WATER QUALITY PREDICTION

SUBMITTED BY (TEAM 4)

SIRI H G
SHRADDHA SHRESTHA
MALLIKARJUN AITHA
MADHAVI KANCHAM

SUBMITTED FOR
DSCI-6001 (FALL 2022) DR. ARDIANA SULA

CONTENTS

1. INTRODUCTION

This project explores the application of machine learning techniques to predict water potability, a critical aspect of public health. Leveraging diverse water quality indicators, including pH, hardness, and chemical concentrations, various models are employed to forecast the safety of drinking water. The study addresses imbalanced datasets, ethical considerations, and feature importance analysis, aiming to contribute to effective water quality management. As access to clean water remains a global concern, the intersection of data science and public health holds the potential for proactive solutions and improved decision-making in ensuring safe drinking water for communities worldwide.

2. EXECUTIVE SUMMARY

The project aimed to develop a predictive model for water potability assessment, leveraging various machine learning algorithms. Through a thorough analysis of water quality data, we sought to enhance the understanding of factors influencing potability and contribute to the improvement of water safety measures. Our water potability prediction project encapsulates a comprehensive exploration of machine-learning techniques to assess and predict the safety of drinking water. This endeavor stands out for its multifaceted approach, encompassing data preprocessing, feature engineering, and the application of various predictive models. Here, we highlight the key aspects that define the significance and uniqueness of our project.

3. BACKGROUND THEORY

The project draws on fundamental principles of machine learning and water quality analysis. Machine learning algorithms, such as Logistic Regression, Decision Trees, and Random Forests, are applied to predict water potability based on key indicators. The study leverages the concept of imbalanced datasets, addressing biases to enhance model performance. Feature importance analysis provides insights into variables influencing predictions. Ethical considerations emphasize transparency and bias mitigation in deploying models for public health. The project forms a bridge between data science and water quality management, aiming to contribute to proactive measures for ensuring safe drinking water As technology advances, this intersection of data science and public health holds promise for proactive water quality management. Leveraging machine learning insights can contribute significantly to ensuring safe drinking water, fostering sustainable practices, and mitigating health risks. The journey from predictive modeling to real-world impact requires collaborative efforts, ongoing refinement, and a commitment to the broader societal well-being.

4. METHODOLOGY

4.1. Data collection

The dataset used in this project is sourced from "water_potability.csv," encompassing a comprehensive set of water quality indicators. The data includes crucial parameters such as pH, hardness, and chemical concentrations. The initial exploration involves understanding the dataset's dimensions, statistical measures, and the identification of missing values. An insightful heatmap visualizes the distribution of missing values, guiding subsequent data handling strategies.

Exploration reveals the existence of missing values in columns, including pH, Sulfate, and Trihalomethanes. To maintain data integrity, outliers in these columns are scrutinized. The exploration culminates in imputing missing values using appropriate statistical measures, ensuring a robust dataset for subsequent analysis.

This meticulous data collection process sets the stage for meaningful insights into water potability, contributing to the project's overarching goal of leveraging machine learning for predictive modeling in water quality management.

- pH
- Hardness
- Solid
- Chloramines
- Sulfate
- Conductivity
- Organic_carbo
- Trihalomethane
- Turbidit
- Potability
-

4.2. Machine Learning Algorithms for Classification

Machine learning algorithms have become increasingly important in the field of healthcare research. These algorithms have the potential to revolutionize the way medical research is conducted and have already started to have an impact in liver disease research. Machine learning algorithms can be used

to analyze large datasets, identify patterns, and make predictions about potential treatments and outcomes. They can also be used to detect potential health risks and detect early signs of disease. As such, the use of machine learning algorithms in healthcare research is rapidly increasing, and they are playing a key role in improving our understanding of both the causes and potential treatments of liver disease. The most used machine learning algorithms for diagnosing and predicting liver disease are K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, and Support Vector Machines (SVM).

This project employs diverse machine learning algorithms to predict water potability based on key features. The algorithms selected for classification tasks include:

4.2.1 Logistic Regression:

Logistic Regression is a supervised learning algorithm commonly used for binary classification problems. Unlike linear regression, which predicts continuous outcomes, logistic regression predicts the probability of an instance belonging to a particular class (0 or 1). The logistic function (sigmoid) is employed to map the output to a probability range between 0 and 1.

In this project, Logistic Regression is suitable for predicting whether water is potable (1) or not (0) based on various water quality indicators. It models the relationship between these indicators and the likelihood of safe drinking water.

