Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL
UNIVERSITY
PRESS

RESEARCH ARTICLE

An Application with Python Software for the Classification of Chemical Data

Gonca ERTÜRK[1], Oğuz AKPOLAT[2]

ABSTRACT

Nowadays, much data can be generated and stored by chemical analyses. It is possible to evaluate these data, to reveal the relationships between them, and to make predictions with new data measured based on these relationships thanks to data mining algorithms. Monitoring the treatment processes and providing the necessary controls for environmental studies are based on the continuous determination of wastewater and activated sludge characteristics. The main criteria for determining the properties of wastewater are biochemical oxygen demand (BOD5), chemical oxygen demand (COD), total organic carbon (TOC), and dissolved oxygen (DO). Among these parameters, BOD5 measurement takes 5 days, while the others can be measured within 1-2 hours at most. Since BOD5 values can be mathematically correlated with other parameters, estimating them in a short time will provide a great advantage in terms of process control. In this study, a data set was created by measuring the specified parameters from 334 samples taken from a treatment plant for statistical evaluation, and the interactions of the parameters in this data set with each other were analyzed by the decision tree method. Thus, by considering the weighted effects of the parameters, it was tried to predict the probable BOD5 value of an unknown sample. The algorithm selected for this data mining study was modeled with PYTHON software and the performance of the algorithm in the estimation of the BOD5 parameter depending on other parameters was examined by extracting decision tree rules.

Keywords: Wastewater, Analysis, Data Mining, Classification, Decision Trees

DOI: 10.26650/JODA.1264915

[1]Muğla Sıtkı Koçman University, Graduate School of Natural and Applied Sciences, Chemistry Department, Mugla, Turkiye
[2]Muğla Sıtkı Koçman University, Science Faculty, Chemistry Department, Mugla, Turkiye

ORCID: G.E. 0000-0002-8821-0330;
O.A. 0000-0002-6623-4323

Corresponding author:
Oğuz AKPOLAT,
Muğla Sıtkı Koçman University, Science Faculty, Chemistry Department, Mugla, Turkiye
E-mail: oakpolat@mu.edu.tr

Submitted: 17.03.2023
Revision Requested: 13.04.2023
Last Revision Received: 01.05.2023
Accepted: 03.05.2023
Published Online: 25.07.2023

Citation: Erturk, G., & Akpolat, O. (2023). An application with python software for the classification of chemical data. *Journal of Data Applications, 1*, 49-68. https://doi.org/10.26650/JODA.1264915

Introduction

One of the areas of chemistry in which a large number of data are produced is environmental chemistry. When examined from an environmental point of view, the largest part of the pollution in wastewater consists of detergents, organic substances, and oils. The analyses used to determine the properties of wastewater is based on chemical ones, where quantitative results can be obtained, rather than biological and physical ones, where qualitative measurements can be performed. Measurements based on quantitative analysis are based on gravimetric, volumetric, or physicochemical methods. The Aeration pool is of great importance for the activated sludge process, determining the characteristics of wastewater and activated sludge refers to analyses such as acidity, temperature, conductivity, dissolved oxygen, oxygen saturation, salinity, electrical conductivity, chemical oxygen demand, suspended solids, total nitrogen, total phosphorus and biological oxygen demand to be performed in samples taken from raw wastewater and treated wastewater coming to treatment plants, as well as measurements made on the design parameters of the sludge samples used for biological treatment. These are ventilation time, suspended solids concentration (MLSS), solid retention time parameter including outlet water quality, temperature and biokinetics as design variables related to sludge production rate and oxygen requirement factors; final precipitation pool surface hydraulic load and solid load, sludge volume index, solid retention time and temperature as design variables related to sludge production rate factors; the MLSS concentration, which includes the sludge recycling rate and the MLSS recycling concentration, as well as the sludge volume index and the final settling pool, are the recycling rate as design variables related to the factors of surface hydraulic load and solid load. The properties of wastewater are classified as physical, chemical, and biological as follows: (Toprak, 2018; Tchobanoglous and Burton, 1991; Eltem, 2001; Wikipedia, 2016; amazon, 2016).

1. Physical Properties of Wastewater: It consists of total solids, odor, temperature, and color.

2. Chemical Properties of Wastewater: Organic substances in wastewater are mostly composed of benzene derivatives, such as proteins, carbohydrates, fats and oils, urea, soap, detergents, and volatile components. Biochemical oxygen demand (BOD_5) is a measure of the amount of dissolved oxygen used by microorganisms for the biochemical oxidation of organic compounds and the most widely used. It is performed by chemical oxygen requirement (COD) test of wastewater to measure the organic matter content. The COD value of wastewater is often higher than the BOD value. In particular, the total organic carbon test (TOC) is applied to measure the total organic carbon content of wastewater at low concentrations. Acidity is important in determining the quality of the inorganic content of wastewater, and pH is most commonly used for this. The others are chloride, alkalinity, nitrogen, phosphorus, sulfur, heavy metals, and toxic compounds and gases.

3. Biological Properties of Wastewater: The apparent group of organisms in domestic wastewater are plants, animals, and microorganisms such as bacteria, algae, fungi, protozoa, and viruses. Coliform bacteria are an indicator of contamination from human waste. Algae also cause taste and smell problems. During the treatment of wastewater, they decompose organic substances with bacteria.

4. Determination of the Properties of Wastewater: Determination of Biochemical Oxygen Demand (BOD_5), Chemical Oxygen Demand (COD), Total Organic Carbon (TOC), and Dissolved Oxygen (DO) amounts are the most basic measurement criteria for the characterization of wastewater.

