©2019 Published in 7th International Symposium on Innovative Technologies in Engineering and Science 22-24 November 2019 (ISITES2019 SanliUrfa - Turkey)
https://doi.org/10.33793/acperpro.02.03.150

Çevresel Veri Problemleri için Veri Madenciliği ile Veri Ön İşleme

*[1]Beytullah Eren ve [2]İpek Aksangür
*[1]Çevre Mühendisliği Bölümü, Mühendislik Fakültesi, Sakarya Üniversitesi, 54187, Sakarya, Türkiye
[2] Fen Bilimleri Enstitüsü, Çevre Mühendisliği Anabilim Dalı, Sakarya Üniversitesi, 54187, Sakarya, Türkiye

Özet

Atık yönetiminin yapıldığı çevresel tesislerin kontrolünde gerçekçi modellere ve doğru tahminlere ihtiyaç vardır. Doğru tahmin modelinin geliştirilmesinin en önemli adımı sağlıklı verin olmasıdır. Çevresel tesislerden temin edilen verilerin ön işlem aşamasında kalitesiz verilerinin temizlenmesi ve eksik verilerin tamamlanması gerekmektedir. Bu çalışmada bir çevresel tesisten Ocak 2016 - Eylül 2018 tarihleri arasında toplanan verilerden veri madenciliği programı ile modelleme öncesi temiz verilerin elde edilmesi hedeflenmiştir. Verilerin temizlenmesi aşamasında; pH, Eİ, AKM, KOİ, $BOİ_5$, Yağ-Gres ve TÇK parametrelerine ait sırasıyla 25, 141, 26, 22, 241, 645 ve 688 adet eksik veri tespit edilmiştir. Eksik veriler ortalama değerler göz önüne alınarak tamamlanmıştır. Sonrasında 10 adet gürültülü veri belirlenmiş ve satır bazlı temizleme yapılmıştır. Sezonluk ortalama değerlerin belirlenmesi için $BOİ_5$ parametresi kullanılmış ve sezonluk ortalama değerler program aracılığı ile hesaplanmıştır. Böylece bir çevresel tesisin ham verilerinin veri madenciliği programları yardımı ile temizlenmesi ve eksik verilerin tamamlanarak modelleme uygulanması için hazır hale getirilmesi sağlanmıştır.

Anahtar Kelimeler: Çevresel veri problemleri, veri madenciliği, RapidMiner Studio, veri ön işleme

Abstract

Realistic models and accurate estimates are needed for the control of environmental facilities where waste management is performed. The most important step in developing an accurate prediction model is clean data. Data from environmental facilities should be cleared during the pre-treatment phase. During the cleaning phase of the data; 25, 141, 26, 22, 241, 645, and 688 missing data were determined for pH, EC, AKM, COD, BOD_5, Oil-Grease and TDS parameters, respectively. The missing data were completed according to the mean values. Then, 10 noisy data were identified and row based cleaning was performed. In order to determine seasonal average values, BOI_5 parameter was studied and seasonal average values were calculated through the program. In this study, it is revealed that the raw data of an environmental facility can be cleaned with data mining programs and made ready for the next stage model application.

Key words: Environmental data problems, data mining, RapidMiner Studio, data preprocessing

*Sorumlu Yazar: Adres: Çevre Mühendisliği Bölümü, Mühendislik Fakültesi, Sakarya Üniversitesi, 54187, Sakarya, Türkiye. E-mail: beren@sakarya.edu.tr, Tel: +902642955642

1. Giriş

Atıklar, uygun teknolojiler (fiziksel, biyolojik ve kimyasal prosesleri içine alan teknolojiler) kullanıldığı taktirde çevreye ve insana zarar vermeden yönetilebilmektedir. Atığa uygun teknolojilerin kullanıldığı bu prosesleri içeren tesislere genel ifade olarak *Çevresel Tesis* adı verilmektedir. Atıksu arıtma tesisleri, proses suyu üretim tesisleri, atık işleme tesisleri, düzenli, geçici depolama ve geri kazanım tesisleri örnek olarak sayılabilmektedir.

