WATER QUALITY ANALYSIS

A PROJECT REPORT

PHASE V

A report submitted in fulfilment of the project

Of

DATA ANALYTICS WITH COGNOS - GROUP 1

In

NAAN MUDHALVAN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COLLEGE OF ENGINEERING GUINDY

ANNA UNIVERSITY

CHENNAI 600025

submission on

30-OCT-2023

Submitted by

NIKHIL PRASANNA A	2021103026
PARTHIBAN R	2021103554
SHANMUGAPRIYAN T	2021103045
SUDEEP R	2021103053
THARUNSRI P	2021103311

ABSTRACT:

This project focuses on water quality analysis, employing machine learning and data-driven methods. The goal is to create a real-time detection system for contaminants. Utilizing a structured design thinking approach, it encompasses data collection, visualization using IBM Cognos, and Python-based machine learning. Insights gained benefit website owners, enhancing user experience by providing real-time water quality information and safety recommendations.

INTRODUCTION:

Water quality analysis is of paramount importance for ensuring the safety of drinking water sources. This documentation outlines a comprehensive project that leverages machine learning and data analysis to achieve this goal. The primary objective is to develop an efficient water quality analysis system that can detect and quantify contaminants in real-time, allowing us to safeguard public health and environmental safety.

We follow a structured design thinking process, beginning with empathy, where we understand the significance of clean and safe drinking water. Defining the project's scope, we aim to create a robust system that encompasses data collection, analysis, and visualization. By implementing this system, we not only fulfil our analysis objectives but also provide valuable insights to website owners.

These insights, once integrated into websites, can significantly improve the user experience by offering real-time information on water quality, recommendations, and safety precautions.

OBJECTIVE:

The project's primary goal is to create a water quality analysis system that uses machine learning to ensure the safety and purity of drinking water sources. This system is essential to safeguard public health by continuously monitoring water quality and detecting contaminants.

DESIGN THINKING PROCESS:

1.Data Collection:

The dataset used in this project comes from Kaggle which is a high-fidelity database with five-star research rating in google scholar. There are nine training variables, such as pH, hardness, solids, chloramines etc… The dataset link is Water Potability Dataset

2. Data Preprocessing:

Once the dataset has been collected, it is important to clean it and prepare it for machine learning. This may involve removing missing values, outliers, and scaling the data.

3. Exploratory Data Analysis (EDA):

Use interactive data visualization tools to create an immersive and user-friendly EDA interface. There are a variety of EDA techniques that we can use, such as histograms, box plots, scatter plots, and heatmaps.

4. Feature Engineering:

This involves creating new features or deriving relevant features from existing ones.

5. Machine Learning:

Split the dataset into training and testing sets for model evaluation.

The model of machine learning can be divided into supervised learning, semi-supervised learning, and unsupervised learning. There are many kinds of algorithms for each learning mode, such as decision trees, naive Bayes, random forests, etc. Each algorithm model has its own advantages and disadvantages.

Among the many machine learning methods, artificial neural networks and support vector machine algorithms became popular in machine learning due to their large processing data and fast calculation speed. Therefore, selecting the above two algorithms to judge the drinking ability of water resources is expected to better achieve the desired purpose.

6. Model Evaluation:

Once the model has been trained, we can evaluate its performance on the testing set. We can use a variety of metrics to evaluate the model's performance, such as accuracy, precision, recall, F1-score, and ROC AUC.

7. Model Tuning:

If the model's performance is not satisfactory, we can try tuning the hyperparameters of the model. Hyperparameters are parameters that control the learning process of the model.

8. Interpretation and Insights:

Once we have trained a satisfactory model, we can interpret the model results to understand which water quality parameters are most influential in determining potability.