4.2.2 Decision Tree Classifier:

Decision Tree is a supervised learning algorithm that can be used for both classification and regression tasks. It works by recursively partitioning the data based on the most significant features, creating a tree-like structure. Each node represents a decision, and each leaf node represents a class label.

Decision Tree Classifier is employed to classify water samples into potable and non-potable categories based on the values of different water quality indicators. It helps to identify the critical features that contribute to the decision-making process.

4.2.3 Random Forest Classifier:

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and merges them to get a more accurate and stable prediction. It operates by aggregating the predictions of multiple trees, reducing overfitting, and improving generalization.

Random Forest Classifier is beneficial for our project due to its ability to handle complex datasets and minimize the risk of overfitting. It combines multiple decision trees to enhance the accuracy of water potability predictions.

4.2.4 K-Nearest Neighbors (KNN):

K-Nearest Neighbors is a simple and intuitive algorithm used for both classification and regression. It classifies an instance based on the majority class of its k-nearest neighbors in the feature space. KNN is a non-parametric and lazy learning algorithm.

In this project, KNN is applied to classify water samples based on the similarity of their feature values. It considers the proximity of instances in the feature space to determine the potability of water.

4.2.5 Support Vector Classifier (SVM):

Support Vector Classifier, or SVM, is a powerful algorithm for classification and regression tasks. It finds a hyperplane that best separates instances of different classes in a high-dimensional space. SVM is effective in scenarios with a clear margin of separation between classes.SVM is utilized in our

project to predict water potability. It aims to find the optimal hyperplane that separates safe drinking water instances from unsafe ones, considering the complex relationships between water quality indicators.

4.2.6 Naive Bayes:

Naive Bayes is a probabilistic algorithm based on Bayes' theorem. It assumes independence among features, even though this assumption might not hold in real-world scenarios. Naive Bayes is widely used in text classification and has applications in various domains. In this project, Naive Bayes is employed to predict water potability based on the probability distribution of features. It calculates the likelihood of water being safe or unsafe given the observed values of water quality indicators.

4.2.7 XGBoost:

XGBoost is an implementation of gradient boosting designed for speed and performance. It is an ensemble learning technique that builds a series of weak learners (usually decision trees) and combines their predictions to create a strong learner.

While XGBoost is imported in the project, it's not explicitly used in the water potability prediction. However, its potential lies in its ability to improve predictive accuracy and handle complex relationships in the data.

In essence, Each algorithm undergoes rigorous evaluation using metrics such as accuracy, precision, recall, and F1-score to assess its effectiveness in predicting water potability. The comparative analysis guides the selection of the most suitable model for deployment in real-world scenarios, contributing to advancements in water quality management.

5. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is a crucial phase that provides insights into the characteristics of the dataset, facilitating informed decisions during the modeling process.

5.1. Basic Exploration:

Dataset Overview: The dataset, sourced from "water_potability.csv," comprises water quality indicators, including pH, hardness, and chemical concentrations.

Dimensions: The dataset contains 3276 entries with 10 columns.

Column Names: Key features include pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and the target variable, "Potability."

5.2. Understanding the Data:

Target Variable: Potability (0: Not Potable, 1: Potable)

Statistical Measures: Descriptive statistics provide an overview of the dataset's central tendencies.

Duplicate Rows: Checked for and identified any duplicated entries in the dataset.

5.3. Detecting Missing Values:

Null Values: Columns like pH, sulfate, and trihalomethanes contain missing values.

Visualization: A heatmap visually represents the distribution of missing values, aiding in identifying patterns.