Çevresel tesisler sürdürülebilir çevrenin korunmasında önemli bir yere sahiptir. Bu noktada çevre tesislerinin işletilmesi, kontrolü ve geliştirilmesi için tesislerin anlık, günlük, dönemsel ve yıllık davranışlarının ve performansının ortaya konulması oldukça önemlidir. Tesise ait çıkış değerlerinin tahmini, özellikle tesis operatörlerinin bir problemle karşılaşmadan önce gerekli önlemleri almasına ve iyileştirmeleri yapabilmesine olanak sağlar. Her tesis bünyesinde barındırdığı prosesler ve işlediği atık türüne bağlı olarak kendine özgüdür. Bir çevre tesisinde elde edilen veriler tesis hakkında sistem performansı için genel bilgi sahibi olmamızı sağlar. Bir çevresel tesise ait giriş ve çıkış değerlerine ait veriler arasındaki ilişki genellikle non-lineer'dir [1,2]. Bu doğrusal olmayan davranış biçimi parametrelerinin tahminini zorlaştırmakta ve karmaşık matematiksel fonksiyonlara ihtiyaç duyulmaktadır. Sistem davranışının gelecekte tahmin edilebilmesi için parametre tahmini yapan bir modelin geliştirilmesine ihtiyaç vardır. En uygun modelin belirlenmesi ve doğru tahminlere ulaşılabilmesi için verilerin sistemi temsil etmeleri ve anlamlı olmaları gerekmektedir.

Bu çalışmada; bir çevresel tesisin sistem performansının ortaya konulması ve makina öğrenmesi ile geleceğe ait veri tahmininin yapılabilmesi için, sistemi temsil eden verilerin, bir veri madenciliği programı ile temiz veri (cleaned data) haline getirilmesi hedeflenmiştir.

2. Materyal/Metot

2.1. Veri Kaynağı

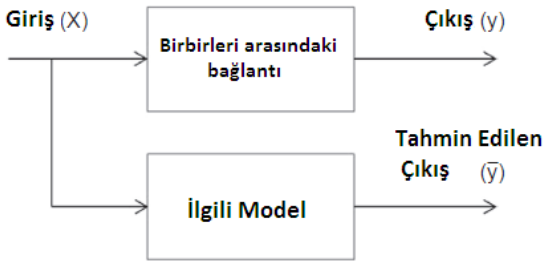
Öncelikle çevresel tesislere ait veriler (günlük, aylık, sezonluk, mevsimsel, yıllık vb.) problemin türü ve çözümüne yönelik belirlenir. Bu çalışmada Ocak 2016- Eylül 2018 tarihleri arasında bir atıksu arıtma tesisinde günlük olarak tesis girişinden alınan numunelerde yapılan analiz çalışmaları ile elde edilmiş veriler kullanılmıştır. Tablo 1'de ilgili verilere ait parametre bilgileri verilmiştir.

Tablo 1. Atıksu arıtma tesisi giriş parametreleri

Parametre	**Birim**
Giriş - pH	-
Giriş - Elektriksel İletkenlik (Eİ)	µs/cm
Giriş - Askıda Katı Madde (AKM)	mg/L
Giriş - Kimyasal Oksijen İhtiyacı (KOİ)	mg/L
Giriş - Biyolojik Oksijen İhtiyacı ($BOİ_5$)	mg/L
Giriş - Yağ-Gres	mg/L
Giriş - Toplam Çözünmüş Katı (TÇK)	mg/L

2.2. Veri Madenciliği ve Yapay Zeka

Bir prosese ait geçmiş verileri kullanarak geleceğe yönelik verilerin tahmin edilebilmesinde *Yapay Zeka (Artifical Intelligent-AI)* uygulamaları önemli bir role sahiptir. En basit ifadeyle yapay zeka, görevleri yerine getirmek için insan zekasını taklit eden ve topladıkları bilgilere göre yinelemeli olarak kendilerini iyileştirebilen sistemler veya makineler anlamına gelir [3]. Makine öğrenmesi ile bir olay/durumu deneyim yolu ile öğrenerek daha önce karşılaşılamayan benzer olaylar için karar verebilme/tahmin yapabilme ve çözüm üretilebilmektedir. Tahmin analitiğinde; girdiler ile çıktılar arasındaki bağlantının saptanmasının ardından gelecek tahminlerinin gerçekleştirilmesi için uygun model oluşturulmaktadır (Şekil 1) [4].



