JOURNAL OF APPLIED ENGINEERING SCIENCES
ISSN: 2247-3769 / e-ISSN: 2284-7197

sciendo

VOL. 14(27), ISSUE 2/2024
ART.NO. 420 pp. 260-267

ESTIMATION OF MANGANESE CONTENT IN POTABLE WATER BY BOOSTING TECHNIQUES

M. Göçer [a,] *, S. B. Coşkun [a], B. Yanık [b]

[a] Kocaeli University, Department of Civil Engineering, Kocaeli, 41001, Turkey, E-mails: * mrtgocer@gmail.com, sb.coskun@kocaeli.edu.tr
[b] Water and Sewerage Administration, Samsun Metropolitan Municipality, Samsun, Turkey, E-mail: bahattinyanik@saski.gov.tr

Received: 29.06.2024 / Accepted: 24.07.2024 / Revised: 15.08.2024 / Available online: 15.12.2024

DOI: 10.2478/jaes-2024-0032

KEY WORDS: machine learning, boosting methods, regression, classification, manganese, potable water.

ABSTRACT:

In this study, boosting machine learning techniques were employed to estimate the concentration of manganese in the potable water of Yuvacık Dam and to predict the days that surpassed and fell below the threshold value of 0.05 mg/L in Turkey. We conducted both regression and classification analyses for the same issue. We also implemented sampling methods when the data distribution in the classification task became imbalanced. We obtained daily measurements for approximately seven years, from 2004 to 2011, to build the dataset, which consisted of seven columns in total. While the XGBoost algorithm forecasted the manganese content in potable water with a mean absolute error of 0.0055, the Light GBM algorithm predicted the days with elevated manganese levels with an accuracy of 0.97. The models' high predictions allow us to adjust the frequency of frequent water sampling and lab analysis and take prompt action during water filtration processes.

1. INTRODUCTION

The quality of natural water has a significant influence on both the environment and the well-being of living organisms. Water is crucial for the survival of living organisms, as it is required directly and indirectly for physiological functions. It is essential that the water used for physiological and metabolic processes be hygienic. While certain minerals and metals in water can be beneficial to human health, those that have detrimental effects on living organisms are also important to consider (Klaassen, 2001). Sometimes, one can find high or trace amounts of these metals and minerals, including elements like iron (Fe), manganese (Mn), copper (Cu), cobalt (Co), and zinc (Zn). Small amounts of some of these metals are necessary for the physiological activities of living organisms (Canpolat and Çalta, 2001).

Potable water commonly contains manganese, a mineral that naturally enters water sources. It can alter the taste and appearance of water. While it is important for human health in small amounts, excessive intake can have harmful effects on the human body (Talaat et al., 2010). Although manganese is abundant in the Earth's crust, water resources typically do not prefer its use (Dönderici et al., 2010).

Manganese infiltrates groundwater over an extended period due to the dissolution of manganese oxides, silicates, and acrybonates in rocks and soils, and it often co-occurs with ferrous water (Kohl and Medlar, 2006). Excessive iron concentrations can have detrimental effects on wildlife and infrastructure, as well as water pipes (Sly et al., 1990).

Conversely, surface water and groundwater frequently contain iron. Elevated levels of iron in the water supply are considered a problem. Iron can alter water's taste and appearance, cause damage to water pipes, and act as a nutrient source for microbial growth, leading to increased bacterial proliferation (Theis and Singer, 1974).

Research on individuals with high levels of Mn and Fe exposure has demonstrated neurotoxic effects and Parkinson-like symptoms (Mergler et al., 1999; Dobson et al., 2004; Holzgraefe et al., 1986; Perl and Olanow, 2007). Research on fetal development during pregnancy has shown that exposure to these metals can result in a decrease in fetal weight and hinder the development of internal organs (Coles et al., 2012).

In general, the concentration of manganese in natural waters can be less than 0.2 mg/l, while it can be up to 10 mg/l in groundwater (Dönderici et al., 2010). According to the World Health Organization (WHO) in 2004, the threshold value for manganese in potable water is 400 µg/L, and the threshold for iron is 0.3 mg/L (WHO, 2002). WHO recommends that all countries establish their own standards and regulations (WHO, 2011). For instance, the US Environmental Protection Agency (US EPA) has set the threshold for manganese in potable water at 500 µg/L (Du, 2004). In Turkey, the recommended value for potable water standards is 20 µg/L, and the maximum allowable value is 50 µg/L (Marmara University, 2022).