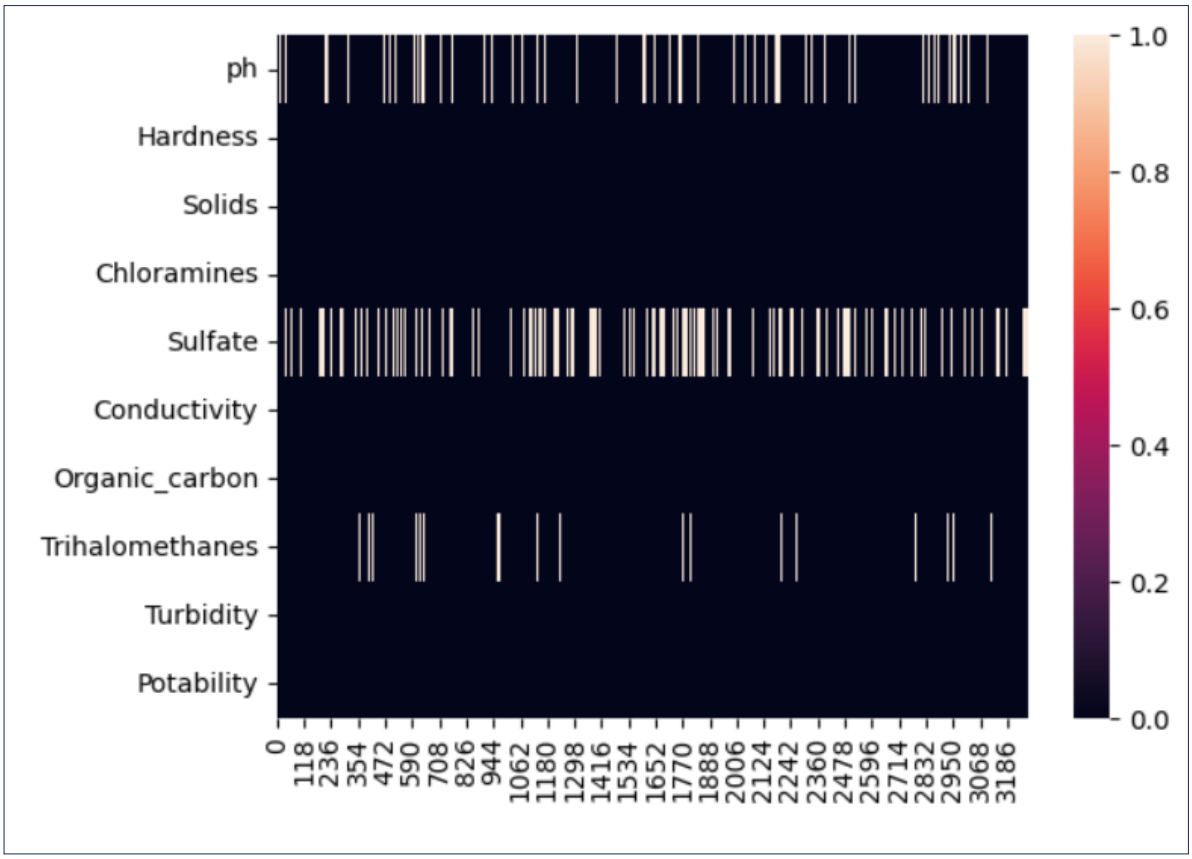

Fig. 5.1. Heat map to visualize missing data

5.4. Handling Missing Values:

Outlier Detection: Outliers in pH, sulfate, and trihalomethanes are examined, considering their impact on data analysis.

Imputation: Missing values are filled using median values for the respective columns.

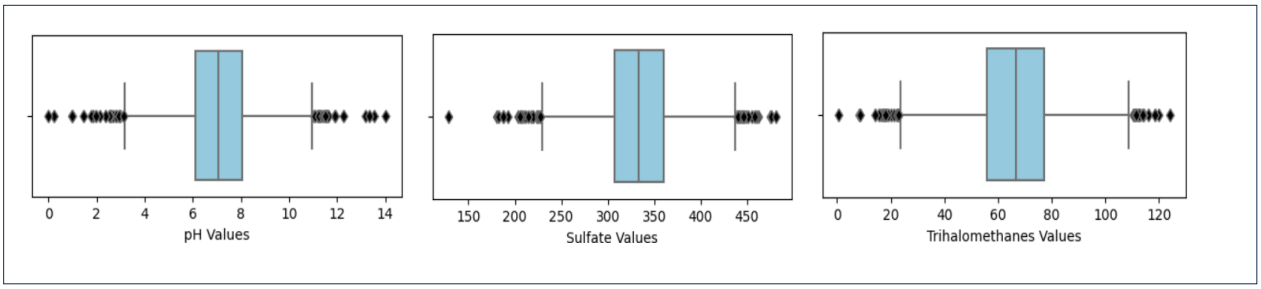

Fig. 5.2. Graph to visualize the outliers

5.5. Target Variable Exploration:

Distribution: Potability distribution visualized through a count plot.

Mean Values: Group the dataset by potability to calculate mean values of other columns for each class.

5.6. Correlation Analysis:

Correlation Matrix: A heatmap illustrates pairwise correlations between numerical columns.

Correlation Coefficients: Insights into linear relationships among features.

