Memba Su Bilimleri Dergisi

https://dergipark.org.tr/tr/pub/memba

# Prediction of Water Quality's pH value using Random Forest and LightGBM Algorithms

İbrahim BUDAK[1*]

[1] Kastamonu University, Data Analysis Monitoring and Evaluation Coordination, Kastamonu / Türkiye
*Corresponding Author: ibudak@kastamonu.edu.tr

**Abstract:** This study aims to compare Random Forest Regression and LightGBM algorithms for the prediction of pH value, which is an important parameter in water quality assessment. The performance of both algorithms is evaluated with metrics such as RMSE, R-squared and AUC (Area Under Curve). The results show that the LightGBM algorithm outperforms Random Forest (0.84) with an AUC value of 0.86 and provides better prediction accuracy, especially on large and complex datasets. These findings demonstrate the applicability of machine learning techniques in environmental monitoring processes and their potential for effective management of water quality. The results highlight the superiority of the LightGBM algorithm in solving environmental problems such as pH prediction, but also provide suggestions for more comprehensive approaches. The application of hybrid modeling techniques, generalizable analyses with datasets from different water sources, and the development of real-time monitoring systems are suggested to extend the findings of the study. This study contributes to the literature by demonstrating the importance of machine learning algorithms in environmental monitoring and water quality management.

**Keywords:** Water quality, pH value prediction, Random forest, LightGBM, Kaggle Dataset

# Random Forest ve LightGBM Algoritmaları Kullanılarak Su Kalitesinin pH Değerinin Tahmini

**Özet:** Bu çalışma, su kalitesinin değerlendirilmesinde önemli bir parametre olan pH değerinin tahmini için Random Forest Regression ve LightGBM algoritmalarını karşılaştırmayı amaçlamaktadır. Kaggle platformundan elde edilen geniş bir veri seti üzerinde gerçekleştirilen analizlerde, her iki algoritmanın performansı RMSE, R-squared ve AUC (Area Under Curve) gibi metriklerle değerlendirilmiştir. Sonuçlar, LightGBM algoritmasının AUC değeriyle (0.86), Random Forest'tan (0.84) daha yüksek performans sergilediğini ve özellikle büyük ve karmaşık veri setlerinde daha iyi bir tahmin doğruluğu sağladığını göstermiştir. Bu bulgular, makine öğrenimi tekniklerinin çevresel izleme süreçlerindeki uygulanabilirliğini ve su kalitesinin etkin bir şekilde yönetilmesindeki potansiyelini ortaya koymaktadır. Elde edilen sonuçlar, pH tahmini gibi çevresel sorunların çözümünde LightGBM algoritmasının üstünlüğünü vurgulamakla birlikte, daha kapsamlı yaklaşımlar için öneriler de sunmaktadır. Hibrit modelleme tekniklerinin uygulanması, farklı su kaynaklarından alınan veri setleriyle genelleştirilebilir analizlerin yapılması ve gerçek zamanlı izleme sistemlerinin geliştirilmesi, çalışmanın bulgularının genişletilmesi adına önerilmektedir. Bu çalışma, çevresel izleme ve su kalitesi yönetiminde makine öğrenimi algoritmalarının önemini bir kez daha ortaya koyarak literatüre katkı sağlamaktadır.

**Anahtar Kelimeler:** Su kalitesi, pH değeri tahmini, Random forest, LightGBM, Kaggle veriseti

**RESEARCH PAPER**

**Citation:** Budak, İ. (2025). Prediction of Water Quality's pH value using Random Forest and LightGBM Algorithms. Memba Water Sciences Journal. 11 (1), 42-49. DOI: 10.58626/menba.1629308

**Submission Date:** 29 January 2025, **Acceptance Date:** 24 March 2025, **Publishing Date:** 27 March 2025

## 1. Introduction

After air, water is the second most important requirement for life to exist. As a result, water quality has been extensively defined in the scientific literature. The most popular definition of water quality is “the physical, chemical and biological properties of water”. Water quality is a measure of the condition of water relative to the requirements of one or more biotic species and/or any human need or purpose (Omer, 2019).