* Corresponding author: Graduate student Murat Göçer, e-mail: mrtgocer@gmail.com

Numerous methods and strategies have been devised for assessing water quality. In recent times, the deployment of artificial intelligence (AI) has gained traction due to its seamless integration into human life (Aldhyani et al., 2020). This has reduced the duration of sampling, laboratory testing, and decision-making, while enhancing the accuracy of outcomes.

The objective of this research was to employ machine learning techniques to forecast the concentration of manganese in Yuvacık Dam, where daily water levels were assessed from 2004 to 2011. By making instantaneous predictions, decisions can be reached regarding when water filtration is necessary, and the need for continuous sampling and laboratory testing can be reduced.

Researchers have conducted several investigations to explore the potential of machine learning and deep learning techniques in the assessment of water quality. A study by Aldhyani et al. (2020) used a deep learning algorithm called autoregressive neural network (NARNET) and a short-term memory (LSTM) to predict water quality. They also used support vector machine (SVM), k-nearest neighbor (K-NN), and Naive Bayes algorithms for machine learning. The goals were to classify water quality (WQC) and predict water quality index (WQI). The researchers reported that the NARNET algorithm demonstrated superior performance for WQI prediction using deep learning models. Moreover, the study achieved 97.01% accuracy for WQC classification with the SVM algorithm. The selection of algorithms and models for the study was based only on their ability to handle the research objectives, without any extraneous influences or considerations.

Khan et al. (2022) utilized principal component analysis (PCA) to identify the most effective features. They subsequently applied various regression and classification techniques and achieved the highest regression R2 result (0.95 with principal component regression (PCR)) and the highest classification result with 100% accuracy using the gradient boosting classifier algorithm.

Wang et al. (2017) conducted a study to examine the correlation coefficients between various features and apply different regression methods. The results indicated that the support vector regressor (SVR) yielded the highest R2 value (0.92).

Yilma, et al. (2018) conducted a study aimed at evaluating the cleanliness index of the Little Akaki River in Ethiopia. The study employed a neural network architecture and utilized an algorithm with 12 inputs. The researchers trained the algorithm with various configurations, including different numbers of neurons (5, 10, 15, 20, and 25) in each layer and different numbers of hidden layers (2–20). The authors achieved the best results with eight hidden layers and 15 neurons and obtained an R2 explainability of 0.93.

Samsudin et al. (2019) used spatial discriminant analysis (SDA) to reduce the number of datasets from 13 to seven parameters. Subsequently, the authors applied artificial neural network (ANN) models to these seven input parameters and achieved the highest R-squared score of 0.90.

Bui (2020) utilized a combination of 12 hybrid models, including random forest (RF), M5P, random tree (RT), and reduced error pruning tree (REPT) techniques. Out of all the models, the bagging random forest hybrid model achieved the highest R2 score of 0.94. Liu et al. (2020) introduced a deep bidirectional stacked simple recurrent unit (Bi-S-SRU) method for unit learning. The study demonstrated that this approach is more precise and efficient than traditional RNN and LSTM methods and attained a prediction accuracy of 94.42%.

In their study, Liao and Sun (2010) investigated Lake Chao in China. The authors found that a decision tree enhanced with the C4.5 parameter outperformed a neural network. The mean absolute error was 0.12, and the mean squared error was 0.24.

In this work, the manganese concentration in the potable water from Yuvacık Dam and the days when it will rise and decrease above Turkey's threshold limit of 0.05 mg/L are predicted using boosting machine learning techniques. For the same problem, classification and regression techniques were used. Sampling techniques were used when the data distribution in the classification problem became unbalanced.

2. DATA SET AND CHARACTERISTICS

The Yuvacik Dam, as seen in Figure 1, is situated in the Marmara region of Turkey, spanning the area between the cities of Kocaeli, Sakarya, and Bursa. The Kirazdere, Kazandere, and Serindere Rivers supply the 258 square kilometer reservoir. The reservoir boasts a useful volume of 51.1 million cubic meters and an annual capacity of 142 million cubic meters of purified water (İzmit Su A.Ş., 2014).