5.7. Skewness Analysis:

Skewness Calculation: Determining the skewness of each column.

Histograms: Visualizing the skewness of specific columns (e.g., Solids and Hardness).

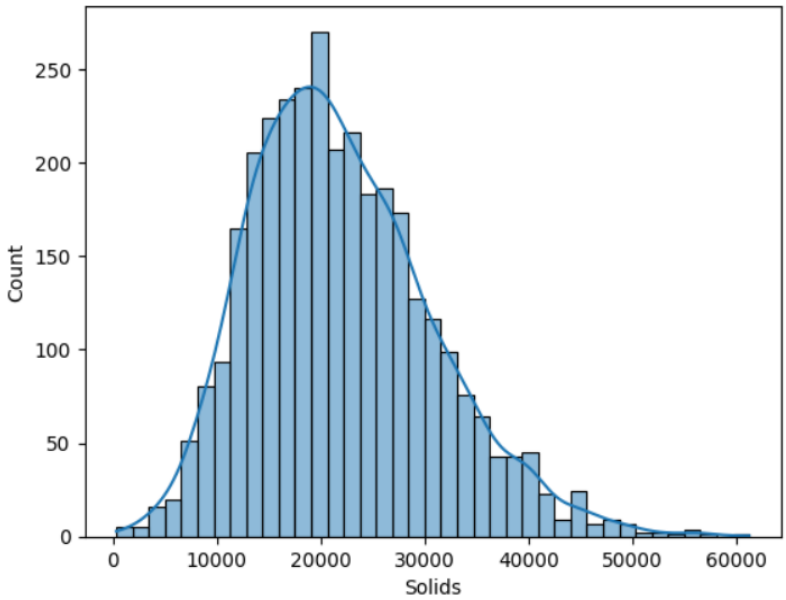

Fig.5.3. Histogram representing positive skewness

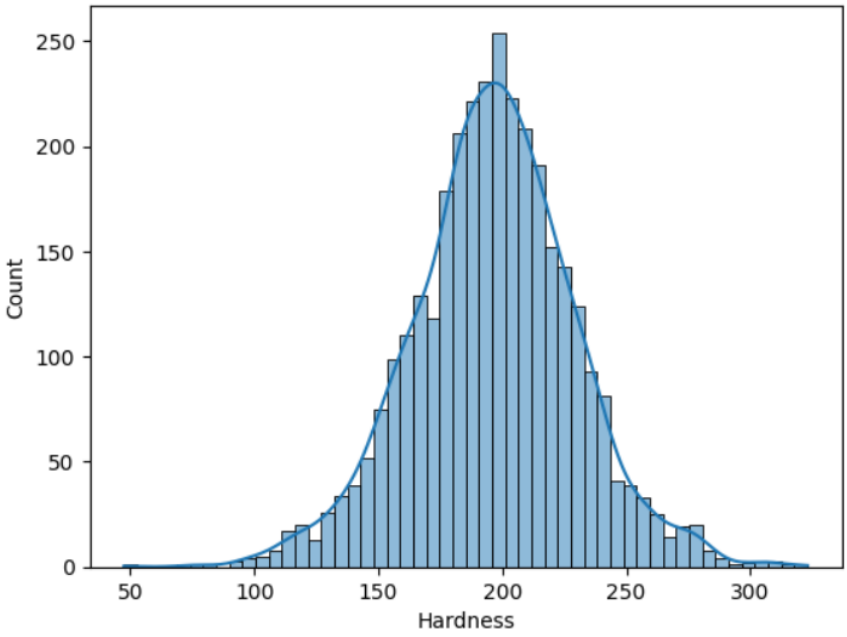

Fig. 5.4. Histogram representing negative skewness

5.8. Data Visualization:

Histograms: Displaying the distribution of each column in the dataset.

