WATER QUALITY PREDICTION FOR CONCRETE MIXING

24EEE431 – ARTIFICIAL INTELLIGENCE AND EDGE COMPUTING

Report

Submitted by

Roll Nos.	Student Names
CB.EN.U4ELC22019	KANISHKA S
CB.EN.U4ELC22037	NIVETHA GK
CB.EN.U4ELC22046	RAHUL B
CB.EN.U4ELC22059	VIKAS K

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING
AMRITA SCHOOL OF ENGINEERING,
AMRITA VISHWA VIDYAPEETHAM,
COIMBATORE - 641112

MARCH,2024

Amrita School of Engineering

Department of Electrical and Electronics Engineering

CONTENTS

ABSTRACT

The quality of water used in concrete mixtures significantly impacts the strength, durability, and overall performance of the final structure. This project leverages machine learning techniques to predict water quality for concrete mixtures based on various physicochemical properties. The dataset consists of key water quality parameters, which undergo preprocessing, including data cleaning, normalization, and exploratory data analysis (EDA) to identify trends and correlations that influence concrete performance.

Several machine learning models, including Logistic Regression, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP), are employed to classify water quality. The dataset is divided into training and testing sets to ensure robust model evaluation. Performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices are used to assess and compare the models. Additionally, hyperparameter tuning is conducted using GridSearchCV to optimize predictive accuracy.

The objective of this study is to determine the most effective model for predicting water quality suitability in concrete mixtures, helping engineers and construction professionals make data-driven decisions. By integrating AI-driven techniques into the assessment process, this project provides an efficient and automated approach to evaluating water quality, ultimately improving construction quality control. The findings can contribute to the optimization of concrete mix designs, ensuring structural integrity and sustainability in the construction industry.

INTRODUCTION

Water quality plays a crucial role in various industrial applications, including construction, where it directly influences the properties of concrete mixtures. The strength, durability, and long-term performance of concrete structures are significantly affected by the chemical and physical characteristics of the water used in the mixing process. Impurities in water, such as high salinity, organic matter, and industrial pollutants, can weaken the cementitious bonds, leading to reduced structural integrity and potential failures over time.

Traditionally, water quality assessment for concrete mixtures has relied on laboratory testing, which can be time-consuming, labor-intensive, and costly. With advancements in machine learning and data-driven approaches, predictive models can offer a faster, more efficient alternative for evaluating water quality. This study aims to develop and implement machine learning models to classify water quality based on key physicochemical properties, providing a reliable method for assessing its suitability for concrete production.

The project involves data preprocessing, exploratory data analysis (EDA), feature selection, and the application of various classification models, including Logistic Regression, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). The models are trained and tested on real-world water quality data to predict whether a given water sample is suitable for concrete mixing. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate model effectiveness.

By integrating artificial intelligence into water quality assessment, this study provides a data-driven approach to ensuring optimal concrete performance. The findings of this research can help engineers, construction companies, and quality control professionals make informed decisions, reducing risks associated with poor-quality water in construction projects. This report presents a detailed analysis of the methodology, results, and implications of using machine learning for water quality prediction in concrete mixtures.

PROBLEM STATEMENT

Water quality plays a crucial role in the concrete mixing process, directly impacting the strength, durability, and overall performance of the final structure. Contaminants such as excessive chloride, organic carbon, solids, sulfate, and turbidity can interfere with the hydration process, weaken the concrete matrix, and lead to structural failures, costly repairs, and reduced lifespan of buildings and infrastructure.

Traditional methods for assessing water quality rely on chemical testing and laboratory analysis, which are not only time-consuming but also prone to human errors, making them unsuitable for real-time decision-making on construction sites. These methods require manual sample collection, transportation, and analysis, leading to delays in quality assurance and increasing the risk of substandard concrete production. Additionally, variations in water quality from different sources add complexity to the assessment process, further emphasizing the need for a more efficient approach.

