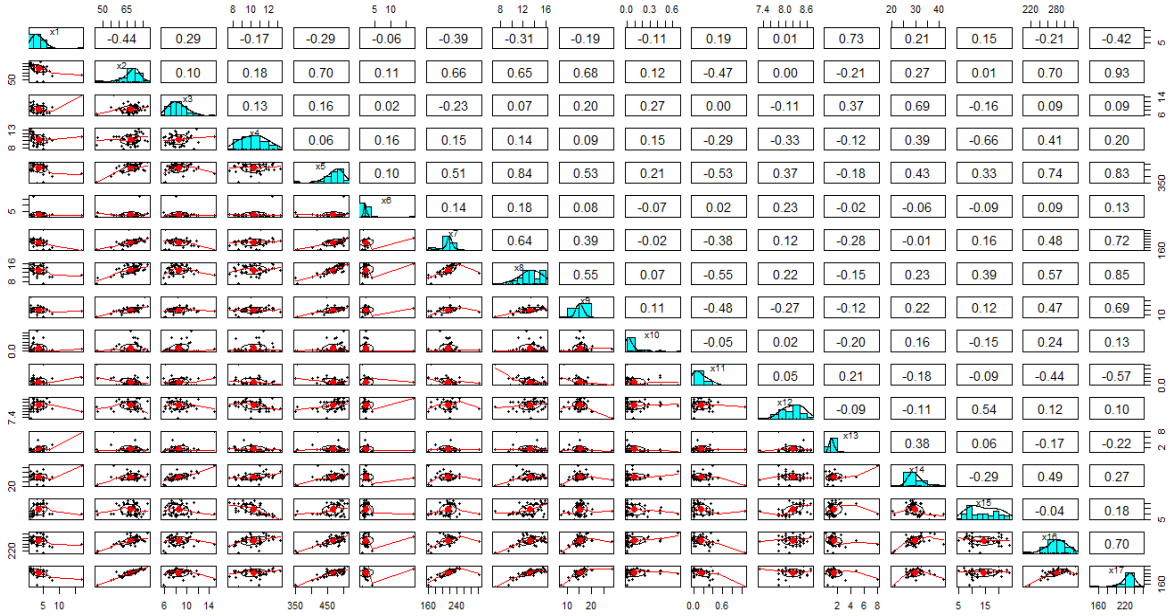
III. karşılaştırma için bağımlı değişken aşağıdaki gibidir:

$$y_i = \begin{cases} 1 & , \text{Sarıyar Baraj Çıkışı} \\ 0 & , \text{Kocaçay Deresi Kayaboğazı Baraj Çıkışı} \end{cases}$$

Çalışmada ele alınan makine öğrenme yöntemlerinin sınıflandırma performansları karşılaştırılacağı için veriler eğitim ve test verisi olarak iki gruba ayrılmıştır. Eğitim verisi ile model parametreleri tahmin edilecek, test verileri ile sınıflandırma performansları değerlendirilecektir. Bu nedenle verilerin 288 tanesi (%75'i) eğitim, 96 sı (%25'i) test verisi olarak ayrılmıştır. Model karşılaştırması için çapraz geçerlilik kriterinden yararlanılacaktır. Çapraz geçerlilik kriterine kullanıldığında veri setinin %5-%33'ünün model doğruluğunun belirlenmesi aşamasında test verisi olarak ayrılması genel kabul görmüş bir orandır.

İlk olarak birinci karşılaştırmada su kalitesinin sınıflandırılmasında ele alınacak verilerin korelasyon değerleri hesaplanmış ve Tablo 1'de verilmiştir.

