Water Quality Analysis

Objective

The primary objective of this project is to analyze water quality data provided by the user and build a predictive model to determine water potability based on key water quality parameters. By analyzing the data and developing a predictive model, the project aims to assist in assessing the quality of water sources and ensuring the safety of drinking water

Introduction

This Python Project code is designed for water quality analysis using a user-provided dataset (in CSV format) and building a predictive model to determine water potability based on water quality parameters. It provides a brief overview of the dataset, explores its characteristics, visualizes data distributions, and uses a Random Forest classifier to predict water potability.

Prerequisites

- Python 3.x installed
- Required libraries: pandas, matplotlib, seaborn, scikit-learn

To install the necessary libraries, you can use pip:

“pip install pandas matplotlib seaborn scikit-learn”

Usage

Loading the Dataset: The user is prompted to provide the path to the water quality dataset in CSV format.

Data Exploration and Analysis:

- Summary Statistics: Summary statistics of the dataset, including mean, standard deviation, and quartiles.
- Missing Values: Checks for missing values in the dataset.
- Histograms: Displays histograms for water quality parameters, comparing potable and non-potable water.
- Correlation Matrix: Visualizes the correlation matrix for feature relationships.

Data Preprocessing:

- Missing values are filled with the mean of each respective column.
- Feature scaling is applied using `StandardScaler`.

Model Training:

- The dataset is split into training and testing sets.
- A Random Forest Classifier is trained on the training data with 100 trees (you can adjust the number of trees as needed).

Model Evaluation:

- The model is used to make predictions on the testing data.
- The accuracy of the model is calculated and displayed.
- A classification report is generated, providing detailed performance metrics.

Design Thinking Process

The project follows a structured approach:

Understanding the Problem: Recognizing the significance of water quality for public health and the need for a reliable method to assess water potability.

1 .Data Gathering: Collecting a user-provided water quality dataset in CSV format.

2 .Data Exploration and Analysis: Analyzing the dataset to gain insights into its characteristics and uncover patterns related to water potability.

3 .Data Visualization: Creating visual representations of data distributions, including histograms and correlation matrices.

4 .Predictive modeling: Building a machine learning model (Random Forest) to predict water potability based on water quality parameters.

5 .Model Evaluation: Assessing the model's performance using accuracy and a classification report.

Development Phases

1. Data Exploration and Analysis

- Summary Statistics: Calculate mean, standard deviation, quartiles, and other basic statistics for the dataset.
- Missing Values: Check for and address any missing values in the data.

2. Data Preprocessing

- Handling Missing Values: Fill missing data with the mean of each respective column.
- Feature Scaling: Standardize features to bring them to a common scale using `StandardScaler`.

3. Data Visualization

- Histograms: Visualize data distributions for each water quality parameter, distinguishing between potable and non-potable water.
- Correlation Matrix: Explore feature relationships using a correlation matrix visualization.

4. Predictive Modeling

- Model Training: Split the data into training and testing sets.
- Random Forest Classifier: Train a machine learning model using Random Forest, an ensemble learning method.
- Model Evaluation: Assess the model's performance in predicting water potability based on the provided data.

Analysis Objectives

The analysis objectives include:

- Providing a comprehensive overview of water quality data.
- Identifying relationships between water quality parameters and water potability.
- Developing a predictive model to assist in determining water potability.

Insights and Assessment

The insights gained from this analysis can help:

- Water authorities and regulatory agencies assess the quality of water sources.
- Ensure the safety of drinking water by identifying potential sources of non-potable water.
- Make informed decisions for water treatment and purification processes.
- Improve public health and environmental quality.

Conclusion

This code serves as a starting point for water quality analysis and prediction based on a provided dataset. By combining data analysis, visualization, and predictive modeling, this project provides a valuable tool for evaluating water quality and ensuring the availability of potable water .It can meet specific requirements and further improve predictive accuracy.