Water quality is a term used to describe the chemical, physical and biological characteristics of water. Moreover, the physical, chemical and biological properties of a water body, i.e. “water quality”, determine the suitability of that water for a particular value. Water quality issues have become a rapidly evolving component of environmental sciences, primarily due to the increasing demand for water resources and amenity value and the complex link between water quality use and ecosystem health. Water quality varies markedly over time and space (Meybeck et al., 2006). Water quality is a critical aspect of environmental health, affecting both aquatic ecosystems and human well-being. Effective monitoring and management of water quality is essential to reduce pollution and ensure safe water supplies (Nayak and Panda, 2024).

Water quality prediction plays an important role in protecting human health, maintaining aquatic ecosystems, supporting sustainable water management practices, and ensuring regulatory compliance in aquatic environments. By analyzing a comprehensive dataset of water quality indicators such as pH, dissolved oxygen, and turbidity, the research employs a variety of ML algorithms, including Random Forest, Support Vector Machines, and Gradient Augmentation Machines. Through rigorous training, validation and optimization, models are evaluated for their accuracy, precision and error rates (Rogers and Ambili, 2024).

This study aims to compare Random Forest and LightGBM algorithms to predict the pH value, which is an important parameter in water quality prediction. Using a large dataset obtained from the Kaggle platform, the study evaluates the effectiveness of these two algorithms in terms of metrics such as accuracy, error rate and generalization performance. Although the analysis reveals the strengths of both algorithms, it shows that LightGBM performs better especially on large and complex datasets. In this context, the study aims both to contribute to the identification of suitable methods for fast and accurate water quality monitoring and to contribute to the literature by comparing the performance of different machine learning algorithms.

## 2. Materials and Methods

This section provides information about Random Forest Regression and LightGBM Regression algorithms, which are machine learning techniques used in the study. The methodological framework of the study is shown in Figure 1.

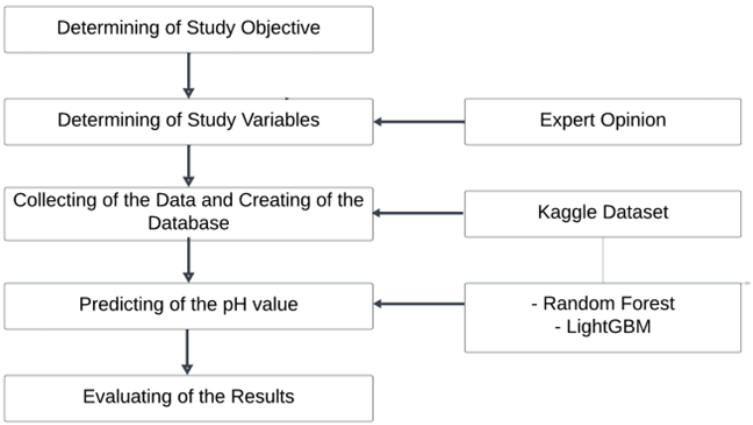


Figure 1. Methodology Framework

### 2.1. Random Forest Regression Algorithm

Random Forest Regression (RFR) is an ensemble learning method that combines multiple decision trees. This method offers high accuracy and generalization capability in regression problems by combining the

predictions of individual decision trees. Each decision tree is constructed using a randomly selected subset of data and a subset of attributes (Breiman, 2001). Each of the decision trees works independently and the predictions are combined by averaging over all trees. This method reduces the tendency of model overfitting and provides more balanced results (Ganapa et al., 2024).

An important advantage of RFR is that it combines both randomness in the data and diversity in feature selection to reduce the correlation between trees. This correlation is reduced by training each tree on only a random subset of the data (bootstrap sampling) and splitting at each node with only a randomly selected subset of attributes (Segal, 2003). This process increases the generalizability of the model both for datasets with high variance and for complex problems with different attribute types (Li et al., 2018).

Another strong feature of the Random Forest algorithm is that it can produce effective results even in data sets with a high number of variables. In particular, the ability to calculate variable importance scores to measure the information contribution of attributes distinguishes this method from other methods that are perceived as “black boxes”. For example, Ganapa et al. (2024) utilized variable importance analysis to evaluate the impact of attributes on price changes when predicting gold prices based on historical data.

