Assessing Water Potability Using Machine Learning

Author: Shunan Zhao, Zhenyi Mei Team: Cicada

1. Executive Summary

This project aimed to uses machine learning to analyze the potability of water. The dataset is from Kaggle, and it contains a series of factors which would affect the potability of water. We analyzed the dataset using machine learning models to predict whether the water is potable or not and create a prediction model.

1.1 Decisions Impacted:

Drinking water companies can optimize the water treatment process through the model to ensure that the water is potable. And environmental agencies can allocate resources efficiently to address water quality issues and ensure compliance with water quality standards.

1.2 Business Value:

Economic Efficiency: Efficient resource allocation can lead to cost savings for water treatment facilities and agencies.

Public Health: Ensuring clean and safe drinking water will directly affect public health and will reduce the risk of waterborne diseases.

1.3 Data asserts:

The dataset revolves around determining the potability of water, based on various physical and chemical properties (pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity). Potability of water, which is binary (0 for non-potable, 1 for potable). We would first deal with the null value then use the dataset to fit different models such as Random Forest, Logistic Regression, ANN, SVM, KNN, and Gradient Boosting Machines, and then get the best model to predict the water potability.

- The data is from Kaggle: **Dataset link:** **https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability**

1.4 Our project in public in GitHub:

https://github.com/mzy199969/ESE527_Project.git

2. Data Preprocessing

2.1 Data Overview

We began our analysis with a dataset comprising various physical and chemical properties of water, such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. The primary objective was to determine the potability of water, represented by a binary target variable. The attached graph is the overview of the dataset.

```
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
```

2.2 Data cleaning

For our dataset, the target is balanced, we have 1278 data which is potable water, and 1998 data show the water is not potable.

```
{'Potable Water Dataset Size': (1278, 10),
 'Non-Potable Water Dataset Size': (1998, 10)}
```

The first step is dealing with the missing values. There are some missing values in ph, Sulfate, and Trihalomethanes. (which is showed in the following graph) Instead of a simplistic median imputation, we divided the dataset based on potability status (0 for non-potable and 1 for potable), then use the median for each potability status to fill the missing values respectively, thereby preserving inherent group characteristics.

	Missing Values	Percentage
ph	491	14.987790
Sulfate	781	23.840049
Trihalomethanes	162	4.945055