Şekil 1. Tahmin analitiği ve modeli [4]

Yapay zeka uygulamalarında en doğru tahmin modelin belirlenmesi için sisteme ait kaliteli verilere ihtiyaç duyulmaktadır. Bir sistemden elde edilen ham verilerin çeşitli veri madenciliği araçları kullanılarak ön işlenmesi (kalitesiz verilerin çıkarılması, eksik verilerin tamamlanması vb.); ileri adım olan tahmin operatörlerinin kullanılması ve başarılı bir modelin oluşturulması için oldukça önemlidir. Yapay zeka çalışmaları çeşitli veri madenciliği tekniklerinin geliştirilmesine neden olmuştur. Bir veri madenciliği uygulaması 5 aşamadan oluşmaktadır (Şekil 2) [5,6,7,8].

Şekil 2. Veri Madenciliği uygulama aşamaları

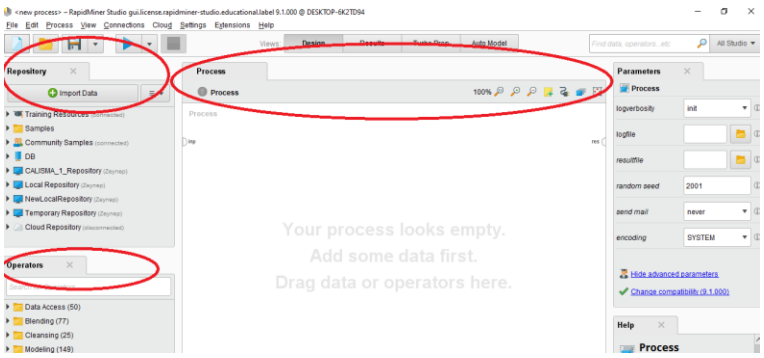
Çalışmada kullanılan veriler Bölüm 2.1 veri kaynaklarında tanımlanmıştır. Çevre tesislerinde veri kaynakları; tesislerden alınan numunelerde yapılan laboratuvar analiz çalışmaları ve anlık ölçüm cihazlarından okunan çeşitli parametrelere ait değerlerdir. Çevresel tesislerden gelen verilerde sıklıkla problemlerle karşılaşılmaktadır. Özellikle belirsiz ve kesin olmayan verilerin (yeterli online analizörlerin tesiste olmaması ve/veya her zaman aktif olamamaları, numune alma sıklığı ya da farklılığı, çok fazla eksik veri… vb.) ön işlemlere tabi tutulması gerekmektedir. Bu çalışmada veri madenciliği uygulama aşamalarından biri olan veri ön işleme çalışması yapılmıştır.

2.1.1. Veri Ön İşleme

Makine öğrenmesinde, denetimsiz öğrenme problemleri için ön işleme adımında veri madenciliğinden yararlanabilmektedir [9]. Veri ön işlemenin (data preprocessing) amacı, makine öğrenmesi için kullanılacak ham verinin içindeki uygun olmayan veya hatalı girilmiş verileri ayıklanması ve eksik verilerin tamamlanmasıdır. Bir veri setinde eksik veriler (missing value) veri madenciliği yöntemleri kullanılarak uygun değerler ile doldurulur (fitting). Veri setinde bir parametreye ait eğer eksik veri çok ise bu parametreye ait verilerin silinmesi (cleaning) gerekmektedir. Ham verilerin ön işlemesinde birçok açık ve kapalı kaynak kodlu program kullanılmaktadır. *Orange, RapidMiner Studio, WEKA, Scriptella ETL, Jhep Work, KNIME, ELKI* gibi açık kaynak programları bu amaçla sıklıkla kullanılmaktadır [9]. Modelleme ile ilgili çalışmalar incelendiğinde; ham verilerin tahmin modeline uygulanmadan önce çoğunlukla veri madenciliği programları ile ön işleme tabi tutulduğu görülmüştür [1;10;11;12] .