As in the rest of the world, all wastewater treatment plants in Türkiye are operated in accordance with the Environment Law and the Water Pollution Control Regulation implemented by the Ministry of Environment and Urbanization. In domestic biological wastewater treatment plants, domestic wastewater from households is treated and restored to nature, and it is also aims to protect the water mass in the basin. Monitoring the treatment processes and providing the necessary controls is only possible by continuously measuring the characteristics of wastewater and activated sludge. Analyses to be made with samples taken from raw wastewater or treated wastewater coming to treatment plants are acidity (pH), temperature (T), conductivity (C), dissolved oxygen (DO), oxygen saturation (SO), salinity (SA), electrical conductivity (EC), chemical oxygen demand (COD), suspended solids (LSS), total nitrogen (TN), total phosphorus (TP) and biological oxygen demand (BOD_5), and similarly analyzes made for activated sludge samples can be listed as aeration time (AT), sludge suspended solids concentration (MLSS), suspended volatile solids concentration (MLVSS), temperature (T), sludge production rate (ASPR) and retention time (RT) including bio-kinetics (BK) and the recycling rate (FBR). The determination of biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total organic carbon (TOC) and dissolved oxygen (DO) amounts are the most basic measurement criteria for the characterization of wastewater in determining its properties.

As 11 of the wastewater parameters counted in one of the studies in recent years can be measured in a one-day study conducted in the laboratory, it is stated that the measurement of the BOD_5 parameter takes a minimum of 5 days. In a laboratory study done for samples taken from a treatment facility for statistical evaluation, a dataset was created by measuring 12 parameters from 334 samples. Over the BOD_5 parameter, the effects of the other parameters in this data set were examined using the decision tree method by the KNIME data mining package. Thus, it was attempted to estimate the possible BOD_5 value of a sample whose result is unknown by considering the weighted effects of parameters whose effects on the BOD_5 parameter are known. This shows us that environmental measurement data can be re-evaluated by data mining methods. From this and similar studies, it is understood that statistical evaluations regarding the measured values and estimates can be made between the parameters from the studies conducted for the activated sludge quality (Güller et al. 2019; Born, 2017; Weka, 2019; Synder and Wyant, 2018; Mukhtarov, 2020).

Data mining is the acquisition of valid and applicable information from data stacks by a dynamic process. In these processes, many different techniques are used, such as classification, clustering, data summarization, learning classification rules, finding dependency networks, variability analysis and abnormal detection. In data mining, classification and curve fitting are defined as prediction methods, while methods such as clustering and association analysis are described as descriptive. Data mining techniques are divided into two different categories such as supervised learning and unsupervised learning. The difference between supervised learning and unsupervised learning is unsupervised learning learns from the data but without reference. Therefore, it is not necessary to create a prior model in unsupervised learning. As classification is an supervised learning technique, clustering is one of the unsupervised. It separates data into some groups called clusters in which objects are similar to each other. The classification method which is one of the main methods of data mining is based on a learning algorithm. It is applied in order to discover hidden patterns in large-scale data. The main classification methods are decision trees, Bayesian classification, artificial neural networks and support vector machines. Classification is done by quickly examining the attributes of a new object and assigning that object to a predefined class. The important thing here is that the characteristics of each class are determined in advance. Clustering is the grouping of data according to their proximity or distance to each other, and there are no predetermined group boundaries, but it can be optimized by giving the number of groups. Data mining software is divided into two groups as commercial and open source and, data mining algorithms are operated directly in some software without coding, or they can be modeled in software that can be coded, such as Python. Python is a widely used, high-level, general-purpose, interpreted, and dynamic programming language. The design philosophy emphasizes the readability of the code, and the syntax allows programmers to express concepts in fewer lines of code than is possible in languages such as C++ or Java. (Silahtaroğlu, 2016; Alan and Karabatak, 2020; Çelik, 2009; Çınar, 2019; Kacur, 2020; Sampaio and Landup, 2022; Robinson, 2022).

During the recovery of wastewater, some tests are performed for the quality of activated sludge by characterizing the wastewater and the acquired water, and only the treatment process can be controlled by these tests. It is clear from the studies conducted that the data stacks obtained by all the physical, chemical and biological analyses performed during these processes can only be examined in detail by data mining techniques that are related to each other. Based on this, estimates of measurement parameters can also be made. Data mining algorithms make it possible to identify the relationships between these data by evaluating them and making predictions with the help of new data measured based on these relationships. In this study, a suitable algorithm for Decision Trees will be selected from the data analysis methods to be applied in the study of wastewater characteristics with this data set, modeled by coding with Python software, and the performance of the algorithm in estimating the BOD_5 parameter depending on other parameters will be examined by extracting decision rules. Pandas, NumPy, Pyplot, and scikit-learn packages will be used as the basis for programming with Python (Nelson, 2022; activestate, 2022; Li, 2017; Sampaio and Landup, 2022; Robinson, 2022; anaconda, 2022).

Material and Method

Data mining models can be grouped under four main headings: prediction, clustering, connection analysis, and difference deviations. Predicting and clustering investigate each record's relationship to others, while objective and temporal connections can be examined in connection analysis. The most well-known classification techniques used for prediction are decision trees, statistical-based algorithms such as Bayesian and Regression, distance-based algorithms, and artificial neural networks. Of these, the classification can be mathematically defined as:

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

Let's have a database and let each t_i show a record.

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

Let m denote the set of classes consisting of classes.