***Tablo 1.*** *I. Karşılaştırmada Ele Alınan Değişkenlere Ait Korelasyon Matrisi*

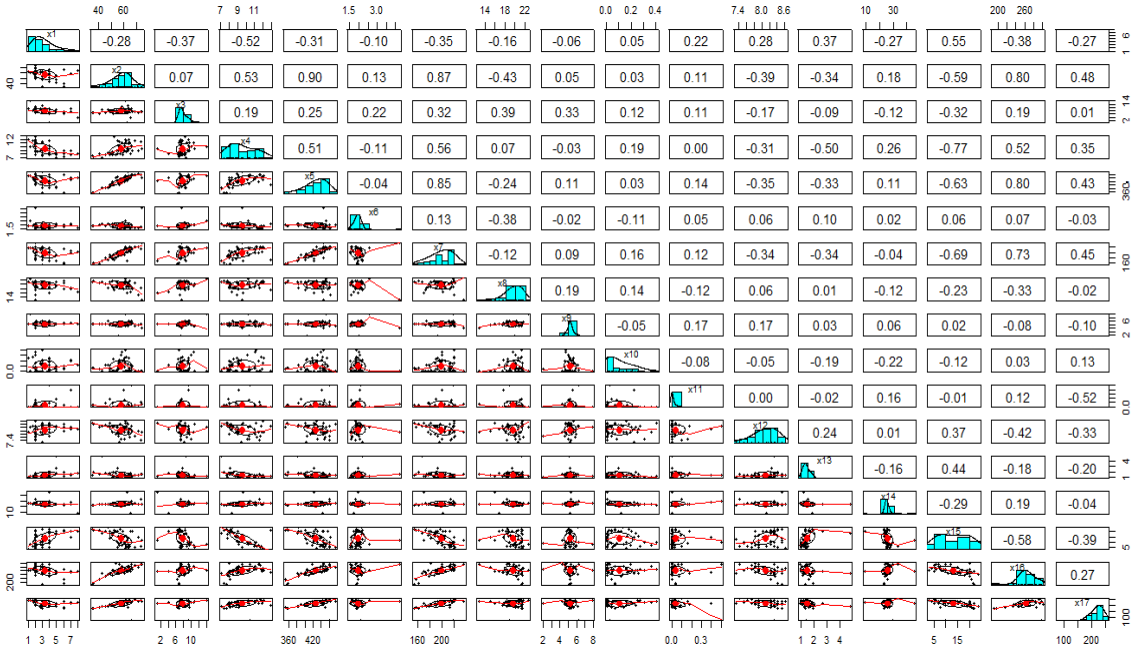



Yüksek derecede korelasyonlar su kalitesinin takip ve tespitini zorlaştıran unsurlardır. Ancak su kalitesi karakteristikleri yapısal olarak korelasyonlu verilerdir ve örnek hacmi

arttırılarak bu korelasyon değerleri azaltılamaz. Tablo 1'e göre, Ca+ ile TH arasında 0,93, EC ile Mg++ arasında 0,84 ve EC ile TH arasında 0,83, Mg++ ile TH arasında 0,85 korelasyon söz konusudur.

İkinci karşılaştırmada ele alınan istasyonlardaki su kalite karakteristiklerine ait korelasyon değerleri de hesaplanmış ve Tablo 2'de verilmiştir.

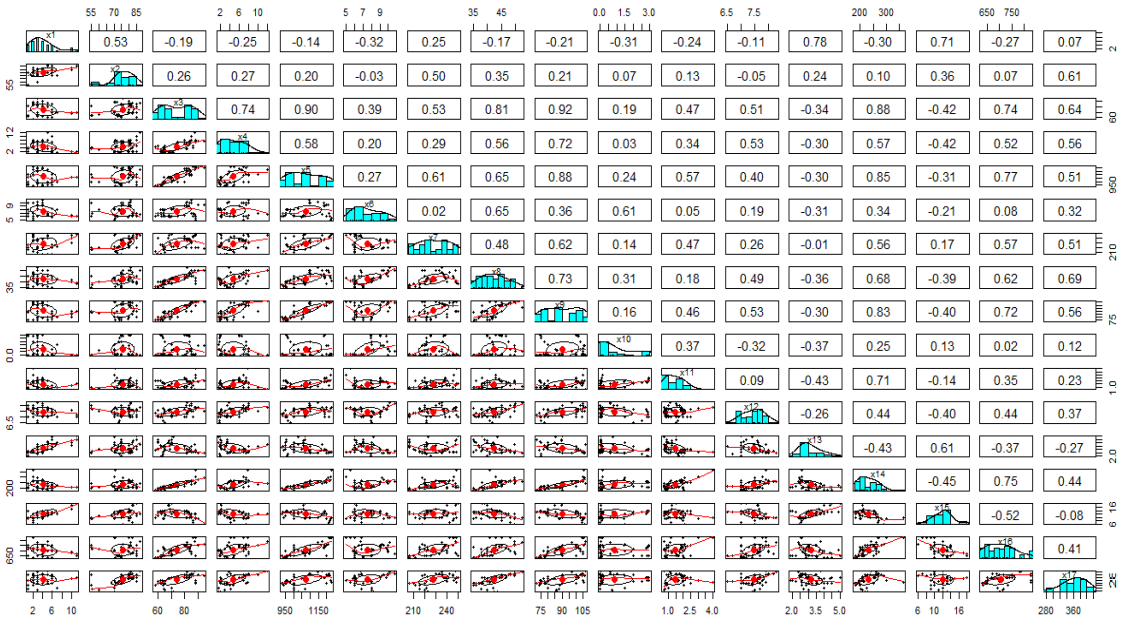
***Tablo 2.*** *II. Karşılaştırmada Ele Alınan Değişkenlere Ait Korelasyon Matrisi*