To address these challenges, an automated machine learning-based system is required to classify water quality efficiently and accurately. By leveraging historical water quality data and applying advanced predictive models, this system can provide real-time assessments of water suitability for concrete mixing. Such an approach will enhance decision-making in construction, ensuring consistent quality, reducing dependency on traditional testing, and ultimately improving the reliability and longevity of concrete structures

METHODOLOGY

A. Data Preprocessing

Before training the model, data preprocessing was performed to enhance the quality of input features. The dataset was examined for missing values, which were handled using mean imputation to ensure completeness. Feature scaling was applied using StandardScaler, which normalizes data by subtracting the mean and scaling to unit variance. This step ensures that all features contribute equally during training. Additionally, correlation analysis was conducted to eliminate redundant features with a correlation coefficient above 0.9, preventing multicollinearity and improving model performance.

B. Exploratory Data Analysis (EDA) and Data Visualization EDA

It was conducted to understand the dataset's structure, distribution, and key patterns. This included:

- **Basic Statistical Analysis:**

  The dataset's structure was examined using df.info(), df.describe(), and df.isnull().sum() to check for missing values, data types, and distributions.

- **Correlation Analysis:**

  A correlation heatmap was generated using seaborn.heatmap() to identify strong relationships between features.

  Highly correlated features (correlation coefficient > 0.9) were eliminated to prevent multicollinearity.

- **Data Distribution Visualization:**
  Histograms and Boxplots:
  seaborn.histplot() and matplotlib.boxplot() were used to visualize feature distributions and detect outliers.

- Outlier Handling Outliers in the dataset were identified and managed using Boxplot Inspection:
  seaborn.boxplot() was used to visualize and detect extreme values.
  Interquartile Range (IQR) Method:
  The IQR formula was applied:
  $IQR=Q3-Q1$
  Any data points outside Q1 - 1.5 × IQR or Q3 + 1.5 × IQR were considered outliers.
  These extreme values were either removed or capped at a threshold based on domain knowledge.

C. Machine Learning Approaches

For Water Quality Classification Various machine learning techniques were employed to classify water quality as Good (1) or Bad (0). Each algorithm was evaluated based on accuracy, precision, recall, and F1-score.

1. Logistic Regression (LR) Logistic Regression is a linear classification algorithm that applies a sigmoid function to predict probabilities between two classes (Good/Bad). The main disadvantage is that it assumes linear separability, which may limit performance on complex datasets.

```
Model: Logistic Regression
Accuracy: 0.8043
Precision: 0.6085
Recall: 0.8283
F1 Score: 0.7016
Confusion Matrix:
[[1036  267]
 [  86  415]]
```

2. Support Vector Machine (SVM) SVM is a supervised learning algorithm that finds an optimal hyperplane to separate data points into distinct classes. Comparitively the accuracy was lower when compared to MLP model

```
Model: SVM
Accuracy: 0.8182
Precision: 0.6358
Recall: 0.8084
F1 Score: 0.7118
Confusion Matrix:
[[1071  232]
 [  96  405]]
```

3. Random Forest (RF) Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions for better accuracy.

```
Model: Random Forest
Accuracy: 0.7140
Precision: 0.4824
Recall: 0.4112
F1 Score: 0.4440
Confusion Matrix:
[[1082  221]
 [ 295  206]]
```

4. k-Nearest Neighbors (k-NN) k-NN is a distance-based algorithm that classifies data points based on the majority class of their nearest neighbors.

```
Model: KNN
Accuracy: 0.7633
Precision: 0.5673
Recall: 0.6228
F1 Score: 0.5937
Confusion Matrix:
[[1065  238]
 [ 189  312]]
```

5. Multi-Layer Perceptron (MLP) - Selected Model MLP is a feedforward artificial neural network with multiple layers of interconnected neurons.

   MLP Architecture: Input Layer: Receives numerical water quality parameters.

   Hidden Layers: Three hidden layers (100, 50, 25 neurons) using ReLU activation

   Output Layer: A single neuron using sigmoid activation for binary classification.

