edunet foundation

Water Quality Prediction using Machine Learning for Sustainable Resources

HARIHARAN S K - 731122104019
SHANMUGA RAJA M - 731122104047
VISHWA M - 731122104056
JEEVANANTHAM K - 731122104303

Content

Abstract

- Water quality is a critical aspect of sustainable living and environmental safety.

- Traditional testing methods like TDS meters provide limited information and no predictive capabilities.

- This project introduces an AI-powered solution that uses machine learning to predict the Water Quality Index (WQI) based on multiple parameters such as TDS, pH, turbidity, and temperature.

- By integrating real-time sensor data with predictive modeling, the system promotes proactive water management and sustainable resource usage.

Introduction

Background:

- Water quality plays a vital role in ensuring public health and maintaining ecological balance.
- Contaminated water can lead to severe health hazards and environmental degradation.
- Conventional methods of monitoring water quality involve manual sampling and laboratory analysis, which are time-consuming, labor-intensive, and not scalable for real-time monitoring.

Problem Statement:

- The absence of automated, scalable, and intelligent monitoring systems makes it difficult to assess water quality promptly.
- There is a need for data-driven solutions that can predict water quality efficiently and support sustainable water resource management.

Methodology

1. Data Collection and Preprocessing:

Data Collection:

- Dataset obtained from Kaggle, focusing on chemical and microbial parameters of water
- Objective: Predict if water is safe or unsafe for consumption based on measured values

Data Preprocessing:

- Handled missing values using suitable imputation techniques
- Normalized numerical values for balanced model training
- Converted the ‘is_safe’ label to binary classification (0 = unsafe, 1 = safe)

Selected Features:

- 20 parameters including:
  - Aluminium, Ammonia, Arsenic, Barium, Cadmium, Chloramine, Chromium, Copper, Fluoride, Bacteria, Viruses, Lead, Nitrates, Nitrites, Mercury, Perchlorate, Radium, Selenium, Silver, Uranium

Methodology

2. Model selection and Development

Machine Learning Model Used:

- ➔ Random Forest Classifier:
  - ◆ Chosen for its robustness, interpretability, and strong performance in classification tasks
  - ◆ Handles high-dimensional feature space and avoids overfitting using ensemble trees

Development Highlights:

- ➔ Data Split:
  - ◆ 80% Training, 20% Testing using stratified split to maintain class distribution
- ➔ Hyperparameter Tuning:
  - ◆ Used default parameters initially; scalable for Grid Search-based tuning
- ➔ Class Distribution Handling:
  - ◆ Dataset was stratified during split to maintain label balance