9. Continuous Improvement:

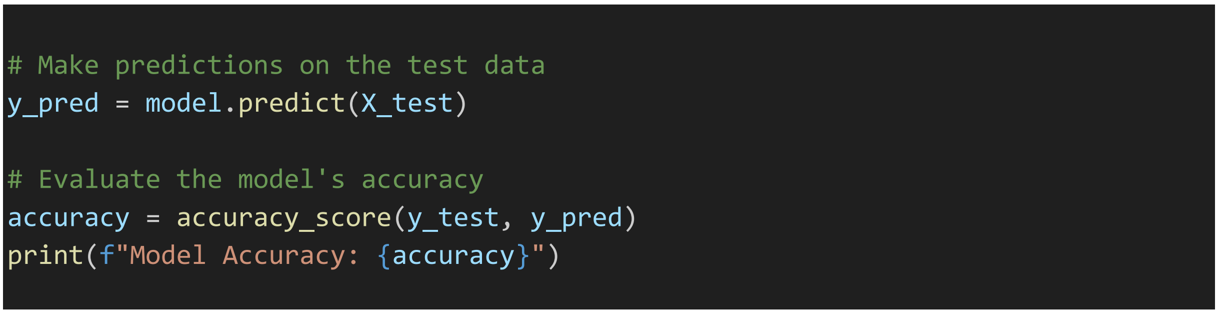
We can implement a feedback loop for continuous model improvement. If possible, connect the model to water treatment facilities and use the feedback from water quality improvements to adapt the model in real-time.

IMPLEMENTATION :

Dataset : " water_potability.csv"

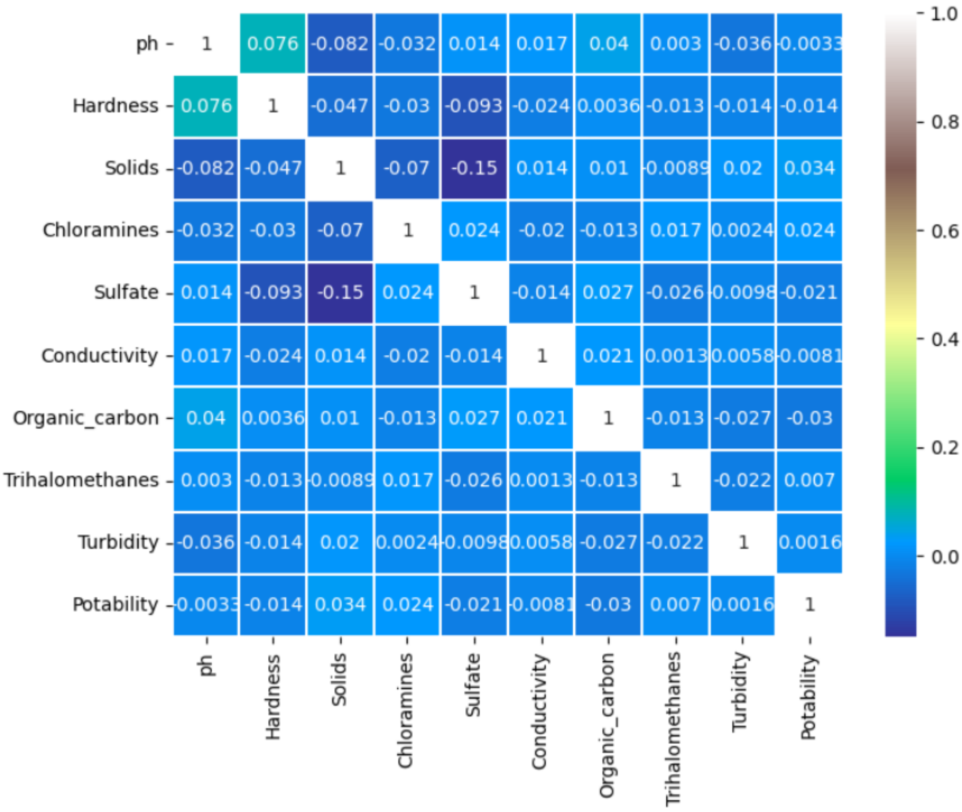
Sample Code :

```
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

# Load the hypothetical water quality dataset
data = pd.read_csv("water_quality_data.csv")

# Data preprocessing (e.g., handling missing values, encoding categorical
data)

# Split the data into features (X) and target (y)
X = data.drop(columns=["Water_Quality"])
y = data["Water_Quality"]

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Create a machine learning model (Random Forest classifier, for example)
model = RandomForestClassifier()

# Train the model on the training data
model.fit(X_train, y_train)
```



```
# Make predictions on the test data
y_pred = model.predict(X_test)

# Evaluate the model's accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Model Accuracy: {accuracy}")
```

Data Analysis Sample Output:



