

# Analyzing the influence of parameters on water quality using logistic regression

*Dmitry Evsyukov*<sup>1</sup>, *Anna Glinskaya*<sup>1\*</sup>, *Anatoly Kukartsev*<sup>1,2</sup>, *Ekaterina Volneikina*<sup>1</sup>, and *Svetlana Kukartseva*<sup>3</sup>

<sup>1</sup>Bauman Moscow State Technical University (BMSTU), 105005 Moscow, Russia

<sup>2</sup>Reshetnev Siberian State University of Science and Technology, 660037 Krasnoyarsk, Russia

<sup>3</sup>Russian State Agrarian University - Timiryazev Moscow Agricultural Academy (RSAU-MAA Named after K.A. Timiryazev, 127550 Moscow, Russia

**Abstract.** This article explores the application of machine learning techniques to analyze and evaluate water quality. In particular, the article focuses on the use of logistic regression to identify and analyze key parameters affecting the potability of water. The application of logistic regression in water quality analysis not only allows us to build models for prediction, but also to formulate recommendations for improving water treatment and monitoring processes. As a result, the resulting data and models can be used to develop strategies to provide safe drinking water, which is important for the health and well-being of the community. Thus, the article proposes a modern approach to analyzing water quality using logistic regression, which allows for a deeper understanding of the relationships between water parameters and its potability, as well as the development of effective methods for water quality management.

## 1 Introduction

Water quality plays a critical role in ensuring the health and well-being of the population. Clean drinking water is essential for preventing many diseases and maintaining a high standard of living [1-4]. With accelerated urbanization, industrial development and climate change, the provision of quality drinking water is becoming increasingly challenging [5]. Therefore, an important aspect of modern science is to study and analyze the parameters affecting water quality in order to develop effective strategies for its monitoring and purification [6-9].

One of the powerful tools that can solve this problem is logistic regression, a machine learning technique that allows modeling and predicting the probability of certain events based on available data [10, 11]. Logistic regression is particularly effective for analyzing binary outcomes, making it an ideal choice for assessing the suitability of water for drinking [12].

The present paper aims to investigate the application of logistic regression to analyze key physical and chemical parameters of water such as pH, hardness, solids, chloramines,

---

\* Corresponding author: [anna\\_glinskaja@rambler.ru](mailto:anna_glinskaja@rambler.ru)

sulfates, conductivity, organic carbon, trihalomethanes and turbidity [13]. The main objectives of the study are:

- Data preprocessing and analysis.
- Construction and training of logistic regression model.
- Assessing the significance of each parameter and its impact on water quality prediction.
- Visualization of results and their interpretation.

It is expected that the results of this study will help identify the most significant parameters affecting water quality and develop recommendations for improving water treatment and monitoring systems [14-16]. This, in turn, will contribute to the improvement of drinking water quality and health of the population.

Thus, the article is an attempt to apply modern machine learning methods to solve one of the most urgent problems of our time - providing quality and safe drinking water [17].

## 2 Materials and method

Datasets containing various physical and chemical parameters of water were used for analysis. The dataset included the following parameters:

1. pH - acidity level of water.
2. Hardness - the hardness of the water.
3. Solids - solids content.
4. Chloramines - the level of chloramines.
5. Sulfate - the sulfate content.
6. Conductivity - the conductivity of the water.
7. Organic Carbon - organic carbon content.
8. Trihalomethanes - level of trihalomethanes.
9. Turbidity - turbidity of the water.
10. Potability - potability of water (binary variable where 1 - water is potable, 0 - not potable).

Before starting to work with the data, data pre-processing was carried out. All data were checked for missing values [18]. Missing values were either filled with median values of the corresponding parameters or removed from the dataset. To improve the accuracy of the model, all numerical data were normalized using StandardScaler from the sklearn.preprocessing library [19]. Data normalization is the process of bringing different scales and units to a single form [20, 21].

The dataset was divided into training and test samples in the ratio of 80:20. The train\_test\_split function from the sklearn.model\_selection library was used for splitting. Splitting the dataset into training and test samples is an important step in the process of building and evaluating machine learning models [22]. The training sample is used to train the model. The model selects parameters and tries to find dependencies in the data. Test sampling: is used to evaluate the model. Test sample contains data that has not been shown to the model during training [23-25]. It helps to understand how well the model can generalize its knowledge to new data.