RapidMiner Studio Version 9.1(https://rapidminer.com/)

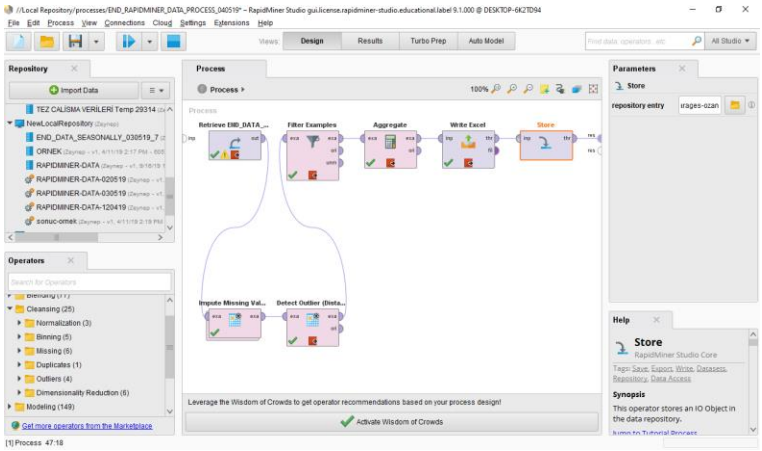
Bu çalışmada açık kaynak kodlu olması ve birçok operatörü (Support Vector Machines (SVM), Discriminant Analysis (DA), Linear Regression (LR), Logistic Regression (LR), Naive Bayes, Decision Tree ve Neural Nets vb.) hazır bulundurması nedeniyle YALE Üniversitesi tarafından geliştirilmiş olan RapidMiner Studio Version 9.1 veri ön işleme programı olarak kullanılmıştır. RapidMiner, Java dilinde yazılmış olup, kendi içerisine Java dili ile kod ekleme imkanı sağlamasının yanında, Python, Weka veya R gibi dillerle/ortamlarla uyumlu olarak çalışabilmektedir [13]. Programda; ön işleme operatörleri olarak; Normalizasyon *(normalization)*, gruplama *(binning)*, eksik veri *(missing)*, tekrarlanan veri *(duplicates)*, aykırı (uç) veri *(outliers)*, boyut indirgeme *(dimensionality reduction)* modülleri hazır halde bulunmakta ve bu modüller kullanılırken olay/durum için gerekli müdahalelere izin vermektedir. Şekil 3'de *RapidMiner Studio Version 9.1.* veri giriş ekranı verilmiştir. Başlangıç ekranında; verilerin içinde bulunduğu Depo (*Repository)*, veri madenciliği operatörlerinin içinde bulunduğu Operatörler *(Operators)* ve işlemlerin yapıldığı Proses *(Process)* kısımları yer almaktadır.



Şekil 3. RapidMiner Studio Version 9.1. başlangıç ekranı

3. Bulgular

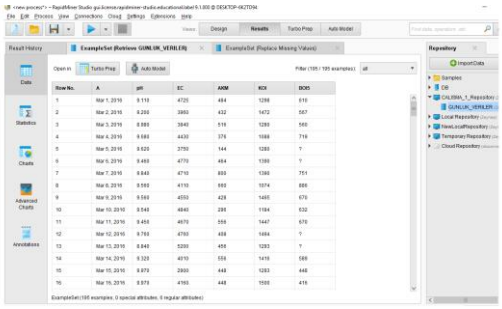
Çalışmada tahmin modelinin geliştirilmesinde kullanılacak ham verinin uygulamaya hazır hale getirilmesi için veri ön işleme yapılmıştır. Toplamda7 parametreye *(pH, Eİ, AKM, KOİ, BOİ$_5$, Yağ-Gres ve TÇK)* ait 730 x 7 adet günlük veri oluşturulmuştur. Tüm veriler kullanılarak her bir parametre için mevsimsel ortalama verilere ulaşılmaya çalışılmıştır. Bu makale çalışması yalnızca BOİ$_5$ parametresinin sezonsal ortalama değerlerini elde etmeyi kapsamaktadır. Programın başında bazı hazır şablonlar bulunmasına rağmen boş (blank) ekranın kullanılması tercih edilmiştir. Çalışmada; Data Access operatörleri (*Retrieve ve Store*), Missing operatörü (*Impute Missing Values*), Outliers operatörü (*Detect Outliner*), Blending operatörü (*Filter Examples*), Table operatörü (*Aggregate*) ve Write operatörü (*Write*) olmak üzere 7 farklı operatör kullanılmıştır. Çalışma 6 adımda gerçekleştirmiştir. Yukarıda belirtilen tüm işlemler için oluşturulan sistemin ekran görseli Şekil 4'te verilmiştir.