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
ph	1	0.076	-0.082	-0.032	0.014	0.017	0.04	0.003	-0.036	-0.0033
Hardness	0.076	1	-0.047	-0.03	-0.093	-0.024	0.0036	-0.013	-0.014	-0.014
Solids	-0.082	-0.047	1	-0.07	-0.15	0.014	0.01	-0.0089	0.02	0.034
Chloramines	-0.032	-0.03	-0.07	1	0.024	-0.02	-0.013	0.017	0.0024	0.024
Sulfate	0.014	-0.093	-0.15	0.024	1	-0.014	0.027	-0.026	-0.0098	-0.021
Conductivity	0.017	-0.024	0.014	-0.02	-0.014	1	0.021	0.0013	0.0058	-0.0081
Organic_carbon	0.04	0.0036	0.01	-0.013	0.027	0.021	1	-0.013	-0.027	-0.03
Trihalomethanes	0.003	-0.013	-0.0089	0.017	-0.026	0.0013	-0.013	1	-0.022	0.007
Turbidity	-0.036	-0.014	0.02	0.0024	-0.0098	0.0058	-0.027	-0.022	1	0.0016
Potability	-0.0033	-0.014	0.034	0.024	-0.021	-0.0081	-0.03	0.007	0.0016	1

Figure 1. Correlation map

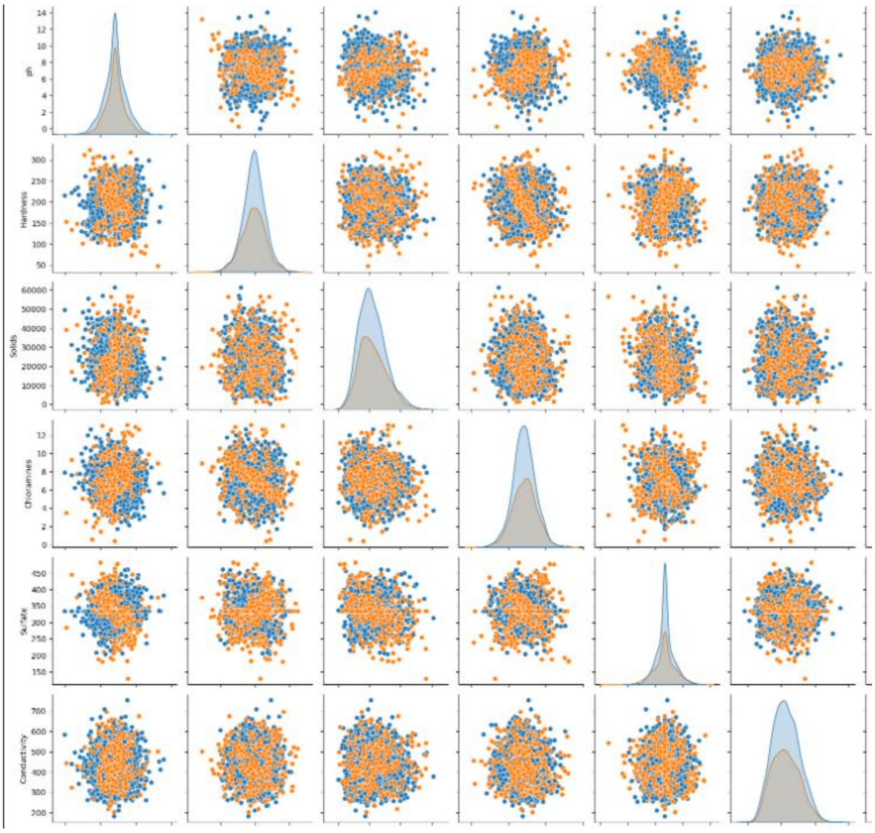


Figure 1. Probability Pairwise

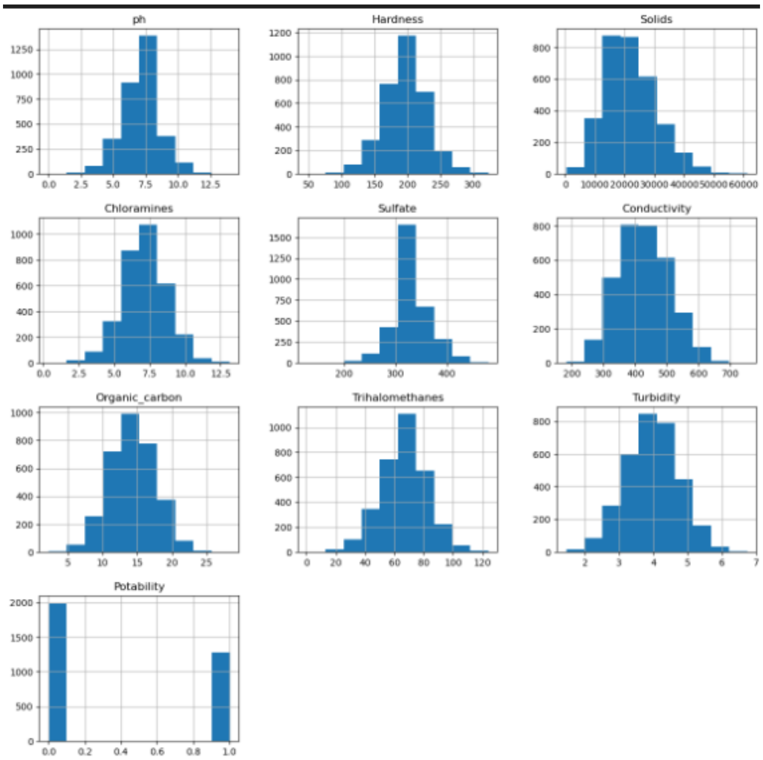


Figure 1. Fig size

DEVELOPMENT PHASES:

1. Analysis Objectives:

- ✓ The key objectives are to detect and quantify contaminants in water sources. This includes heavy metals, bacteria, and chemical pollutants.
- ✓ The system should provide real-time or timely results to protect consumer health and environmental safety.

2. Data Collection Process:

- ✓ To achieve the analysis objectives, data is collected from various water sources. This data includes information on water quality parameters such as pH, turbidity, chemical compositions, and the presence of bacteria.
- ✓ Data is recorded at regular intervals and stored in a structured database for analysis.

3. Data Visualization using IBM Cognos:

- ✓ IBM Cognos is employed to create interactive and informative data visualizations.
- ✓ Dashboards and reports are designed to present water quality trends in a clear and accessible manner.
- ✓ Geographical maps are used to visually identify contamination hotspots, helping users understand the spatial distribution of water quality issues.

4. Python Code Integration:

- ✓ Machine learning models are developed using Python, with popular libraries like Scikit-Learn.
- ✓ These models are trained to predict water quality based on historical data.
- ✓ The Python code is integrated into the system to provide real-time analysis and predictions.

5. Improving User Experience:

- ✓ The insights gained from water quality analysis can be shared with website owners.Website owners can use this data to inform the public about water quality issues, promoting transparency and safety.
