Avrupa Bilim ve Teknoloji Dergisi
Sayı 51, S. 1-13, Ağustos 2023

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 51, pp. 1-13, August 2023

Research Article

Karar Ağaçları ve Yapay Sinir Ağlarının Karşılaştırılması: Kimyasal Verilerin Tahmini Üzerine Bir Örnek Çalışma

Oğuz Akpolat[1*], Gonca Ertürk[2]

[1*] Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, Kimya Bölümü, Muğla, Türkiye (ORCID: 0000-0002-6623-4323), oakpolat@mu.edu.tr
[2] Muğla Sıtkı Koçman Üniversitesi, Fen Bilimleri Enstitüsü, Kimya ABD, Muğla, Türkiye, (ORCID: 0000-0002-8821-0330), goncaerturk@posta.mu.edu.tr

(İlk Geliş Tarihi 23 Şubat 2022 ve Kabul Tarihi 15 Şubat 2023)

DOI: 10.31590/ejosat.1073201)

ATIF/REFERENCE: Akpolat, O. & Ertürk, G. (2023). Karar Ağaçları ve Yapay Sinir Ağlarının Karşılaştırılması: Kimyasal Verilerin Tahmini Üzerine Bir Örnek Çalışma. *Avrupa Bilim ve Teknoloji Dergisi*, (51), 1-13.

Öz

Atık suların özelliklerinin belirlenmesinde biyokimyasal oksijen ihtiyacı (BOD_5), kimyasal oksijen ihtiyacı (COD), toplam organik karbon (TOC) ve çözünmüş oksijen (DO) miktarlarının tayini atık suyun karakterizasyonu açısından en temel ölçüm kriterleridir. Biyolojik oksijen ihtiyacı (BOD_5), atık su arıtma tesislerine gelen ham atık su veya arıtılmış atık sudan alınan örneklerle yapılacak olan asitlik (pH), sıcaklık (T), iletkenlik (C), çözünmüş oksijen (DO), oksijen doygunluğu (SO), tuzluluk (SA), elektriksel iletkenlik (EC), kimyasal oksijen ihtiyacı (COD), sıvıda askıda katı madde (LSS), toplam azot (TN), toplam fosfor (TP) analizleri ile birlikte aynı anda gerçekleştirilir. Ancak, bunlardan BOD_5'in tamamlanması en az 5 gün sürerken diğer test sonuçları en çok bir gün süre almaktadır. Daha önce yapılan yukarıdaki parametrelerin ölçüldüğü bir çalışmada 334 adet örneğe ilişkin veri setinde veri setinde bulunan bu parametrelerinin karar ağacı yöntemiyle KNIME veri madenciliği paketinden yararlanarak BOD_5 parametresine etkileri irdelenmiştir. Böylece BOD_5 parametresine etkileri bilinen parametrelerin ağırlıklı etkileri dikkate alınarak sonucu bilinmeyen bir örneğin muhtemel BOD_5 değerinin tahminine çalışılmıştır. Bu çerçevede yapılmış olan bu çalışmada da bu veri seti esas alınarak, veri madenciliği yöntemlerinden Karar Ağaçları ve Yapay Sinir Ağları hem yapısal hem de sonuçlar açısından ayrıntılı olarak incelenmiştir. Her iki yöntemin sonuçları karşılaştırıldığında, kutulanmış (Binned) değerlerin bulunduğu sınıflar arasında dağılımların yakın ancak kaymalar içerdiği görülmektedir. Sınıf sayıları arttırıldığında bu kaymaların kısmen de olsa giderilebileceği unutulmamalıdır. Ayrıca bu sonuçlar gelecek çalışmalarda hem **(*Karar Ağaçları*)** için gruplama sayısı, kazanç gibi hem de (**Yapay Sinir Ağları**) için ağ katman sayısı ve kazanç oranı gibi parametreler değiştirilerek optimize edilebilir.

Anahtar Kelimeler: Atık Su, Aktif Çamur, Optimizasyon, Parametre, Karar Ağacı, Yapay Sinir Ağları.

Comparison of Decision Trees and Artificial Neural Networks: A Case Study on Prediction of Chemical Data

Abstract

In determining the properties of wastewater the amounts of biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total organic carbon (TOC) and dissolved oxygen (DO) are the most basic measurement criteria for characterization of wastewater. Biological oxygen demand analysis (BOD_5), together with the analysis of acidity (pH), temperature (T), conductivity (C), dissolved oxygen (DO), oxygen saturation (SO), salinity (SA), electrical conductivity (EC), chemical oxygen demand (COD), suspended solids in liquid (LSS), total nitrogen (TN) and total phosphorus (TP) made for the samples taken from the raw waste water coming to waste water treatment plants or treated waste water, lasts at least 5 days, as all others less than a day. In a study in which the above parameters were measured before, the effects of these parameters in the data set of 334 samples on the BOD5 parameter were investigated by using the decision tree method by the KNIME data mining package. Thus, taking into account the weighted effects of the parameters whose effects on the BOD_5 parameter are known, the probable BOD_5 value of an unknown sample has been estimated. In this study, based on this data set, Decision Trees and Artificial Neural Networks, which are among the data mining methods, were examined in detail in terms of both structural and results. When the results of both methods are compared, it could be seen that the distributions among the classes in binned values are close at 95 % confidence interval, minimum. It should be kept in mind that these shifts could be partially eliminated when the number of classes is increased. In addition, these results can be optimized in future studies by changing parameters such as the number of groupings or gain for (Decision Trees), and such as network layer number and gain rate for (Artificial Neural Networks).

Keywords: Waste Water, Activated Sludge, Optimization, Parameter, Decision Tree, Artificial Neural Networks.

* Sorumlu Yazar: oakpolat@mu.edu.tr

1. Giriş

Bilgisayarların hayata katılmasıyla, artık ister ticari, ister yönetimsel, ya da ister çevre, sağlık, endüstriyel ya da akademik araştırma faaliyetleri olsun elde edilen bilgiler sayısal ortamda kayıt altına alınmaya ve veri tabanlarında tutulmaya başlanmıştır. Tüm bu veriler, veri tabanlarında bekleyen değerli madenler gibi bilgiye dönüşmeyi beklemektedirler. En genel yaklaşımla veri madenciliği daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin veri tabanlarından elde edilmesi ve bilgilerin ışığı altında işletmelerin açısından geleceğe yönelik kararların alınması veya denenen ve ölçülebilen fiziksel, kimyasal ya da biyolojik olayların anlamlı korelasyonlarının saptanması, örüntü ve eğilimlerin keşfedilmesi ve onlara ilişkin tahminler yapılabilmesi olarak tanımlanmaktadır. Veri madenciliğinin uygulama alanlarına gelince, pazarlama alanında müşterilerin demografik özellikleri ile satın alma alışkanlıklarını belirlenmesi, devamlılığı ve yenilerinin kazanılması ile satış tahminleri, bankacılıkta risk yönetimi ve kredilendirme ilişkilerinin saptanması en tanınınmış olanlarıdır. Yine iletişimde ses ve görüntü kirliliklerinin temizlenmesinde, biyolojide DNA analizleri ile milyonlarca gen arasında hastalıklara sebep olan gen sıralamalarının tanımlanmasında ve tıpta birçok hastalığın önceden güvenilir oranda tespit edilerek tedavisinde sıkça kullanılmaktadır. Benzer olarak kimya ve çevre alanında da çok fazla veriyi barındıran analizlerin ve iklim değişikliklerinin tahmini, çevresel kirlenmenin izlenmesi gibi ve pek çok organik ve kimyasal ürünün kümelenerek orijinlerini belirlenmesi ancak veri madenciliği yöntemleri sayesinde mümkün olabilmektedir. Günümüzde ister doğrudan ölçümsel isterse de deneysel olsun artık kimya ile ilgili olarak da çok miktarda veri elde edilmekte ve bu veriler kolayca da dijital ortamlarda depolanabilmektedir. Böylece çok miktarlarda üretilen verilerin içerisinden bu veri kütlesinin özelliklerini taşıyan örnek kümeleri seçmek yerine; doğrudan elde edilmiş olan veri yığını, ana kütle gibi değerlendirilir. Ve bu verilere ilişkin ana kütle dışındaki tahminlerde bulanabilmek için geliştirilen sınıflandırma, kümeleme ve ilişkilendirme yöntemleri de veri madenciliğinin alanını oluşturur. Yine burada da ana kütlenin özelliklerini belirleyebilmek için gerekli olan ortalama, standart sapma veya onun karesi olan varyans gibi değerlerin söz konusu veri grubu için de tanımlanması gerekli olacaktır. Örneğin, bir yıl veya daha uzun süreli olarak bir arıtma tesisinde atık suyun arıtılması süresince alınan kirlilik, sıcaklık, pH, BOD_5, COD gibi günlük veya saatlik ölçümler ile onların analiz sonuçları bir veri ana kütlesi olarak değerlendirilebilmekte ve veri madenciliği yöntemleri kullanılarak ölçülen bu parametreler ilişkilendirilebilmekte ve geleceğe yönelik uzun vadeli tahminler kolayca yapılabilmektedir (Brereton, 2018, Özdemir, vd., 2012, Güller, vd., 2019, Silahtaroğlu, 2013).