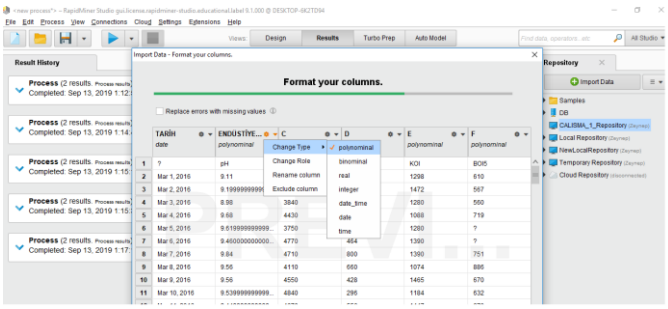
Şekil 4. Veri ön işleme için oluşturulan sistem

- *Retrieve END_DATA (Veri Girişi)*: Sisteme ham verilerin girişi ve her bir parametrenin (kolonun) rol ve tiplerinin atanması (Şekil 5-6) işlemi,
- *Impute Missing Values operatörünün sisteme eklenmesi*: Replace by Average yöntemi ile ortalama değerler göz önüne alınarak eksik verilerin tamamlanması (Şekil 7-8) işlemi,
- *Detect Outliner operatörünün sisteme eklenmesi*: Distance yöntemi ile True/False atamasının yapılması ve bir verinin çevresindeki ilk 10 komşu veri baz alınarak gürültülü verilerin (noisy data) belirlenmesi (Şekil 9-10) işlemi,
- *Filter Examples operatörünün sisteme eklenmesi*: Satır bazlı silme ile gürültülü verilerin temizlenmesi (Şekil 11-12) işlemi,
- *Aggregate operatörünün sisteme eklenmesi*: Sezonsal ayrımın yapılabilmesi için belirlenen parametre (BOİ$_5$) bazında ortalama sezonsal değerlerin elde edilmesi (Şekil 13-14) işlemi,
- *Write Excel ve Store operatörlerinin sisteme eklenmesi*: Sistemden temizlenmiş verilerin excel olarak dışarıya aktarılması ve kaydetme (Şekil 15) işlemi,

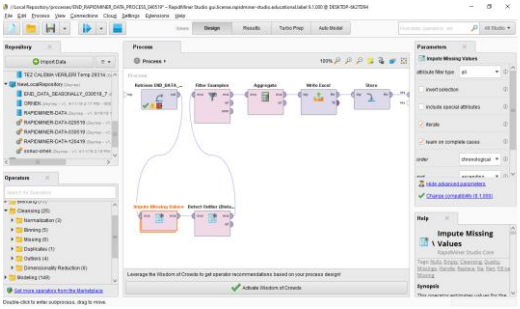
kullanılarak temizlenmiş veriler elde edilmiştir.



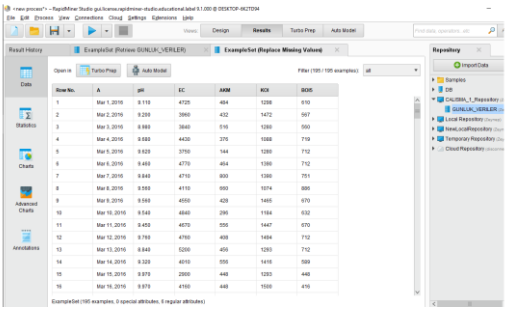
Şekil 5. Veri girişi ekranı (Günlük veriler)



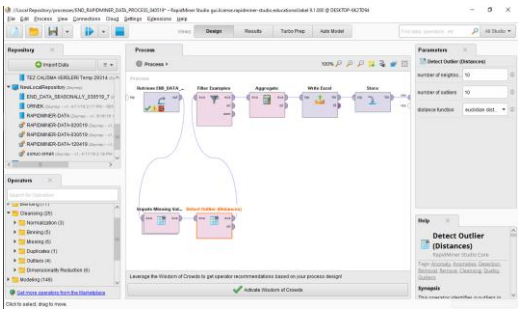
Şekil 6. Rol ve türlerin atanması



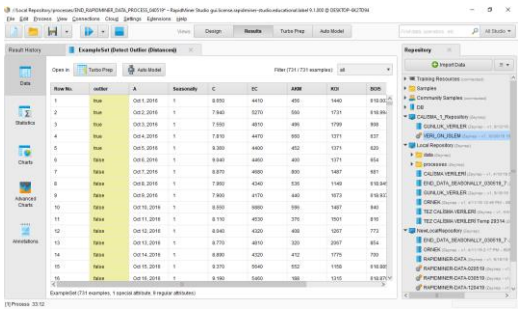
Şekil 7. Replace Missing Values operatörü ile eksik verilerin tamamlanması ekranı ve belirlenen kriterler