Furthermore, the RFR algorithm has minimal requirements on data preprocessing and parametric assumptions. This allows the user to spend less time on model tuning and allows the algorithm to deal with complex data structures. For example, in Segal (2003), it is reported that RFR gives better results than other machine learning models even with minimal settings in experiments on various datasets.

One of the most remarkable features of this algorithm is that it provides an Out-of-Bag (OOB) error estimate, which is a general performance metric. By testing each decision tree on data not selected during bootstrap sampling, a realistic assessment of the generalization performance of the model can be made. OOB error estimation also eliminates the need for an additional test data set to measure the accuracy of the model (Li et al., 2018).

*Mathematical Model*

1. Prediction Function:

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} h_k(x) \quad (1)$$

Equation (1) calculates the prediction $\hat{y}$ by averaging the predictions of $K$ independent decision trees (Breiman, 2001).

2. Variable Selection and Randomization:
   - At each node, a random subset ($m$) of attributes is evaluated and the best split point is selected from this subset.
   - $m$ is usually $\sqrt{p}$ or $\log_2(p) + 1$, depending on the total number of features ($p$) (Segal, 2003).
3. Out-of-Bag (OOB) Error Prediction:

$$\text{OOB } Error = \frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}}(\text{y}_\text{i} - \hat{\text{y}}\text{OOB}, \text{i})^2 \quad (2)$$

In Equation (2), $N$ denotes the total number of data points, and for each data point, only the estimates generated in bootstrap
samples that do not include that point are used (Li et al., 2018).

### 2.2. LightGBM Algorithm

Light Gradient Boosting Machine (LightGBM) is a gradient boosting algorithm developed by Microsoft. This algorithm is especially optimized for large datasets and datasets with high-dimensional features (Gao and Balyan, 2022). LightGBM differs from traditional gradient boosting algorithms (GBDT) by offering faster training time, low memory usage, and high accuracy (Liang et al., 2019).

*Mathematical Model*

1. Objective Function: Since LightGBM is a gradient boosting framework, the objective function is expressed as the sum of the loss function ($L$) and an adjustment term ($\Omega$) (Yang et al., 2021):

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i\ ) + \sum_{t=1}^{T} \Omega(f_t) \quad (3)$$

Equation (3):

- $L(y_i, \hat{y}\ )$: The loss between the true value ($y_i$) and the predicted value ($\hat{y}_i$),
- $\Omega(f_t) = \gamma\text{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$: The adjustment term controls the complexity of the Tree,

- $T$: Number of leaves of the tree, $w_j$: Leaf weights (Gao and Balyan, 2022).

2. Gradient Boosting: LightGBM adds each new decision tree in a way that minimizes the prediction errors of the existing model (Liang et al., 2019):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \quad (4)$$

Equation (4):

- $\hat{y}_i^{(t)}$: Prediction at $t$-th iteration,
- $\eta$ : Learning rate,
- $f_t(x_i)$ : Prediction made by the decision tree at the $t$-th iteration (Yang et al., 2021).

3. Histogram Based Division: LightGBM selects the split points by dividing the values of the features into histogram buckets.

This reduces memory usage and computational cost (Gao and Balyan, 2022):

$$Gain = \frac{1}{n_l + n_r}\left[\frac{(G_l)^2}{H_l + \lambda} + \frac{(G_r)^2}{H_r + \lambda} - \frac{(G_l + G_r)^2}{H_l + H_r + \lambda}\right] - \gamma \quad (5)$$

Equation (5):

- $G_l + G_r$ : Gradient sums on the left and right node,
- $H_l + H_r$ : Hessian sums on the left and right node,
- $\lambda$ : L2 regulation parameter (Liang et al., 2019).

4. Leaf-Wise Growth Strategy: LightGBM uses a leaf-based growth strategy instead of the traditional level-based growth (Yang et al., 2021). This allows more complex patterns to be captured and generally provides higher accuracy:

$$BestLeaf = argmax_l Gain\ (l) \quad (6)$$

The leaf-based strategy ensures that the leaf with the highest gain is expanded at each iteration. LightGBM also supports parallel processing, significantly reducing processing time when working with large-scale datasets (Gao and Balyan, 2022).