Arıtım süreçlerin izlenebilmesi ve gerekli kontrollerin sağlanabilmesi ancak atık su ve aktif çamur karakteristiklerinin sürekli olarak belirlenmesine dayanmakta olup, atık su arıtma tesislerine gelen ham atık su veya arıtılmış atık sudan alınan örneklerle yapılacak olan asitlik (pH), sıcaklık (T), iletkenlik (C), çözünmüş oksijen (DO), oksijen doygunluğu (SO), tuzluluk (SA), elektriksel iletkenlik (EC), kimyasal oksijen ihtiyacı (COD), sıvıda askıda katı madde (LSS), toplam azot (TN), toplam fosfor (TP) ve biyolojik oksijen ihtiyacı (BOD_5) gibi yapılan analizler ile biyolojik arıtma için kullanılan aktif çamur örneklerine ait prosesin tasarım parametrelerine ilişkin yapılan ölçümlere ilişkindir. Bunlar da havalandırma süresi (AT), çamurun ölçülecek olan sıvıda askıda katı madde konsantrasyonu (LSS), sıcaklık (T), çamur üretim hızı (ASPR) ve biyo-kinetiği (BK) içeren katı alıkonma süresi (RT) ile geri çevrim oranıdır (FBR) Atık suların özelliklerinin belirlenmesinde Biyokimyasal oksijen ihtiyacı (BOD_5), kimyasal oksijen ihtiyacı (COD), toplam organik karbon (TOC) ve çözünmüş oksijen (DO) miktarlarının tayini atık suyun karakterizasyonu açısından en temel ölçüm kriterleridir. Son yıllarda yapılan çalışmalardan birinde sayılan atık su parametrelerden 11 tanesi laboratuvarda yapılan bir günlük çalışmada ölçülebilirken BOD_5 parametresinin ölçümünün 5 gün sürdüğü belirtilmekte ve istatistiksel değerlendirme için bir arıtma tesisinden alınan laboratuvar çalışmasında, 334 adet numuneden 12 parametrenin ölçümü yapılarak bir veri seti oluşturulmuştur. Bu veri setinde bulunan parametrelerinin karar ağacı yöntemiyle KNIME veri madenciliği paketinden yararlanarak BOD_5 parametresine etkileri irdelenmiştir. Böylece BOD_5 parametresine etkileri bilinen parametrelerin ağırlıklı etkileri dikkate alınarak sonucu bilinmeyen bir numunenin muhtemel BOD_5 değerinin tahminine çalışılmıştır. Bu da bize veri madenciliği çerçevesinde çevre ölçüm verilerinin yeniden değerlendirilebileceğini göstermektedir. Yine bu çalışmaya benzer olarak aktif çamur kalitesi için yapılan çalışmalardan da ölçülen değerlere ilişkin istatistiksel değerlendirmeler ve parametreler arasında tahminlemeler yapılabileceği anlaşılmaktadır. Bu çalışmanın sonuçlarına bakıldığında, incelenen 334 örneklemin BOD_5 (***Biyolojik Oksijen İhtiyacı***) değeri dağılımı %53 oranında 100'den düşük bulunmuştur. BOD_5 değeri 100-200 arasında olanların oranı %15,3 iken 450-550 arasında olanların oranı %12,6'dır. Oluşturulan karar ağacının kök dağılımından ise, BOD_5 değerini en fazla etkileyen değişken COD (***Kimyasal Oksijen İhtiyacı)*** olduğu anlaşılmaktadır. Kimyasal oksijen ihtiyacı 214.93 değerinden küçük ve eşit ise BIO_5 değeri 0-100 arası bir değere ulaşmaktadır. Bunun rastlanma sıklığı (%98,6) dır. COD 214.93 değerinden büyük olduğu durumlarda BOD_5 hiçbir şekilde 200 değerini aşmamıştır. BOD_5 in 100 – 200 arasında olma olasılığı da yalnızca %1,4'dür (Güller, vd., 2019, Doğan, 2017; Qiao, ve Han 2014, Silahtaroğlu, 2016, Jiawei, vd., 2012, https://www.cs.waikato.ac.nz/ml/weka/index.html, 2019, https://www.knime.com, 2021). Bu çalışmalardan atık suyun geri kazanılması sırasında atık suyun ve kazanılmış suyun karakterizasyonu ile aktif çamurun kalitesi için bazı testler yapıldığı ve bu testler ile ancak arıtım prosesinin kontrol edilebileceği anlaşılmaktadır. Yine bu çalışmalardan bu işlemler sırasında yapılan tüm fiziksel, kimyasal ve biyolojik analizler ile elde edilen veri yığınlarının ancak veri madenciliği teknikleri ile ayrıntılı olarak incelenebileceği, birbirleriyle ilişkilendirilebileceği ve buradan yola çıkılarak ölçüm parametrelerine ilişkin tahminlemeler de de bulunulabileceği açıktır. Bu çerçevede yapılacak olan bu çalışmada, atık su ve aktif çamur karakteristiklerinin incelenmesinde uygulanacak veri analiz yöntemlerinden Karar Ağaçları ve Yapay Sinir Ağları bir çevresel veri ölçüm seti esas alınarak hem yapısal hem de sonuçlar açısından ayrıntılı olarak karşılaştırılacaktır. Bu çalışma KNIME veri madenciliği platformu üzerinde gerçekleştirilmiştir.

2. Materyal ve Metot

2.1. Kısaca Veri Madenciliği

Kavramsal anlamda anlamda veri, kayıt altına alınmış her türlü olay, durum ya da fikirdir. Veri madenciliği ise, büyük miktarda veriden önceden bilinmeyen, faydalı, kullanışlı, anlamlı bilgilerin keşfedilme sürecidir. Veri madenciliği, diğer bir adla veri tabanında bilgi keşfi; çok büyük veri hacimleri arasında tutulan, anlamı daha önce keşif edilmemiş potansiyel olarak faydalı ve anlaşılır bilgilerin çıkarıldığı ve arka planda veri tabanı yönetim sistemleri, istatistik, yapay zekâ, makine öğrenme, paralel ve dağıtık işlemlerin bulunduğu veri analiz tekniklerine verilen addır. Bu süreçte kümeleme, veri özetleme sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması, değişkenlik analizi ve anormalin tespiti gibi farklı birçok teknik kullanılmaktadır. Sınıflandırma ise, yeni bir nesnenin niteliklerini incelemek ve bu nesneyi önceden tanımlanmış bir sınıfa atamaktır. Burada önemli olan, her bir sınıfın özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır. Kümeleme ise verilerin birbirine yakınlığı veya uzaklığına göre gruplandırılması olup önceden belirlenmiş grup sınırları yoktur ancak grup sayısı verilerek optimize edilebilir. Verilerin sınıflandırılma süreci iki adımdan oluşur. Bunlar:

1- Veri kümelerine uygun bir model ortaya konur. Söz konusu model veri tabanındaki alan isimleri kullanılarak gerçekleştirilir. Sınıflandırma modelinin elde edilmesi için veri tabanından bir kısım eğitim verileri olarak kullanılır. Bu veriler veri tabanından rastgele seçilir.
2- Test verileri üzerinde sınıflandırma kuralları belirlenir. Ardından söz konusu kurallar bu kez test verilerine dayanarak sınanır.

Kredi başvurusu değerlendirmesi, kredi kartı harcamasında sahtekarlık olup olmadığına karar verme, hastalık teşhisi, ses tanıma, karakter tanıma, gazete haber ve yazılarını konularına göre ayırma, kullanıcı davranışları belirleme gibi pek çok alanda sınıflandırma teknikleri kullanılmaktadır. Kümeleme, veri tabanlarındaki verilerin gruplar veya kümeler altında toplanarak, benzer özelliklere sahip nesnelerin bir araya gelmesini sağlayan bir veri madenciliği tekniğidir. Kümeleme, verilerin kendi aralarındaki benzerliklerin göz önüne alınarak gruplandırılması işlemidir. Birliktelik analizi ise, bir işlem kaydında bir elemanın meydana gelme olasılığını, diğer elemanların meydana gelme olasılıklarından tahmin etmek için kuralların bulunmasıdır. Birliktelik kuralı, geçmiş verilerin analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen bir yaklaşımdır Örneğin, alışverişlerde ekmek, süt veya ekmek, süt, çocuk bezi ve çikolata ya da ekmek, süt, çocuk bezi ve kola gibi ürün gruplarını oluşturulması birliktelik analizi örnekleridir (Silahtaroğlu, 2016).

2.2. Veri Madenciliği Yazılımları ve KNIME

Veri madenciliğine ilişkin sınıflandırma teknikleri ve algoritmalarından bahsetmeden önce veri madenciliği işlemlerin gerçekleştirilebileceği yazılımlar üzerinde durulmalıdır. Bunlar ticari ve açık kaynak kodlu olmak üzere 2 gruba ayrılmakta olup açık kaynak kodlu olarak hazırlanan KNIME- Konstanz Information Miner (http://www.knime.org/) veri madenciliği platformu Konstanz Üniversitesi görsel veri madenciliği araştırma grubu tarafından "Eclipse Rich Client Platform" üzerinde geliştirilmiştir ve genişletilebilme özelliklerine sahip kullanıcılara bir yazılım geliştirme kiti sunarak, kullanıcıların kendi modüllerini yazabilmelerini sağlayan tek uygulamadır. Kurulum şartı olmayan ve txt uzantılı metin dosyalarından veya .arff, .table veya pek çok formattan veri alabilme ve veri tabanları ile veri alışverişi yapabilme özelliklerine sahiptir. Bu çalışmada KNIME veri madenciliği platformu üzerinde gerçekleştirilmiştir. Burada ilk yapılması gerekenin kayıp verilerin tahmini, gürültülü verilerin temizlenmesi ve birleştirme, boyut indirgeme, sıkıştırma ve kesikli hale getirme gibi ön işlemlerin yapılması gerektiği unutulmamalıdır. Normalizasyon en önemli veri dönüştürme işlemlerinden birisidir. En sık karşılaşılanı da 0-1 aralığı normalizasyonu için "min-max" yöntemidir. Veri madenciliği modelleri tahminleme, kümeleme ve bağlantı analizi ile fark sapmaları gibi üç ana başlık altında toplanabilir. Bunlardan tahminleme ve kümeleme her bir kaydın diğerleriyle olan ilişkisini araştırırken, bağlantı analizinde hem nesnel hem de zamansal bağlantılar incelenebilmektedir. Bu yöntemler üzerine pek çok algoritma geliştirilmiş olup, bu algoritmalar tahminleyici, tanımlayıcı veya her ikisini de içerebilirler. Bu algoritma, yöntem ya da modellerin tamamı birer bilgisayar programı olup programa tanıtılan verilerin bir kısmıyla öğrenip sonuçlar ve kurallar çıkartarak kalan kısmıyla da bu öğrenme test edilerek doğrulama yaparlar. Daha sonrada doğrulanmış bu sonuç ve kurallara bakarak yeni gelen bir verinin sonucunu tahmin etmeye çalışırlar. Bu alanda geliştirilen pek çok algoritma mevcut olup, bu çalışmada sınıflandırma yöntemi olarak karar ağaçları ve yapay sinir ağları teknikleri birbiriyle karşılaştıılmıştır. (Silahtaroğlu, 2016, Jiawei, vd., 2012).