Tablo 2'ye göre, Ca+ ile EC arasında 0,90, Ca+ ile M-Al arasında 0,87, Ca+ ile TDS ile arasında 0,80 korelasyon söz konusudur. Bu durumda Ca+'nun EC, M-Al ve TDS ile arasında yüksek düzeyde çoklu iç ilişki olduğu görülmektedir. Benzer şekilde EC'nin Ca+ ile 0,90, M-Al 0,85 ve TDS ile 0,80 korelasyonlu olduğu görülmektedir. Diğer su kalite karakteristikleri arasında da benzer şekilde yüksek korelasyonlar mevcuttur.

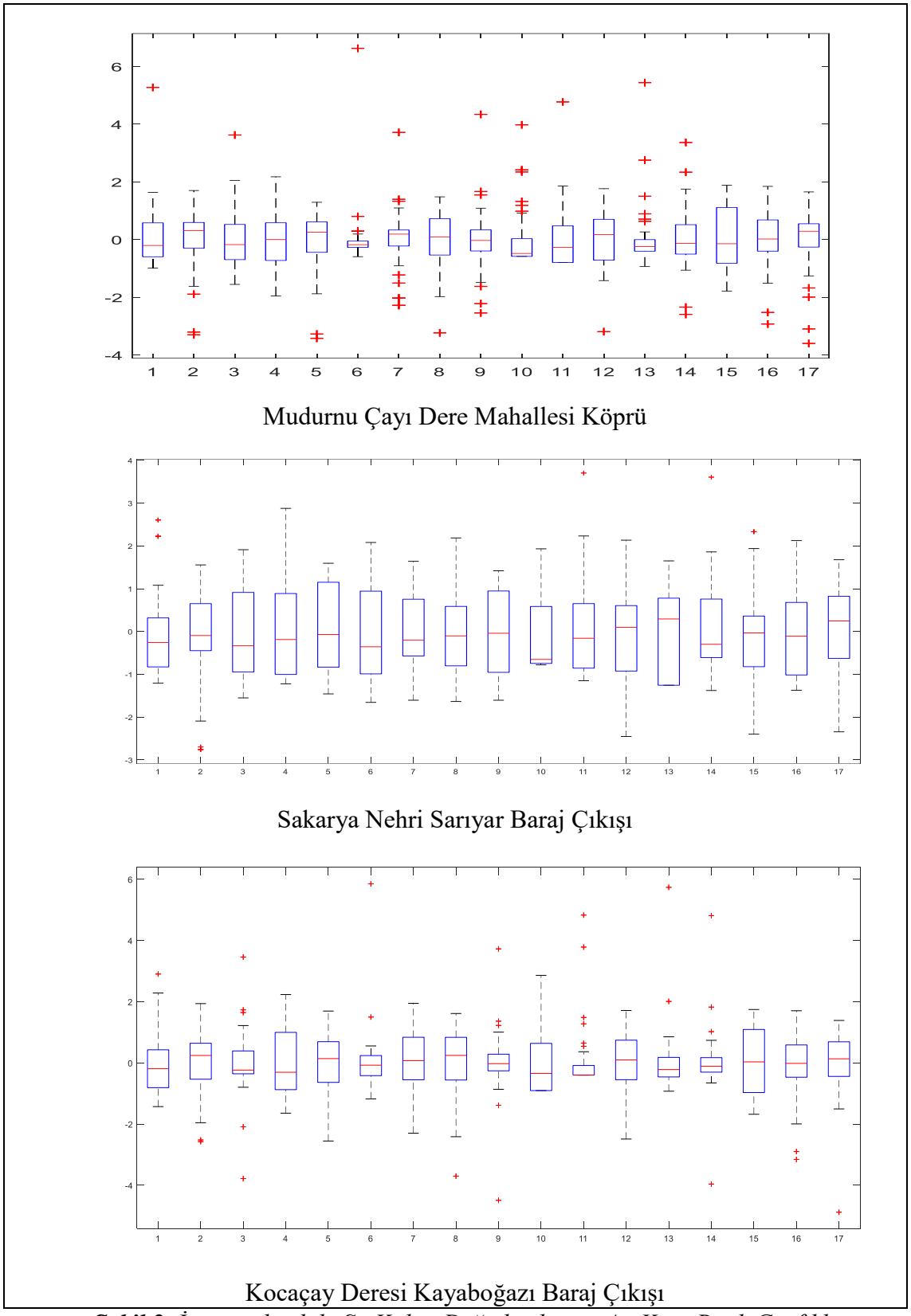
Son olarak üçüncü karşılaştırmada ele alınan istasyonlardaki ölçümler için korelasyon analizi gerçekleştirilmiş ve Tablo 3'te elde edilen sonuçlar verilmiştir.