$f\mathbf{:}D\rightarrow C$ and each t_i should belong to a class. Here C_j is a separate class, and each class contains its own records. So, it can be shown in the form:

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

Classification can also carry a class (discrete) and continuous value of the dependent variable with the class we have or its statistical definition. In this respect, it approaches regression or multi-term regression. Classification can also be defined as a supervised learning approach in which hidden patterns within a certain range are revealed. The most common of these algorithms are ID3 and C4.5 (Silahtaroğlu, 2016). Normalization is one of these algorithms' most frequently used data transformation processes. With Min-Max normalization (Eq. 1), which is the most used of data normalization techniques, the original data are converted to the new data in range by a linear transformation. This data range is usually 0-1.

$$Newdata=\{(Rawdata\text{-minRawdata})/(maxRawdata\text{-minRawdata})\} \qquad \textit{Eq. 1}$$

The principles of the decision tree method and the steps of the decision tree algorithm are given below:

Basics of Decision Tree Method

1. Identification of the problem.
2. Drawing/structuring the decision tree.
3. Assigning the probabilities of the occurrence of events.
4. Calculation of the expected return (or benefit) for the corresponding chance point-backward, the transaction.
5. Assignment of the highest expected return (benefit) to the relevant decision point-backward, comparison.
6. The submission of the proposal is based on its principles.

The Steps of The Decision Tree Algorithm

1. The learning set T is created.
2. The attribute that best separates the samples in the set T is determined.
3. A node of the tree is created with the selected attribute, and child nodes or leaves of the tree are created from this node. Determine the instances of the subset of child nodes.
4. for each sub-dataset created in step three:

- If the samples all belong to the same class
- If there is no qualification to divide the samples
- If there is no sample with the remaining attribute value, the process is terminated. In the other case, the process is continued from the second step to separate the subset of data.

The decision tree can be easily encoded in any programming language using IF-ELSE expressions. Decision tree classification is a classification method that creates a model in the form of a tree structure consisting of decision nodes and leaf nodes according to the property and goal. The decision tree algorithm is improved by dividing the data set into small and even smaller pieces. A decision node may contain one or more branches. The first node is called the root node. A decision tree can consist of both categorical and numerical data. The randomness, uncertainty, and probability of an unexpected situation occurring in the formation of any situation are defined by entropy, and if all the samples are regular/homogeneous, their entropy becomes zero. Here entropy is defined as in Eq. 2:

$$E(S) = \sum_{i \quad 1}^{c} -p_i log_2 p_i$$ *Eq. 2*

Entropy is not calculated only on the target. In addition, entropy can also be calculated on properties. But when calculating entropy on properties, it is taken into account in the target. In this case, entropy is defined as in Eq. 3:

$$E(T,X) = \sum_{ceX} P(c)E(c)$$ *Eq. 3*

Information gain (Gain) is based on subtracting all entropy after dividing a dataset on a feature (Eq. 4). If the entropy is small, the importance of the feature increases for the Decision Tree algorithm ID3. On the other hand, as it gets closer to 1 the importance of the feature decreases. However, in information gain, the situation is the opposite, and in this respect, it can be thought of as the inverse of entropy. While constructing the Decision Tree, the feature with the highest information gain is selected.

$$Gain(T,X)=Entropy(T)-Entropy(T,X)$$ *Eq. 4*

Overfitting is an important problem for decision tree models and many other prediction models. Overfitting occurs when the training set continues to reduce errors in a way that affects the learning algorithm. To avoid overfitting in a decision tree construction, two approaches are usually used:

- Pre-pruning: Stopping the growth of the tree before the containment process.
- Post-pruning: first creating the whole tree and then removing the unnecessary parts from the tree.

Due to the difficulty in determining when pruning will be done in practice, the first approach is hardly used. The second approach is much more successful. Attention should be paid to the following steps in this approach.

- A different dataset than the training data is used to decide on the pruning process. This data set is called the validation dataset. The validation dataset is used to decide on unnecessary nodes.
- After obtaining a decision tree, using statistical methods such as error estimation and significance testing (Chi-Square Testing), it is decided whether there will be pruning

and expansion (expanding – adding new nodes to the tree) on the training data.

- The Minimum Distance Description Principle is a measure between the Decision tree and the training dataset. When the size (tree) + Size (non-classifiable tree) is minimized, the tree growth is stopped.

The data tested in the decision tree modeling performed using the Python programming language in this section were measured for 334 days in a wastewater treatment plant and presented in Table 1 (Güller et al., 2019**)**. The measured values were chemically defined as acidity (pH: -), Temperature (TemperC: °C), Total Phosphate (TotalphosmgPL: mg/L), Suspended Solids (CVLmgPL: mg/L), Chemical Oxygen Demand (CODmgPL: mg/L), and Biological Oxygen Demand (BOD_5mgPL: mg/L).

Table 1. *Analytical values of chemical substances measured in wastewater samples.*

Label	Ph	TemperC	TotPhosmgPL	CVLmgPL	CODmgPL	BOD_5mgPL
1	7.3	8.7	17.7	310	920	19.41
2	7.55	9.7	15.9	150	495	169
3	7.47	10.3	11.6	180	401	209
4	8.03	9.7	5.2	130	433	272
Experimental data file in dimension of 334*7 (# **Data: EnvirodataI.txt)**						
331	8	17	1.95	154	474	120
332	7.77	17.2	0.55	8.4	142.65	36.4
333	7.76	30	0.31	42	162.12	45.6
334	7.41	24.4	3.43	33	153.5	38.4

The Python codes required for the analysis of the algorithm written for decision trees in data evaluation (Environment_Classification_00_Decission_Tree_Performance.py) and its numerical output are given as **Appendix 1**. The graphs of the program outputs are given too in Figure 1, Figure 2, Figure 3, and Figure 4 as “Distribution of samples according to BOD_5 values”, “Display of sample distributions in box graphics”, “Histograms of distributions related to chemical measurement values”, “Correlation of chemical measurement values to each other” and “Decision tree of samples classified according to BOD_5 values”, respectively.