2.3. Sıflandırma

En çok bilinen veri madenciliği tekniklerinden biri olup, matematiksel olarak aşağıdaki gibi tanımlanabilir;

$$D = \{t_1, t_2, \dots, t_n\}$$

Bir veri tabanı olsun ve her bir t_i bir kaydı göstersin.

$$C = \{C_1, C_2, \dots, C_m\}$$

M adet sınıftan oluşan sınıflar kümesini göstersin,

$$f: D \to C$$

Her bir t_i bir sınıfa ait olmalıdır. Burada C_j ayrı bir sınıftır ve her bir sınıf kendine ait kayıtları içerir.

$$C_j = \{t_i \,/\, f(t_i) = C_j\,, 1 \le i \le n, ve\ t_i \in D$$

Şeklinde gösterilebilir. Sınıflandırma da elimizdeki sınıf veya istatistiksel tanımıyla bağımlı değişken sınıfsal (kesikli) ve sürekli değer taşıyabilir. Bu açıdan regresyon veya çok terimli regresyona yaklaşır. Sınıflandırma belirli aralıklar içerisindeki gizli örüntülerin ortaya çıkarıldığı bir denetimli öğrenme yaklaşımı olarak da tanımlanabilir.

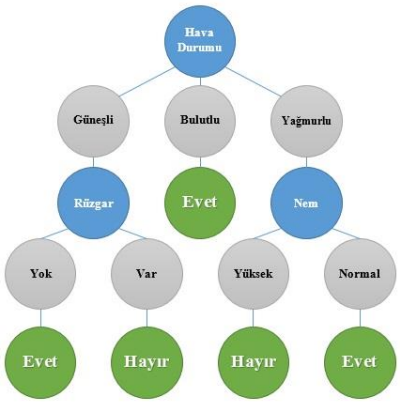
2.3.1. Karar Ağaçları

Bir işletme yönetimi tarafından tercihlerin, risklerin, kazançların, hedeflerin tanımlanmasında yardımcı olabilen ve birçok önemli yatırım alanlarında uygulanabilen, birbirini izleyen şansa bağlı olaylarla ilgili olarak çıkan çeşitli karar noktalarını incelemek için kullanılan bir tekniktir. En çok karşılaşılan algoritmalar ID3 ve C4.5 tir. Buradaki açıklamalar ID3 algoritması üzerinden yapılmıştır. Bir karar ağcı yapısının daha iyi anlaşılabilmesi aşağıdaki geleneksel örnek sıklıkla verilmektedir ((https://erdincuzun.com/makine_ogrenmesi/.../ (2020); http://mail.baskent.edu.tr/ ~20410964/DM_8.pdf (2020)). Bu örnek, hava durumu, sıcaklık, nem ve rüzgar faktörlerine bağlı olarak bir futbol maçının oynanıp oynanmaması üzerinedir. Karar ağacı yönteminin esasları ve karar ağacı algoritmasının adımları aşağıda verilmiştir.

Karar ağacı yönteminin esasları
1. Sorunun tanımlanması
2. Karar ağacının çizilmesi / yapılandırılması
3. Olayların oluşma olasılıklarının atanması
4. Beklenen getirinin (veya faydanın) ilgili şans noktası için hesaplanması- geriye doğru, işlem
5. En yüksek beklenen getirinin (faydanın) ilgili karar noktasına atanması- geriye doğru, karşılaştırma
6. Önerinin sunulması

Karar ağacı algoritmasının adımları
1. T öğrenme kümesini oluştur
2. T kümesindeki örnekleri en iyi ayıran niteliği belirle
3. Seçilen nitelik ile ağacın bir düğümünü oluştur ve bu düğümden çocuk düğümleri veya ağacın yapraklarını oluştur. Çocuk düğümlere ait alt veri kümesinin örneklerini belirle
4. 3. adımda yaratılan her alt veri kümesi için
◦ Örneklerin hepsi aynı sınıfa aitse
◦ Örnekleri bölecek nitelik kalmamışsa
◦ Kalan niteliklerin değerini taşıyan örnek yoksa işlemi sonlandır. Diğer durumda alt veri kümesini ayırmak için 2. adımdan devam et.

Karar ağacını IF-ELSE IF-ELSE ifadeleri kullanılarak rahatlıkla herhangi bir programlama dilinde kodlanabilir. Örneğin Hava Durumu üç IF şart içerir. 2. IF şartı "Bulutlu" seçilmişse "Futbol Oyna" için olumlu sonuç alınmış olur. Temelde karar ağaçları – sınıflama, özellik ve hedefe göre karar düğümleri (decision nodes) ve yaprak düğümlerinden (leaf nodes) oluşan ağaç yapısı formunda bir model oluşturan bir sınıflandırma yöntemidir. Şekil 1'de görselleştirilmiş olarak verilen hava durumuna ilişkin karar ağacı görülmektedir.



Şekil 1 Karar ağacının görselleştirilmesi
(Figure 1 Visualization of the decision tree)

Karar ağacı algoritması, veri setini küçük ve hatta daha küçük parçalara bölerek geliştirilir. Bir karar düğümü bir veya birden fazla dallanma içerebilir. İlk düğüme kök düğüm (root node) denir. Bir karar ağacı hem kategorik hem de sayısal verilerden oluşabilir. Herhangi bir durumun oluşmasında rastgeleliği, belirsizliği ve beklenmeyen durumun ortaya çıkma olasılığı Entropy ile tanımlanır ve eğer örneklerin tamamı düzenli / homojen ise entropisi sıfır olur. Eğer değerler birbirine eşit ise entropi 1 olur. Örneğin Futbol Oyna hepsi "Evet" veya "Hayır" olsa entropi sıfır olurdu. Entropi formülasyonu

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Entropi sadece hedef üzerine hesaplanmaz. Ayrıca özellikler üzerine entropi hesaplanabilir. Fakat özellikler üzerine entropi hesaplanırken hedefte göz önüne alır. Bu durumda entropi formülü:

$$E(T, X) = \sum_{c \in X} P(c)E(c)$$

Bilgi kazanımı (Gain), bir veri setini bir özellik üzerinde böldükten (Örneğin **E** (FutbolOyna, HavaDurumu)) sonra tüm entropiden (**E**(FutbolOyna)) çıkarmaya dayanır. Entropinin küçük değer içermesi durumunda özelliğin önemi Decision Tree algoritması ID3 için

artmaktadır. Diğer taraftan 1'e yaklaştıkça özelliğinin önemi azalır. Ancak bilgi kazanımında olay tam tersidir ve bu açıdan entropinin tersi gibi düşünülebilir. Decision Tree inşa edilirken en yüksek değerleri bigi kazanımı'na sahip özellik seçilir. Bu özellik seçildikten sonra özelliğin değerlerine bakılarak en yüksek information gain'e sahip alan seçilir. Aşırı Uyum (Overfitting) karar ağacı modelleri ve diğer pek çok tahmin modeli için önemli bir sorundur. Öğrenme algoritmasını etkileyecek şekilde eğitim seti hataları azaltmaya devam edildiğinde overfitting olur. Bir karar ağaç inşasında overfitting'ten kaçınmak için genelde iki yaklaşım kullanılır.

- Pre-pruning: Sınırlandırma işleminde önce ağacın büyümesini durdurmak.
- Post-pruning: öncelikle tüm ağacı oluşturup daha sonra ağaçtaki gereksiz kısımları çıkarmak.

Uygulamada ne zaman pruning (budama) işleminin yapılacağını belirlemedeki zorluk sebebiyle ilk yaklaşım pek kullanılmaz

2.3.2. Yapay Sinir Ağları

Sezgisel algoritmalar herhangi bir amaca ulaşmak ya da hedefe varmak için çeşitli alternatif hareketlerden etkili olanla karar verebilmek için tanımlanan kriterler ve bilgisayar algoritmaları ağıdır. Bunlar çözüm uzayında çözüme yakınsaması ispat edilemeyen yöntemler olup yakınsama özelliğine sahiptirler ancak kesin çözüm yerine onun yakınındaki bir çözümü garanti edebilirler. Bu tür yöntemlere gereksinim duyulmasının en büyük nedeni matematiksel algoritmaların karmaşıklığı yerine daha basit olmaları, kesin çözüm bulma işlemlerinin tanımlanamadığı problemlere uygun olmaları, öğrenme amaçlı ve kesin çözüm bulma işleminin bir parçası olmalarıdır (Silahtaroğlu, 2016, Jiawei, vd., 2012). Basitliği, çözüm kalitesi ve çözümleme zamanı bir algoritmanın değerlendirilmesi için başlıca kriterdir. Bu nedenle bütün algoritmalarda olduğu gibi sezgisel algoritmalar için de bu kriter çok önemlidir ve başlangıç değerlerine bağlı olarak bölgesel optimum çözümleri üretilirken bu da önem kazanmaktadır. Ancak bu dezavantajın yanında insan makine etkileşimine yatkınlıkları ve ısıl bir işlem olan soğumanın doğal adımları, bir soyun genetik yapısındaki iyileşme ya da bir karınca kolonisinin bir hedefe en kısa sürede varmasında grubun üyelerinin diğerleri için iz olarak bıraktıkları bir kokunun algılanması gibi olay akışının optimize edilebilmesini sağlayan mekanizmaları esas alması, matematiksel ifade edilememesi gibi güçlükler içeren olayların optimizasyonunda büyük kolaylıklar getirdiği de unutulmamalıdır.

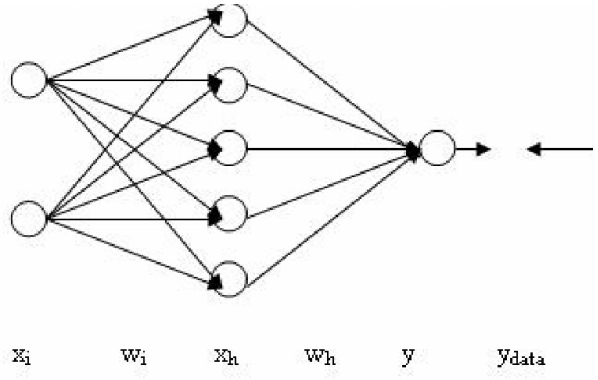
Yapay zekâ yaklaşımlarından olan yapay sinir ağlarının bazı modelleri ve bulanık programlama da optimizasyon amaçlı kullanılabilmektedir. Günümüzde pek çok alanda yapay zekâ teknolojisi ürünleri boy gösterirken özellikle otomasyon sistemleri yapay zekâ teknolojisi ile donatılarak bilgisayarların hesaplama ve değerlendirme güçlerine karar verme gücüde eklenmiştir. Böylece sistemlerin fonksiyonel özellikleri arttırılmakta zeki etmenler yaşamda daha fazla yerini almaktadır. Bunlardan yapay sinir ağları bilgisayarın öğrenmesini sağlamaktadır. Makine öğrenmesi kısmen de olsa canlı düşünce sistemine anoloji (benzeşim) kurularak zaman içinde davranışların iyileştirilmesi olarak tanımlanmaktadır. Bunun için değişik öğrenme paradigmaları geliştirilmiştir ve bu paradigmalar başlıca 3 strateji üzerine kurulmuştur. Bunlar:

1. Öğretmenli öğrenme
2. Destekli öğrenme
3. Öğretmensiz öğrenme

Bu stratejilere dayanarak geliştirilmiş öğrenme kuralları vardır ve bu kurallara dayanarak hazırlanan yapay zekâ algoritmaları kullanılmaktadır. Bu modellerden en basiti sadece girdi ve çıktı katmanlarından oluşan ve girdi çıktı ilişkisinin doğrusal olduğu tek katmanlı algılayıcılardır (X and Y). Eğer girdi çıktı ilişkisi doğrusal değilse çok katmanlı algılayıcılar kullanılmalıdır (X and Y ya da X or Y). Burada örnek olarak çok katmanlı geriye doğru hesaplamalı yapay sinir ağ modeli üzerinde durulacaktır.