Logistic regression was selected as the primary method for analyzing and predicting the potability of water. Logistic regression is a powerful statistical method that models the probability of a binary outcome based on one or more independent variables [26-28].

The following metrics were used to evaluate the accuracy of the model:

- Accuracy - the proportion of correct predictions.
- F1 Score - the harmonic mean between precision and recall.
- Confusion Matrix - A table showing the number of correct and incorrect predictions for each class.

– Classification Report - includes metrics for precision, completeness, and F1 Score for each class.

To visualize the results, feature importance graphs and error matrices were constructed using matplotlib and seaborn libraries.

### 3 Results

This section presents the results of analyzing the effects of parameters on water quality using a logistic regression model. The following metrics and visualizations allow us to evaluate the performance of the model and determine the significance of the various parameters.

The logistic regression model was trained on a training sample and tested on a test sample. The main performance metrics include accuracy and F1 Score Figure 1.

```

Accuracy: 0.5707196029776674
F1-score: 0.0
Classification Report:

```

	precision	recall	f1-score	support
0	0.57	1.00	0.73	231
1	0.00	0.00	0.00	172
accuracy			0.57	403
macro avg	0.29	0.50	0.36	403
weighted avg	0.33	0.57	0.42	403

**Fig. 1.** Basic Metrics and Classification Report.

The main metrics of model evaluation provide insight into how well the model performs on the prediction task:

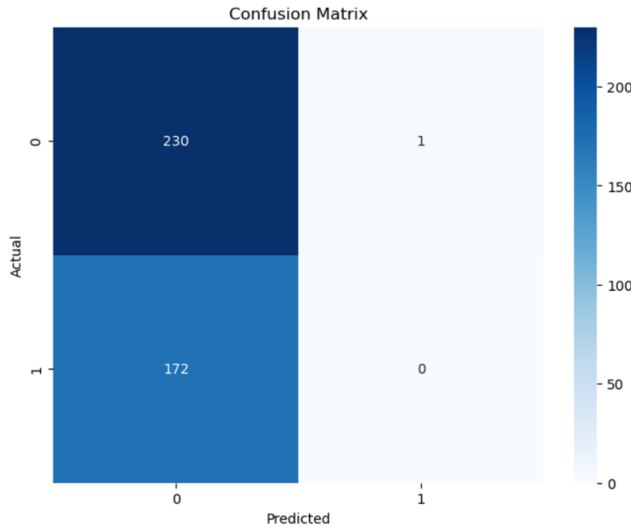
- Accuracy measures the proportion of correct predictions among all predictions. An Accuracy value of 0.5707 means that the model correctly predicted about 57.07% of all examples in the test sample.
- F1 Score is the harmonic average between Precision (Precision) and Completeness (Recall). This metric is useful when classes are unbalanced because it accounts for both false positive and false negative predictions. An F1 Score: 0.0 indicates that the model failed to correctly predict any positive cases (class 1). This can occur when either the accuracy or completeness for class 1 is zero.

It follows that the model has an average performance. A zero F1 Score indicates a serious problem with class 1 prediction. This means that the model either does not predict class 1 at all or does so infrequently that it is insignificant compared to class 0 prediction. From the classification report, we can see that Precision and Recall for class 1 are zero. This could mean that all class 1 examples were misclassified as class 0.

The Classification Report provides detailed information about Precision, Recall and F1 Score for each class:

- The model shows high Recall for class 0 (99.57%), which means that it almost always correctly identifies all class 0 cases.
- A low Precision value for class 0 (57.21%) indicates that the model makes many false positive errors (predicts class 0 when in fact it does not).
- Zero Precision, Recall, and F1-score values for class 1 indicate that the model does not predict this class at all.
- The overall accuracy of the model (57.07%) is not high, and the zero F1-score values for class 1 indicate serious problems with predicting this class.

The error matrix (Confusion Matrix) provides important information about the quality of model predictions by showing the number of correct and incorrect predictions for each class Figure 2.



**Fig. 2.** Error Matrix.

The error matrix shows:

1. Class 0:

- True Positive (TP): 230

- This is the number of correct predictions when the model correctly predicted Class 0.
- Of the 231 Class 0 examples in the test sample, the model correctly predicted 230.

- False Positive (FP): 1

- This is the number of false positive predictions when the model predicted class 0 but it was actually class 1.
- Out of 172 instances of class 1 in the test sample, the model incorrectly predicted 1 instance as class 0.