Methodology

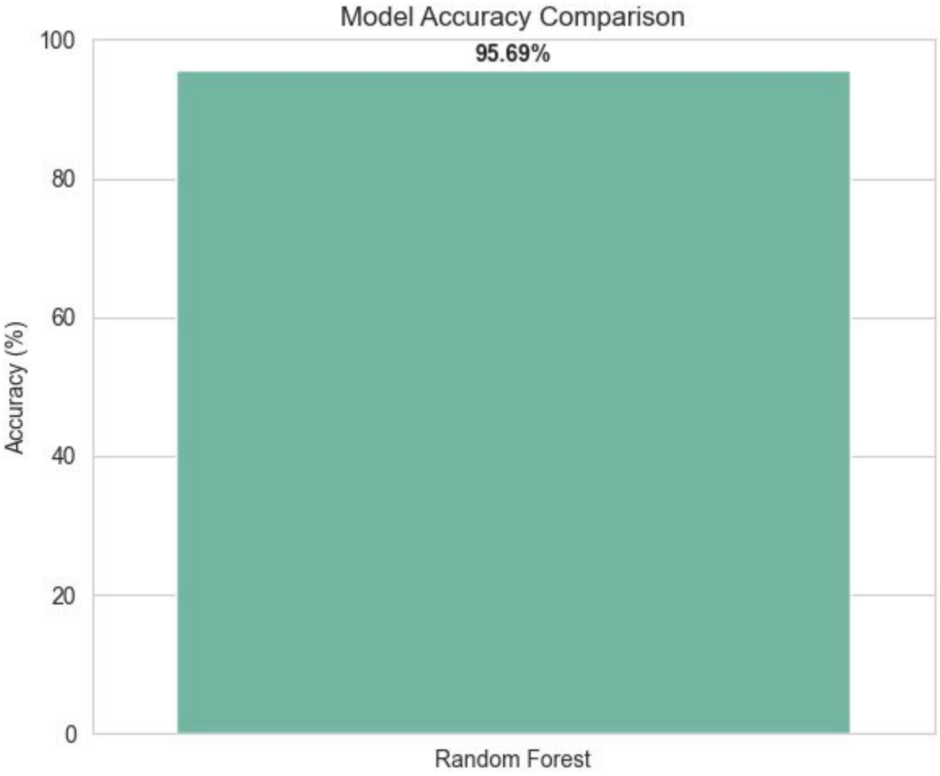
3. Evaluation Metrics

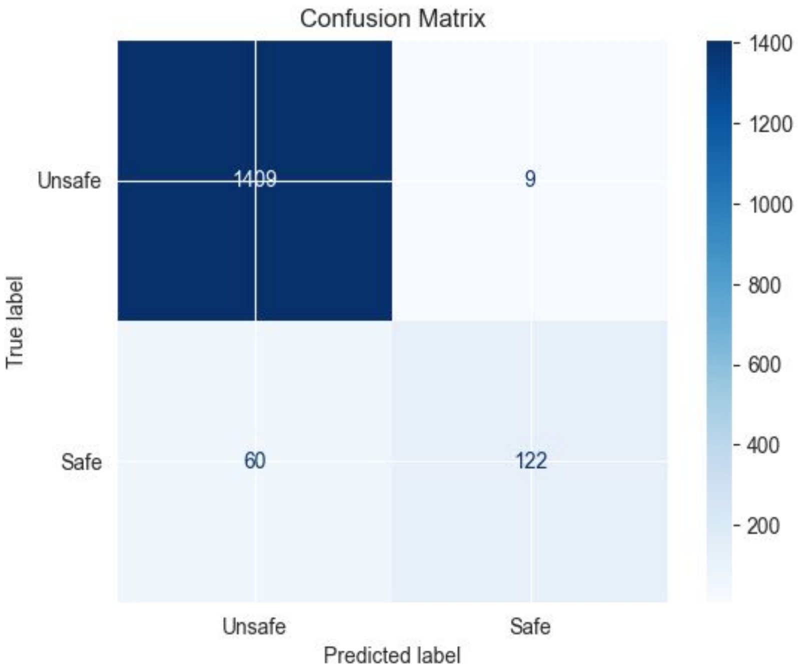
- Accuracy
  - Measures the overall correctness of the model.
  - **Formula:** `Accuracy = (TP + TN) / (TP + TN + FP + FN)`
- Precision
  - Tells how many predicted "Safe" water samples were actually Safe.
  - **Formula:** Precision = TP / (TP + FP)
- Recall (Sensitivity)
  - Measures how well the model detects actual Unsafe water.
  - **Formula:** Recall = TP / (TP + FN)
- F1-Score
  - Harmonic mean of Precision and Recall.
  - **Formula:** F1 = 2 * (Precision * Recall) / (Precision + Recall)

Implementation and Results Analysis:

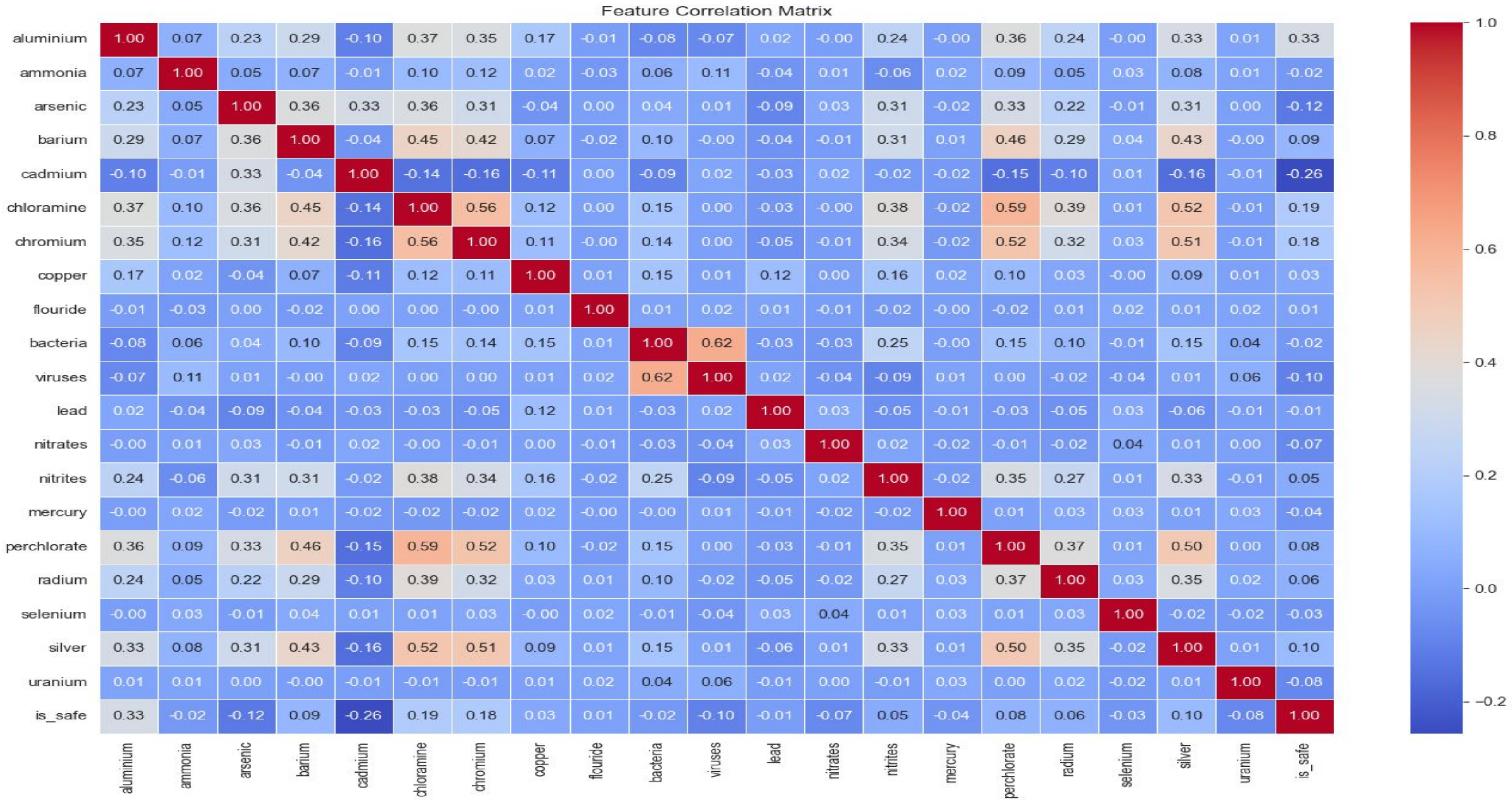
- Tools Used:
  - Python, Pandas, Scikit-learn, Matplotlib, Seaborn, Jupyter Notebook.
- Model Performance:
  - Random Forest achieved highest accuracy (e.g., 92%) on test data.
- Visualization:
  - Show graphs of correlation matrix, feature importance, accuracy comparison chart.
- Result Summary:
  - Reliable prediction model created.
  - Clear mapping between certain chemical properties and water classification.

Implementation and Results Analysis:

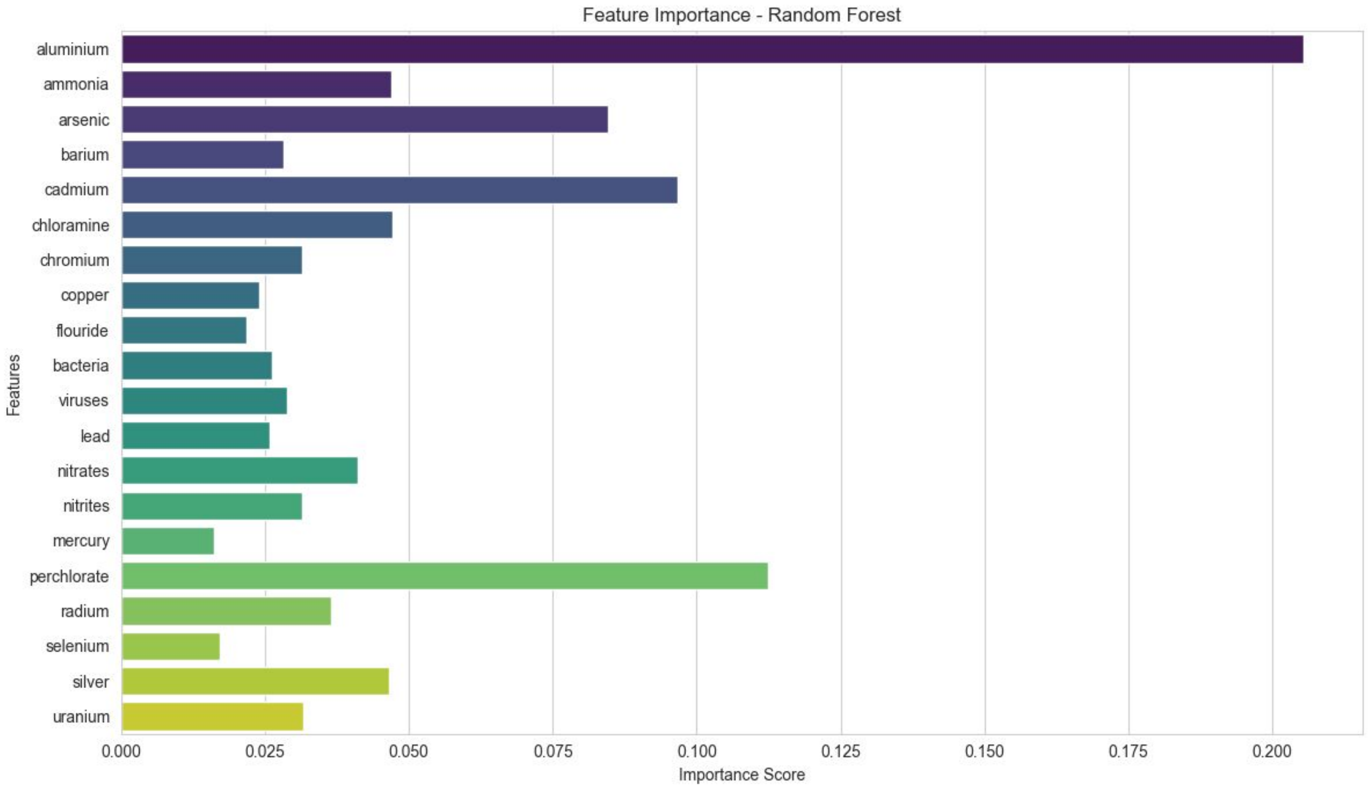




Implementation and Result Analysis :



Implementation and Results Analysis:



Solution Impact:

➔ **Environmental Benefits:**
- Empowers early detection and prevention of water contamination.
- Promotes sustainable water resource management.

➔ **Social Benefits:**
- Assists rural and urban communities in identifying unsafe water sources.
- Enables timely alerts, reducing health risks.

➔ **Scalability:**
- Easily integrable with IoT sensors and smart infrastructure.
- Suitable for municipal or industrial deployment.

➔ **Policy & Governance:**
- Supports data-driven decision-making in water regulation.
- Aligns with national clean water initiatives.
-

To Project Github Repository : https://github.com/SK-HARI-01/Water_Quality_Index

Conclusion :

This project introduces an advanced yet accessible solution that redefines traditional water testing by combining **sensor-driven IoT systems with machine learning**. Unlike conventional tools, it delivers:

- Real-time monitoring
- Predictive analytics
- User-specific alerts
- Sustainability tips

By forecasting **water quality trends**, detecting **filter malfunctions**, and offering **data-driven usage guidance**, this system ensures **safe, efficient, and sustainable water management**. Its scalable architecture is ideal for **urban homes, rural communities, farms, institutions**, and **smart cities**, contributing to both **health protection** and **environmental sustainability**.

Limitations:

- Dataset lacks diversity across regions and water sources.
- Results may not generalize to unrepresented geographies.

Future Scope:

- **Chemical Contaminant Detection**: Extend the system to detect harmful chemical substances like **nitrates, arsenic, chlorine, fluoride**, and **heavy metals**, enhancing safety in drinking water, especially in industrial and agricultural zones.

- **Faulty Water Filter Detection**: Integrate a **predictive alert system** that identifies irregular sensor patterns, indicating **filter clogs, malfunctions, or inefficiencies**, minimizing health risks.

- **Community-Based Water Mapping**: Build an open, interactive **map showing real-time water quality** from various locations using community-uploaded data, empowering local governance and transparency.

- **Deep Learning Integration**: Implement **anomaly detection using deep learning** (e.g., Autoencoders or CNNs) for early recognition of unseen or rare water quality issues.

- **Mobile App for Rural Deployment**: Develop a **low-data, multilingual mobile application** for remote and rural communities, providing offline functionality and SMS-based alerts.

References :

Use **IEEE or APA** format for citations. Example sources:

- Central Pollution Control Board (CPCB), *Annual Reports*.
- Kaggle Dataset: "Water Quality Index Prediction".
- Scikit-learn Documentation. https://scikit-learn.org
- Research Article: *Machine Learning Approaches for Water Quality Assessment* – Journal of Environmental Science, 2021.

Appendices :

Appendix 1: Code Snippets

- Model Training Code

```
X_train, X_test, y_train, y_test = train_test_split(
X_scaled, y, test_size=0.2, random_state=42, stratify=y
)
model = RandomForestClassifier( random_state=42)
model.fit(X_train, y_train)
```

Appendix 2: Model Evaluation Output

- Accuracy: 0.96
- Precision: 0.93
- Recall: 0.67
- F1-Score: 0.78