Results

In this section, the outputs and results of the decision tree application selected for solving classification problems have also been examined, and the numerical and % distribution of BOD_5 samples are given in Table 2.

Table 2. *BOD5 percent distributions of all data in the specified intervals.*

BOD_5 Intervals	Numeric value	Distribution (%)
0-250	279	72
251-500	70	18
501-750	30	8
751-999	5	2
0-999	384	100

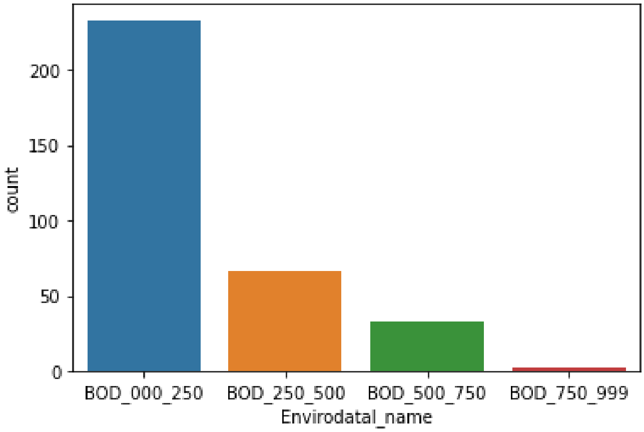

Figure 1. *Distribution of samples according to BOD_5 values.*

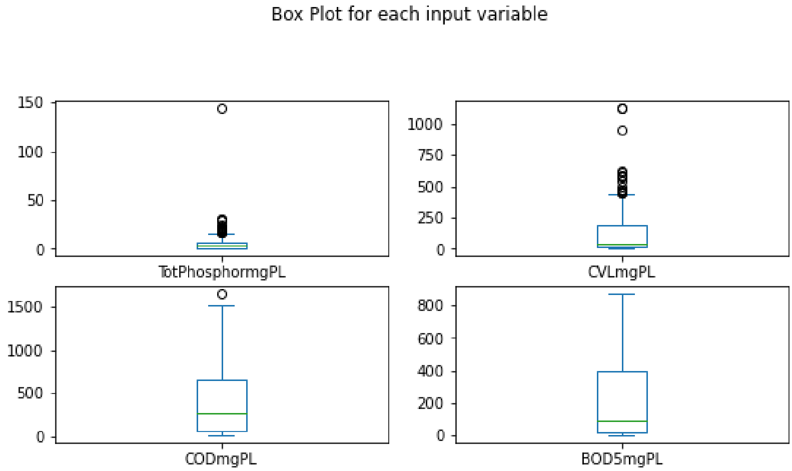

Figure 2. *Showing the sample distributions in box graphics.*

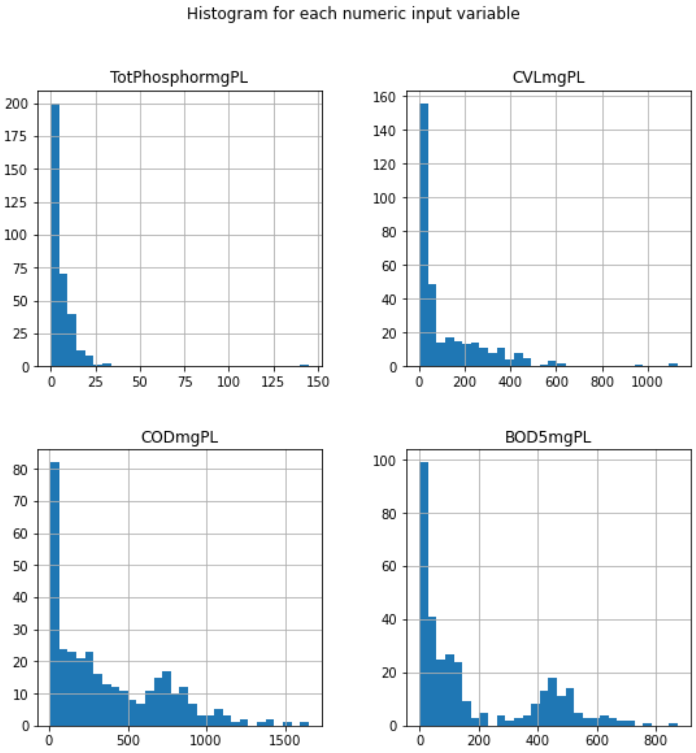

Figure 3. *Histograms of distributions related to chemical measurement values.*

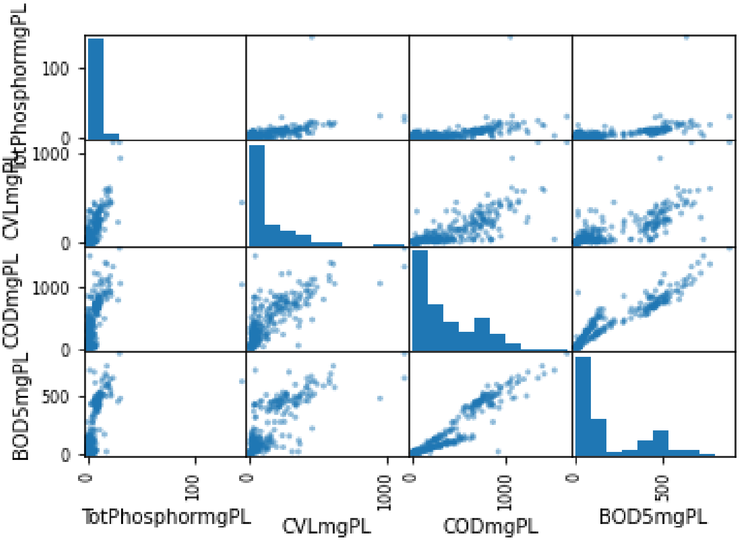

Figure 4. *Correlations of chemical measurement values with each other.*

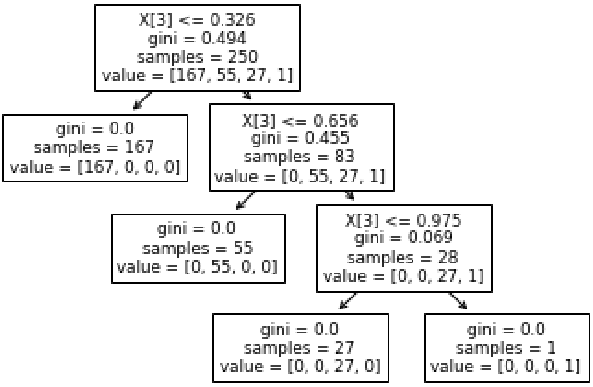

Figure 5. *Decision tree of samples classified according to BOD_5 values.*

The distribution of the Confusion matrix related to the test data taken from the decision tree is given in Table 3, and the accuracy value of this distribution is calculated as 100%.

Table 3. *The Confusion matrix for the test data, which accounts for 20% of the total data.*

	Bin1	**Bin2**	**Bin3**	**Bin4**
Bin1	66			
Bin2		11		
Bin3			6	
Bin4				1

Conclusion

Monitoring the treatment processes and providing the necessary controls is only possible by continuously measuring the characteristics of wastewater and activated sludge. Analyses to be made with samples taken from raw wastewater or treated wastewater coming to treatment plants are acidity (pH), temperature (T), conductivity (C), dissolved oxygen (DO), oxygen saturation (SO), salinity (SA), electrical conductivity (EC), chemical oxygen demand (COD), suspended solids (LSS), total nitrogen (TN), total phosphorus (TP) and biological oxygen demand (BOD_5), and similarly analyses made for activated sludge samples can be listed as aeration time (AT), sludge suspended solids concentration (MLSS), suspended volatile solids concentration (MLVSS), temperature (T), sludge production rate (ASPR) and retention time (RT) including bio-kinetics (BK) and recycling rate (FBR). Determination of biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total organic carbon (TOC) and dissolved oxygen (DO) amounts are the most basic measurement criteria for the characterization of wastewater in determining its properties.