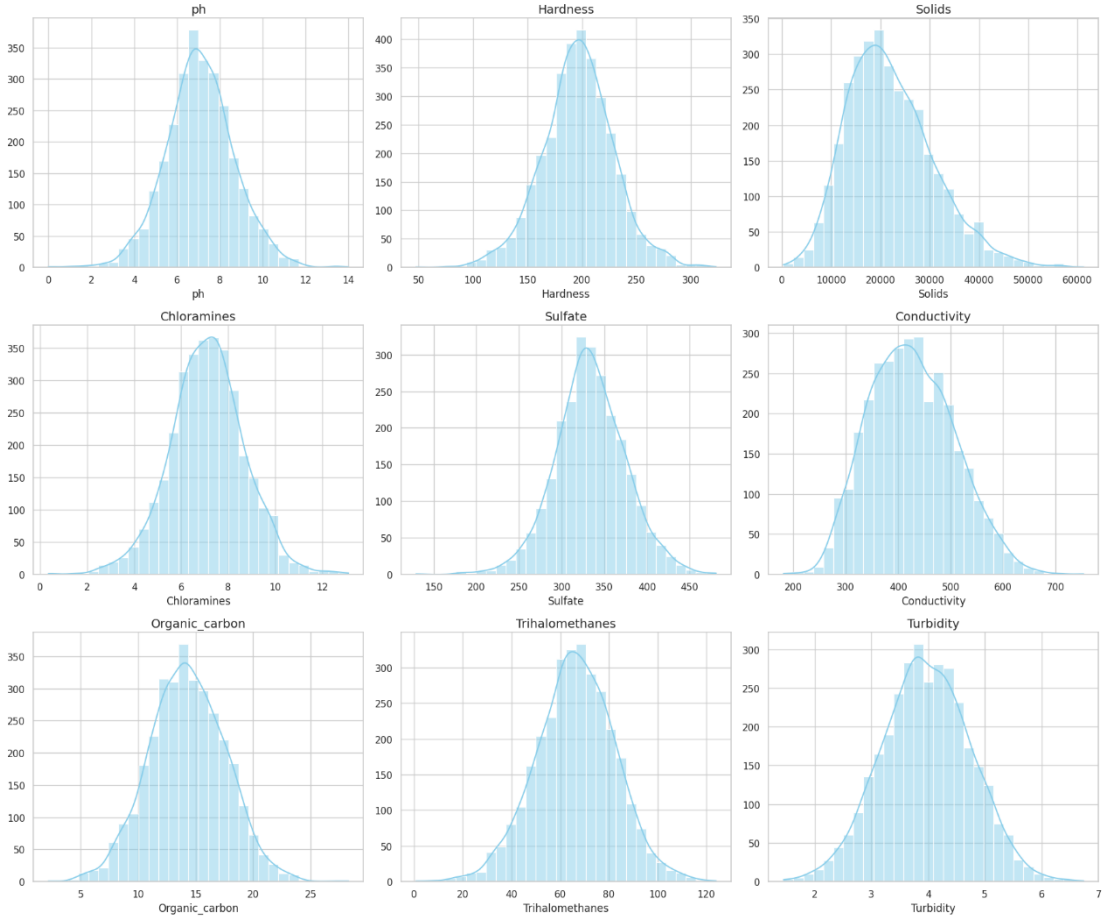
Upon the completion of missing value treatment, we applied the Mahalanobis distance for outlier detection. We chose this method due to its effectiveness in considering the covariance among different variables. However, the analysis at both 95% and 99% confidence levels indicated no significant outliers in either group.

```
(0,
 Empty DataFrame
 Columns: [ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes,
Turbidity, Mahalanobis]
 Index: [])
```

2.3 Visualization and Statistical Analysis