Figure 1. Location of Yuvacik Dam

The dataset under consideration consists of 2881 rows and seven columns, containing daily measurements of dam water collected from January 1, 2004, to November 20, 2011. The dataset records data on water turbidity, manganese content, iron content, the useful volume of the dam, and environmental factors like air temperature and precipitation over approximately seven years, in addition to the measurement date. Table 1 provides the names and descriptions of the columns.

VARIABLE	EXPLANATION
Date	Date of measurement (DD/MM/YYYY)
Turbidity	Turbidity of water at the measurement date (NTU)
Manganese	Manganese content at the measurement date (mg/L)
Temperature	Air temperature at the measurement date (C degree)
Reservoir Volume	Reservoir volume at the measurement date (M m^3)
Iron	Iron content at the measurement date (mg/L)
Precipitation	Precipitation at the measurement date (mm)

Table 1. Yuvacik Dam data set descriptions

The dataset under consideration did not contain any null values, and thus, imputation was not necessary. The average manganese content was 0.0332 mg/L, while the average iron content was 0.0666 mg/L. The maximum and minimum manganese values were 0.683 and 0.001 mg/L, respectively, and for iron, they were 1.55 mg/L and 0.0 mg/L. As shown in Figure 2, the values' distributions were generally concentrated between 0 and 0.2 and were right-skewed.

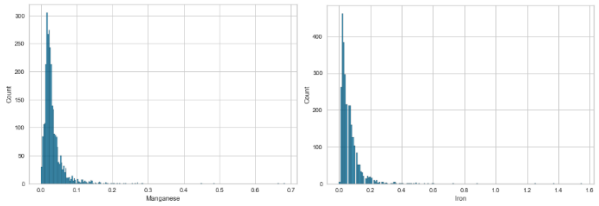

Figure 2. Distribution of manganese and iron content

The graph in Figure 3 demonstrates that the measured manganese values have a seasonal pattern with respect to the measurement dates. While certain days recorded extremely high values, the overall manganese levels were relatively low in the middle of the year, with a tendency to rise at the start and end of the year.

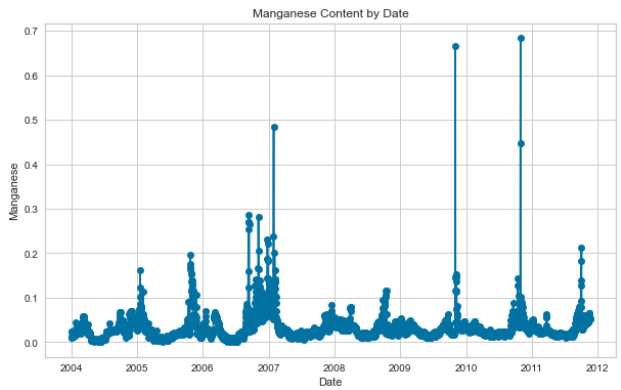

Figure 3. Relationship between time and manganese values

As shown in Figure 4, the date-iron relationship is similar to the date-manga relationship. In particular, iron values were lowest in the middle of the year and highest at the beginning and end.

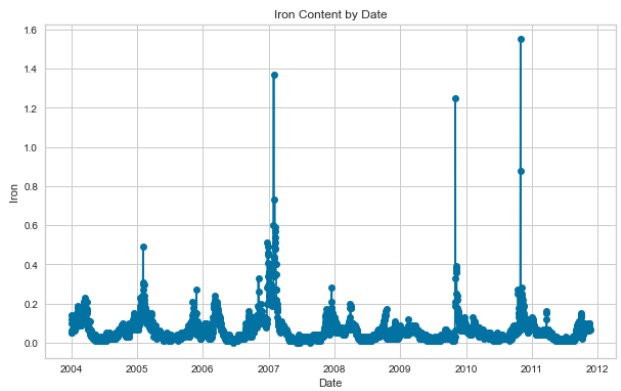

Figure 4. Relationship between time and iron values

After examining the striking resemblance between manganese and iron, as shown in Figure 5, the correlation between the two variables proved to be highly positive and significant.