2.3.3. Çok Katmanlı Geriye Doğru Hesaplamalı Yapay Sinir Ağı

Yapay sinir ağları biyolojik sinir ağlarından esinlenerek geliştirilmiş bir bilgi işleme sistemidir. Modellemede ve optimizasyonunda ölçülen iki büyüklük arasındaki korelasyonlar çok önemlidir. Örneğin bir biyokimyasal reaksiyonda substrat konsantrasyonu (kg ton^{-1}) ile spesifik büyüme hızı (h^{-1}) arasındaki ilişki kuramsal açıdan polinomial olarak ilişkilendirilebilir. Ancak non lineer ilişkileri içeren bu tür karmaşık ifadelerin çözümlenmesinde farklı bir yol olarak Yapay Sinir Ağları (Artificial Neural Network -ANN) tekniği kullanılabilir. Bu konunun daha detaylı anlaşılması için 2 düğümlü açık ve 5 düğümlü kapalı olmak üzere iki tabakalı bir giriş ve giriş verilerinin değerlendirilip sonuçlandırıldığı tek düğümlü bir çıkıştan oluşan üç tabakalı bir ileri beslemeli ağdan oluşan network Şekil 2'de verilmiştir.



Şekil 2 Çok katmanlı yapay sinir ağ modeli
(Figure 2 Multilayer neural network model)

Burada ağırlıklara bağlı olarak girdi denklemi

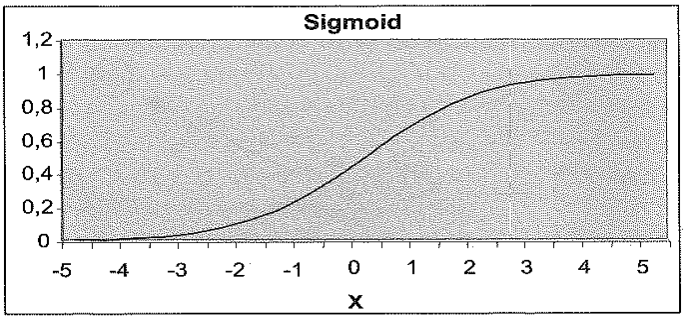
$$\mathbf{y\text{-}girdi}=\sum \mathbf{w_i.x_i}$$

ve bağımlı değişkeni x ağırlıkları w göstermek üzere giriş tabakasında ortalama sinyal denklemi:

$$\mathbf{S_i} = \mathbf{w_i}\ \mathbf{x_i}$$

olarak yazılır. Çözümlemede düğüm noktaları için Şekil 3'de gösterilen sigmoidal tip cevap (respons) fonksiyonu f($\mathbf{S_i}$) kullanılır.

$$\mathbf{X}_h = \frac{1}{1+\exp(-\mathbf{S_i})}$$



Şekil 3 Sigmoidal ateşleme fonksiyonu
(Figure 3 Sigmoidal ignition function)

1. İfadenin doğruluk derecesi ile Kazanç/Maliyet oranı ve **2.** Pratik çalışmalarda kullanım kolaylığı'dır.

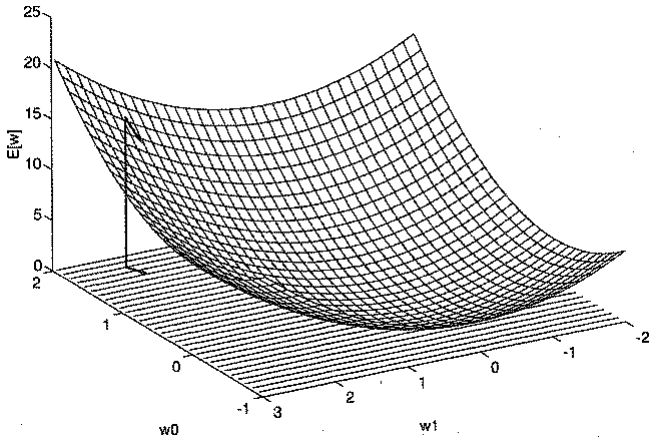
Ağın çalışmasında kullanılan tekniklerden en tanınmış ve basit olanı geri hesaplamada hatayı azaltma tekniği olarak adlandırılır.

$$Err = y_{data}\text{-}y$$

Burada hatayı en küçük kareler yöntemine göre minimize etmek amaçlanır.

$$K = (y_{data}\text{-}y)^2 \rightarrow \min$$

Şekil 4'de yapay sinir ağı hatalarının minimize edilmesi çizilmiştir.



Şekil 4 Yapay sinir ağı hatalarının minimizasyonu
(Figure 4 Minimization of neural network errors)

Burada parametrelerin kısmi türevleri alınarak:

$$\frac{\partial K}{\partial w} =$$

$$\frac{\partial K}{\partial w} = c\ Err \frac{\partial y}{\partial w} = 0 \quad \text{Burada } c\neq 0 \text{ dır.}$$

$$\frac{\partial y}{\partial w_h} = x_h \qquad \text{ve}$$

$$\frac{\partial y}{\partial w_i} = = w_h\ x_h^2\ x_i \exp(-w_i\ x_i) \text{ yazılır.}$$

Buradan kazanç vektörü

$$W_{h,new} = W_{h,old} + gain*Err*\frac{\partial y}{\partial w_h}$$

$$W_{i,new} = W_{i,old} + gain*Err*\frac{\partial y}{\partial w_i}$$

olarak tanımlanır ve bu yeni tanımlarla hesaplama zinciri tekrarlanır (Lübbert, A., vd., 2000).

2.4. Problemin KNIME'da modellemesi

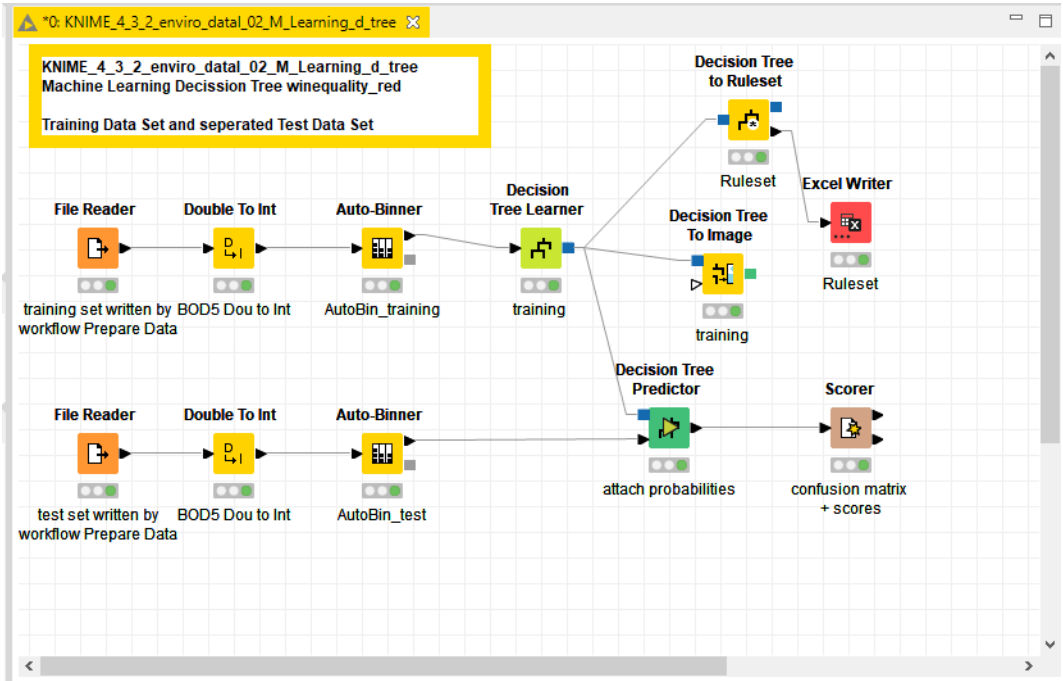
Bu kısımda daha önce yapılan bir çalışmanın **(Güller, v.d., 2019)** KNIME platformunda Karar Ağaçları Yöntemiyle değerlendirilen 334 adet evsel nitelikli atık suya ilişkin 7 adet parametrenin analiz sonuçlarını gösteren veri seti Yapay Sinir Ağları Yöntemiyle de tekrar değerlendirilerek iki yöntem karşılaştırılmıştır. İncelenen 334 örneklemin BOD_5 (***Biyolojik Oksijen İhtiyacı***) değeri dağılımı %53 oranında 100'den düşük olarak bulunmuştur. BOD_5 değeri 100-200 arasında olanların oranı %15,3 iken 450-550 arasında olanların oranı %12,6'dır. Bu çalışmadan BOD_5 değerini en fazla etkileyen değişkenin COD (***Kimyasal Oksijen İhtiyacı)*** olduğu anlaşılmaktadır. Kimyasal oksijen ihtiyacı 214.93 değerinden küçük ve eşit ise BOD_5 değeri 0-100 arası bir değere ulaşmaktadır. Bunun rastlanma sıklığı (%98,6) dır. COD 214.93 değerinden büyük olduğu durumlarda BOD_5 hiçbir şekilde 200 değerini aşmamıştır. BOD_5 in 100 – 200 arasında olma olasılığıysa sadece %1,4'dür.