D. Hyperparameter Tuning

To improve model performance, GridSearchCV was used to optimize hyperparameters:

Multi-Layer Perceptron (MLP): Tuned Parameters:

```
Fitting 3 folds for each of 96 candidates, totalling 288 fits
✅ Best Parameters: {'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (64, 32), 'learning_rate': 'constant', 'solver': 'sgd'}
🎯 Final Tuned Model Accuracy: 0.82
```

Hyperparameter tuning significantly improved classification accuracy, especially for MLP.

E. Model Selection and Evaluation:

Each model was trained and evaluated using accuracy, precision, recall, and F1-score. MLP outperformed all other models in handling non-linearity and feature interactions.

F. Final Model Justification

MLP was selected as the best-performing model due to:

- Higher Classification Accuracy
- Better Handling of Non-Linear Patterns
- Adaptive Learning via Adam Optimizer

G. Water Quality Prediction using GUI

To enhance usability, a Graphical User Interface (GUI) was developed for user-friendly water quality prediction.

Implementation Details: Technology Used: Python with Tkinter for GUI development.

User Input Fields: pH, alkalinity, turbidity, chloride content, hardness, and other water quality parameters.

The model predicts water quality classification and displays the result.

Output Display:

The GUI shows whether the water is Good (1) or Bad (0) for concrete mixing. Thresholds for Classification: The GUI classifies water quality based on the following thresholds:

Good Water (Label = 1):

- Chloride: 500 - 6984 mg/L (Mean: 1745)
- Organic Carbon: 50 - 598 mg/L (Mean: 164)
- Solids: 502 - 11991 mg/L (Mean: 2088)
- Sulphate: 20 - 799 mg/L (Mean: 268)
- Turbidity: 11 - 5971 NTU (Mean: 1438)
- pH: 2.00 - 11.99 (Mean: 4.75)

Bad Water (Label = 0):

- Chloride: 500 - 6999 mg/L (Mean: 3994)
- Organic Carbon: 50 - 599 mg/L (Mean: 356)
- Solids: 502 - 11999 mg/L (Mean: 6110)
- Sulphate: 21 - 799 mg/L (Mean: 541)
- Turbidity: 16 - 5999 NTU (Mean: 3538)
- pH: 2.00 - 11.99 (Mean: 8.19)

Output Display:

If input values fall within the Good Water thresholds → Displays "Water is Suitable for Concrete Mixing".If input values fall within the Bad Water thresholds → Displays "Water is Not Suitable for Concrete Mixing". This automated system eliminates manual calculations, making water quality assessment efficient and accessible for engineers and construction professionals.