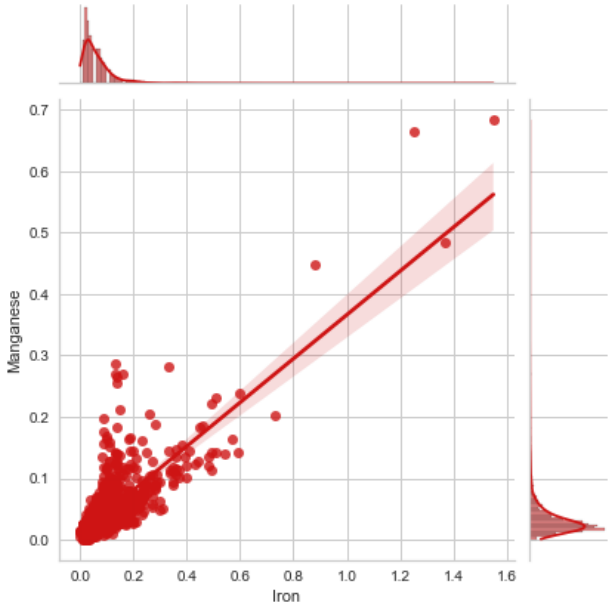

Figure 5. Manganese and iron relationship

The following section presents the methods and techniques utilized to address the problem.

3. METHODOLOGY

All dataset variables were quantitative, no categorical or discrete variables. The date variable, formatted as datetime64, was the only exception. The remaining variables were all in float64 format. Consequently, no categorical variable engineering techniques, such as transformation or one-hot encoding, were employed.

The selection of manganese as the dependent variable transformed the problem into a regression issue, necessitating the application of regression models to forecast the water's Mn value.

The aforementioned passage established the threshold value for manganese for each country. This study presents a novel property that is based on this threshold value. Turkey has designated a maximum acceptable manganese level of 0.05 mg/L as the threshold value. We formulated a new column by assigning a value of 1 to days with measurements below this threshold and 0 to days with measurements above it. We then employed classification algorithms to predict the values of the newly created column.

In Figure 6, the data for the new feature, Threshold_005, reveals that 2477 days recorded values below 0.05 mg/L, while 404 days recorded values above 0.05 mg/L.

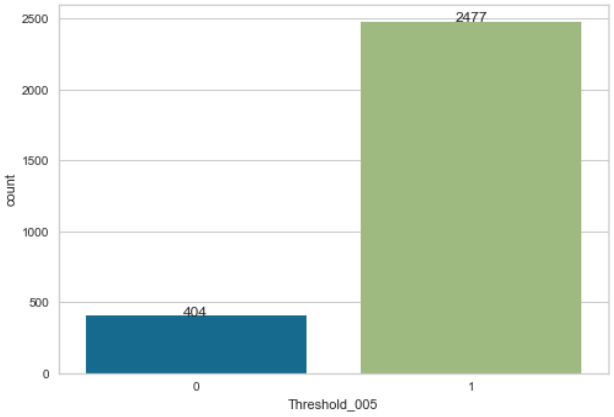

Figure 6. Distribution of the Threshold_005 feature

The data in Figure 6 shows a significant imbalance in the newly formed column, which leads the algorithms to learn the majority class more often. To overcome this issue, several approaches such as undersampling, oversampling, and SMOTE balancing have been employed. Additionally, sophisticated boosting techniques, including Light GBM, Catboost, Gradient Boosting Machines, and XGBoost have been used as efficient algorithms for regression and classification.

3.1 Boosting Methods

The foundations of the boosting method were established by Schapire (1990) in (Kearns, 1988) and Kearns and Valiant (1989), who positively answered the question "Is it possible to obtain a strong learner from a weak learner?". Freund and Schapire (1995) developed the initial adaptive AdaBoost algorithm, which forms the foundation of contemporary algorithms.

In this research, Gradient Boosting Machines (GBM), extreme gradient boosting (XGBoost), Light GBM, and Categorical Boosting (CatBoost) algorithms were utilized to address both regression and classification challenges.

