Project Report

Global Water Scarcity & Freshwater Quality Analysis Using Data Mining Techniques

---

1. Project Overview

Water scarcity and water quality are critical global challenges. This project uses **data mining techniques** to analyze:

1. **Freshwater quality at the sample level** – classifying water as **safe or unsafe**.
2. **Global water stress at the country level** – grouping countries based on freshwater availability.

By combining **classification** and **clustering**, this project extracts actionable insights for environmental analysis, policy-making, and sustainability planning.

---

2. Objectives

- Classify water samples into **potable or non-potable** using physicochemical parameters.
- Analyze **global water stress** at the country level.
- Group countries based on **water stress severity** (low, moderate, high).
- Extract **data-driven insights** using data mining techniques.
- Prepare a **portfolio-ready, reproducible project** demonstrating practical application of data mining.

---

3. Datasets

3.1 Water Potability Dataset (Kaggle)

- Physicochemical properties of water samples.
- Features: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity.
- Target variable: Potability (0 = Unsafe, 1 = Safe).

3.2 Global Water Stress Dataset (World Bank)

- Freshwater withdrawal as a % of available freshwater.
- Year used: **2022**.
- Metadata included for indicator definition and country information.

---

4. Tools & Technologies

- **Programming:** Python
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- **Environment:** Jupyter Notebook / Python Script

---

5. Methodology

5.1 Data Loading & Cleaning

- Imported datasets using Pandas.
- Filled missing values in water quality dataset with **column mean**.
- Selected **latest year (2022)** for water stress data.

5.2 Exploratory Data Analysis (EDA)

- Visualized water quality distributions (histograms, correlation heatmap).
- Checked global water stress distribution (boxplots).

5.3 Feature Scaling

- Normalized features using **Min-Max Scaler** for consistent model performance.

5.4 Classification (Random Forest)

- Random Forest model to predict water potability.
- Split data: **80% train / 20% test**.
- Handled **class imbalance** using class_weight='balanced'.

5.5 Model Evaluation

- Accuracy score, classification report, and **confusion matrix**.
- **Feature importance** to identify most influential parameters.

5.6 Clustering (K-Means)

- Elbow method to determine optimal clusters (k=3).
- Grouped countries into:
    - Low Water Stress
    - Moderate Water Stress
    - High Water Stress
- Evaluated clustering using **Silhouette Score**.

---

6. Results

6.1 Water Quality Classification

- Random Forest **accuracy:** ~X% (replace with your output).
- Confusion matrix shows correct classification of potable vs non-potable samples.

6.2 Global Water Stress Analysis

- Countries grouped into 3 clusters based on stress percentage.
- Silhouette Score confirms good cluster separation.
- Top 15 countries with **highest water stress** include:

  [Kuwait,United Arab Emirate, Saudi Arabia, Libya, Qatar]