RESULTS

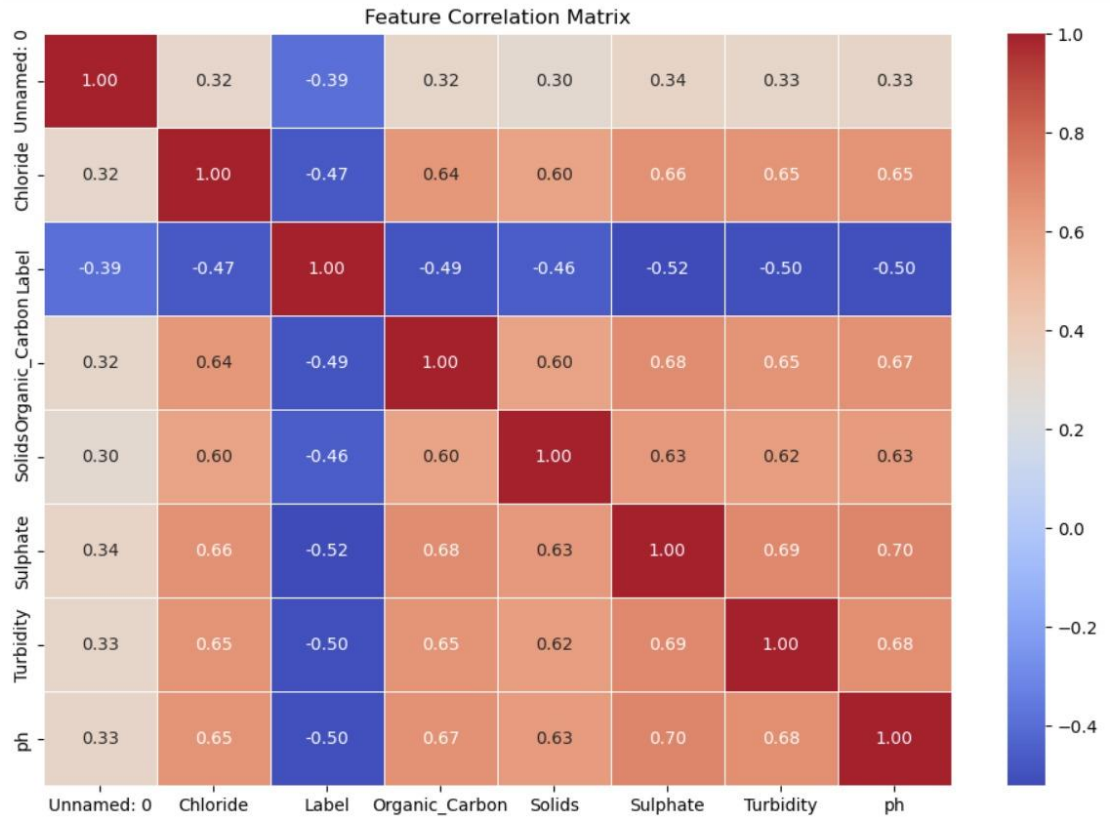

Fig 1-Feature Correlation Matrix

This heatmap shows water quality correlations, where red indicates a strong positive correlation and blue represents a negative correlation. Chloride, Sulphate, and Turbidity are highly correlated, with pH showing strong correlations with Sulphate (0.70) and Turbidity (0.68). The water quality label negatively correlates with most features, suggesting higher chemical concentrations often indicate poorer water quality.

