Major Project-1 Report

Water Quality Test Using Machine Learning Classification Algorithms

---

1. Introduction

Water quality plays a crucial role in public health, environmental safety, and sustainable living. Contaminated water can cause severe diseases such as cholera, diarrhea, typhoid, and chemical poisoning. Therefore, detecting whether water is **safe (potable)** or **unsafe (non-potable)** is extremely important.

Traditional laboratory testing methods are reliable but:

- Time-consuming
- Require expert supervision
- Need expensive laboratory equipment

To overcome these challenges, this project proposes a **Machine Learning-based Water Quality Classification System** that can automatically predict whether water is safe to drink using physicochemical properties of water. This makes water testing **faster, cost-effective, and smarter**.

---

2. Project Objective

The main objectives of this project are:

1️⃣ Understand important water quality indicators such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity.
2️⃣ Perform **Exploratory Data Analysis (EDA)** to study feature behavior and dataset patterns.
3️⃣ Preprocess the dataset by:

- Handling missing values
- Scaling features
- Managing class imbalance

4️⃣ Train multiple machine learning classification models:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)
- XGBoost Classifier

5️⃣ Evaluate models using:

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix
- ROC Curve
- ROC-AUC Score

6️⃣ Compare models and select the **best model** for final deployment.
7️⃣ Build a **final prediction pipeline** to classify new unseen water samples.

---

3. Dataset Description

The dataset used in this project is the **Water Potability Dataset** from Kaggle.

- Total Records: **3276**
- Total Features: **10**
- Target Variable:
    - 0 = Not Potable (Unsafe Water)
    - 1 = Potable (Safe Water)

Dataset Features

Feature	Meaning
pH	Acidity / alkalinity of water
Hardness	Water hardness level
Solids	Total dissolved solids
Chloramines	Disinfectant level
Sulfate	Sulfate concentration
Conductivity	Electrical conductivity
Organic Carbon	Organic contamination level
Trihalomethanes	Chemical compounds present
Turbidity	Water clarity

Feature	Meaning
Potability	Target output (0 or 1)

---

4. Project Workflow

1️⃣ Dataset Acquisition
2️⃣ Exploratory Data Analysis
3️⃣ Data Preprocessing
4️⃣ Model Building
5️⃣ Model Evaluation
6️⃣ Best Model Selection
7️⃣ Final Prediction Pipeline

---

5. Exploratory Data Analysis (EDA)

5.1 Missing Values

The dataset contained missing values in multiple features.
To handle them, **Median Imputation** was applied because:

- Median is robust against outliers
- Provides stable results compared to mean

5.2 Class Distribution

Dataset is **imbalanced**:

- More unsafe water samples
- Fewer safe water samples

To handle imbalance effect:

- Stratified Train-Test Split was used
- Robust models like SVM, Random Forest, and XGBoost were applied

---

6. Data Preprocessing

✔ Handled missing values
✔ Separated features (X) and target (y)
✔ Train-Test Split (80% Train, 20% Test)
✔ Applied **StandardScaler()**
Scaling is necessary because:

- Different features have different ranges
- SVM and Logistic Regression perform better on scaled data

---

7. Machine Learning Models Used

Following models were trained:

1️⃣ Logistic Regression
2️⃣ Decision Tree Classifier
3️⃣ Random Forest Classifier
4️⃣ Support Vector Machine (SVM)
5️⃣ XGBoost Classifier

---

8. Model Evaluation Results

ROC-AUC Scores

Logistic Regression → 0.548

Decision Tree → 0.559

Random Forest → 0.643

XGBoost → 0.639

SVM → 0.648 ⭐ Highest

Performance Comparison

XGBoost

Accuracy: 0.649
Precision: 0.579
Recall: 0.371
F1 Score: 0.452

SVM

Accuracy: 0.670 ⭐ Highest
Precision: 0.704 ⭐ Highest
Recall: 0.269
F1 Score: 0.389

Random Forest

Accuracy: 0.660
Precision: 0.632
Recall: 0.308
F1 Score: 0.414

Decision Tree

Moderate but unstable performance

Logistic Regression

Failed to classify class 1 → Very poor recall

---

9. Best Model Selection

🎯 **Final Best Model: Support Vector Machine (SVM)**

Reasons

- ✔ Highest Accuracy
- ✔ Highest Precision
- ✔ Highest ROC-AUC score
- ✔ Best separation between safe vs unsafe water
- ✔ Stable + consistent performance
- ✔ Works great on scaled numeric data

---

10. Final Prediction Pipeline

A complete prediction pipeline was created using:

- StandardScaler
- SVM Classifier
- Pipeline Integration
- Model Saving using joblib
- Final prediction function for new data samples

This allows real-time water quality prediction.

---

11. Expected Output

User inputs 9 water features → Pipeline processes → Model predicts:

- "Water is SAFE to drink"
  or
- "Water is NOT SAFE to drink"

Along with **probability score**.

---

12. Challenges Faced

- Dataset imbalance
- Many missing values
- Handling chemical parameter variation
- Choosing right model
- Balancing recall vs precision

---

13. Future Improvements

- 🔹 Use deep learning models
- 🔹 Apply SMOTE oversampling to improve recall
- 🔹 Deploy as Web App using Flask / Streamlit
- 🔹 Integrate IoT sensors for real-time monitoring
- 🔹 Improve dataset quality using larger samples

---

14. Conclusion

This project successfully developed a Machine Learning-based system to predict water potability using chemical properties. Multiple models were tested and evaluated using standard performance metrics.

Among all tested models, **Support Vector Machine (SVM)** performed the best with the highest accuracy, precision, and ROC-AUC score. The final pipeline is capable of predicting water safety efficiently and can be used as a smart decision-support tool.

This project demonstrates how Machine Learning can significantly assist in improving environmental safety and public health.

---

15. Deliverables

- ✔ Jupyter Notebook
- ✔ Clean Processed Dataset
- ✔ Final Trained Model File
- ✔ Project Report
- ✔ PPT Presentation

---

16. Learning Outcomes

Through this project, the following skills were gained:

- Data preprocessing & cleaning
- Exploratory data analysis

- Machine learning model training
- Evaluation & comparison
- Understanding ROC & AUC
- Pipeline creation & deployment readiness