As 11 of the wastewater parameters counted in one of the studies in recent years can be measured in a one-day study conducted in the laboratory, in a study conducted by Güller et al. (2019), it is stated that the measurement of the BOD_5 parameter takes a minimum of 5 days. For this work, in a laboratory study done for samples taken from a treatment facility for statistical evaluation, a dataset was created by measuring 12 parameters from 334 samples. Over the BOD_5 parameter, the effects of the other parameters in this data set were examined using the decision tree method by the KNIME data mining package. Thus, it was attempted to estimate the possible BOD_5 value of a sample whose result is unknown by considering the weighted effects of parameters whose effects on the BOD_5 parameter are known. This shows us that environmental measurement data can be re-evaluated by data mining. These studies in this area are quite new and show promise in the evaluation of environmental data stacks.

In this study, the four selected parameters considered to be much more effective from the data set given above, were chemically defined as acidity (pH), Temperature (°C), Total Phosphate (mg/L), Suspended Solids (mg/L), Chemical Oxygen Demand (mg/L), and Biological Oxygen Demand (mg/L). For this study a suitable algorithm was selected as a decision tree from the data analysis methods to be applied and modeled by coding with Python software, and the performance of the algorithm in estimating the BOD_5 parameter depending on other parameters was examined by extracting decision rules. Pandas, NumPy, Pyplot, and scikit-learn packages were used as the basis for programming with Python.

When the results obtained from the data set showing the analysis results of 4 parameters related to 334 domestic qualified wastewaters taken from an earlier study and evaluated by the decision tree method encoded by the Python algorithm were examined in this study, it was found that the BOD_5 value distribution of 334 samples was below 250 by 72%. The proportion of those with a BOD_5 value between 251-500 is 18%, while the proportion of those between 500-750 is 6%. As Figure 4 is examined, it is understood that the variable that affects the BOD_5 value the most is COD (Chemical Oxygen Demand). As compared to those of the study published by Güller et al. (2019) performed by KNIME software it is understood that the results of this study is very close to theirs, as expected.

Ethics Committee Approval: Authors declared that this study does not require ethics committee approval.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- O.A., G.E.; Data Acquisition- O.A.; Data Analysis/ Interpretation- O.A., G.E.; Drafting Manuscript- O.A., G.E.; Critical Revision of Manuscript- O.A.; Final Approval and Accountability- O.A., G.E.; Material and Technical Support- O.A.; Supervision- O.A.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Activestate. (2022). *How to Classify Data In Python using Scikit-learn.* Retrieved May 3, 2023, from https://www.activestate.com/resources/quick-reads/how-to-classify-data-in-python/

Alan, A., & Karabatak, B. (2020). *Veri Seti - Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi*, *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 32(2), 531-540.

