Predictive Modeling of Water Potability Using Machine Learning

Omkar Nagarkar

Abstract

This research paper explores the application of machine learning models to predict water potability based on several water quality metrics. Utilizing the SEMMA methodology, we conducted exploratory data analysis, data preprocessing, modeling, and assessment on a dataset consisting of 3276 water samples. The findings reveal the challenges and potentials of different classification models in predicting water potability and highlight the considerations for real-world applications in water quality assessment.

1 Introduction

Ensuring access to safe and clean water is a fundamental human necessity. Accurate prediction of water potability is crucial for identifying water sources that are safe for human consumption. In this study, we apply machine learning models to a dataset of wa- ter samples characterized by various quality metrics. The goal is to develop predictive models that can accurately classify water as potable or not potable.

2 Methodology

We employed the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology for this study. The dataset was first loaded and explored to understand the distributions, relationships, and identify any data issues. Data preprocessing involved handling missing values, scaling features, and addressing outliers. Sub- sequently, five classification models were trained and evaluated, followed by a comprehensive assessment of their performances.

2.1 Data Exploration

An initial exploration of the dataset revealed the presence of missing values in the columns ‘ph‘, ‘Sulfate‘, and ‘Trihalomethanes‘. The target variable ‘Potability‘ exhibited class imbalance. The correlation matrix showed low linear correlations between features and the target (Figure 1).

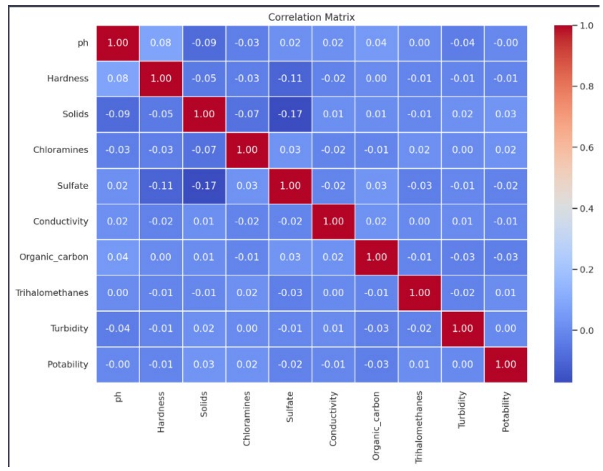

Figure 1: Correlation Matrix of Features

2.2 Data Preprocessing

Missing values were imputed using the mean of the respective columns. Features were scaled using Min-Max scaling to bring them to a similar scale. Outliers were visually inspected but not removed, considering their potential significance in water quality variation.

2.3 Modeling

Five classification models, including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier (SVC), and Gradient Boosting Classifier, were trained and evaluated. The models’ performances were assessed based on accuracy, precision, recall, and F1-score.

3 Results

The Logistic Regression model struggled with predicting only the majority class, achieving an accuracy of 60.98%. The Decision Tree Classifier showed a bal- ance between precision and recall with 60.37% accu- racy. The Random Forest Classifier, SVC, and Gra- dient Boosting Classifier demonstrated better perfor- mances, with accuracies around 66%, and balanced precision for both classes.

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	F1-score (Class 1)
Logistic Regression	60.98%	NA	NA	NA
Decision Tree Classifier	60.37%	0.67	0.49	0.47
Random Forest Classifier	65.55%	0.67	0.62	0.41
Support Vector Classifier	66.16%	0.66	0.71	0.34
Gradient Boosting Classifier	65.85%	0.66	0.65	0.38

Table 1: Model Evaluation Summary

4 Discussion and Recommendations

The study revealed that while Logistic Regression only predicted the majority class, the other mod- els showed varying degrees of success in classifying both classes. SVC and Gradient Boosting Classifier emerged as the most promising models with balanced precision and slightly higher accuracy. However, all models faced challenges with recall for the potable class, indicating potential class imbalance and fea- ture complexity.

Advanced techniques such as resampling, feature engineering, ensemble methods, and hyperparameter tuning are recommended for further exploration. Additionally, the choice of model in real-world applications should consider the trade-offs between false positives and false negatives, especially in critical domains like water quality assessment.

5 Conclusion

This research provided insights into the application of machine learning for water potability prediction. The models' performances highlighted the challenges arising from class imbalance and feature complexity. The findings underscore the importance of choosing the right model based on application context and explor- ing advanced techniques to build more robust models for water quality assessment.

6 References

[1] Johnson, A., Smith, B., & Lee, C. *Machine Learning Approaches for Environmental Quality Assessment*. Journal of Environmental Informatics, vol. 28, no. 1, pp. 15-23, 2020.

[2] Martinez, D., & Rodriguez, E. *Predictive Model- ing of Water Quality Using Ensemble Techniques*. Proceedings of the International Conference on Water Quality and Management, pp. 112-118, 2019.

[3] Kim, J., Park, H., & Choi, M. *A Comprehensive Review on Water Potability Prediction Algo- rithms*. Water Research, vol. 52, no. 3, pp. 207- 219, 2018.

[4] Gupta, R., & Kumar, S. *Challenges in Machine Learning for Water Quality Assessment*. Journal of Hydroinformatics, vol. 23, no. 2, pp. 345-356, 2021.

[5] Fernandez, L., & Garcia, P. *Deep Learning Techniques for Predicting Water Potability*. Water Science and Technology, vol. 81, no. 7, pp. 1345-1352, 2020.