2.4.1. Verilerin Hazılanması ve İstatistiksel Değerlendirme

Veriler atık su örneklerine ilişkin kimyasal analiz sonuçları olarak (enviro_dataI.xls- MUSKİ Atık Su Verileri) Güller, v.d., 2019' den alınmıştır. Yapılan uygulamaların tasarım sayfaları ve kullanılan Node ların yapılması gereken ayarları ve program çıktıları ile verilerin "öğrenme ve test" ("training" ve "test") olmak üzere bölünmesi verilere uygulanan istatistiksel analizler için hazırlanan KNIME tasarımı KNIME ara yüzeyinde gerçekleştirilmiştir. Öncelikle **csv Uzantılı verilerin okunmasında Dialog Box ayarları**, verilerin izlenmesinde **Interactive Table ayarları, Interactive Table çıktıları, Scatter Plot ayarları, Scatter Plot çıktıları, Interactive Histogram ayarları ve Statistics ayarları** ile verilerin bölünmesinde **Partitioning ayarları, Partitioning çıktıları** ve **csv Writer ile verilerin dosyaya yazdırılmasında Dialog Box ayarları** yapılmıştır. Bu kısımda "bölünme" ("**Partitioning**") %80 oranında ve "rastgele" ("**random")** olarak seçilmiştir (**KNIME Analytics Platforrm, Version 4.3.2**, 2021)

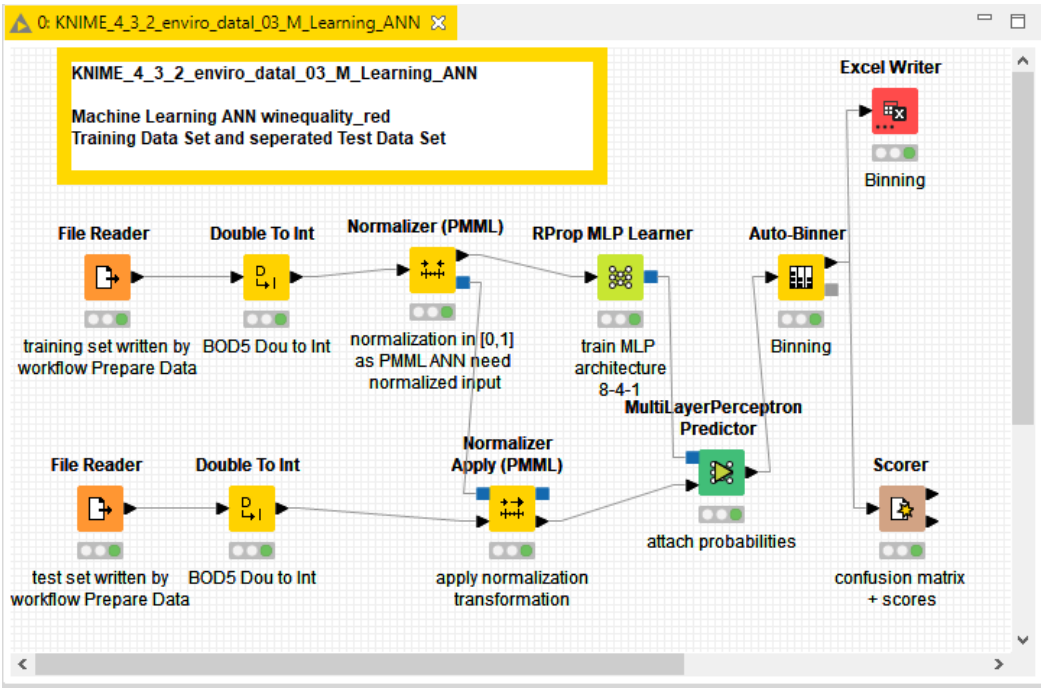
2.4.2. Atık Su Örnekleri için Karar Ağaçlarının KNIME'da Uygulaması

Bu kısımda sınıflandırma problemlerinin çözümünde seçilen karar ağacı yöntemlerinden SLIQ algoritması Gini indeksi (Entropi hesabı için) ile kullanılmaktadır. Uygulamada Budama (Pruning) yapılmamıştır. Şekil 5 atık su örneklerinin kimyasal analiz sonuçlarının değerlendirilmesine ilişkin hazırlanan KNIME tasarımının ara yüzey görüntüsüdür. Burada yapılan ayarlar sırasıyla; "öğrenme seti" ("training set") için csv uzantılı verilerin okunmasında "iletişim kutusu" ("Dialog Box") ayarları olup, verilerin dönüştürülmesinde, **Auto-Binner ayarları,** uygulama için sırasıyla **Decission Tree Learner Dialog Box** ve **Decission Tree PredictorDialog Box** ayarları, sonuçların değerlendirilmesinde **Scorer** ayarları ve sonuçların kurallara dönüştürülmesinde **Decission Tree to Ruleset** ayarları'dır



Şekil 5 Atık su örneklerinin değerlendirilmesine ilişkin hazırlanan KNIME Karar ağacı tasarımı
(Figure 5 KNIME Decision tree design for the evaluation of wastewater samples)

Bu kısımda da sınıflandırma problemlerinin çözümünde ileri beslemeli, geri hata yayımlı çok katmanlı yapay sinir ağı uygulanmıştır. Şekil 6. atık su örneklerinin kimyasal analiz sonuçlarının değerlendirilmesine ilişkin hazırlanan KNIME tasarımının ara yüzey görüntüsüdür. Burada yapılan ayarlar sırasıyla; training set için csv uzantılı verilerin okunmasında Dialog Box ayarları, verilerin dönüştürülmesinde, **Normalizer (train)** ayarları **ve Normalizer Apply (train ve test**) ayarları**,** uygulama için **Learner (train) ayarları** ve **MultiLayer Perception Predictor** ayarları, sonuçların değerlendirilmesinde **Auto-Binner ayarları** ve **Scorer – Confusion matrix** ayarları ve son olarak sınıflandırma verilerinin dosyaya yazdırılmasında **Excel Writer** ayarları olarak seçilmiştir.



Şekil 6 Atık su örneklerinin değerlendirilmesine ilişkin hazırlanan KNIME Yapay sinir ağı tasarımı
(Figure 6 KNIME Artificial neural network design for the evaluation of wastewater samples)

3. Araştırma Sonuçları ve Tartışma

Arıtım süreçlerin izlenebilmesi ve gerekli kontrollerin sağlanabilmesi ancak atık su ve aktif çamur karakteristiklerinin sürekli olarak belirlenmesine dayanmakta olup, atık su arıtma tesislerine gelen ham atık su veya arıtılmış atık sudan alınan örneklerle yapılacak olan asitlik (pH), sıcaklık (T), iletkenlik (C), çözünmüş oksijen (DO), oksijen doygunluğu (SO), tuzluluk (SA), elektriksel iletkenlik (EC), kimyasal oksijen ihtiyacı (COD), askıda katı madde (LSS), toplam azot (TN), toplam fosfor (TP) ve biyolojik oksijen ihtiyacı (BOD) gibi yapılan analizler ile biyolojik arıtma için kullanılan aktif çamur örneklerine ait prosesin tasarım parametrelerine ilişkin yapılan ölçümlere hızı (ASPR) ve biyo-kinetiği (BK) içeren katı alıkonma süresi (RT) ile geri çevrim oranıdır (FBR) Atık suların özelliklerinin belirlenmesinde Biyokimyasal oksijen ihtiyacı (BOD_5), kimyasal oksijen ihtiyacı (COD), toplam organik karbon (TOC) ve çözünmüş oksijen (DO) miktarlarının tayini atık suyun karakterizasyonu açısından en temel ölçüm kriterleridir. Son yıllarda yapılan çalışmalardan birinde sayılan atık su parametrelerden 11 tanesi laboratuvarda yapılan bir günlük çalışmada ölçülebilirken BOD_5 parametresinin ölçümünün 5 gün sürdüğü belirtilmekte ve istatistiksel değerlendirme için bir arıtma tesisinden alınan laboratuvar çalışmasında, 334 adet numuneden 12 parametrenin ölçümü yapılarak bir veri seti oluşturulmuştur. Bu veri setinde bulunan parametrelerinin karar ağacı yöntemiyle KNIME veri madenciliği paketinden yararlanarak BOD_5 parametresine etkileri irdelenmiştir. Böylece BOD_5 parametresine etkileri bilinen parametrelerin ağırlıklı etkileri dikkate alınarak sonucu bilinmeyen bir numunenin muhtemel BOD_5 değerinin tahminine çalışılmıştır.

3.1. Karar Ağaçlarına İlişkin Uygulamanın Sonuçları

Bu kısımda sınıflandırma problemlerinin çözümünde seçilen karar ağacı uygulamasının çıktıları ve sonuçları verilmiş olup, **Decission Tree Laerner**'dan elde edilen **Örneklerin BOD5 dağılımı** Şekil 7.'de**, Karar Ağacı'nın Görünümü** Şekil 8.'de, **Kuralları** Şekil 9'da **ve Decission Tree Laerner-Predictor'** a ilişkin **Scorer-Confusion Matrix** de Şekil 10'da sunulmuştur.

