Executive Summary

This project develops a machine learning classification system to predict water potability from physicochemical properties using the Kaggle Water Potability Dataset with 3,276 samples and 9 input features plus a binary potability label. Four classification algorithms were trained and evaluated: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). The best-performing model on the held-out test set was the SVM, achieving about 67% accuracy and a ROC-AUC of around 0.65, while the Random Forest achieved about 66% accuracy and a similar ROC-AUC. These results show that machine learning can provide an automated screening tool for water safety, although performance remains moderate and should complement, not replace, laboratory testing.

Introduction

Access to safe drinking water is critical for public health, but conventional laboratory-based water quality testing can be slow, costly, and resource-intensive. This project explores whether supervised machine learning models can predict whether a water sample is potable using only physicochemical measurements. The objective is to build, evaluate, and compare several classification models and to assess whether their performance is sufficient for use as a decision-support tool in water quality assessment.

Dataset and Features

The project uses the Kaggle Water Potability dataset, which contains 3,276 water samples with 9 numeric features and a binary target variable “Potability” (0 = not safe, 1 = safe). The features are pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity, all representing standard physicochemical indicators of water quality. The target classes are imbalanced, with more non-potable than potable samples, which influences model training and evaluation.

Data Quality and Preprocessing

Exploratory data analysis showed missing values in several numeric features, particularly pH, Sulfate, and Trihalomethanes, as well as skewed distributions and outliers typical of environmental data. Median imputation was applied to all numeric features to handle missing values, and a StandardScaler was used within a preprocessing pipeline to normalize feature ranges before model training. The data was then split into training and test sets using an 80–20 stratified split to preserve the original class distribution.

Methodology

The modeling pipeline consisted of a shared preprocessing step (median imputation and scaling) followed by four separate classifiers: Logistic Regression, Decision Tree, Random Forest, and SVM with probability estimates enabled. Each model was trained on the training set and

evaluated on the test set using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC, with ROC curves plotted for visual comparison. Finally, a full-data Random Forest model was trained and saved as a deployment artifact for potential future use.

Model Performance

The table below summarizes the actual test performance recorded in the notebook:

Model	Accuracy (test)	ROC-AUC (test)
Logistic Regression	≈ 0.61	≈ 0.55
Decision Tree	≈ 0.60	≈ 0.57
Random Forest	≈ 0.66	≈ 0.65
SVM	≈ 0.67	≈ 0.65

Water_Quality_Classification.pdf

The SVM achieved the highest test accuracy at about 67%, with a ROC-AUC slightly below 0.65, making it the best-performing model in this experiment. The Random Forest was a close second, with approximately 66% accuracy and a ROC-AUC of about 0.65, while Logistic Regression and the Decision Tree lagged behind in both metrics.

Best Model Analysis

Although the SVM achieved the highest accuracy, both SVM and Random Forest showed similar ROC-AUC values, indicating comparable discrimination between potable and non-potable samples. Random Forest has the additional advantage of providing feature importance, which helps interpret which physicochemical variables most influence predictions, even if exact importance percentages are not explicitly printed in the notebook. Confusion matrices for both models show a tendency to correctly classify non-potable water more often than potable water, reflecting the underlying class imbalance and a conservative bias in predicting safe water.

Challenges and Limitations

The main challenges include class imbalance, moderate overall accuracy, and limited feature scope. The dataset contains substantially more non-potable than potable samples, which can bias models toward predicting the majority class and reduce recall for potable water. In addition, the dataset size (3,276 samples) is modest for complex models, and the features are purely physicochemical measurements without temporal, geographic, or microbiological information, which constrains maximum achievable performance.

Recommendations and Future Work

Future work should explore systematic hyperparameter tuning, such as grid search for SVM and Random Forest, to see whether performance can be improved beyond the current 66–67% accuracy range. Techniques such as SMOTE or other resampling strategies could be used to mitigate class imbalance, and additional engineered features (for example, pH deviation from neutrality or binned hardness levels) might capture more nuanced relationships. Incorporating domain expertise from water quality specialists and extending the dataset with temporal and spatial information could further improve model robustness and real-world utility.

Conclusion

This project demonstrates that standard supervised learning algorithms can use physicochemical water quality indicators to predict potability with moderate but meaningful performance. On the held-out test set, the best models (SVM and Random Forest) reach around 66–67% accuracy and ROC-AUC values near 0.65, indicating better-than-random classification but leaving room for improvement. As a result, these models are best viewed as decision-support tools that can prioritize samples for laboratory testing rather than as stand-alone certification mechanisms for drinking water safety.