For the whole coding part, I assure that AI is not used in this assignment.

Problem statement

We aim to forecast the spatio-temporal water quality, specifically the "power of hydrogen (pH)" value for the next day, based on historical data of various water measurement indices. The dataset comprises daily samples from 36 sites, focusing on pH measurements in Georgia, USA. The input features include 11 common indices, such as dissolved oxygen volume, temperature, and specific conductance (details available in the dataset). The target variable for prediction is pH, water, unfiltered, field, and standard units (Median).

There are two major water systems of interest: one centered around Atlanta and the other along the eastern coast of Georgia. This spatial distribution introduces dependencies between different locations, which is crucial for accurate forecasting. To address this challenge, we will employ machine learning techniques, encompassing data preprocessing, feature engineering, and model development.

Explanation of code and understanding

To address the mentioned problem, we have employed the XGBoost algorithm for modeling. XGBoost is a gradient-boosting tree method based on decision tree construction. Gradient boosting trees are an ensemble learning technique that progressively builds multiple decision trees to enhance the model's predictive performance. Each tree aims to correct the prediction errors of the previous tree. Additionally, we have visualized the selected features of the model. Visualizing feature importance is a method for understanding the contribution of each feature (input variable) to the model's predictions. This visualization provides insights into which features have the most significant impact on the model's performance, aiding in feature selection and explaining model behavior.

Finally, we evaluated the model's performance using the Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R-squared (R2) coefficient. MSE, which measures the average squared difference between predicted and actual values, assesses the model's fit by calculating the average of these squared differences. A lower MSE indicates a better fit of the model to the actual values. MAE, on the other hand, quantifies the average absolute difference between predicted and actual values. A smaller MAE suggests a better fit of the model to the actual values. The coefficient of determination (R2) is a measure of model fit, indicating the proportion of the variance in the dependent variable explained by the model. R-squared typically ranges from 0 to 1, with values closer to 1 signifying a better fit of the model to the data.

We used the first three locations as calculating examples. The prediction results and figures got are saved in the prepared dataset. When we need to analyze data from other locations, we only need to enter the data corresponding to that location. In future web apps, the application of this strategy is that when the user clicks the button of the location they want to know about, the corresponding data will be entered into the program, thereby producing the corresponding results.

product overview

App 1:

App 1 is a REST API developed to provide software developers with access to water quality data for integration into their web applications. The API is designed to facilitate the retrieval and utilization of historical and real-time water quality information, including parameters such as pH levels and various water quality indices. It serves as a valuable resource for developers seeking to enhance their applications with environmental monitoring and management capabilities.

The primary audience for App 1 is software developers who are proficient in web application development and require water-quality data in their projects. These developers may work in various sectors, including environmental monitoring, water resource management, or any field where access to water quality information is essential. They have a technical background and are experienced in API integration.

The main features of the app include:

1. Data Access: App1 allows developers to access a wide range of water quality data, such as pH levels, dissolved oxygen volume, temperature, and specific conductance, from various monitoring sites.
2. Integration: Developers can easily integrate the API into their web applications, allowing them to retrieve real-time and historical water quality data without the need to collect or maintain the data themselves.
3. Customization: The API provides flexibility for developers to tailor the data to their specific needs, allowing them to create applications that serve diverse purposes, from environmental research to water quality monitoring.
4. Real-time Updates: Developers can access up-to-date water quality information, enabling their applications to reflect the most current environmental conditions.

App 2:

We have developed a water quality prediction application (APP) aimed at assisting water resource managers and decision-makers in better-utilizing water quality data to forecast the pH value of water bodies for the next day. This APP is based on historical data, including various water quality measurement indices from 36 different locations, with a primary focus on the pH values in the USA. Users can easily access data from different locations by clicking on various buttons, allowing them to take timely actions as needed. Additionally, the XGBoost algorithm designed enhances predictive performance, as well as the provided feature of visualizations helping users understand the contributions of different indices to the model's performance. Finally, the Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R-squared (R2) coefficient are designed to assess the model’s performance to ensure a good fit of the model to actual values.

The main features of this APP include:

1. Viewing water quality data from different locations.
2. Predicting pH values for the next day using historical data.

3. Visualizing important features to understand key contributing factors to predictions.
4. Evaluating model performance for accuracy.
5. Assisting water resource managers and decision-makers in better managing water quality data to take appropriate actions.

Persona

For App2, the target audience is water resource managers and decision-makers in the field of environmental monitoring and management. We can create a persona to represent this audience:

Background:
Sarah Waterford is a seasoned water resource manager with over 15 years of experience in the field. She holds a degree in environmental science and has a deep understanding of water quality parameters and their implications. Sarah is responsible for overseeing water quality management in a large region of Georgia, USA.

Goals and Needs:
Sarah's primary goal is to ensure the safety and quality of water bodies under her jurisdiction. She needs a tool that can help her make data-driven decisions regarding water quality. Quick access to accurate water quality predictions, especially pH levels, is crucial for her daily tasks. She requires a user-friendly platform that allows her to view data from multiple monitoring sites and assess potential issues.

Challenges:
Managing water quality data from multiple locations can be overwhelming, and she needs a solution that simplifies this process. Accurate forecasting of pH levels is essential to take timely actions and prevent water quality issues. Sarah needs an easy-to-use interface as she may not have a technical background in data analysis or modeling.

How App2 Helps Sarah:
App2 provides Sarah with a user-friendly interface to access and visualize water quality data from various sites in Georgia. The pH prediction feature helps her anticipate changes in water quality and make informed decisions. Visualizations of feature importance assist her in understanding the key factors affecting pH predictions. The performance evaluation metrics (MAE, MSE, R2) ensure that she can rely on the model's accuracy.

Overall, App2 caters to Sarah Waterford's needs as a water resource manager, helping her effectively manage water quality and make informed decisions to maintain the safety and quality of water bodies in her region.

Source Code Control

https://github.com/LastTreasure/COMP0035-CW1-Final.git