3.2. Yapay Sinir Ağlarına İlişkin Uygulamanın Sonuçları

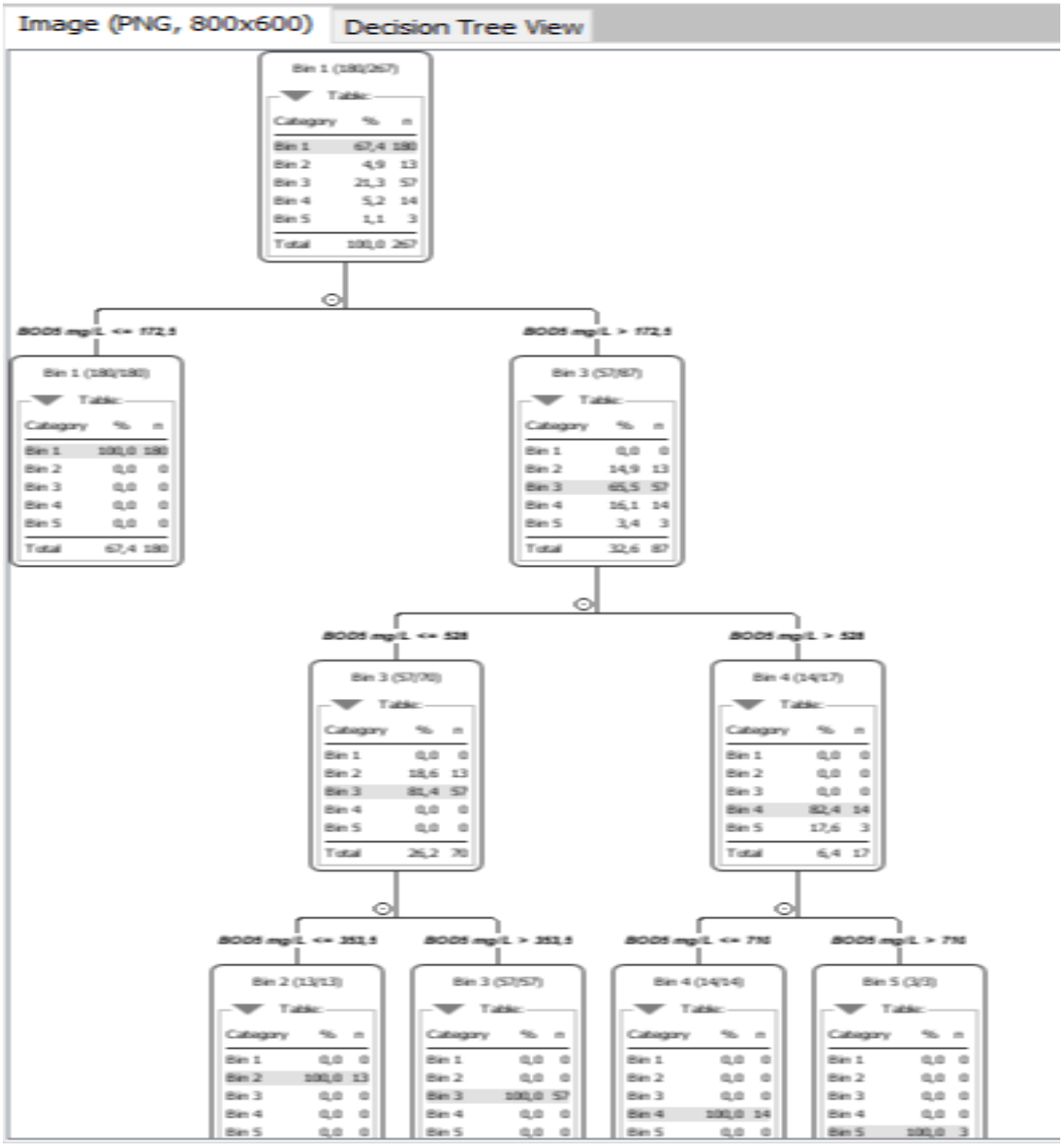
Bu kısımda da sınıflandırma problemlerinin çözümünde seçilen yapay sinir ağı uygulamasının uygulamasının çıktıları ve sonuçları incelenmiş olup, **Artificial Neural Network**'dan elde edilen **quality** ve **predicted quality (hesaplanmış**) değerleri Şekil 11'de verilmiş olup Karar Ağaçları'na benzer olarak 5 kutuda (sınıfta) toplanmış değerlerin karşılaştırılabilmesi için buna ilişkin hazırlanan Scorer-Confusion Matrix de Şekil 12 'de ve Artificial Neural Network'da training için bulunmuş hata grafiği Şekil 13'de sunulmuştur.

(177/334)

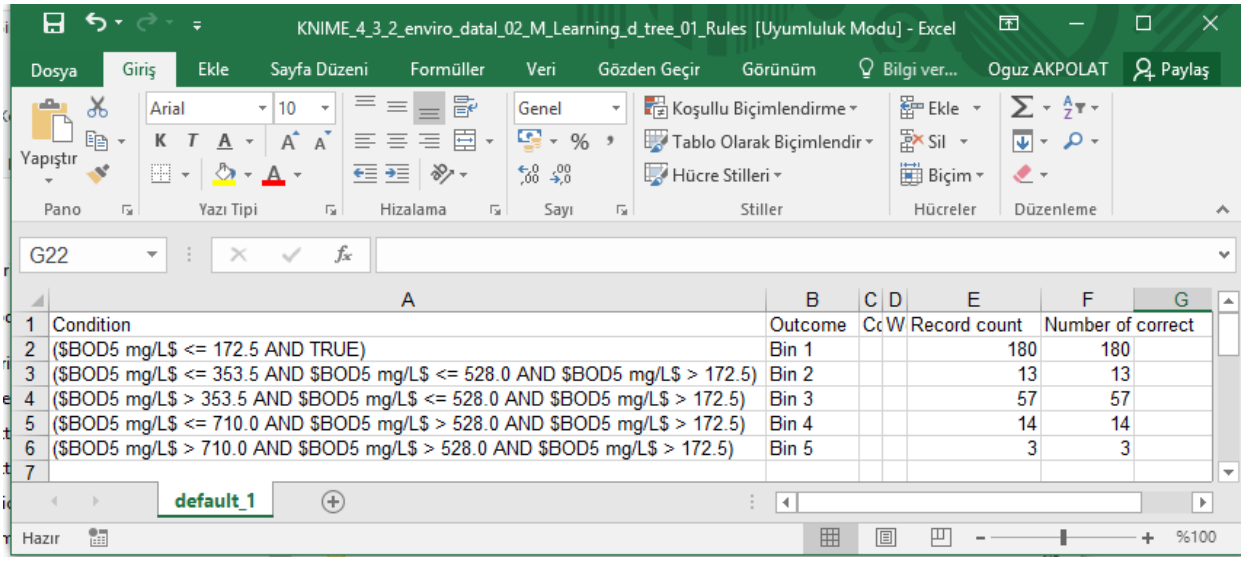
Table:

Category	%	n
100	53.0	177
100 200	15.3	51
200 – 350	4.2	14
350 – 450	9.3	31
450 550	12.6	42
550 750	5.1	17
750 950	0.6	2
Total	100.0	334

Şekil 7 Örneklerin BOD5 dağılımı
(Figure 7 BOD5 Distribution of samples)

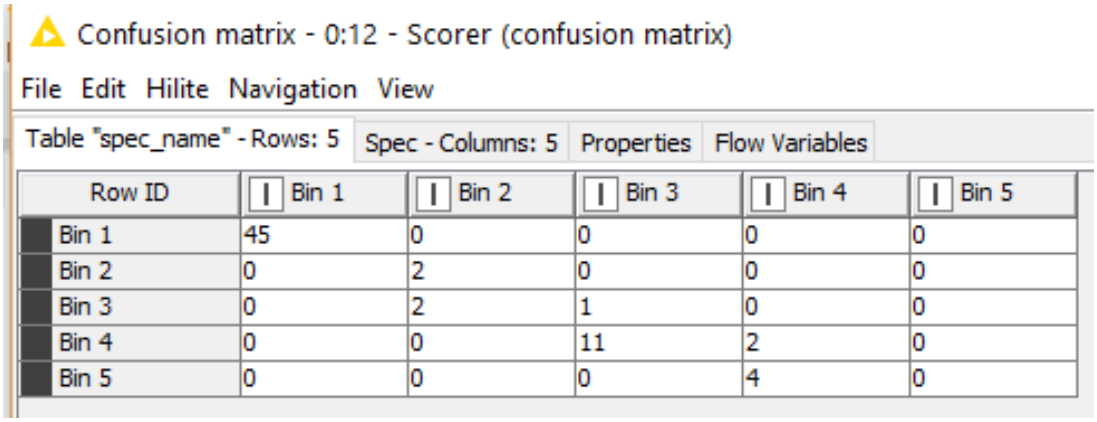


Şekil 8 Decission Tree Laerner'dan elde edilen Karar Ağacı'nın görünümü (Training)
(Figure 8 View of Decision Tree from Decision Tree Learner (Training)



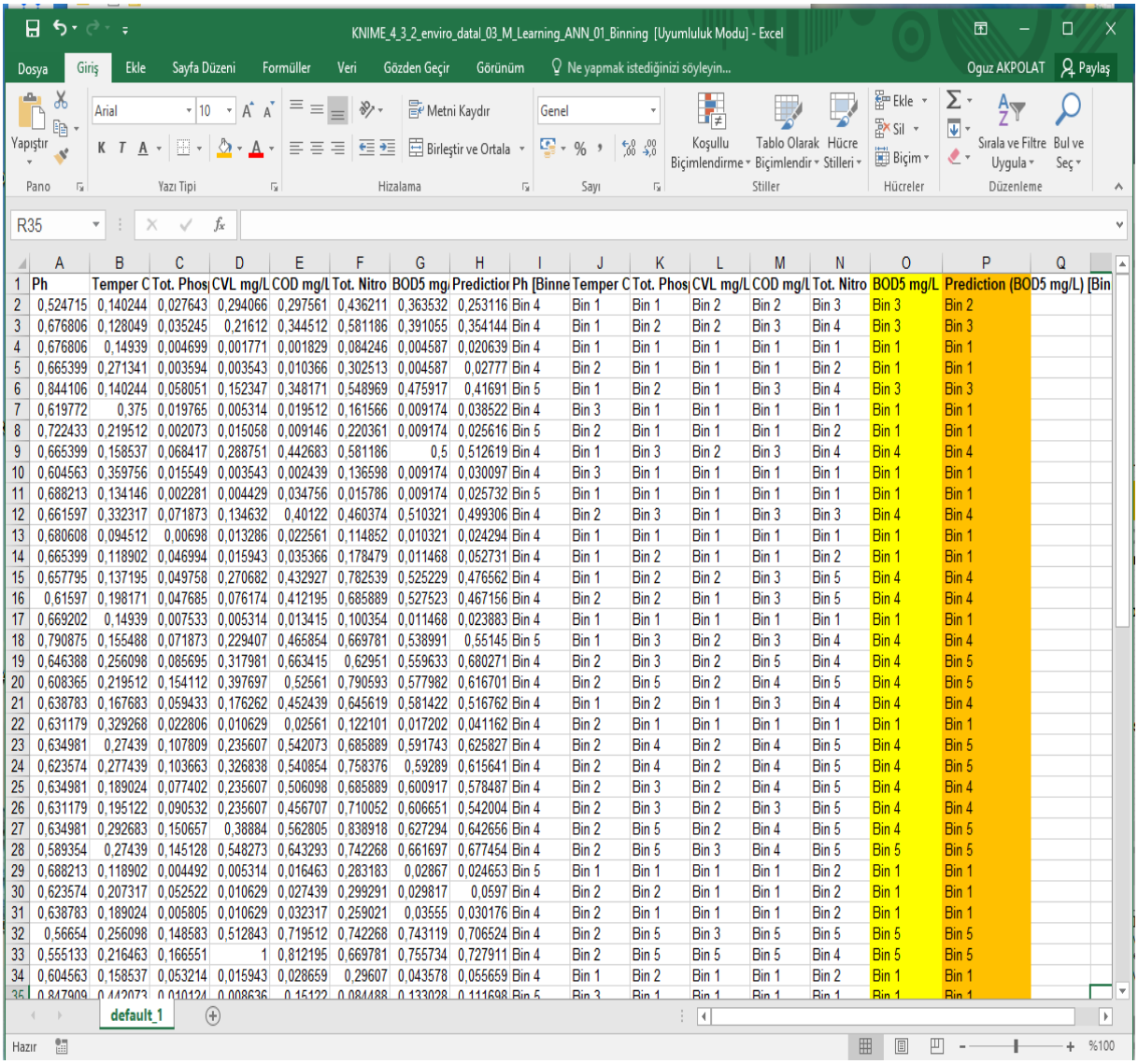
	A	B	C	D	E	F	G
1	Condition	Outcome	Cc	W	Record count	Number of correct	
2	($BOD5 mg/L$ <= 172.5 AND TRUE)	Bin 1			180	180	
3	($BOD5 mg/L$ <= 353.5 AND $BOD5 mg/L$ <= 528.0 AND $BOD5 mg/L$ > 172.5)	Bin 2			13	13	
4	($BOD5 mg/L$ > 353.5 AND $BOD5 mg/L$ <= 528.0 AND $BOD5 mg/L$ > 172.5)	Bin 3			57	57	
5	($BOD5 mg/L$ <= 710.0 AND $BOD5 mg/L$ > 528.0 AND $BOD5 mg/L$ > 172.5)	Bin 4			14	14	
6	($BOD5 mg/L$ > 710.0 AND $BOD5 mg/L$ > 528.0 AND $BOD5 mg/L$ > 172.5)	Bin 5			3	3	
7							