Amazon, (2016). Retrieved May 3, 2023, from https://www.amazon.com/Hach-8505700-Measurement-Luminescent-Dissolved/dp/B00R3EGHJ4

Anaconda. (2022). *anaconda/packages/python.* https://anaconda.org/anaconda/python/anaconda/packages/ (python3.10.6)

Çelik, M. (2009). *Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama* [Yüksek Lisans Tezi]. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Ana Bilim Dalı.

Çınar, A. (2019). Veri Madenciliğinde Sınıflandırma Algoritmalarının Performans Değerlendirmesi ve R Dili ile Bir Uygulama, *Marmara Üniversitesi Öneri Dergisi*, 14(51), 90-111.

Doğan, O. (2017). Ücretsiz Veri Madenciliği Araçları ve Türkiyede Bilinirlikleri Üzerine Bir Araştırma, *Ege Stratejik Araştırmalar Dergisi*, 8(1), 77-93.

Eltem, R. (2001). *Atık Sular ve Arıtım*, Ege Üniversitesi Fen Fakültesi Yayınları, 172

Güller, S., Silahtaroğlu, G. ve Akpolat, O. (2019). Analysis waste water characteristics via data mining: A Muğla province case and external validation. *Communications in Statistics Case Studies Data Analysis and Applications*, 5(3), 200-213. https://dx.doi.org/10.1080/23737484.2019.1604192

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

Kacur, T., M. (2020). *Atık Su ve Aktif Çamur Karakteristiklerinin Tahmininde Karar Ağaçları ve Yapay Sinir Ağlarının Karşılaştırılması* [Yüksek Lisans Tezi]. Muğla Sıtkı Koçman Üniversitesi Çevre Bilimleri Ana Bilim Dalı.

Li, S. (2017). *Solving A Simple Classification Problem with PYTHON — Fruits Lovers' Edition*. Retrieved May 3, 2023, from https://towardsdatascience.com/solving-a-simple-classification-problem-with-python-fruits-lovers-edition-d20ab6b071d2

Meyers, D.N., & Wilde, F. D. (2012). *USGS TWRI Book 9–A7* (Third Edition), http://water.usgs.gov/owq/FieldManual/Chapter7/NFMChap7.pdf

Mukhtarov, M. (2020). *Atık Su ve Aktif Çamur Karakteristiklerinin Sınıflandırılması ve Uygulanan Analiz Yöntemlerinin Değerlendirilmesi* [Yüksek Lisans Tezi]. Muğla Sıtkı Koçman Üniversitesi Çevre Bilimleri Ana Bilim Dalı.

Nelson, D. (2022). *Overview of Classification Methods in PYTHON with Scikit-Learn*. Retrieved May 3, 2023, from https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/

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.

Robinson, S. (2022). *Decision Trees in PYTHON with Scikit-Learn.* Retrieved May 3, 2023, from https://stackabuse.com/decision-trees-in-python-with-scikit-learn/

Sampaio, C., & Landup, D. (2022). *Linear Regression in PYTHON with Scikit-Learn*. Retrieved May 3, 2023, from https://stackabuse.com/linear-regression-in-python-with-scikit-learn/

Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları*, (2. Baskı). Papatya Yayıncılık.

Synder, R., & Wyant, D. (2018). *Activated Sludge Process Control Training Manuel, DEO, Water Resources Division*. Retrieved May 3, 2023, from https://www.michigan.gov/documents/deq

Tchobanogluos, G., & Burton, F. L. (1991). *Wastewater Engineering Treatment, Disposal, and Reuse*, McGraw-Hill Book Co.

Toprak, H. (2018). *Aktif Çamur Sürecinin Tanımı*. Retrieved May 3, 2023, from http://web.deu.edu.tr/atiksu/ana58/aktifkurs.doc

Weka. (2019). *Weka*. Retrieved May 3, 2023, from https://www.cs.waikato.ac.nz/ml/weka/index.html

Wikipedia, (2016). *Biochemical oxygen demand*. Retrieved May 3, 2023, from https://en.wikipedia.org/wiki/Biochemical_oxygen_demand