3.1.1 Gradient boosting machines: Gradient boosting machines (GBMs) are one of the earliest modern boosting algorithms. These machines function iteratively, aiming to formulate a robust ensemble of learners by consecutively training single decision trees that are initially weak learners (Friedman, 2001). The first tree, a weak learner, creates a new column of errors during its initial training. The second tree, also a weak learner, utilizes this column of errors in its training. The second and all subsequent trees' objective is to reduce the values in this error column. As a result, a community of weak learners transforms into a strong community of learners.

3.1.2 Light GBM: Ke et al. (2017) developed the Light Gradient Boosting Machine (GBM) algorithm from the original GBM algorithm. Light GBM has gained recognition for its low memory requirements and rapid processing capabilities, differentiating itself from other algorithms through a slight alteration in its operational logic. Unlike other boosting algorithms that progress level by level, Light GBM advances leaf by leaf. In other words, while other boosting algorithms persist in allocating leaves without regard to the information contained in prior leaves, Light GBM exclusively operates on leaves that retain data and consume memory to assign them. This approach allows for more efficient operation and maintains a high success rate (Weng, 2019).

3.1.3 eXtreme Gradient Boost: The XGBoost algorithm, like LightGBM, was developed as an improved version of the GBM algorithm (Chen and Guestrin, 2016). Widely regarded as one of the most powerful algorithms in the boosting family, it operates iteratively, constructing a strong learner from a group of weak learners. This process culminates in the completion of the model learning process, which focuses on the errors of interconnected decision trees (Mitchell and Frank, 2017).

3.1.4 Categorical Boost: Dorohush et al. (2017) developed the CatBoost algorithm, which benefits from its utilization of GPU resources during resource consumption, resulting in a faster execution time. Furthermore, this algorithm does not necessitate the transformation of categorical data, enabling it to process such data directly, thus providing a considerable advantage to users during the data preprocessing step.

3.2 Data Balancing Methods

Imbalanced datasets are a common challenge in machine learning applications. When machine learning algorithms endeavor to learn from data, a preponderance of examples in a particular class prompts the algorithms to prioritize that class. As a result, the algorithms fail to fully understand the relationships within the minor class. Moreover, imbalanced datasets tend to skew success metrics when evaluating algorithms. Given the excessive emphasis on the major class, success metrics are unduly optimistic for that class, thereby masking the prediction of failure in the minor class. Thus, it is essential to choose the appropriate success metric for imbalanced datasets and to strive for data balance as much as possible. A variety of techniques exist for balancing data, such as undersampling, oversampling, and synthetic minority oversampling (SMOTE).

3.2.1 Undersampling: Undersampling involves deliberately reducing the number of observations in the majority class to match the number of observations in the minority class, thereby achieving class balance (Weiss and Provost, 2001; Laurikkala, 2001).

3.2.2 Oversampling: The main difference between undersampling and oversampling is that oversampling copies the minor class's number of observations at random and multiplies it by the major class's number of observations until the dataset is balanced (Nitesh et al., 2002).

3.2.3 Smote: Oversampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), are designed to increase the number of observations in a minor class. Unlike traditional oversampling methods, SMOTE generates new synthetic observations rather than duplicating existing ones. This process involves creating new data points between the two closest observations (Shelke et al., 2017).

We use sampling techniques to balance datasets that vary at different rates. This research employed methods that achieved data balance at a 50%–50% ratio. This equalized the number of observations for the major and minor classes in the dataset.

We divided the entire dataset into training and test sets at a ratio of 80% to 20%. Sampling techniques were only applied to the training set, and there was no synthetic or duplicate data from the test set present in the training dataset to prevent data leakage.
Hyperparameter optimization was carried out using the grid search method on the training set. The grid search method is a widely used optimization technique that attempts all given hyperparameter combinations and determines the optimal hyperparameters through trial and error (Liashchynskyi and Liashchynskyi, 2019).

A five-fold cross-validation approach was implemented to prevent overfitting during model training. We fixed this approach for all algorithms at a value of k=5.

3.3 Evaluation Metrics

This research implemented both regression and classification methods, and consequently employed both solutions' success metrics.

The evaluation of regression models utilized the root mean squared error (RMSE) and mean absolute error (MAE) as metrics, as given in Eqs. (1-2)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \quad (2)$$

where n is the number of observations, y_i is the actual value and $\hat{y_i}$ is the predicted value.

