Predictive Analytics for Water Potability: A Comparative Study of SDG 6 Indicators Across Nations

Overview

This project aims to predict water potability using advanced machine learning techniques while conducting a comparative analysis of **Sustainable Development Goal (SDG) 6** indicators across regions. The research highlights regional disparities in water quality and provides data-driven insights to policymakers for enhancing global water sustainability efforts.

Key Features

- Machine Learning Models for Predicting Water Potability.
- Comparative Analysis of SDG 6 Indicators.
- Data Visualizations for Regional Disparities in Water Quality.
- Tools to Identify Key Predictors Influencing Water Potability.

Table of Contents

Project Objectives

1. **Predictive Modeling**
   Build machine learning models such as Gradient Boosting, SVM, Decision Tree, and Random Forest to predict water potability based on parameters like pH, turbidity, and hardness.
2. **Comparative Analysis**
   Evaluate **SDG 6 Indicators** to analyze regional disparities and progress in clean water accessibility.
3. **Policy Insights**
   Provide actionable insights and recommendations for targeted interventions and sustainable water resource management.

Dataset Description

The dataset includes the following water quality parameters:

- pH
- Hardness
- Solids
- Chloramines
- Sulfate
- Conductivity
- Organic Carbon
- Trihalomethanes
- Turbidity

Target variable: **Potable (1 for drinkable, 0 for non-drinkable)**

Methodology

1. Data Preprocessing

- Cleaning missing values and outliers.
- Feature scaling using normalization/standardization.
- Splitting the dataset into **training** and **testing** sets.

2. Model Development

- Algorithms used:
  - **Support Vector Machine (SVM)**
  - **Decision Tree**
  - **Gradient Boosting**
  - **Random Forest**
- Evaluation metrics: Accuracy, Precision, Recall, Sensitivity, and Specificity.

3. Visualization & Analysis

- Correlation Heatmaps for SDG Goals.
- World Maps for Safely Managed Services.
- Comparative plots for algorithm performance.

4. Deployment

- Best-performing model integrated as a web-based tool for real-time water quality monitoring.

Results

- **Gradient Boosting** achieved the highest accuracy: **91.67%**.
- Key predictors: **Sulfate**, **Chloramines**, and **pH**.

- Significant disparities in water quality noted between **developed** and **developing regions**.

Metric	Accuracy	Kappa	Sensitivity	Specificity	F1-Score
Random Forest	0.875	0.7333	0.9	0.8333	0.9
Gradient Boosting	0.9167	0.8182	0.9667	0.8333	0.9355
Decision Tree	0.7083	0.44	0.6	0.8889	0.72
SVM	0.8958	0.7701	0.9667	0.7778	0.9206

Technologies Used

- **Programming Language**: Python
- **Libraries**:
  - `scikit-learn` (for ML models)
  - `matplotlib` & `seaborn` (for visualizations)
  - `pandas` & `numpy` (for data processing)

How to Run

1. Clone the repository:

   ```
   bash
   Copy code
   git clone https://github.com/Yuthish3/CSE-3505-J-Component.git
   cd CSE-3505-J-Component
   ```

2. Install dependencies:

   ```
   bash
   Copy code
   pip install -r requirements.txt
   ```

3. Run the main script:

```bash
Copy code
python main.py
```

4. View results:
   - Predictions in the console.
   - Visualizations saved in the `outputs/` folder.

Contributors

- **Yuthish Kumar V** (21MIA1023)
- **Goutham S** (21MIA1014)
- **Harish A S** (21MIA1021)
- **Visva R** (21MIA1064)

License

This project is licensed under the MIT License.

For more details, check the full report here.