Appendix 1. Python Commands for Classification and Performance Evaluation

```
print(‘# Environment_Classification_00_Decission_Tree_Performance’)
# Data: EnvirodataI.txt
attribures = [“EnvirodataI_label”, “EnvirodataI_name”, “pH”, “TemperC”,
“TotPhosphormgPL”, “CVLmgPL”, “CODmgPL”, “BOD5mgPL”]
target_attribute = [“BOD5mgPL”]
# (‘Downloading Required Libraries ‘)
import pandas as pd
import numpy as np

envirodata = pd.read_table(‘envirodataI.txt’)
print(envirodata.head())
print(envirodata.shape)
print(envirodata[‘EnvirodataI_name’].unique())
print(envirodata.groupby(‘EnvirodataI_name’).size())
import seaborn as sns

sns.countplot(envirodata[‘EnvirodataI_name’], label=”Count”)
# Distribution Measures
import matplotlib.pyplot as plt

envirodata.drop(‘EnvirodataI_label’, axis=1).plot(kind=’box’, subplots=True,
layout=(4, 2), sharex=False, sharey=False, figsize=(9, 9), title=’Box Plot
for each input variable’)
plt.savefig(‘envirodata_box’)
plt.show()

import pylab as pl

envirodata.drop(‘EnvirodataI_label’, axis=1).hist(bins=30, figsize=(9, 9))
pl.suptitle(“Histogram for each numeric input variable”)
plt.savefig(‘envirodata_hist’)
plt.show()

import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix

# Selection of three numerical features
attribute = [“TotPhosphormgPL”, “CVLmgPL”, “CODmgPL”, “BOD5mgPL”]
# Plot the scatter matrix
# Depending on the features
scatter_matrix(envirodata[attribute])
plt.show()
# Statistical Study
X = envirodata[attribute]
y = envirodata[‘EnvirodataI_label’]
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
print(“X_train”, X_train, “X_test”, X_test, “y_train”, y_train, “y_test”,
y_test)
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
print(“X_train”, X_train, “y_train”, y_train)
# Models
# Decission Tree
from sklearn.tree import DecisionTreeClassifier

clf = DecisionTreeClassifier().fit(X_train, y_train)
print(‘Accuracy of Decision Tree classifier on training set: {:.2f}’
      .format(clf.score(X_train, y_train)))
print(‘Accuracy of Decision Tree classifier on test set: {:.2f}’
      .format(clf.score(X_test, y_test)))
# Confussion Matrix for Decission Tree
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

pred = clf.predict(X_test)
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))
# Creating a Decision Tree
print(‘# Preparation of Decision Tree ‘)
X = [[0, 0], [1, 1]]
Y = [0, 1]
clf = clf.fit(X, Y)
clf.predict([[2., 2.]])
print(‘# Transforming the Data Set)’)
X, y = X_train, y_train
# print(“X”,X,”y”,y)
from sklearn import tree

print(‘# Creating a Decision Tree ‘)
clf.fit(X, y)
tree.plot_tree(clf)
print(“#Performance for Decision Tree “)
CMAT = [[66, 0, 0, 0],
        [0, 11, 0, 0],
        [0, 0, 6, 0],
        [0, 0, 0, 1]]
print(‘CMAT’, CMAT)
print(‘Interpretation of the Confussion Matrix’)
print(CMAT)
print(‘Lines Prediction(T) Columns Real(G)’)
print(‘High:”H”, Low:”L”, Middle:”M”, VeryLow:”V”’)
print(‘TOTAL:”TOT”’)
CMAT00_TH_GH = CMAT[0][0]
CMAT01_TH_GL = CMAT[0][1]
CMAT02_TH_GM = CMAT[0][2]
CMAT03_TH_GV = CMAT[0][3]
CMAT10_TL_GH = CMAT[1][0]
CMAT11_TV_GL = CMAT[1][1]
CMAT12_TL_GM = CMAT[1][2]
CMAT13_TL_GV = CMAT[1][3]
CMAT20_TM_GH = CMAT[2][0]
CMAT21_TM_GL = CMAT[2][1]
CMAT22_TM_GM = CMAT[2][2]
CMAT23_TM_GV = CMAT[2][3]
CMAT30_TV_GH = CMAT[3][0]
CMAT31_TV_GL = CMAT[3][1]
CMAT32_TV_GM = CMAT[3][2]
CMAT33_TV_GV = CMAT[3][3]
print(CMAT00_TH_GH, CMAT01_TH_GL, CMAT02_TH_GM, CMAT03_TH_GV)
print(CMAT10_TL_GH, CMAT11_TV_GL, CMAT12_TL_GM, CMAT13_TL_GV)
print(CMAT20_TM_GH, CMAT21_TM_GL, CMAT22_TM_GM, CMAT23_TM_GV)
print(CMAT30_TV_GH, CMAT31_TV_GL, CMAT32_TV_GM, CMAT22_TM_GM)
#
TOT_TH_G = CMAT00_TH_GH + CMAT01_TH_GL + CMAT02_TH_GM + CMAT03_TH_GV
TOT_TL_G = CMAT10_TL_GH + CMAT11_TV_GL + CMAT12_TL_GM + CMAT13_TL_GV
TOT_TM_G = CMAT20_TM_GH + CMAT21_TM_GL + CMAT22_TM_GM + CMAT23_TM_GV
TOT_TV_G = CMAT30_TV_GH + CMAT31_TV_GL + CMAT32_TV_GM + CMAT33_TV_GV
TOT_T = TOT_TH_G + TOT_TL_G + TOT_TM_G + TOT_TV_G
N = TOT_T  #
TOT_GH_T = CMAT00_TH_GH + CMAT10_TL_GH + CMAT20_TM_GH + CMAT30_TV_GH
TOT_GL_T = CMAT01_TH_GL + CMAT11_TV_GL + CMAT21_TM_GL + CMAT31_TV_GL
TOT_GM_T = CMAT02_TH_GM + CMAT12_TL_GM + CMAT22_TM_GM + CMAT32_TV_GM
TOT_GV_T = CMAT03_TH_GV + CMAT13_TL_GV + CMAT23_TM_GV + CMAT33_TV_GV
TOT_G = TOT_GH_T + TOT_GL_T + TOT_GM_T + TOT_GV_T
#
Total_Accurate_Forecast = CMAT00_TH_GH + CMAT11_TV_GL + CMAT22_TM_GM +
CMAT33_TV_GV
# Accuracy= Total_Accurate_Forecast /N
Accuracy = Total_Accurate_Forecast / N
print(“N=”, N, “Accuracy=”, Accuracy)
```

Appendix 1. Output:

Environment_Classification_00_Decission_Tree_Performance.py the outputs of the named program are given below:

```
EnvirodataI_label EnvirodataI_name ... CODmgPL BOD5mgPL
0 1 BOD_000_250 ... 449.50 107.6
1 1 BOD_000_250 ... 393.38 100.0
2 1 BOD_000_250 ... 371.90 108.0
3 1 BOD_000_250 ... 560.41 155.0
4 1 BOD_000_250 ... 350.00 165.0
Selected: [5 rows x 7 columns] Total:(334, 7)
[‘BOD_000_250’ ‘BOD_250_500’ ‘BOD_500_750’ ‘BOD_750_999’]
EnvirodataI_name
BOD_000_250 233
BOD_250_500 66
BOD_500_750 33
BOD_750_999 2

X_train TotPhosphormgPL CVLmgPL CODmgPL BOD5mgPL
278 5.100 245.00 575.00 341.0
92 0.800 66.00 227.71 59.6
312 11.900 413.00 807.00 519.0
234 4.785 76.20 950.00 455.0
216 2.940 15.85 230.00 122.8
.. ... ... ... ...
323 144.700 453.00 1051.00 629.0
192 1.110 15.00 45.00 10.0
117 0.790 5.00 63.00 12.0
47 1.420 35.60 308.29 58.0
172 1.280 17.00 51.00 16.0
Training set (%80): [250 rows x 4 columns]

X_test TotPhosphormgPL CVLmgPL CODmgPL BOD5mgPL
166 0.830 13.00 46.00 12.00
78 0.790 12.00 284.81 70.80
15 5.795 17.95 402.00 195.20
221 1.140 27.00 26.00 19.00
194 2.260 98.00 260.00 61.20
.. ... ... ... ...
171 0.320 13.60 43.61 8.92
8 2.145 81.10 420.00 195.12
223 2.975 29.80 274.00 125.20
236 6.950 38.60 604.00 315.00
156 1.090 7.00 32.00 10.00
Test set (%20): [84 rows x 4 columns]
```

```
for Training set y_train
278 2
92 1
312 3
234 2
216 1
..
323 3
192 1
117 1
47 1
172 1
```

```
for Test set y_test
166 1
78 1
15 1
221 1
194 1
..
171 1
8 1
223 1
236 2
156 1
```

```
X_train
[[3.52453352e-02 2.57112750e-01 3.73918829e-01 4.46833054e-01]
[5.52868003e-03 6.84931507e-02 1.42854291e-01 7.79677013e-02]
[8.22391154e-02 4.34141201e-01 5.28276780e-01 6.80159396e-01]
[3.30684174e-02 7.92413066e-02 6.23419827e-01 5.96266779e-01]
[2.03178991e-02 1.56480506e-02 1.44377911e-01 1.60811661e-01]
[9.60608155e-03 1.28556375e-01 3.76067864e-01 1.80736158e-01]
[4.07740152e-02 1.47523709e-02 3.06054558e-02 1.81942114e-02]
[1.03662751e-01 3.88830348e-01 5.88157019e-01 6.77537752e-01]
......................................................................... . .
[2.14236351e-03 4.32033720e-02 9.92149035e-02 5.96161913e-02]
[3.04077402e-02 2.95047418e-02 1.26413839e-02 2.08158557e-02]
[4.90670352e-03 2.25500527e-02 8.67198935e-02 2.93361997e-02]
[7.53282654e-02 2.46575342e-01 5.58882236e-01 6.35591443e-01]
[4.90670352e-02 1.95995785e-01 4.73719228e-01 6.13307466e-01]
[7.87836904e-02 3.85669125e-01 5.54890220e-01 6.44767198e-01]
[5.25915688e-02 9.35721812e-02 1.00465735e-01 9.81543624e-02]
[6.91085003e-03 2.55005269e-02 1.76533599e-01 9.46151426e-02]
[4.56116102e-03 5.75342466e-02 1.66573520e-01 9.02894295e-02]
[7.46371804e-02 2.23393045e-01 5.30272788e-01 6.04131711e-01]
[1.00000000e+00 4.76290832e-01 6.90618762e-01 8.24349832e-01]
[7.67104354e-03 1.47523709e-02 2.12907518e-02 1.29509228e-02]
[5.45957153e-03 4.21496312e-03 3.32667997e-02 1.55725671e-02]
[9.81340705e-03 3.64594310e-02 1.96467066e-01 7.58703859e-02]
[8.84588804e-03 1.68598525e-02 2.52827678e-02 2.08158557e-02]]

y_train
278 2
92 1
312 3
234 2
216 1
..
323 3
192 1
117 1
47 1
172 1

Accuracy of Decision Tree classifier on training set: 1.00
Accuracy of Decision Tree classifier on test set: 1.00
[[66 0 0 0]
[ 0 11 0 0]
[ 0 0 6 0]
[ 0 0 0 1]]
precision recall f1-score support
1 1.00 1.00 1.00 66
2 1.00 1.00 1.00 11
3 1.00 1.00 1.00 6
4 1.00 1.00 1.00 1
accuracy 1.00 84
macro avg 1.00 1.00 1.00 84
weighted avg 1.00 1.00 1.00 84
CMAT [[66, 0, 0, 0], [0, 11, 0, 0], [0, 0, 6, 0], [0, 0, 0, 1]]
Interpretation of the ‘Confussion Matrix print(CMAT)
Rows Estimate (T) Columns Real (G)
High:”H”, Low:”L”, Middle:”M”, VeryLow:”V”
Rows Estimate (T) Columns Real (G)
TOTAL:”TOT”
66 0 0 0
0 11 0 0
0 0 6 0
0 0 0 6
N= 84 Accuracy= 1.0
```