### 2.3. Evaluation Criteria

For validation, the dataset (training and test data set) should be separated. RMSE (Root Mean Square Error Squared) was used to evaluate the algorithms used in the datasets (Karaatlı et al., 2012). The mathematical formulation of the RMSE value is given in Equation (7).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (7)$$

The mean model prediction error, expressed in units of the relevant variable, is known as the RMSE. The direction of the mistakes has no impact on this metric, which can also vary from 0 to ∞. Lower values are therefore preferable. R-squared, the second metric, shows how much of the total variance can be accounted for by the model. R-squared is shown in Equation (8).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{x}_i)^2} \quad (8)$$

where $x_i$ shows the predicted value for the $i$ th observation, $y_i$ represents the observed value for the $i$ th observation, $\bar{x}_i$ shows the average of predicted values, and $n$ represents the number of observations (Nguyen et al., 2021). Models with higher R-squared values were more successful than those with lower R-squared values.

### 2.4. Implementation

In this section, firstly, information about the data set used in the study is given and then the pH value of water quality is estimated
using Random Forest Regression and LightGBM algorithms from machine learning techniques.

### 2.5. Data Set

In this study used for water quality, the Kaggle platform, which contains various data sets in many fields, was utilized. This dataset consists of a total of 15000 data, including 10000 training and 5000 test data

published in Kaggle under the name “waterdataset”. The data set consists of pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes and turbidity indicators. The data are defined as eight independent variables and one (pH) dependent variable (Kaggle, 2024).

### 2.6. Prediction of Water Quality’s pH Value

In order for the variables in the study to be processed a certain way, the values were first normalized between 0.0 and 1.0. The normalized indicators were coded as independent and independent variables. Training (66.6%) and test sets (33.3%) of data were separated. The high-level programming language Python is used to present the algorithm outputs as RMSE and R-squared. The Random Forest Regression and LightGBM Regression methods were used in the analyses to estimate the pH value of the water quality.

| pH (observed) | Random Forest | LightGBM |
|---|---|---|
| 0.458 | 0.405 | 0.378 |
| 0.006 | 0.314 | 0.230 |
| 0.682 | 0.585 | 0.521 |
| 0.563 | 0.343 | 0.569 |
| 0.336 | 0.598 | 0.580 |
| 0.018 | 0.294 | 0.220 |
| 0.983 | 0.481 | 0.454 |
| 0.366 | 0.471 | 0.443 |
| 0.388 | 0.449 | 0.451 |
| 0.011 | 0.119 | 0.080 |

Table 1. shows 10 randomly selected normalized observed pH values and the predicted value obtained using Random Forest Regression and LightGBM algorithms.

