Water Quality Test Prediction for Concrete Mixing

Machine Learning Project Report

Anurag Raj Putheril

23BEC0310

---

1. Introduction

Concrete is one of the backbones of modern infrastructure, and water is a fundamental component in the concrete mixing process. The quality of water used has a direct impact on the strength, durability, and overall integrity of the final concrete product. Poor water quality can lead to expensive repairs or replacements.

The aim of this project is to develop a machine learning (ML) model that accurately classifies water quality for concrete mixing based on site-specific measurements. Such a system ensures that only water meeting quality standards is used, thus enhancing structural longevity and reducing operational risks.

2. Problem Statement

To predict whether available water is suitable for concrete mixing based on the following parameters:

- pH
- Chloride
- Organic Carbon
- Solids
- Sulphate
- Turbidity

The goal is to design and evaluate Machine Learning classifiers that can:

- Predict water quality accurately (good/bad)
- Help automate real-time site decisions on water usage

3. Dataset

The model uses a historical dataset containing:

- **Input features:** pH, Chloride, Organic Carbon, Solids, Sulphate, Turbidity
- **Target:** Water quality label (good/bad)

4. Models Developed

The following machine learning classifiers were implemented and evaluated:

- Logistic Regression
- Random Forest Classifier
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)

5. Evaluation Metrics

The models were assessed using the following metrics:

- Accuracy
- Precision
- Recall (Sensitivity)
- F1-score
- Confusion matrix

6. Results

Metrics	Logistic Regression	Random Forest	KNN	SVM
Accuracy	0.858	0.774	0.856	0.8575
Precision	0.7603	0.6957	0.7598	0.7594
Recall (Sensitivity)	0.8699	0.6280	0.8598	0.8669
F1-Score	0.8101	0.6601	0.8067	0.8096
Specificity	0.8531	0.8524	0.8539	0.8524

7. Analysis and Discussion

- SVM and Logistic Regression offer the best trade-off between all the metrics. Both models are highly reliable for this task.
- Random Forest's low recall means it would miss detecting poor-quality water more often — risky in construction contexts where false negatives could have serious consequences.
- The high recall of SVM and Logistic Regression (0.8669) is valuable, as it minimizes the chances of incorrectly classifying poor water as good (critical for safety).

8. Conclusion

- The project successfully implemented machine learning classifiers to predict water quality for concrete mixing.
- The SVM and Logistic Regression were the most suitable, providing high accuracy, precision, and recall.
- These models can be integrated into real-time systems to assist engineers on construction sites in making informed decisions about water usage.

9. Future Work

- **Model improvement**: Explore neural networks for potentially higher accuracy.
- **Feature expansion**: Include additional water quality parameters (e.g., hardness, alkalinity) for even better predictions.
- **Deployment**: Develop a app or IoT integration for field usage.

10. References

- Dataset on water quality characteristics (as provided/attached)
- Scikit-learn documentation