To assess the classification metrics, it is essential to initially determine the confusion matrix. Figure 7 illustrates the confusion matrix.

		Predicted Values: 1	Predicted Values: 0
Actual Values	1	True Positive	False Negative
	0	False Positive	True Negative

Figure 7. Confusion Matrix

According to the confusion matrix, classification metrics, including Precision, Recall, F1 Score, and Accuracy, were computed as follows.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

$$F1\ Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (5)$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \quad (6)$$

4. RESULTS

Table 2 presents the outcomes of the regression algorithms. Based on these outcomes, the XGBoost algorithm emerged as the most effective in terms of both RMSE and MAE.

Model	RMSE	MAE
Light GBM	0,0172	0,0073
XGBoost	0,0102	0,0055
CatBoost	0,0208	0,0135
Gradient Boost	0,0119	0,0066

Table 2. Performance of the regression algorithms

Table 3 presents a comparison of the classification algorithms' performances. The F1 Score values represent the number of cases above and below the threshold, respectively. For the base dataset, the Light GBM algorithm demonstrates the highest performance. However, the base dataset's unbalanced nature may make its accuracy metric misleading. It would be better to stabilize the dataset and reevaluate the metric.

In this case, the oversampling technique resulted in the highest performance when compared to other techniques.

Analyzing the F1 score revealed that the base dataset's performance for days with measurements above the manganese threshold value was 0.87. Applying the oversampling technique increased this performance to 0.89.

		Models			
Datasets	Metrics	Light GBM	XG Boost	Cat Boost	Gradient Boost
Base	Accuracy	0.97	0.95	0.95	0.95
	F1 Score	0.87-0.98	0.83-0.97	0.81-0.97	0.79-0.97
Undersample	Accuracy	0.91	0.90	0.90	0.92
	F1 Score	0.73-0.94	0.73-0.94	0.74-0.94	0.77-0.95
Oversample	Accuracy	0.97	0.96	0.96	0.96
	F1 Score	0.89-0.98	0.87-0.98	0.85-0.98	0.86-0.98
SMOTE	Accuracy	0.96	0.95	0.95	0.95

	F1 Score	0.85-0.97	0.84-0.97	0.83-0.97	0.85-0.97

Table 3. Performance of the classification algorithms

As depicted in Figure 8, the key characteristics that significantly impact the independent variable threshold value comprise Turbidity, Dam Useful Volume, and Temperature columns. These variables are of paramount importance in determining the concentration of manganese in water.

Therefore, by analyzing these three characteristics, we can reliably predict the level of manganese in the water, regardless of its low or high quantity.

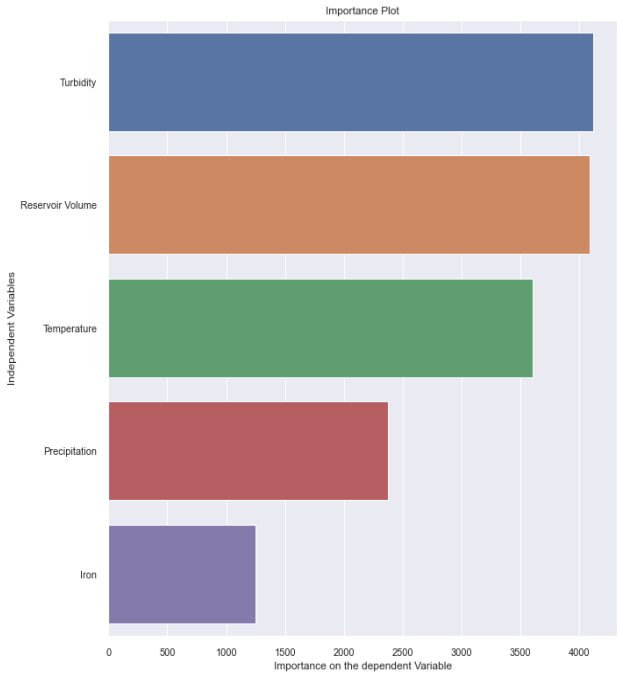

Figure 8. Order of importance of independent variables

5. CONCLUSION

In this study, incremental machine learning techniques are used to predict the manganese concentration in the potable water from Yuvacık Dam and to predict the days exceeding and falling below the threshold value of 0.05 mg/L in Turkey. Both regression and classification methods were employed for the same issue. When the data distribution in the classification task became imbalanced, sampling methods were implemented.