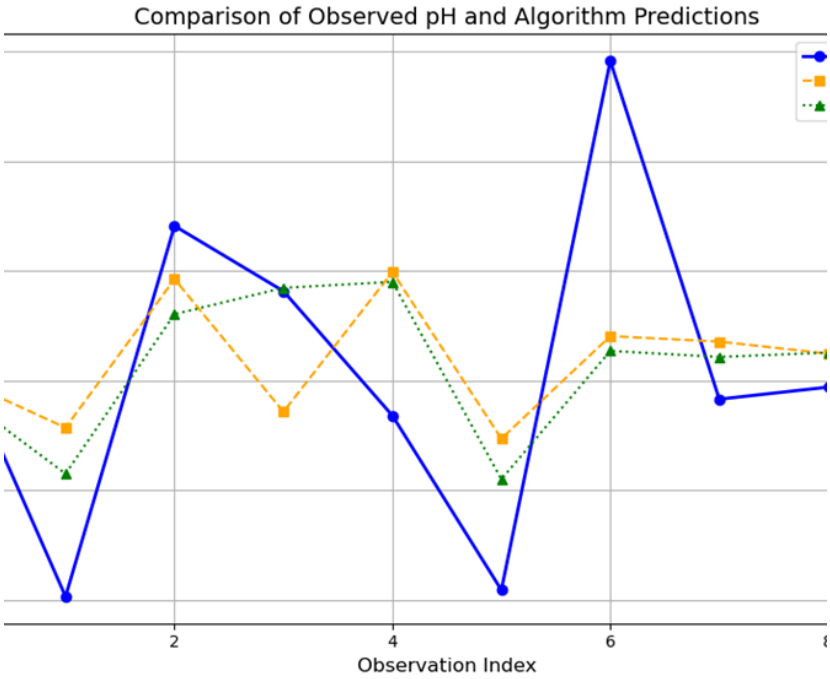


Figure 2. Comparison of algorithm results with randomly observed pH Value

The pH value observation and prediction results of the algorithms in Table 1 are visualized in Figure 1. In Figure 1, the blue line shows the observed values, while the green and yellow lines show the prediction values of the pH value. Accordingly, it shows which algorithm predicts which indicator value more successfully on randomly selected data. Table 2 shows the error value of all test data (5000).

Table 2. Comparison of Error Values of Algorithms

| | Random Forest | LightGBM |
|---|---|---|
| RMSE | 0.233 | 0.207 |
| R-squared | 0.346 | 0.481 |

In Table 2, RMSE is 0.233, and R-squared is 0.346 for the prediction using Random Forest algorithm, while RMSE is 0.207, and R-squared is 0.481 for the prediction using LightGBM algorithm. Therefore, when the performance values of the algorithms are compared, it is seen that the LightGBM algorithm has more successful results. In Figure 3, the ROC curve compares the performances of the algorithms.

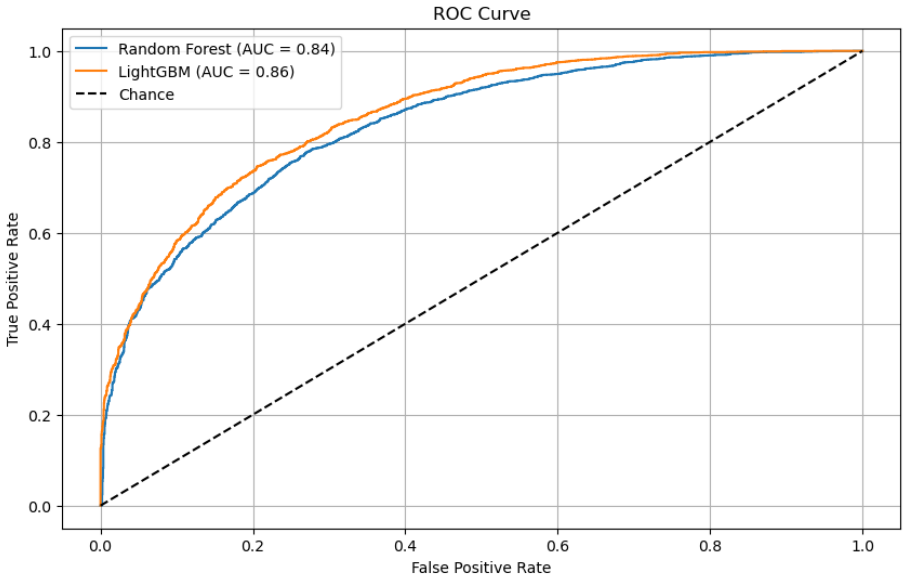


Figure 3. ROC Curve of Random Forest Regression and LightGBM Algorithms

Figure 3 shows the ROC curve of the performance of the algorithms when all data is analyzed. True Positive Rate (TPR) (Vertical Axis) refers to the rate at which the model correctly predicts true positives. These metric measures sensitivity. False Positive Rate (FPR) (Horizontal Axis) refers to the rate at which the model incorrectly predicts true negatives as positives. ROC Curve (colored lines): Visualizes the relationship between both metrics (TPR and FPR). The curve shows the performance of the algorithms at different thresholds. AUC (Area Under Curve), the area under the curve, numerically summarizes the overall performance of the algorithms. AUC expresses how much better a model is than a random guess. Chance (Dashed Line) represents the performance of a random forecast (FPR = TPR line). The AUC value of LightGBM is higher compared to Random Forest (0.86 > 0.84). This indicates that the LightGBM algorithm has better overall prediction performance and optimizes the data better.