Şekil 9 Decission Tree Laerner'dan elde edilen Karar Ağacı'nın kuralları (Training)
(Figure 9 Rules of Decision Tree (Training) obtained from Decision Tree Learner)



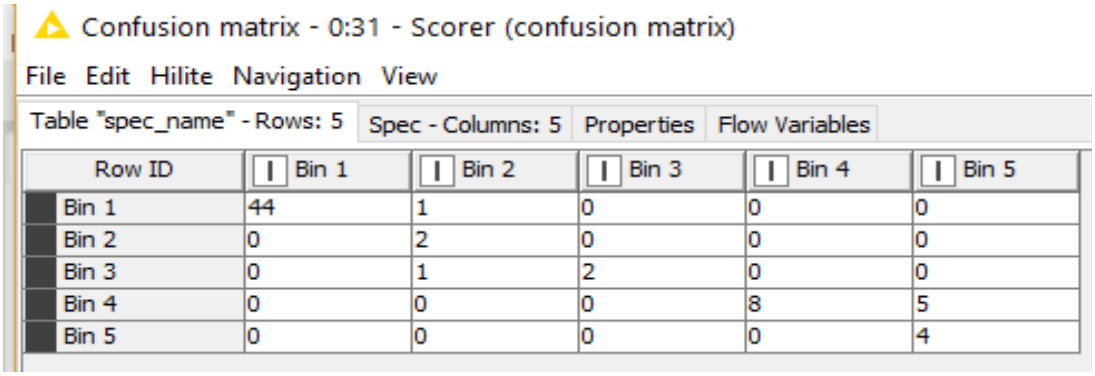
Row ID	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Bin 1	45	0	0	0	0
Bin 2	0	2	0	0	0
Bin 3	0	2	1	0	0
Bin 4	0	0	11	2	0
Bin 5	0	0	0	4	0

Şekil 10 Decission Tree Laerner-Predictor' a ilşkin Scorer-Confusion Matrix
(Figure 10 Scorer-Confusion Matrix for Decision Tree Learner-Predictor)



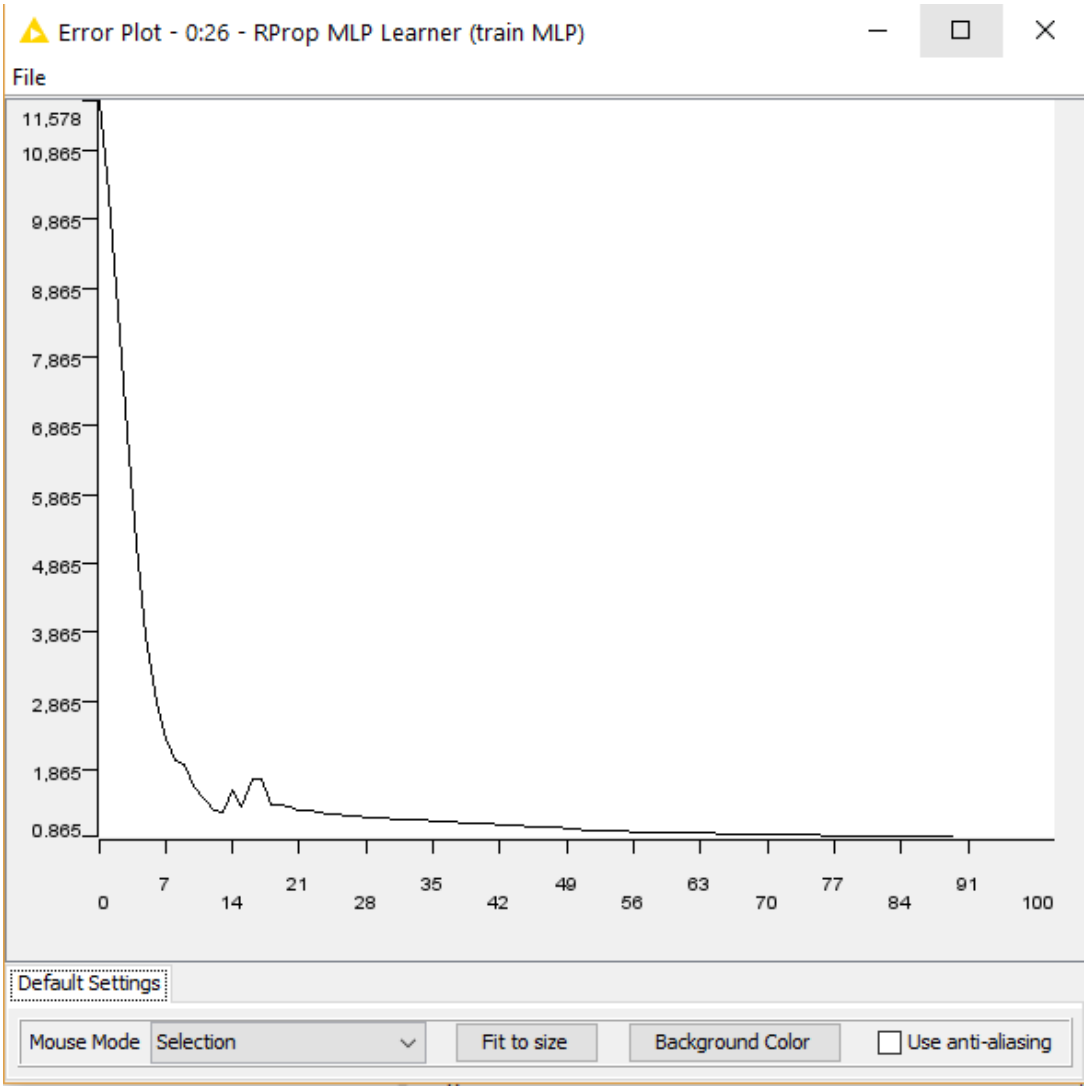
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Ph	Temper C	Tot. Phos	CVL mg/L	COD mg/L	Tot. Nitro	BOD5 mg/	Prediction	Ph [Binne	Temper C	Tot. Phos	CVL mg/L	COD mg/L	Tot. Nitro	BOD5 mg/L	Prediction (BOD5 mg/L)	[Bin
2	0,524715	0,140244	0,027643	0,294066	0,297561	0,436211	0,363532	0,253116	Bin 4	Bin 1	Bin 1	Bin 2	Bin 2	Bin 3	Bin 3	Bin 2	
3	0,676806	0,128049	0,035245	0,21612	0,344512	0,581186	0,391055	0,354144	Bin 4	Bin 1	Bin 2	Bin 2	Bin 3	Bin 4	Bin 3	Bin 3	
4	0,676806	0,14939	0,004699	0,001771	0,001829	0,084246	0,004587	0,020639	Bin 4	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	
5	0,665399	0,271341	0,003594	0,003543	0,010366	0,302513	0,004587	0,02777	Bin 4	Bin 2	Bin 1	Bin 1	Bin 1	Bin 2	Bin 1	Bin 1	
6	0,844106	0,140244	0,058051	0,152347	0,348171	0,548969	0,475917	0,41691	Bin 5	Bin 1	Bin 2	Bin 1	Bin 3	Bin 4	Bin 3	Bin 3	
7	0,619772	0,375	0,019765	0,005314	0,019512	0,161566	0,009174	0,038522	Bin 4	Bin 3	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	
8	0,722433	0,219512	0,002073	0,015058	0,009146	0,220361	0,009174	0,025616	Bin 5	Bin 2	Bin 1	Bin 1	Bin 1	Bin 2	Bin 1	Bin 1	
9	0,665399	0,158537	0,068417	0,288751	0,442683	0,581186	0,5	0,512619	Bin 4	Bin 1	Bin 3	Bin 2	Bin 3	Bin 4	Bin 4	Bin 4	
10	0,604563	0,359756	0,015549	0,003543	0,002439	0,136598	0,009174	0,030097	Bin 4	Bin 3	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	
11	0,688213	0,134146	0,002281	0,004429	0,034756	0,015786	0,009174	0,025732	Bin 5	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	
12	0,661597	0,332317	0,071873	0,134632	0,40122	0,460374	0,510321	0,499306	Bin 4	Bin 2	Bin 3	Bin 1	Bin 3	Bin 3	Bin 4	Bin 4	
13	0,680608	0,094512	0,00698	0,013286	0,022561	0,114852	0,010321	0,024294	Bin 4	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	
14	0,665399	0,118902	0,046994	0,015943	0,035366	0,178479	0,011468	0,052731	Bin 4	Bin 1	Bin 2	Bin 1	Bin 1	Bin 2	Bin 1	Bin 1	
15	0,657795	0,137195	0,049758	0,270682	0,432927	0,782539	0,525229	0,476562	Bin 4	Bin 1	Bin 2	Bin 2	Bin 3	Bin 5	Bin 4	Bin 4	
16	0,61597	0,198171	0,047685	0,076174	0,412195	0,685889	0,527523	0,467156	Bin 4	Bin 2	Bin 2	Bin 1	Bin 3	Bin 5	Bin 4	Bin 4	
17	0,669202	0,14939	0,007533	0,005314	0,013415	0,100354	0,011468	0,023883	Bin 4	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	
18	0,790875	0,155488	0,071873	0,229407	0,465854	0,669781	0,538991	0,55145	Bin 5	Bin 1	Bin 3	Bin 2	Bin 3	Bin 4	Bin 4	Bin 4	
19	0,646388	0,256098	0,085695	0,317981	0,663415	0,62951	0,559633	0,680271	Bin 4	Bin 2	Bin 3	Bin 2	Bin 5	Bin 4	Bin 4	Bin 5	
20	0,608365	0,219512	0,154112	0,397697	0,52561	0,790593	0,577982	0,616701	Bin 4	Bin 2	Bin 5	Bin 2	Bin 4	Bin 5	Bin 4	Bin 5	
21	0,638783	0,167683	0,059433	0,176262	0,452439	0,645619	0,581422	0,516762	Bin 4	Bin 1	Bin 2	Bin 1	Bin 3	Bin 4	Bin 4	Bin 4	
22	0,631179	0,329268	0,022806	0,010629	0,02561	0,122101	0,017202	0,041162	Bin 4	Bin 2	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	
23	0,634981	0,27439	0,107809	0,235607	0,542073	0,685889	0,591743	0,625827	Bin 4	Bin 2	Bin 4	Bin 2	Bin 4	Bin 5	Bin 4	Bin 5	
24	0,623574	0,277439	0,103663	0,326838	0,540854	0,758376	0,59289	0,615641	Bin 4	Bin 2	Bin 4	Bin 2	Bin 4	Bin 5	Bin 4	Bin 5	
25	0,634981	0,189024	0,077402	0,235607	0,506098	0,685889	0,600917	0,578487	Bin 4	Bin 2	Bin 3	Bin 2	Bin 4	Bin 5	Bin 4	Bin 4	
26	0,631179	0,195122	0,090532	0,235607	0,456707	0,710052	0,606651	0,542004	Bin 4	Bin 2	Bin 3	Bin 2	Bin 3	Bin 5	Bin 4	Bin 4	
27	0,634981	0,292683	0,150657	0,38884	0,562805	0,838918	0,627294	0,642656	Bin 4	Bin 2	Bin 5	Bin 2	Bin 4	Bin 5	Bin 4	Bin 5	
28	0,589354	0,27439	0,145128	0,548273	0,643293	0,742268	0,661697	0,677454	Bin 4	Bin 2	Bin 5	Bin 3	Bin 4	Bin 5	Bin 5	Bin 5	
29	0,688213	0,118902	0,004492	0,005314	0,016463	0,283183	0,02867	0,024653	Bin 5	Bin 1	Bin 1	Bin 1	Bin 1	Bin 2	Bin 1	Bin 1	
30	0,623574	0,207317	0,052522	0,010629	0,027439	0,299291	0,029817	0,0597	Bin 4	Bin 2	Bin 2	Bin 1	Bin 1	Bin 2	Bin 1	Bin 1	
31	0,638783	0,189024	0,005805	0,010629	0,032317	0,259021	0,03555	0,030176	Bin 4	Bin 2	Bin 1	Bin 1	Bin 1	Bin 2	Bin 1	Bin 1	
32	0,56654	0,256098	0,148583	0,512843	0,719512	0,742268	0,743119	0,706524	Bin 4	Bin 2	Bin 5	Bin 3	Bin 5	Bin 5	Bin 5	Bin 5	
33	0,555133	0,216463	0,166551	1	0,812195	0,669781	0,755734	0,727911	Bin 4	Bin 2	Bin 5	Bin 5	Bin 5	Bin 4	Bin 5	Bin 5	
34	0,604563	0,158537	0,053214	0,015943	0,028659	0,29607	0,043578	0,055659	Bin 4	Bin 1	Bin 2	Bin 1	Bin 1	Bin 2	Bin 1	Bin 1	
35	0,847909	0,442073	0,010124	0,008636	0,15122	0,084488	0,133028	0,111698	Bin 5	Bin 3	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	Bin 1	