***Tablo 3.*** *III. Karşılaştırmada Ele Alınan Değişkenlere Ait Korelasyon Matrisi*




Tablo 3 incelendiğinde, $Cl^-$ ile EC arasında 0,90, Mg++ arasında 0,81, Na+ arasında 0,92 ve $SO_4$ arasında 0,83 korelasyon söz konusudur. Benzer şekilde EC ile Na+ arasında 0,88 ve $SO_4$ arasında 0,85 korelasyon vardır. Bu değişkenlerin birlikte değişimlerinin oldukça güçlü olduğu tespit edilmiştir.

Daha sonra 3 istasyondaki veri setleri için aykırı değer analizi gerçekleştirilmiştir. Aykırı değerler su kalitesindeki önemli değişikliklerin göstergeleridir. Verilerin ölçü birimleri farklılık gösterdiğinden ele alınan yöntemlerinin uygulanmasında standartlaştırılmış gözlemler kullanılmıştır. Aykırı değer tespiti için her değişkene ait çizilen kutu bıyık grafikleri Şekil 2'de verilmiştir. Kutu bıyık grafiklerinde aykırı değer olarak belirlenen gözlemler kırmızı renkle gösterilmiştir. Şekil 2 incelendiğinde Mudurnu Çayı Dere Mahallesi Köprü istasyonundaki ölçümlerde çok sayıda aykırı değer tespit edilmiştir. Daha sonra Kocaçay Deresi Kayaboğazı Baraj Çıkışı istasyonuna ait verilerde de aykırı değerlerin çok sayıda olduğu görülmüştür.




***Şekil 2.*** *İstasyonlardaki Su Kalite Değişkenlerine Ait Kutu Bıyık Grafikleri*

Çalışmanın devamında ilk olarak her bir karşılaştırma için eğitim verileri (288 veri) kullanılarak denetimli makine öğrenme yöntemlerinin parametre tahminleri elde edilmiştir. Daha sonra bu tahminler kullanılarak test verilerine (96 veri) ilişkin performans değerleri hesaplanmış ve Tablo 4'te verilmiştir.