## 3. Discussion

In this study, the analysis of different machine learning algorithms used for pH prediction is compared with the methods in the literature. In particular, the performance of Random Forest and LightGBM algorithms are compared with the models used in other studies and the differences and similarities are discussed.

Iyer et al. (2023) used SVM, Random Forest and Decision Tree algorithms to predict water quality. In this study, the superior performance of the Random Forest algorithm was particularly emphasized. Similarly,

the Random Forest algorithm was used in this study and its effect on pH estimation was analyzed. However, this study evaluated the performance of the algorithms in more detail by comparing the Random Forest method with LightGBM. In addition, in the model of this study, ROC analysis of the models was performed using metrics such as AUC, which provided a more comprehensive performance evaluation.

In the study by Stackelberg et al. (2021) pH predictions were made using the Boosted Regression Tree (BRT) method. The study specifically analyzed the impact of hydrogeological factors on model performance. In this study, such detailed hydrogeological factors were not included, but pH prediction performance was optimized with the LightGBM algorithm. In contrast to Stackelberg, the model of this study uses the Kaggle dataset and conducts analyses based on a broader data base. This difference reflects the diversity of the purpose and approach of the two studies.

Tziachris et al. (2020) compared machine learning models and hybrid geostatistical methods in pH prediction. The study showed that Gradient Boosting and Random Forest algorithms were prominent. Similarly, Random Forest and LightGBM algorithms were compared in this study. However, this study does not include geostatistical methods and only compares machine learning algorithms. In this respect, the two studies emphasize different methodological approaches.

The study by Koranga et al. (2022) evaluated eight regression and nine classification algorithms for predicting water quality in Nanital Lake. The study emphasized that Random Forest and SVM were effective. Our study, on the other hand, focused specifically on Random Forest and LightGBM algorithms. In addition, the performance advantage of the LightGBM algorithm on high-dimensional datasets is discussed in detail in this study. While Koranga's work presents a wider range of algorithms, this study focuses on an in-depth comparison of the two algorithms.

## 4. Conclusion

This study demonstrated the effectiveness of Random Forest Regression and LightGBM algorithms for the prediction of pH, an important parameter in water quality assessment. Both algorithms were evaluated on a comprehensive dataset and LightGBM outperformed Random Forest in metrics such as RMSE and R-squared. These results show that LightGBM is a powerful and effective tool for working with large-scale and high-dimensional data, offering significant potential for real-time monitoring and management of water quality. Through the use of machine learning models, this research contributes to the literature on the application of advanced computational techniques in environmental monitoring.

In addition to the findings obtained in the study, combining different modeling techniques with the use of hybrid approaches can lead to higher accuracy rates. In particular, the integration of geostatistical methods and machine learning algorithms may allow for a more comprehensive assessment of environmental factors. Including additional environmental parameters such as temperature, precipitation and water flow rate in the dataset can improve pH prediction performance and improve the generalization capability of the model. The use of longer-term datasets would be useful to measure the performance of the algorithms in assessing seasonal variations and long-term trends. The advantages of the LightGBM algorithm, such as low memory usage and fast training time, can make a significant contribution to the development of systems that monitor water quality in real time. Furthermore, comparing other state-of-the-art algorithms, such as XGBoost or deep learning-based methods, with the models used in this study may provide an opportunity to evaluate the modeling performance in a broader perspective. In order to increase the generalizability of the study, it is recommended to conduct analyses on data sets from different water sources and to develop user-friendly software tools for public institutions and environmental experts in line with the findings obtained. Such applications can contribute to developing faster and more effective solutions to water quality problems.

## 5. Compliance with Ethical Standard

### a) Authors' Contributions

Conflict of Initials of Author : Designed the study and collected and analyzed the data.

### b) Conflict of Interest

The author(s) declare that there is no conflict of interest.