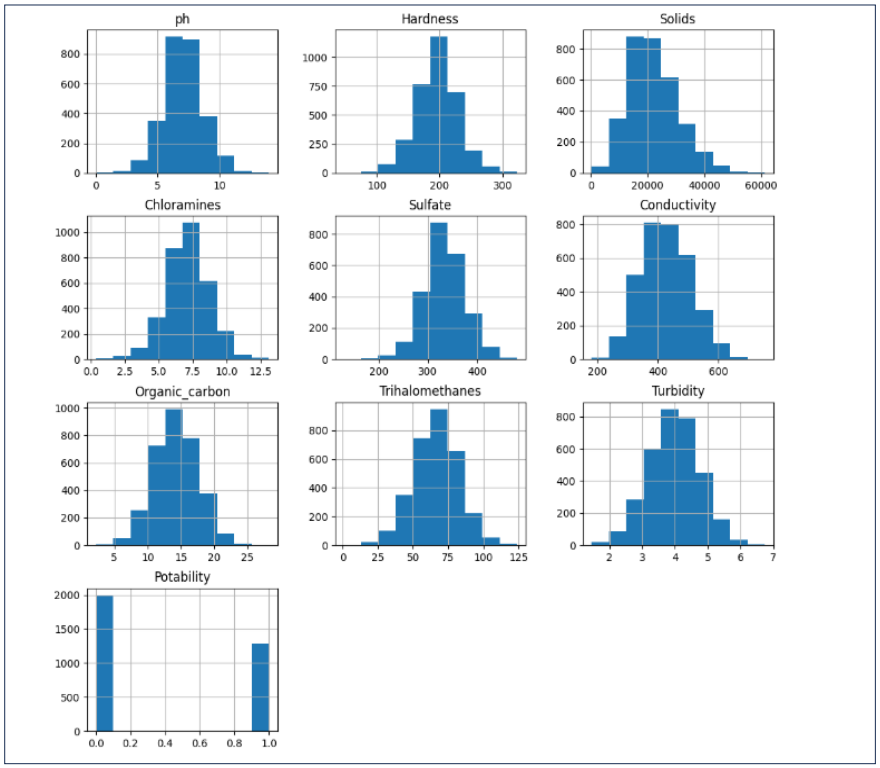

Fig. 5.5. Distribution of each column

Exploratory Data Analysis sets the foundation for subsequent modeling steps, guiding preprocessing decisions and uncovering patterns crucial for predictive modeling of water potability.

6. RESULTS

The results highlight variations in model performance, emphasizing the need for continuous refinement and exploration of ensemble methods. Each model's strengths and weaknesses contribute valuable insights for further optimization and real-world deployment in water quality management.

SVM stands out for its robust performance, making it a promising model for water potability prediction. It effectively identifies both potable and non-potable water samples. The SVM model demonstrates superior accuracy and a balanced precision-recall trade-off compared to other models. Its potential for accurate water quality classification positions it as a recommended choice for real-world applications in ensuring safe drinking water. Further optimization and fine-tuning of the SVM model can lead to even more reliable predictions.

6.1.Classification report

	precision	**recall**	**f1-score**	**support**
0	0.71	0.66	0.68	993
1	0.56	0.61	0.58	685
accuracy			0.64	1678
Macro avg	0.64	0.64	0.64	1678
Weighted avg	0.65	0.64	0.65	1678

Table 6.1. Classification report for SVM

7. APPLICATION DEPLOYMENT

The demo application has been deployed using Flask as It is a lightweight framework that offers hassle-free development, flexibility, and freedom to choose libraries and extensions, and Offers a built-in development server and fast debugger.

7.1 Model Serialization:

After selecting the best-performing machine learning model, serialize it using library joblib. Serialization converts the model into a format that can be easily stored and loaded.

7.2 Creating Flask App:

Initialized a Flask application by defining routes and handling user requests. Flask follows the Model-View-Controller architecture, where the machine learning model acts as the model, and the Flask app serves as the controller and view.

7.3. Web Form for Input:

Design a simple web form using HTML and integrate it into the Flask app. This form will collect input features required for predicting water potability.

7.4. Request Handling:

Define a route in Flask to handle incoming requests from the web form. Extract the input data from the request and preprocess it as required by the machine learning model.

7.5. Model Inference:

Load the pre-trained machine learning model and use it to make predictions on the preprocessed input data.

7.6. Displaying Results:

Display the prediction results on a new webpage or provide them as a response to the user's request. This step completes the feedback loop, and users can see the model's predictions.

7.7. Styling and UX:

Enhance the user interface by adding stylesheets (CSS) for better aesthetics and usability. Consider user experience (UX) principles to ensure a seamless interaction.

7.8. Deployment to a Web Server:

Choose a web server (e.g., Gunicorn) to deploy the Flask app. Platforms like Heroku, AWS, or Azure can host the application.