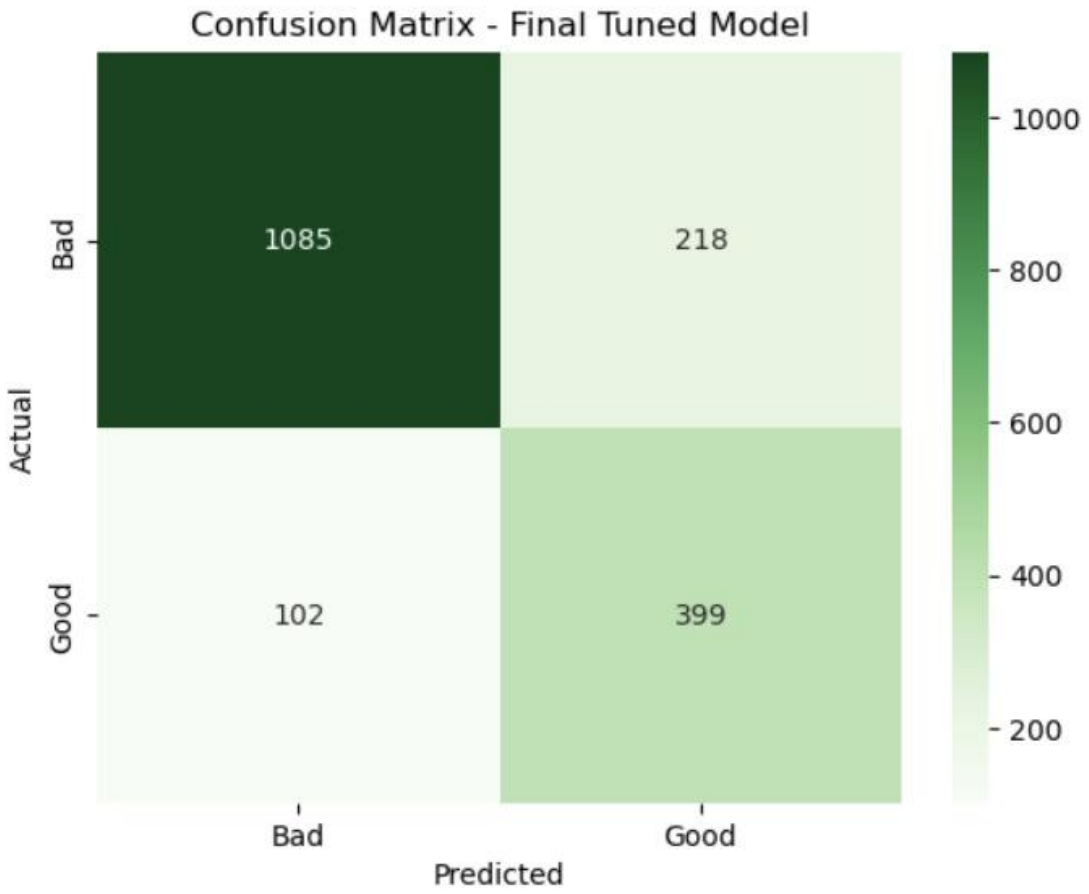

Fig 2-Confusion Matrix - Final Tuned Model

This confusion matrix represents the performance of the optimized water quality prediction model. It correctly classifies 399 Good and 1085 Bad water samples, with 218 False Positives and 102 False Negatives. Compared to the basic model, it improves Bad water classification by reducing false positives from 230 to 218, enhancing reliability for water quality assessment in concrete mixing.

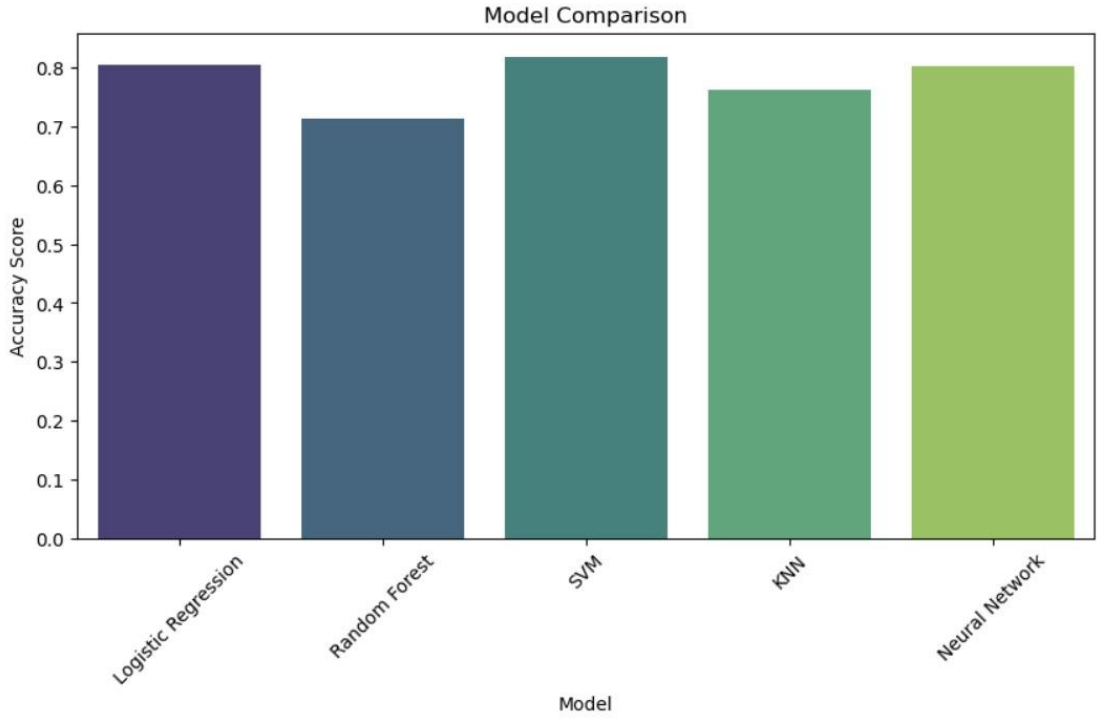

Fig 3-Model Comparison

Neural Network (~80%) is competitive with SVM and Logistic Regression and has the highest accuracy

- Accuracy: 0.8226
- Precision: 0.6467
- Recall (Sensitivity): 0.7964
- Specificity: 0.8327
- F1-Score: 0.7138