## 6. References


Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Elsenety, M. M., Mohamed, M. B. I., Sultan, M. E., & Elsayed, B. A. (2022). Facile and highly precise pH-value estimation using common pH paper based on machine learning techniques and supported mobile devices. Scientific Reports, 12(22584). https://doi.org/10.1038/s41598-022-27054-5

Ganapa, J. R., Choudari, S., & Rao, M. K. (2024). Gold price prediction using random forest regression. Educational Administration: Theory and Practice, 30(1), 1052–1055. https://doi.org/10.53555/kuey.v30i1.5928

Gao, B., & Balyan, V. (2022). Construction of a financial default risk prediction model based on the LightGBM algorithm. Journal of Intelligent Systems, 31(767–779). https://doi.org/10.1515/jisys-2022-0036

Iyer, S., Kaushik, S., & Nandal, P. (2023). Water quality prediction using machine learning. Manav Rachna International Journal of Engineering and Technology, 10(1), 59-68. https://doi.org/10.58864/mrijet.2023.10.1.8

Kaggle, https://www.kaggle.com/datasets/somasreemajumder/waterdataset , (30.12.2024).

Karaatlı, M., Helvacıoğlu, Ö. C., Ömürbek, N., & Tokgöz, G. (2012). Yapay sinir ağlari yöntemi ile otomobil satiş tahmini. Uluslararası Yönetim İktisat ve İşletme Dergisi, 8(17), 87-100.

Koranga, M., et al. (2022). Machine learning algorithms for water quality prediction for Nanital Lake, Uttarakhand. International Journal of Advanced Research, 10(2), 103-114.

Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chand, J. C.-W., van den Bossche, P., Van Mierlo, J., & Omar, N. (2018). Random forest regression for online capacity estimation of lithium-ion batteries. Applied Energy, 232, 197–210. https://doi.org/10.1016/j.apenergy.2018.09.182

Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., & Chen, Z. (2019). Product marketing prediction based on XGBoost and LightGBM algorithm. AIPR 2019, ACM. https://doi.org/10.1145/3357254.3357290

Meybeck, M., Peters, N. E., & Chapman, D. V. (2006). Water quality. Encyclopedia of hydrological sciences.

Nayak, B., & Panda, P. K. (2024). A Comprehensive Review of Water Quality Analysis. International Journal of Image and Graphics, 2650033.

Nguyen, X. C., Nguyen, T. T. H., La, D. D., Kumar, G., Rene, E. R., Nguyen, D. D., ... & Nguyen, V. K. (2021). Development of machine learning-based models to forecast solid waste generation in residential areas: A case study from Vietnam. Resources, Conservation and Recycling, 167, 105381.

Omer, N. H. (2019). Water quality parameters. Water quality-science, assessments and policy, 18, 1-34.

Rogers III, O. N., & Ambili, P. S. (2024). Water Quality Predıction with Machine Learning Algorithms. EPRA International Journal of Multidisciplinary Research (IJMR), 10(4), 82-86.

Segal, M. R. (2003). Machine learning benchmarks and random forest regression. Biostatistics Division, University of California, San Francisco.

Song, J., Liu, G., Jiang, J., Zhang, P., & Liang, Y. (2021). Prediction of protein–ATP binding residues based on ensemble of deep convolutional neural networks and LightGBM algorithm. International Journal of Molecular Sciences, 22(939). https://doi.org/10.3390/ijms22020939

Stackelberg, P. E., Belitz, K., Brown, C. J., Erickson, M. L., Elliott, S. M., Kauffman, L. J., Ransom, K. M., & Reddy, J. E. (2021). Machine learning predictions of pH in the glacial aquifer system, northern USA. Groundwater, 59(3), 352-368. https://doi.org/10.1111/gwat.13063

Tziachris, P., Aschonitis, V., Chatzistathis, T., Papadopoulou, M., & Doukas, I. D. (2020). Comparing machine learning models and hybrid geostatistical methods using environmental and soil covariates for soil pH prediction. International Journal of Geo-Information, 9(4), 276. https://doi.org/10.3390/ijgi9040276.

Yang, Y., Wu, Y., Wang, P., & Xu, J. (2021). Stock price prediction based on XGBoost and LightGBM. E3S Web of Conferences, 275 (01040). https://doi.org/10.1051/e3sconf/20212750104