WATER POTABILITY PREDICTOR

KARTHIK KEMIDI, RAJESH KURVA & K THARUN KUMAR REDDY

Abstract

This study determines whether water is safe for consumption and compares machine learning models like Decision Tree, Random Forest, XGBoost, KNN, SVM, and Gaussian Naive Bayes to identify the most efficient technique for predicting water quality. Machine learning techniques can be used for the analysis and prediction of water quality by considering various parameters like pH value, turbidity, hardness, conductivity, and dissolved solids.

Keywords: Decision Tree, Random Forest, KNN, SVM, XGBoost, Gaussian Naive Bayes, Performance Metrics.

Introduction

The study revolves around learning as the ability of machines to learn from experience by their inherent intelligence; because that is the human machine. Current advances in machinery brought forward by improved learning techniques have indeed brought machine learning avenues into great prowess.

In this study, having applied data mining approaches, we analyzed and predicted potability of water as the valuable insights into its consumption safety.

Materials & Methods

The research aimed to predict water potability using machine learning techniques. It utilized a dataset of 3,276 water samples with key parameters like pH, hardness, TDS, conductivity, chloramines, and sulfates to classify water as potable or non-potable. The study ensured efficient classification aligned with public health and environmental standards.
Key steps in the methodology are summarized as follows:

- **Data Preprocessing:** The dataset was cleaned by handling missing values, selecting relevant features, and using a correlation matrix to ensure data quality and efficiency.
- **Training and Testing:** The dataset was split into training and testing sets, and multiple machine learning models (Decision Tree, Random Forest, SVM, KNN, and XGBoost) were evaluated for accuracy and performance.
- **Model Evaluation and Parameters:** Performance metrics like accuracy, precision, recall, and F1-score were used to assess the models, with Random Forest showing the best performance. Key parameters such as pH, TDS, and conductivity were used for classification, following WHO standards.

This methodology provided a reliable framework for water quality classification using machine learning.

References

[1] D. Poudel, D. Shrestha, S. Bhattarai, and A. Ghimire. Comparison of machine learning algorithms in statistically imputed water potability dataset. *Journal of Innovations in Engineering Education*, 5, 2022.

[2] X. Wang, Y. Li, Q. Qiao, A. Tavares, and Y. Liang. Water quality prediction based on machine learning and comprehensive weighting methods. *Entropy*, 25, 2023.

Contact Information

Project Supervisor Mr. Panigrahi Srikanth

Project Guide Mr. M. Vishnu Chaitanya

WebPage https://water-potability-predictor.netlify.app/

Email kemidikarthik2004@gmail.com

Results

The Random Forest Classifier emerged as a robust model with strong performance across key metrics like accuracy, precision, recall, and F1-score. It achieved approximately 70% accuracy, demonstrating balanced precision and recall, which is vital for reliable evaluations. Its ensemble nature effectively mitigates overfitting, enhancing generalizability to unseen data. Moreover, the model handles imbalanced datasets proficiently, ensuring consistent results across diverse scenarios.

```
Random Forest Results:
Accuracy: 0.6784
Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      0.86      0.77       412
         1.0       0.61      0.38      0.47       244

    accuracy                           0.68       656
   macro avg       0.65      0.62      0.62       656
weighted avg       0.67      0.68      0.66       656

Confusion Matrix:
[[353  59]
 [152  92]]
```

Figure 1: Performance of Random Forest Classifier

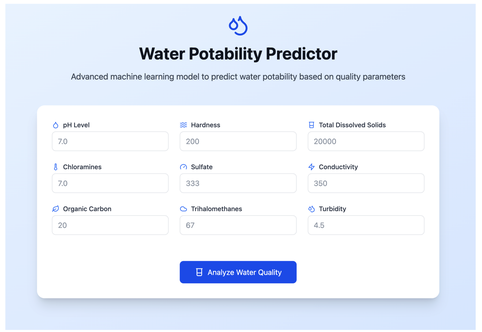

Figure 2: WebPage

The Water Potability Predictor evaluates drinkability using key parameters like pH, Hardness, TDS, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity. With these inputs, the model reliably predicts water quality, offering a quick and effective solution for assessing potability.Once these values are entered, the model processes the data and provides a prediction about the water's suitability for consumption.

Conclusion

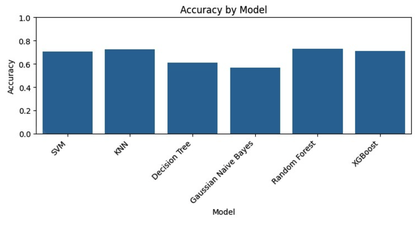

Figure 3: Various Models Accuracies

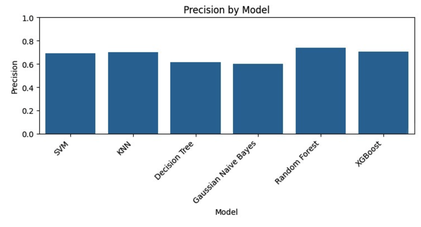

Figure 4: Various Models Precision

The project "Water Potability Predictor" demonstrates the following key findings:

- Several machine learning models were tested, highlighting the importance of data-driven solutions in classifying water samples as potable or non-potable.
- The Random Forest classifier was the most effective due to its balanced performance across multiple metrics, including accuracy, precision, recall, and F1-score.
- Random Forest's robustness and versatility in handling various forms of data make it suitable for real-world water quality assessment applications.
- The study emphasizes the potential of machine learning to transform water quality monitoring, making it more efficient and reliable.