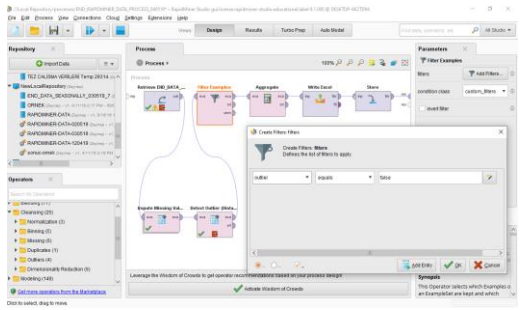
Şekil 8. Eksik verilerin tamamlandığı ekranı



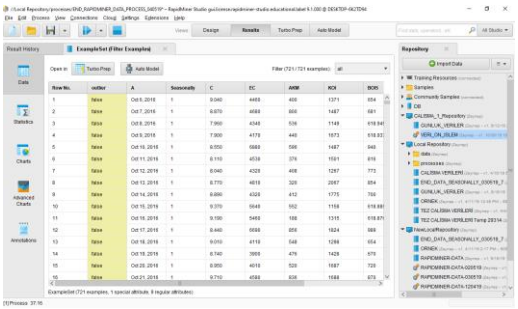
Şekil 9. Detect Outliner ekranı ve belirlenen kriterler



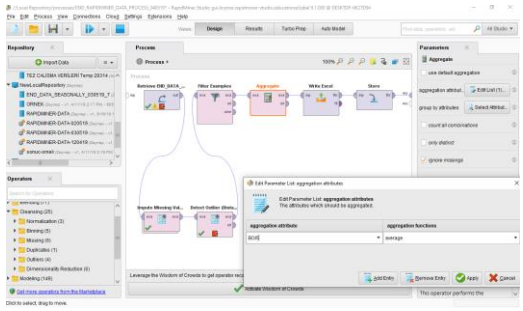
Şekil 10. Detect Outliner sonuç ekranı



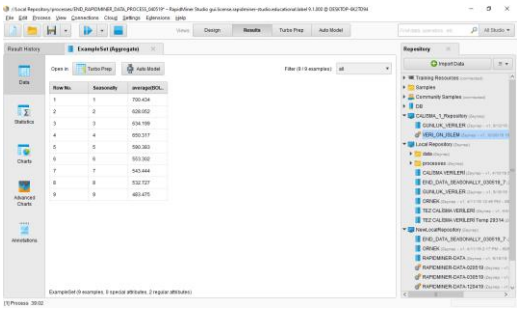
Şekil 11. Filter Examples ekranı ve belirlenen kriterler



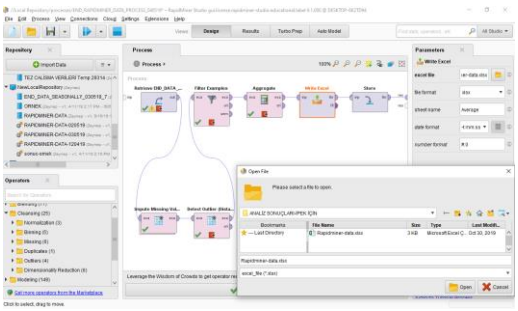
Şekil 12. Filter examples sonuç ekranı



Şekil 13. Aggregate ekranı ve belirlenen kriterler



Şekil 14. Aggregate sonuç ekranı



Şekil 15. Write Excel ve Store operatörleri ekranı

4. Sonuçlar ve Tartışma

Çevresel tesislerine ait ham veriler kullanılarak modelle yöntemi ile parametre tahminlerinin yapılabilmesi için ham verilerin veri madenciliği yöntemleri kullanılacak ön işleme tabi tutulması ve sağlıklı bir şekilde temizlenmesi ve eksik verilerin tamamlanması gerekmektedir. Bu çalışmada çevresel tesislere ait ham verilerin ön işlemesinde veri madenciliği programlarından açık kaynak kodlu *RapidMiner Studio* programı kullanılmıştır. Programın tercih edilmesinin nedenlerinden biride modelleme çalışmalarının bir sonraki adımı olan modellin uygulanması aşamasında kullanılabilecek birçok tahmin modelini içeren operatörlere (k-NN, Naive Bayes, Karar Ağacı, Yapay Sinir Ağları, Regresyon, Destek Vektör Makinaları vb) sahip olmasıdır.