***Tablo 4.*** *Eğitim Kümesine İlişkin Performans Değerleri*

| Karşılaştırma | Yöntem | Doğruluk | Kesinlik | Duyarlılık | F1-score |
|---|---|---|---|---|---|
| I. Karşılaştırma | DVM | 0,90104 | 0,93750 | 0,87379 | 0,90452 |
| | Rassal Orman | 0,80208 | 0,79167 | 0,80851 | 0,80000 |
| | Lojistik Regresyon | 0,88021 | 0,90625 | 0,86139 | 0,88325 |
| | LDA | 0,84375 | 0,86458 | 0,83000 | 0,84694 |
| | | | | | |
| II. Karşılaştırma | DVM | 0,87500 | 0,90625 | 0,85294 | 0,87879 |
| | Rassal Orman | 0,73438 | 0,73958 | 0,73196 | 0,73575 |
| | Lojistik Regresyon | 0,81771 | 0,83333 | 0,80808 | 0,82051 |
| | LDA | 0,79167 | 0,80208 | 0,78571 | 0,79381 |
| | | | | | |
| III. Karşılaştırma | DVM | 0,91667 | 0,95833 | 0,88462 | 0,92000 |
| | Rassal Orman | 0,83333 | 0,81250 | 0,84783 | 0,82979 |
| | Lojistik Regresyon | 0,90625 | 0,94792 | 0,87500 | 0,91000 |
| | LDA | 0,89063 | 0,92708 | 0,86408 | 0,89447 |

İlk karşılaştırmada ele alınan istasyonlar arası su kalite değişiminin tespiti için doğrusal DVM kullanıldığında, test verileri yardımıyla hesaplanan doğruluk değeri 0,90104'tür. Benzer şekilde kesinlik değeri 0,93750, duyarlılık değeri 0,87379 ve F1 skoru 0,90452'dir. Bu kriterler veri yapısına duyarlı olarak yorumlanabilir. Ele alınan istasyonlarda eşit sayıda gözlemler incelendiğinden veri yapısı dengelidir. Bir başka ifade ile 1 ve 0 olarak kodlanan gözlem sayıları eşittir. Bu nedenle karşılaştırmada dengeli veri durumunda etkin sonuçlar veren doğruluk kriterinin kullanılması uygun olacaktır. İlk iki istasyon arasındaki su değişiminin makine öğrenmesi yardımıyla tespiti için rassal orman algoritması kullanılırsa doğruluk değeri 0,80208, lojistik regresyon kullanılırsa 0.88021 ve LDA kullanılırsa 0.84375'tir. Doğruluk değeri en yüksek olan yöntemin sınıflandırma amacıyla kullanılması daha uygundur. F1 skoru kesinlik ve duyarlılık ölçütlerinin birlikte değerlendirilmesini sağlamaktadır. Özellikle dengeli olmayan veri yapılarında F1 skoru ölçüt olarak tercih edilmektedir. İlk karşılaştırma için F1 skorları incelendiğinde doğruluk kriterine benzer olarak denetimli öğrenme yöntemlerinin performans sırası en yüksekten başlayarak DVM, lojistik regresyon, LDA ve rassal orman şeklindedir.

İkinci karşılaştırma için benzer şekilde test verileri üzerinde 4 farklı makine öğrenme yöntemi için performans kriterleri belirlenmiştir. Bu karşılaştırmada da veri seti dengeli olduğundan doğruluk kriteri baz alınmıştır. Doğruluk kriterine göre ikinci karşılaştırmada ele alınan iki istasyondaki su kalitesi değişimi için en etkili yöntemin doğrusal DVM olduğu belirlenmiştir. Doğrusal DVM için doğruluk değeri 0.87500, rassal orman için 0.73438, lojistik regresyon için 0.81771 ve LDA için 0.79167'dir.

Son olarak üçüncü karşılaştırmada ele alınan istasyonlar arasında su kalitesi değişiminin tespiti için denetimli öğrenme yöntemlerinin performansı şu şekilde belirlenmiştir. Doğrusal DVM için doğruluk değeri 0.91667 ve F1 skoru 0.9200'dır. Lojistik regresyon için doğruluk değeri 0.90625 ve F1 skoru 0.91'dir. En düşük performans değeri rassal orman algoritmasına aittir. Rassal orman için doğruluk değeri 0.83333 ve F1 skoru 0.82979 olarak belirlenmiştir.

