Enhancing Water Potability Prediction Machine Learning Model using Standardization and Normalization

Aditya Rawat

1. Introduction

Availability of safe drinking water is very important for the health. Testing water quality using limited features like – pH or amount of solids present in water etc. Using Machine Learning Algorithms we can combine multiple features to predict the water potability by analysing features like - pH, hardness, solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes and turbidity. This research explores the application of various Machine Learning Algorithms – Random Forest Classifier, Gradient Boosting, Support Vector Machine, Logistic Regression, Neural Network and Decision Tree, to classify water samples as potable or non-potable. A very important part of the research involves the examining of pre-processing techniques, such as Standardization and Normalization, on improving the model’s accuracy.

The result shows that the Support Vector Machine and Random Forest Classifier achieved highest accuracy of 69.51% and 68.29%, despite not matching the highest reported accuracies in the existing studies of 88% using Decision Tree [1] and 81% using Random Forest Classifier [2]. This research underscores the importance of pre-processing and highlights the strengths of pre-processing. It provides a foundation for further exploration, including advanced techniques and larger datasets, to enhance accuracy.

1. Literature Review

The traditional method of finding water quality involves laboratory testing of properties such as pH, hardness, solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes and turbidity. While the traditional method was highly accurate, these methods were expensive, time-consuming and often impractical for large – scale or real-time analysis. Machine Learning helps us automate the prediction of water quality.

After going through the several researches on the water potability prediction using Machine Learning, with different methodologies resulting in varying performance. Research conducted in Russia shows that the Random Forest Classifier, achieving an accuracy of 69.36% with the addition of SMOTE for class imbalance the accuracy deducted to 67.07%. This highlights that balancing the data using SMOTE deducts the accuracy of the Random Forest Classifier [3].

A study in Bulgaria shows that the optimized Decision Trees achieved the accuracy of 88%, while the Support Vector Classifier and Random Forest achieved the accuracy of 83% and 81%, respectively. [1] Instead of using SMOTE, which was used in a research achieved the accuracy of 81%, the Support Vector Classifier and Random Forest outperform without the use of SMOTE. Whereas some other studies, showed the slightly lower accuracy rates of 74% and 70% with Random Forest Classifier. This shows that while the Random Forest Classifier is a robust model, performance can vary depending on data quality, pre-processing and feature selection.

Machine learning (ML) has been increasingly used to predict water potability, addressing challenges associated with traditional water quality assessment methods. Numerous studies have explored the potential of different ML algorithms and pre-processing techniques, with varying degrees of success. Research has shown that pre-processing techniques like feature scaling, normalization, and data balancing significantly impact the performance of ML models in water potability prediction.

In one study, researchers used the Random Forest (RF) classifier to predict water potability, achieving 69.36% accuracy and a ROC-AUC of 0.63. By addressing class imbalance using SMOTE (Synthetic Minority Oversampling Technique), the accuracy dropped slightly to 67.07%, but the ROC-AUC improved to 0.64, highlighting the trade-offs involved in balancing imbalanced datasets​ (Drinking water portability…). Similarly, Patel et al. utilized SMOTE to balance their dataset and tested multiple algorithms, including RF and XGBoost, achieving an accuracy of 81%. Their study emphasized the importance of explainable AI techniques, such as Local Interpretable Model-agnostic Explanations (LIME), to improve model interpretability and reliability ​(Patel’s research).

Another study applied dimensionality reduction using Principal Component Analysis (PCA) to enhance the performance of several classifiers. Without PCA, Support Vector Machines (SVM) achieved the highest accuracy of 69%, while other algorithms, including XGBoost, KNN, Gaussian Naive Bayes, and RF, ranged between 62% and 68%. Remarkably, applying PCA increased accuracy across all classifiers to nearly 100%, showcasing the transformative potential of dimensionality reduction ​(Optimizing machine learn…).

In Bulgaria, researchers tested modified machine learning models and achieved high accuracy rates, with Decision Trees (DT) reaching 88%, SVM 83%, and RF 81%. Their study highlights the efficacy of tree-based models when optimized for structured datasets ​(Drinking water portability…). Another noteworthy contribution from Patel et al. demonstrated the consistent performance of ensemble methods like RF and XGBoost when combined with robust pre-processing techniques​ (Patel’s research).

The literature underscores the importance of pre-processing steps, such as normalization, dimensionality reduction, and data balancing, in enhancing model performance. Moreover, studies show that while high accuracy rates are achievable, the choice of algorithm and pre-processing strategy depends heavily on dataset characteristics. This research builds on these findings by focusing on SVM and RF, emphasizing the role of pre-processing and exploring ways to address dataset-specific challenges.

Proposed System

1. Predicting Water Potability Using a Machine Learning Model
2. Conclusion and Future Work