Current research estimates the daily manganese levels in water with a margin of error of 0.0055 mg/L. This low margin of error enhances the efficiency and effectiveness of laboratory processes and facilitates prompt decision-making in filtration systems. In the filtration process, the predictions made to determine the days when the manganese levels are above or below the threshold value are 97% accurate. This high level of accuracy provides a substantial advantage in water quality control by enabling more precise and rapid management of filtration systems.

Three key variables, namely Turbidity, Dam Useful Volume, and Temperature, can be easily measured and integrated into machine learning models on a daily basis.

By incorporating these data points, the model can predict the manganese level of water with precision, and the continuous updating of the data flow allows it to improve its performance and make more accurate predictions. This approach can play a critical role in the development of water management strategies and the more efficient use of water resources. The integration of these variables can increase the likelihood of proactively addressing potential problems in water treatment plants, as well as improve the overall quality of the water supply. This integrated approach is an essential step in preventing potential environmental and health risks by ensuring continuous monitoring of water quality.

References

Aldhyani, T. H., Al-Yaari, M., Alkahtani, H., and Maashi, M. 2020. Water quality prediction using artificial intelligence algorithms. *Applied Bionics and Biomechanics*, https://doi.org/10.1155/2020/6659314.

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., and Kazakis, N., 2020. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, 721, 137612, https://doi.org/10.1016/j.scitotenv.2020.137612.

Canpolat, Ö., and Çalta M., 2001. Keban Baraj Gölü'nden (Elazığ) yakalan Acanthobrama Marmid (Heckel, 1843)'de bazı ağır metal düzeylerinin belirlenmesi (Determination of some heavy metal levels in Acanthobrama Marmid (Heckel, 1843) caught from Keban Dam Lake (Elazığ)). *Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi* 13(2), pp. 263-268.

Chawla, N. V., Bowyer, K. W., O. Hall, L. O., and Kegelmeyer, W. P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, pp. 321-357.

Chawla, N. V., 2010. Data mining for imbalanced datasets: an overview. *Data mining and knowledge discovery handbook*, pp. 875-886.

Coles, C., Crawford, J., McClure, P. R., Roney, N., and Todd, G. D., 2012. "Toxicological profile for manganese". Georgia, USA https://www.atsdr.cdc.gov/toxprofiles/tp151.pdf (view at 12 May 2024).

Dobson, A. W., Erikson K.M., Aschner M., 2004. Manganese neurotoxicity. *Ann NY Acad Sci*, 1012, pp. 115–128.

Dorogush, A. V., Ershov, V., and Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Dönderici, Z. S., Dönderici A., and Başarı F., 2010. An investigation on physical and chemical quality of spring waters. *Türk Hijyen ve Deneysel Biyoloji Dergisi* 67(4), pp. 167-172.

Du, J., 2004. *Potable Water Health Advisory For Manganese*. US Environmental Protection Agency, Washington, pp. 36.

Freund, Y., and Schapire, R. E., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting.

In: *European conference on computational learning theory*. Springer, Berlin Heidelberg, pp. 23-37.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189-1232.

Holzgraefe, M., Poser, W., Kijewski, H., and Beuche, W., 1986. Chronic enteral poisoning caused by potassium permanganate: a case report. *Journal of Toxicology: Clinical Toxicology*, 24(3), pp. 235-244.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W.,Ye, Q., and Liu, T. Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, pp. 3146–3154.

Kearns, M., 1998. Thoughts on hypothesis boosting. Unpublished manuscript, 45, pp. 105.

Kearns, M., and Leslie V., 1994. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1), pp. 67-95.

Khan, M. S. I., Islam, N., Uddin, J., Islam, S., and Nasir, M. K., 2022. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University-Computer and Information Sciences*, 34(8), pp. 4773-4781.

Klaassen, C. D., 2006. Heavy metals and heavy-metal antagonists. *Goodman and Gilman's The Pharmacological Basis of Therapeutics*. 11th edition, New York: McGraw-Hill, pp. 1753-1775.