2. Class 1:

- False Negative (FN): 172

- This is the number of false negative predictions where the model did not predict class 1, but it was actually class 1.
- Of the 172 Class 1 examples in the test sample, the model failed to correctly predict any of them (all 172 examples were incorrectly assigned to Class 0).

- True Negative (TN): 0

- This is the number of correct predictions where the model correctly predicted class 1.
- Of the 172 Class 1 examples in the test sample, the model failed to correctly predict any, so TN is 0.

- The model barely recognizes class 1, as evidenced by the zero Precision, Recall, and F1-score for this class. The model has a high True Positive (230) and low False Positive (1) for class 0, indicating its high ability to recognize class 0.

- ROC curve (Receiver Operating Characteristic Curve) and AUC value (Area Under the Curve) are important tools to evaluate the quality of classification models. These metrics

provide insight into how well the model distinguishes between positive and negative classes. ROC Curve: This is a graph that shows the ratio of True Positive Rate (the proportion of true positive predictions, also known as completeness or Recall) to False Positive Rate (the proportion of false positive predictions) at different classification thresholds. The ROC curve illustrates how the quality of the model changes as the classification threshold changes.

– AUC (Area Under the Curve): This is a numerical metric that represents the area under the ROC curve. The value of AUC ranges from 0 to 1, where:

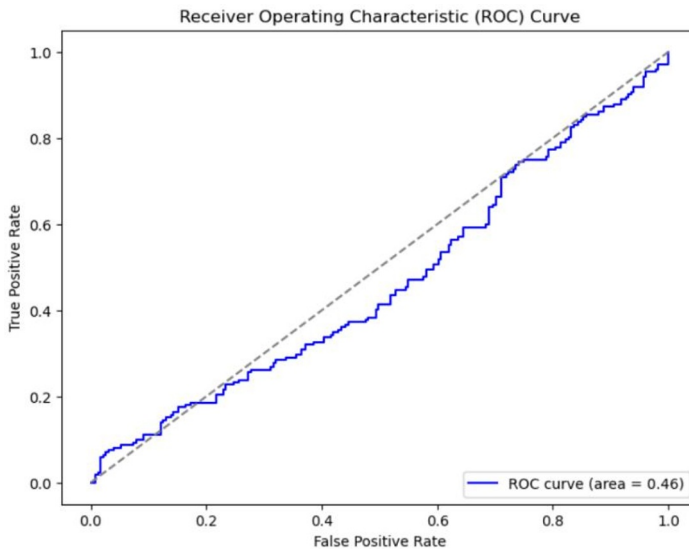
1. AUC = 1 indicates a perfect model that always distinguishes classes correctly.
2. AUC = 0.5 indicates a model that is no better than random guessing.
3. AUC < 0.5 indicates a model that is worse than random guessing [29].

In Figure 3 of the ROC curve, the performance of the logistic regression model can be observed.

The blue line in the graph represents the performance of the model at different thresholds. The gray dashed line represents the random guess with AUC = 0.5. The model has an AUC value of 0.46. This value is below 0.5, indicating that the model is not only unable to discriminate between classes effectively, but also performs worse than random guessing. This model makes more errors in classifying positive and negative classes [29, 30].

The ROC curve and the AUC value provide a clear indication of the poor performance of the logistic regression model:

- The low AUC value (0.46) indicates that the model performs poorly on the task of class distinction.
- The ROC curve plot shows that at any classification thresholds, the model cannot effectively distinguish between positive and negative examples.



**Fig. 3.** ROC Curve.

## 4 Discussion

The results of the analysis showed that the logistic regression model has significant limitations in predicting water quality, especially in classifying water suitable for drinking (Class 1). In this section, we discuss the reasons for the poor performance of the model and possible ways to improve it. Reasons for poor model performance may include:

1. Unbalanced dataset:

- There is a strong class imbalance in the dataset, where there are significantly more instances of class 0 (unfit for drinking) than instances of class 1 (fit for drinking). This results in the model being trained predominantly on class 0 instances and unable to effectively recognize class 1.
  - The low Precision, Recall and F1-score for class 1 confirm this problem.
2. Lack of informative features:
    - Current attributes may not contain sufficient information to accurately distinguish between classes. It is important to consider adding additional attributes that may better characterize water quality.
  3. Inappropriate model parameters:
    - Logistic regression model parameters may not be optimal for this data set. Further tuning of hyperparameters is required to improve performance.
  4. Overtraining or undertraining of the model:
    - The model may be too simple (undertraining) or too complex (overtraining), which also affects its ability to predict both classes.