Ayrıca aykırı değerlerin çok olduğu Mudurnu Çayı Dere Mahallesi Köprü ile Kocaçay Deresi Kayaboğazı Baraj Çıkışı istasyonları arasında su kalitesi değişiminin denetimli makine öğrenme yöntemleri yardımıyla izlenmesi konusunda performansların, diğer karşılaştırmalara göre daha düşük olduğu tespit edilmiştir. Sakarya Nehri Sarıyar Baraj Çıkışı istasyonundaki su kalite karakteristiklerinin ise daha az aykırı değer içermesi, diğer istasyonlara göre su kalitesinin tespitini olumlu olarak etkilemiştir.

## SONUÇ

Su kalitesinin tespiti ve takibi geleneksel olarak zaman alıcı laboratuvar analizleriyle yapılmaktadır. Ancak günümüzde akıllı sistemlerin hızlı gelişimi ve uyarlanması ile daha hızlı ve güvenilir bir şekilde yapılabilmektedir. Birçok ülkede gelecekte karşılaşılabilecek su kıtlığı sorununun çözümü için akıllı sistemlerle su kalitesinin takibi yapılmaya başlanmıştır. Akıllı sistemler istatistiksel öğrenme algoritmalarını kullanmaktadır. Denetimli makine öğrenme yöntemleri de sınıflandırma alanında yaygın olarak kullanılan istatistiksel öğrenme yöntemleridir.

Bu çalışmada su kalitesinin değerlendirilmesi için kullanılabilecek 4 farklı denetimli makine öğrenme yöntemi tanıtılmıştır. Daha sonra Sakarya havzası üzerindeki istasyonlar arası su kalitesi değerlendirilmesinde bu yöntemlerin performansları incelenmiştir. Sakarya Nehri üzerindeki 3 istasyon ikişerli olarak karşılaştırılmıştır. İstasyondan istasyona su kalitesi değişiminin tespiti için en yüksek doğrulukla sonuç veren denetimli makine öğrenme yönteminin DVM olduğu belirlenmiştir. İlk karşılaştırmada DVM göre doğruluk oranı 0.90, ikinci karşılaştırmada 0.87 ve son karşılaştırmada yaklaşık 0.92 olarak hesaplanmıştır. Ayrıca veri setlerinde bulunan aykırı değerlerin sınıflandırma performansları üzerinde olumsuz etkilerinin olduğu tespit edilmiştir. Birinci ve üçüncü istasyon arasındaki mesafe diğer mesafelere göre daha kısadır. Konum olarak daha yakın iki istasyon arasındaki su kalitesi değişiminin doğru sınıflandırma değerinin de düştüğü tespit edilmiştir. Bunun nedeni yakın iki istasyondan alınan su numunelerinin benzer kirleticilere maruz kalması olabilir.

Küresel iklim krizinin derinleştiği günlerde, kıt kaynakların etkin kullanılması önem arz etmektedir. İçilebilir su kaynakları da kıt kaynaklar arasında ilk sırada yer almaktadır. Bu çalışma ile sürdürülebilir su kalite yönetimi için denetimli öğrenme algoritmaların etkin sonuçlar verdiği belirlenmiştir. Nehir suyu kalite izleme çalışmalarında suyun sınıflandırılmasında özellikle DVM tercih edilebilir.

## KAYNAKLAR


[1] Orsam Su Araştırmaları Programı, :http://www.orsam.org.tr/tr/SuKaynaklari/MerakEdilenler.aspx?SoruID=2

[2] http://cevreonline.com/su.htm

[3] İ. Yüksel, M. Sandalcı, G. Çeribaşı ve Ö. Yüksek, «Küresel Isınma ve İklim Değişikliğinin Su Kaynaklarına Etkileri,» %1 içinde *7. Kıyı Mühendisliği Sempozyumu*, Trabzon, 2011.

[4] WHO (World Health Organization); UNICEF (United Nations International Children's Emergency Fund). (2015): Progress on Sanitation and Drinking Water -2015 Update and MDG Assessment. World Health Organization, 90. https://doi.org/10.1007/s13398-014-0173-7.2

[5] Asadollah, S.B.H.S., Sharafati, A., Motta, D., Yaseen, Z.M. (2021): River water quality index prediction and uncertainty analysis: A comparative study of machine learning models, Journal of Environmental Chemical Engineering, 9, 1-14.