*RapidMiner Studio ile v*erilerin ön işlemesi aşamasında; pH, Eİ, AKM, KOİ, $BOİ_5$, Yağ-Gres ve TÇK parametrelerine ait sırasıyla 25, 141, 26, 22, 241, 645, ve 688 adet eksik veri tespit edilmiştir. Eksik veriler ortalama değerler göz önüne alınarak program aracılığı ile tamamlanmıştır. Sonraki aşamada 10 gürültülü veri tespit edilmiş ve satır bazlı temizleme yapılmıştır. Sezonsal ortalama değerlerin belirlenmesi için BOI_5 parametresi kullanılmış ve sezonluk ortalama değerler program aracılığı ile hesaplanmıştır.

Bu çalışma ile çevresel bir tesiste ölçülen parametrelerden elde edilen ham verilerin ön işlemesinde *RapidMiner Studio* programının kullanılabileceği görülmüştür. İlgili tesise ait verilerin modellenmesinde uygun modelin belirlenmesi, modelin uygulanması ve yorum/ değerlendirme aşamaları bir sonraki çalışma konusu olacaktır.

Referanslar

[1] Çelik H., Yurtay N., Sertkaya C., Wastewater Effluent Prediction Based on Decision Tree. Digital Proceeding Of THE ICOEST'2013 - , Cappadocia C.Ozdemir, S. Şahinkaya, E. Kalıpcı, M.K. Oden (editors), p. 138-148 Nevsehir, Turkey, June 18 – 21, 2013.
[2] Oke I.A., Lukman S. Amoko J.S., Fehintola E.O. An evaluation of solutions to moment method of biochemical oxygen demand kinetics. Nigerian Journal of Technolojy, 2018:37:1-12. Print ISSN: 0331-8443, Electronic ISSN: 2467-8821 web: http://dx.doi.org/10.4314/njt.v37i1.1
[3] Anonim 2019. Erişim Tarihi: 30.10.2019. https://www.oracle.com/tr/artificial-intelligence/what-is-artificial-intelligence.html
[4] Kotu V., Deshpande B. Predictive Analytics and Data Mining Concepts and Practice with RapidMiner. Introduction, Massachusetts: Elsevier Inc:2015, p.4 ISBN:978-0-12-801460-8
[5] Kim, C., Son, H., Kim, C., "Automated construction progress measurement using a 4D building information model and 3D data", Automation in Construction (2013)
[6] Kriegel, H.,P., Kröger, P., Sander, J., Zimek, A., "Density-based clustering", John Wi ley & Sons, Inc ., Volume 1, pp.231-240. (2011)
[7] MacQueen, J.,M., "Some methods for classification and analysis of multivariate observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297. (1967)
[8] Durap A., Doğan Y., İnşaat Mühendisliğinde Bilişim Kavramı ve Veri Madenciliği Algoritmaları ile Bir Uzman Sisteminin Oluşturulması. May 2015 DOI: 10.13140/RG.2.1.1700.2403
[9] Teker A. Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları Akademik Bilişim 11 - XIII. Akademik Bilişim Konferansı Bildirileri 2 - 4 Şubat 2011 İnönü Üniversitesi, Malatya
[10] Ribeiro D., Sanfins A., Belo O., Wastewater Treatment Plant Performance Prediction with Support Vector Machines. P. Perner (Ed.): ICDM 2013, LNAI 7987, pp. 99–111, 2013.
© Springer-Verlag Berlin Heidelberg 2013.
[11] Korhonen P., Kaila J. 2015. Waste Container Weighing Data Processing to Create Reliable Information of Household Waste Generation. Waste Management 39 (2015) 15–25 https://doi.org/10.1016/j.wasman.2015.02.021
[12] Qiu Y, Li J, Huang X. Shi H. A Feasible Data-Driven Mining System to Optimize Wastewater Treatment Process Design and Operation. Water 2018, 10, 1342; doi:10.3390/w10101342
[13] Şeker S.E., Erdoğan D. 2016. Rapid Miner. Bilgisayar Kavramları Yayınları S. 2016, s. 1. ISBN: 9781536530544