- Classification Report:

	precision	recall	f1-score	support
Bad	0.91	0.83	0.87	1303
Good	0.65	0.80	0.71	501
accuracy			0.82	1804
macro avg	0.78	0.81	0.79	1804
weighted avg	0.84	0.82	0.83	1804

- Confusion Matrix:

```
[[1085  218]
 [ 102  399]]
```

Fig 4-Best model Evaluation metrics

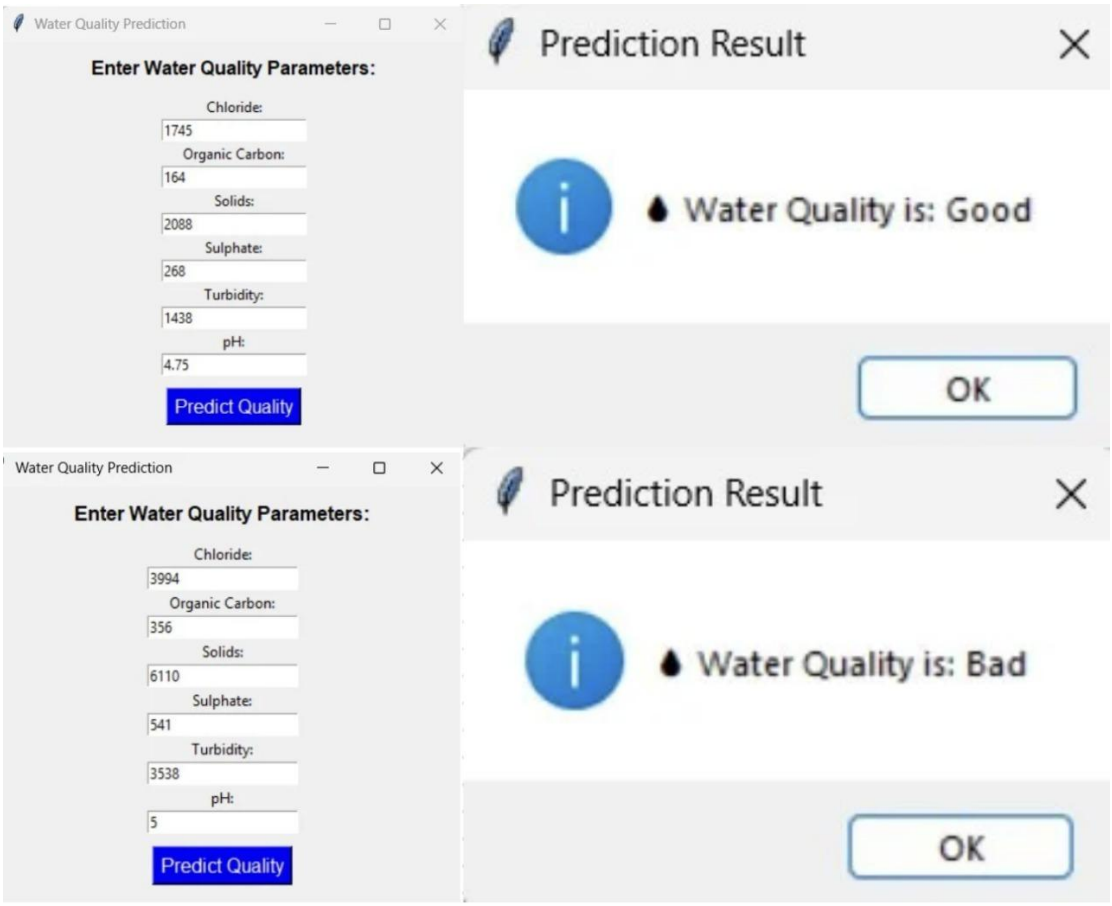

Fig 5- Water Quality Prediction Results

This water quality prediction application evaluates parameters like Chloride, Organic Carbon, Solids, Sulphate, Turbidity, and pH to determine suitability. Higher contaminant levels lead to a "Bad" quality result, while optimal values indicate "Good" water. This tool aids in assessing water safety for concrete mixing and environmental use.