## 5 Conclusion

The application of the logistic regression method to the considered dataset containing various physical and chemical water parameters demonstrated several key findings regarding its effectiveness in predicting water quality.

Logistic regression, while being a simple and interpretable method, was not sufficiently effective for this task. Key metrics such as accuracy (57.07%) and F1 Score for drinking water class (0.0) indicate the inability of the model to accurately classify water samples. This is particularly evident in the low Precision and Recall values for the drinking water class, indicating a high level of error in predicting this class.

Overall, logistic regression showed limited performance for this dataset, but the suggested areas for improvement offer the potential to significantly improve the quality of the model. This will allow for more accurate and reliable monitoring of water quality, which is important for public safety and health.

## References

1. Tynchenko V. et al. E3S Web of Conferences **458**, 01011 (2023)
2. Kukartsev V. V. et al. E3S Web of Conferences **460**. 07003 (2023)
3. Malozyomov B. V. et al. Energies **16**. 13. 5046 (2023)
4. Kukartsev V. et al. E3S Web of Conferences **458**, 01010 (2023)
5. Filina O. A. et al. Energies **17**. 1. 17 (2023)
6. Boychuk I.P., Grinek A.V., Martyushev N.V., et.al., (2023). Energies, **16(24)**, 8101.
7. Golik V. I. et al. Materials **16**. 21. 7004 (2023)
8. Bosikov I.I. et al. Fire **6**. 3. 95 (2023)
9. Malozyomov B.V. et al. Energies **16**. 9. 3909 (2023)
10. Strateichuk D.M. et al. Crystals **13**. 5. 825 (2023)
11. Bashmur K.A. et al. Sustainability **14**. 20. 13083 (2022)
12. Kolenchukov O.A. et al. Energies **15**. 22. 8346 (2022)
13. Tynchenko Ya.A., Kukartsev V.V., Gladkov A.A., Panfilova T.A., Sustainable Development of Mountain Territories **16**, 1, 56–69 (2024)

14. Kukartsev V.V., Kravtsov K.I., Tynchenko Ya.A., Panfilova T.A., Sustainable Development of Mountain Territories **15(3)**: 784-797 (2023)
15. Yelemessov K.K. et al., Sustainable Development of Mountain Territories **15(2)**: 450-461 (2023). DOI 10.21177/1998-4502-2023-15-2-450-461
16. Kolenchukov O.A. et al. SOCAR Proceedings **No.1** (2023) 123-130  
<http://dx.doi.org/10.5510/OGP20230100814>
17. Tynchenko V.V. et al. Mathematics **12**. 2. 276 (2024)
18. Brigida V. et al. (2024). Resources **13(2)**, 33.
19. Malozyomov B. V. et al. World Electric Vehicle Journal, **15(2)**, 64 (2023)
20. Golik V.I. et al. MIAB. Mining Inf. Anal. Bull. (**11-1**): 175-189 (2023)
21. Panfilova T.A. et al. MIAB. Mining Inf. Anal. Bull. (**11-1**): 239-251 (2023)
22. Sokolov A.A. et al. MIAB. Mining Inf. Anal. Bull. (**11-1**): 278-291 (2023)
23. Gutarevich V.O. et al. MIAB. Mining Inf. Anal. Bull. (**11-1**): 72-87 (2023)
24. Degtyareva K. et al. *Finding patterns in employee attrition rates using self-organizing Kohonen maps and decision trees*. In 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) (pp. 1-6). IEEE (2023)
25. Gladkov A. et al. *Development of Requirements for AIS Aimed at Controlling High Turnover*. In 2023 IEEE International Conference on Computing (ICOCO) (pp. 232-236). IEEE (2023)
26. Degtyareva K. et al. *Analyzing Credit Card Defaulters: A Comparative Study Using Kohonen Maps, Neural Networks, and Decision Trees*. In 2023 International Conference on Information Technology and Computing (ICITCOM) (pp. 152-157). IEEE (2023)
27. Orlov V. et al. E3S Web of Conferences **460**. 07002 (2023)
28. Kravtsov K. et al. E3S Web of Conferences **458**, 09022 (2023)
29. Gutarevich V.O. et al., Applied Sciences **13**. 8. 4671 (2023)
30. Malozyomov B.V. et al. Energies **16**. 11. 4276 (2023)