- ✓ Websites can offer real-time quality updates, recommend safe water sources, and provide safety guidelines, enhancing the user experience.Ultimately, the project

contributes to public health by ensuring that users have access to vital water quality information.

EXPLANATION: ENHANCING WATER QUALITY ANALYSIS

- **Data Collection Process:**

The first crucial phase of this project involves the systematic collection of data from various water sources. This data comprises a spectrum of water quality parameters, such as pH levels, turbidity, chemical compositions, and the presence of harmful bacteria. A structured database is employed to organize and store this data. Collecting data at regular intervals ensures the availability of a comprehensive dataset, which serves as the foundation for subsequent analysis.

- **Data Visualization using IBM Cognos:**

IBM Cognos is a powerful tool for translating raw data into visually comprehensible information. Through the creation of interactive dashboards, reports, and data visualizations, water quality trends become evident to stakeholders. Geo-mapping is also employed to pinpoint contamination hotspots, making it easier to identify areas with potential water quality issues. These visualizations not only enhance understanding but also facilitate informed decision-making.

- **Python Code Integration for Analysis:**

Machine learning models, developed using Python and libraries like Scikit-Learn, play a pivotal role in this project. These models are trained using historical water quality data, allowing for real-time analysis of water samples. The Python code is seamlessly integrated into the system, ensuring that the analysis is conducted promptly and accurately.

- **Improving User Experience:**

By sharing the insights gained through water quality analysis with website owners, we extend the benefits of this project to a broader audience. Website owners can leverage this data to enhance the user experience by providing real-time water quality updates, recommendations for safe water sources, and safety guidelines to the public. This way, users can make informed decisions about their water consumption and protect their health. By enhancing the user experience in this manner, website owners contribute to public health and provide a valuable service to their audience.

CONCLUSION:

In conclusion, this project integrates data-driven solutions to improve water safety, demonstrating the potential for positive change and more transparent access to clean drinking water.