Şekil 11 Artificial Neural Network'den elde edilen kutulanmış (binned) değerler
(Figure 11 With binned value obtained from Artificial Neural Network)



Row ID	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Bin 1	44	1	0	0	0
Bin 2	0	2	0	0	0
Bin 3	0	1	2	0	0
Bin 4	0	0	0	8	5
Bin 5	0	0	0	0	4

Şekil 12 Artificial Neural Network'den elde edilen kutulanmış ve yeniden kutulanmış değerler için Scorer-Confusion Matrix
(Figure 12 Scorer-Confusion Matrix for binned and re-binned values from Artificial Neural Network)



Şekil 13 Artificial Neural Network'da training için bulunmuş hata grafiği
(Figure 13 Error graph found for training in Artificial Neural Network)

4. Sonuç

Son yıllarda yapılan çalışmalardan birinde sayılan atık su parametrelerden 11 tanesinin karar ağacı yöntemiyle KNIME veri madenciliği paketinden yararlanarak BOD_5 parametresine etkileri irdelenmiştir. Böylece BOD_5 parametresine etkileri bilinen parametrelerin ağırlıklı etkileri dikkate alınarak sonucu bilinmeyen bir numunenin muhtemel BOD_5 değerinin tahminine çalışılmıştır. Ardından da daha önce yapılan bur çalışmada KNIME platformunda Karar Ağaçları Yöntemiyle değerlendirilen 334 adet evsel nitelikli atık suya ilişkin etkin olduğu düşnülen 7 adet parametrenin analiz sonuçlarını gösteren veri seti Yapay Sinir Ağları Yöntemiyle de tekrar değerlendirilerek iki yöntem karşılaştırılmıştır. İncelenen 334 örneklemin BOD_5 (*Biyolojik Oksijen İhtiyacı*) değeri dağılımı, %53 oranında 100'den düşük olarak bulunmuştur. BOD_5 değeri 100-200 arasında olanların oranı %15,3 iken 450-550 arasında olanların oranı %12,6'dır. Bu çalışmadan BOD_5 değeri ile en fazla etkileşen değişkenin COD (*Kimyasal Oksijen İhtiyacı)* olduğu anlaşılmaktadır. Kimyasal oksijen ihtiyacı 214.93 değerinden küçük ve eşit ise BOD_5 değeri 0-100 arası bir değere ulaşmaktadır. Bunun rastlanma sıklığı (%98,6) dır. COD 214.93 değerinden büyük olduğu durumlarda BOD_5 hiçbir şekilde 200 değerini aşmamıştır. BOD_5 in 100 – 200 arasında olma olasılığıysa sadece %1,4'tür (Güller, v.d., 2019).

Her iki yöntemin sonuçları karşılaştırılmak için Tablo 1'de verilen aşağıdaki sonuçlar incelendiğinde, Kutulanmış (Binned) değerlerin bulunduğu sınıflar arasında dağılımların yakın ancak kaymalar olduğu görülmektedir. Bununla birlikte sınıf sayıları arttırıldığında bu kaymaların kısmen de olsa giderilebileceği unutulmamalıdır. Ayrıca bu sonuçlar gelecek çalışmalarda hem (***Karar Ağaçları***) için gruplama sayısı, kazanç gibi parametreler hem de (Parantez içindeki değerler **Yapay Sinir Ağları** nı göstermektedir) için ağ katman sayısı ve kazanç oranı gibi parametreler değiştirilerek en iyilenebilir (optimize edilebilir). Buradan da bu çalışma için karar ağaçlarının daha iyi sonuçlar verdiği söylenebilir.

Tablo 1. Atık su örneği için (***Karar Ağaçları***) ve (**Yapay Sinir Ağları**)ının karşılaştırılması
*(Table 1. Comparison of (**Decision Trees**) and (**Artificial Neural Networks**) for wastewater sample)*

	Bin1	**Bin2**	**Bin3**	**Bin4**	**Bin5**
Bin1	***45*** (44)	(1)			
Bin2		***2*** (2)			
Bin3		***2*** (1)	***1*** (2)		
Bin4			***11***	***2*** (8)	(5)
Bin5				***4***	(4)

Kaynakça

Brereton, R. G., (2016), Chemometrics: Data Driven Extraction for Science, 2nd Edition, Wiley Pub.

Doğan, O., (2017), Ücretsiz Veri Madenciliği Araçları ve Türkiyede Bilinirlikleri Üzerine Bir Araştırma, Ege Stratejik Araştırmalar Dergisi Cilt 8, Sayı 1

Güller, S., Silahtaroğlu, G., Akpolat, O., (2019), Analysis waste water characteristics via data mining: A Muğla province case and external validation, Communication in Statistics: Case Studies, Data Analysis and Applications, Vol.5, No. 3, 200-213.

https://www.muski.gov.tr/aritmaveicmesuyutesislerimiz.aspx,, *(2020)*

https://www.cs.waikato.ac.nz/ml/weka/index.html, (2019)

https://erdincuzun.com/makine_ogrenmesi/.../, (2020), Decision Tree (Karar Ağacı): ID3 Algoritması–Classification

http://mail.baskent.edu.tr/~20410964/DM_8.pdf, (2020), Karar Ağacı (Decision Karar Ağacı (Decision tree) nedir?

Jiawei, H., Kamber, M., Pei, J., (2012), Data Mining; Concepts and Technics, Morgan Kaufmann Publishers, Elsevier Inc.,

KNIME Anaytics Platforrm, Version 4.3.2, (2021), https://www.knime.com

Lübbert, A., Simutis, R., Volk, N., Galvanuskas, V., (2000), Biochemical Process Optimization and Control. Hands-on Course, Martin Luther University, Germany.

Özdemir, D., (Eğitmen), Taner, M. S., Ertaş, H.,, (2012), Kemometri Eğitimi Ders Notları, İzmirYüksek teknoloji Enstitüsü, 01-03 Temmuz, Akdeniz Üniversitesi ve Kimya Eğitim Akademisi, Türkiye.

Qiao, J., Li, W., Han, H., (2014), Soft Computing of Biochemical Oxygen Demand Using an Improved T–S Fuzzy Neural Network, Chinese Journal of Chemical Engineering, 22, 1254–1259

Silahtaroğlu, G., (2016), Veri Madenciliği kavram ve Algoritmaları, II. Baskı, Papatya Yayıncılık