Kohl, P. M., and Steven J. M., 2006. Occurrence of manganese in potable water and manganese control. *American Water Works Association*.

Laurikkala, J., 2001. Improving identification of difficult small classes by balancing class distribution. In: *Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, Proceedings 8*. Springer, Berlin Heidelberg, pp. 63-66.

Liao, H., and Sun, W., 2010. Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method. *Procedia Environmental Sciences*, 2, pp. 970-979.

Liashchynskyi, P., and Liashchynskyi, P., 2019. Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv:1912.06059*.

Liu, J., Yu, C., Hu, Z., Zhao, Y., Bai, Y., Xie, M., and Luo, J., 2020. Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network. *IEEE Access*, 8, pp. 24784-24798.

Marmara University, Doğa Bitkileri ve Su Ürünleri Araştırma ve Uygulama Merkezi, 2022. İçme suyu kabul edilebilir değerler (Potable water acceptable values), İstanbul, Türkiye https://dobisu.marmara.edu.tr/orta-menu/yararli-bilgiler/icme-suyu-kabul-edilebilir-degerler (view at 12 May 2024).

Mergler D., Baldwin M., Belanger S., Larribe F., Beuter A., Bowler R., Panisset M., Edwards R., De Geoffroy A., Sassine M. P., Hudnell K., 1999. Manganese neurotoxicity a continuum of dysfunction: results from a com- munity based study. *Neurotoxicology*, 20(2-3), pp. 327.

Mitchell, R., and Frank, E., 2017. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science 3:e127* https://doi.org/10.7717/peerj-cs.127 (view at 12 May 2024).

Perl, D. P., and Olanow, C. W., 2007. The neuropathology of manganese-induced Parkinsonism. *Journal of Neuropathology & Experimental Neurology*, 66(8), pp. 675-682.

Samsudin, M. S., Azid, A., Khalit, S. I., Sani, M. S. A., and Lananan, F., 2019. Comparison of prediction model using spatial discriminant analysis for marine water quality index in mangrove estuarine zones. *Marine Pollution Bulletin*, 141, pp. 472-481.

Schapire, R. E., 1990. The strength of weak learnability. *Machine learning*, 5, pp. 197-227.

Shelke, M. S., Deshmukh, P. R., and Shandilya, V. K., 2017. A review on imbalanced data handling using undersampling and oversampling technique. *Int. J. Recent Trends Eng. Res*, 3(4), pp. 444-449.

Sly, L. I., Hodgkinson, M. C., and Arunpairojana, V., 1990. Deposition of manganese in a potable water distribution system. *Applied and environmental microbiology*, 56(3), pp-628-639.

Talaat, H., Montaser Y. G., Eman M. K., Enas M. A., and Awad E. M., 2010. Simultaneous removal iron and manganese from ground water by combined photo-electrochemical method. *Journal of American Science*, 6(12), pp. 1-7.

Theis, T. L., and Singer, P. C., 1974. Complexation of iron (II) by organic matter and its effect on iron (II) oxygenation. *Environmental Science & Technology*, 8(6), pp. 569-573.

Tianqi, C., and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794.

Wang, X., Zhang, F., and Ding, J., 2017. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Scientific reports*, 7(1), pp. 12858.

Weiss, G. M., and Provost, F., 2001. The effect of class distribution on classifier learning: an empirical study. *Rutgers University*, 2001.

Weng, T. Y., Liu, W. Y., and Xiao, J., 2020. Supply chain sales forecasting based on lightGBM and LSTM combination model. *Industrial Management & Data Systems*, 120(2), pp. 265-279.

World Health Organization, 2002. *Guidelines for potable-water quality*. World health organization, Geneva, pp. 303-304.

World Health Organization, 2011. *Guidelines for potable-water quality*. World health organization, Geneva, pp. 31-32.

Yilma, M., Kiflie, Z., Windsperger, A., and Gessese, N., 2018. Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa,

Ethiopia. *Modeling Earth Systems and Environment*, 4, pp. 175-187.

İzmit Su A.Ş., 2014. Yuvacık Barajı (Yuvacık Dam), İzmit, Türkiye https://www.izmitsu.com.tr/sayfa.asp?ID=15&PID=2&SID=13 (view at 12 May 2024).