[6] Bilali, A., Taleb, A. (2020): Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment, Journal of the Saudi Society of Agricultural Sciences, 19, 439-451.

[7] Yıldız, D., Özgüler, H. (2020): Yapay Zekâ ve Su Yönetimi, Su Politikaları Derneği, Rapor No: 30. Ankara.

[8] Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M, Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., 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, 1-8.

[9] Tousi, E.G., Duan, J., Gundyb, P.M., Bright, K., Gerba, C.P. (2021): Evaluation of E. coli in sediment for assessing irrigation water quality using machine learning, Science of the Total Environment, 799, 1-13.

[10] Ahmed, A.N., Othman, F.B., Afanb, H.A., Ibrahim, RK., Fai, C.M., Hossain, D., Ehteram, M., Elshafie, A. (2019): Machine learning methods for better water quality prediction, Journal of Hydrology, 578, 1-18.

[11] Castrillo, M., García, A.L. (2020): Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods, Water Research, 172, 1-9.

[12] Lu, H., Ma, X. (2020): Hybrid decision tree-based machine learning models for short-term water quality prediction, Chemosphere, 249, 1-12.

[13] Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., Kazakis, N. (2020): Improving prediction of water quality indices using novel hybrid machine-learning algorithms, Science of the Total Environment, 721, 1-16. of

[14] Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C. (2019): A Survey Machine Learning Methods Applied To Anomaly Detection On Drinking-Water Quality Data, Urban Water Journal, 16 (3), 235–248

[15] Shah, M.I., Javed, M.F., Abunama, T. (2021): Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques, Environmental Science and Pollution Research , 28, 13202–13220.

[16] Muharemi, F., Logofătu, D., Leon, F. (2019): Machine learning approaches for anomaly detection of water quality on a real-world data set, Journal of Informatıon And Telecommunıcatıon,3 (3), 294–307.

[17] 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, 1-14.

[18] Haghiabi, A.H., Nasrolahi, A.H., Parsaie, A. (2018): Water Quality Prediction Using Machine Learning Methods, Water Quality Research Journal, 53(1), 3-13.

[19] Meltek, S., Kayaalp, K. (2020): Makine Öğrenmesinde,Teoriden Örnek Matlab Uygulamalarına Kadar Destek Vektör Makineleri, İksad Publications.

[20] Breiman, L. (2001): Random forests. Machine Learning, 45(1), 5-32.

[21] Hastie, T., Tibshirani, R., & Friedman, J. (2009): The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
[22] Dangeti, P. (2017): Statistics for machine learning. Packt Publishing Ltd.
[23] Klecka, W. (1980): Discriminant Analysis. Sage Publications, London.
[24] Johnson, R. Wichern, D. (2002): Applied Multivariate Statistical Analysis, *Prentice Hall, Upper Saddle River, NJ, 5. ed edition.*
[25] Vapnik, V. (1998): Statistical Learning Theory. Wiley-Interscience, New York.
[26] Koranga, M., Pant, P., Pant, D., Bhatt, A.K., Pant, R.P., Ram, M., Kumar, T. (2021): SVM Model to Predict the Water Quality Based on Physicochemical Parameters, International Journal of Mathematical, Engineering and Management Sciences, Vol. 6, No. 2, 645-659.
[27] Leonga, W.C, Bahadorib, A., Zhangc, J., Ahmad, Z. (2021): Prediction of Water Quality Index (WQI) Using Support Vector Machine (SVM) And Least Square-Support Vector Machine (LS-SVM), International Journal of River Basin Management, VOL. 19, NO. 2, 149-156.
[28] Uğuz, S. (2021): Makine Öğrenmesi Teorik Yönleri ve Pyhton Uygulamaları ile Bir Yapay Zekâ Ekolü, Nobel Yayıncılık, 2. Baskı.
[29] Özer, Ç., Köklü, R. (2019): Aşağı Sakarya Nehri Su Kalitesinin Sulama Suyu Açısından Değerlendirilmesi. Artvin Çoruh Üniversitesi Doğal Afetler Uygulama ve Araştırma Merkezi, Doğal Afetler ve Çevre Dergisi5(2): 237-246.