Fig.7.1. Application Deployment

The code is hosted on GitHub:

Projects-UNH/Water-Quality-Prediction (github.com)

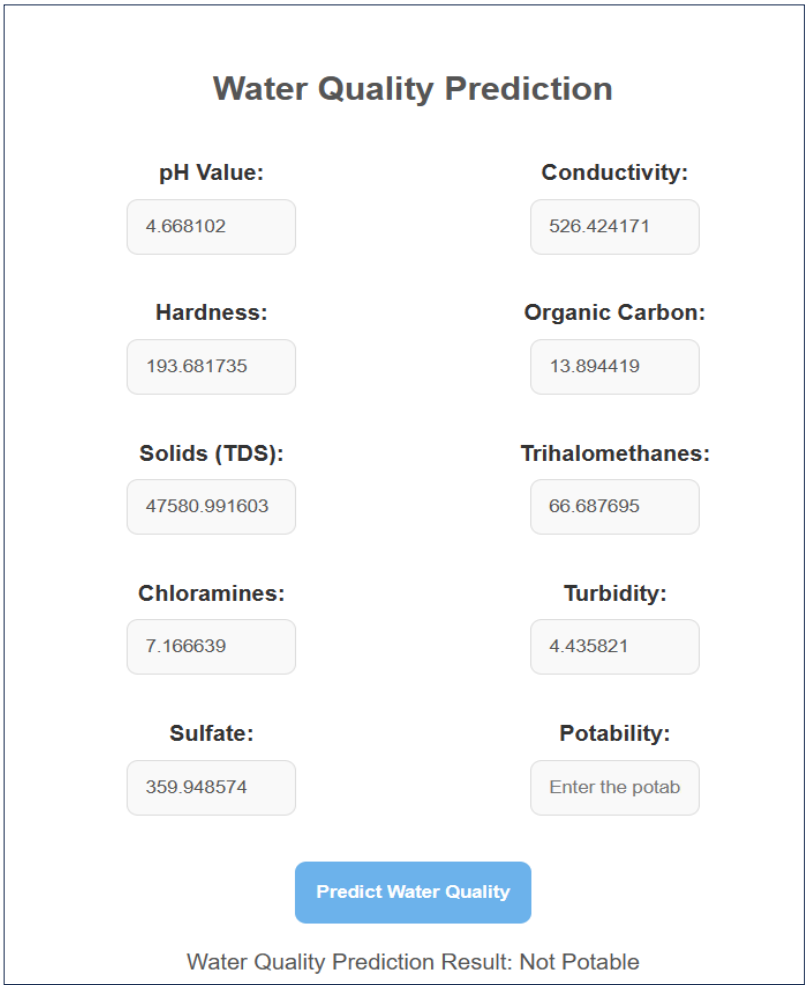

Fig. 7.2. Application Demo

8. CONCLUSION

In conclusion, the water potability prediction project represents a significant step towards ensuring the safety of drinking water through advanced machine learning techniques. The exploration, analysis, and implementation of various models have provided valuable insights into the complexity of assessing water quality and the challenges associated with predictive modeling in this domain. The water potability prediction project not only addresses the pressing issue of water safety but also exemplifies the power of machine learning in solving real-world challenges. As technology continues to advance, the synergy between data science and environmental science becomes increasingly pivotal for creating a sustainable and healthy future.

In essence, this project stands as a testament to the potential of data-driven solutions in safeguarding the most fundamental resource—water. By combining technological innovation with domain knowledge, we pave the way for smarter, more informed decisions that impact the well-being of communities and the environment.

9. FUTURE IMPROVEMENTS

Continuous refinement and optimization of the SVM model can enhance its predictive capabilities. Fine-tuning hyperparameters and exploring advanced SVM configurations may lead to improved accuracy.

Integrating educational components into the application can empower users with knowledge about water quality indicators and the significance of the model predictions. This fosters informed decision-making regarding water consumption.

Collaborating with water authorities and environmental agencies can contribute to a more comprehensive understanding of water quality dynamics. This collaboration can provide access to diverse datasets and domain expertise.

10. REFERENCES

[1] https://www.linkedin.com/learning/data-science-foundations-fundamentals-14537508/data-preparation?u=2359714

[2] https://www.linkedin.com/learning/hands-on-introduction-data-engineering/hands-on-data-engineering?u=2359714

[3] https://www.coursera.org/learn/python-for-data-visualization

[4] https://www.linkedin.com/learning/data-science-foundations-fundamentals-14537508/data-preparation?u=2359714

[5] https://www.linkedin.com/learning/hands-on-introduction-data-engineering/hands-on-data-engineering?u=2359714

[6] https://www.coursera.org/learn/python-for-data-visualization