To gain deeper insights into our dataset, we employed various visualization and statistical techniques. These methods were instrumental in understanding the distribution of different features and their interrelationships, which are crucial for informed feature engineering and model selection.

- **Histograms and Distribution Analysis**



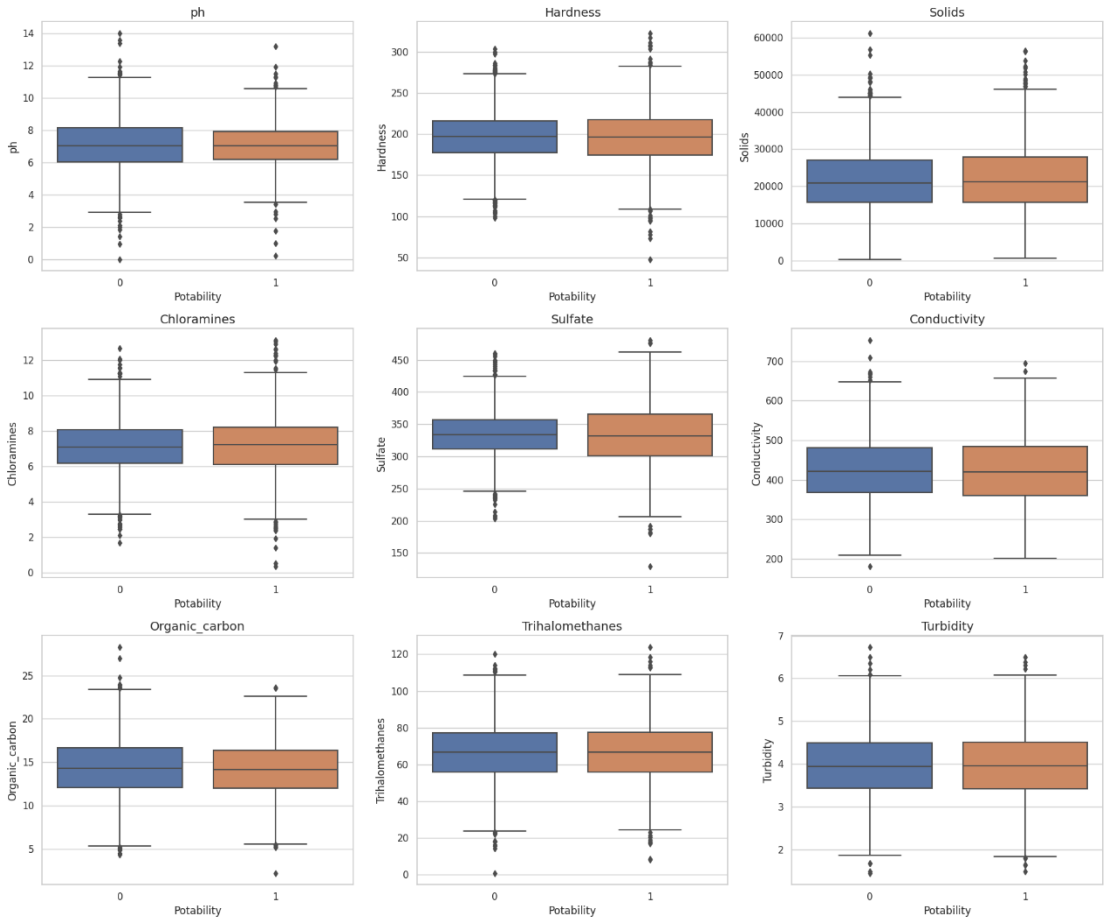
We used histograms supplemented with Kernel Density Estimation (KDE) to visualize the distribution of each feature. Key observations included:

• pH: The distribution appeared slightly skewed to the right, indicating variability in water's acidity levels.

• Solids: This feature showed a near-normal distribution, suggesting a standard range of dissolved solids in most water samples.

• Sulfate and Chloramines: Exhibited skewed distributions, hinting at varied treatment levels across samples.

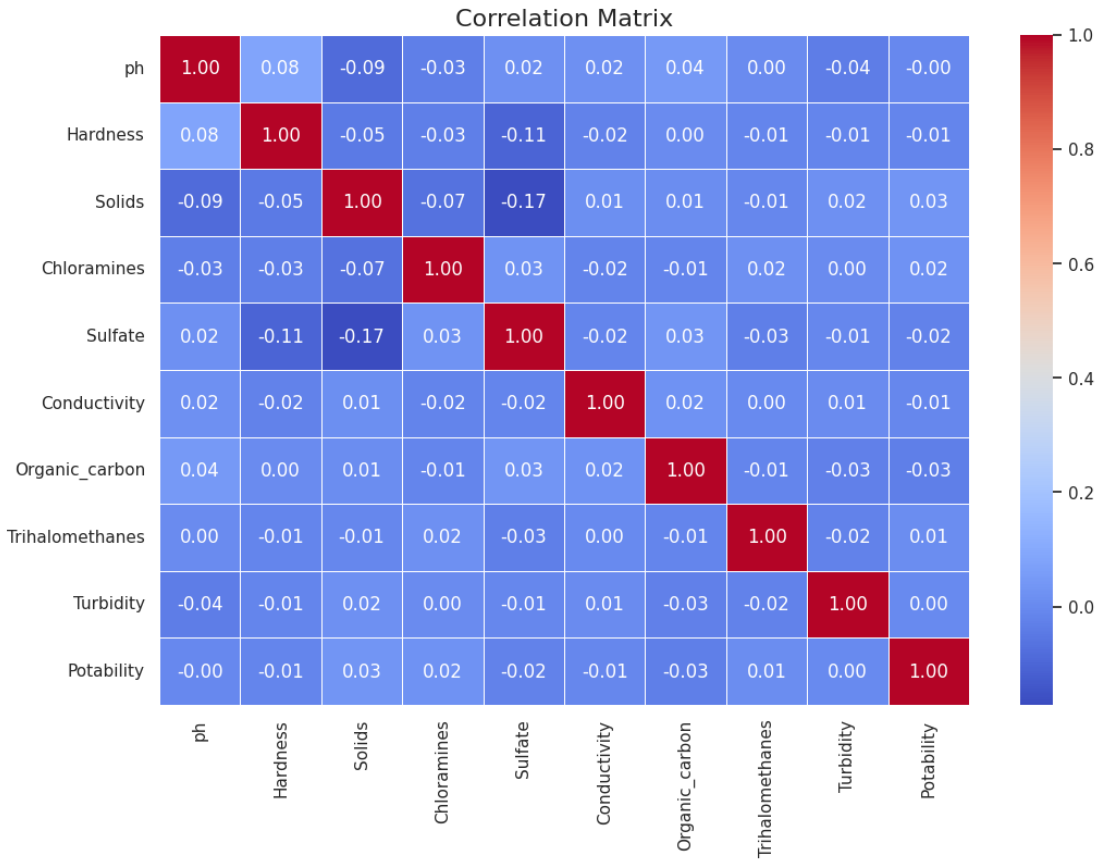
Boxplots for Comparative Analysis



Boxplots were utilized to compare features across the two potability classes (0 for non-potable and 1 for potable). Significant findings were:

- Hardness and Turbidity: Similar distributions in both potability classes, indicating these features might not be strong differentiators for potability on their own.
- Sulfate and pH: Showed noticeable differences between potable and non-potable water, suggesting their potential significance in predicting water potability.

Correlation Matrix



A heatmap of the correlation matrix was created to explore the relationships between features. Most features exhibited weak to moderate correlations with each other, implying a low degree of redundancy. No strong positive or negative correlations were observed, indicating the potential independence of these features. The lack of high correlation suggested that most features contribute unique information, which is beneficial for building predictive models.

2.4 Feature Engineering

Our approach to feature engineering was driven by practical considerations and a deep dive into the dataset's characteristics. Here's a breakdown of our key steps and the rationale behind them:

- **pH Value Categorization**

The Insight: We recognized that the impact of pH on water's potability isn't straightforward. Water quality can vary significantly across different pH levels, and this relationship isn't necessarily linear.

Our Approach: To better capture this complexity, we categorized pH values into three distinct groups: acidic, neutral, and alkaline. By doing so and applying one-hot encoding, we converted these categories into binary features. This step was crucial to help our models more effectively differentiate between these varied acidity levels.

- **Transforming 'Solids'**

The Observation: We noticed a right-skewed distribution in the 'Solids' feature, which

represents the concentration of dissolved solids in water. Such skewness can pose challenges, particularly for models that assume feature normality.

Our Solution: To address this, we applied a log transformation to 'Solids', a proven method for normalizing skewed distributions. This transformation was key to making the feature more model friendly.

- **Creating an Interaction Feature**

The Realization: In water, organic carbon and trihalomethanes often occur together as byproducts of organic matter. We hypothesized that their combined presence could be more indicative of water quality than considering each separately.

Our Action: We engineered a new feature by multiplying 'Organic_carbon' and 'Trihalomethanes'. This interaction term was intended to capture any combined effects on potability, offering a nuanced perspective to our models.

- **Implementing Polynomial Features**

The Need: Some relationships in our dataset, particularly involving chemical properties like hardness and sulfate, seemed complex and potentially non-linear.

Our Method: To unearth these hidden dynamics, we selected pivotal features such as pH, hardness, chloramines, and sulfate for polynomial transformation. Generating second-degree polynomial features allowed us to explore these non-linear interactions, potentially uncovering new predictive insights.

3. Modeling approach

3.1 Overview of Methodology

The analytical approach undertaken in this project was meticulously designed to bridge the gap between raw data interpretation and actionable insights regarding water potability. Our methodology encompassed a comprehensive spectrum of analytical techniques, ranging from descriptive analysis to predictive modeling, and culminated in prescriptive suggestions grounded in data-driven insights.

3.2 Descriptive Analysis

The initial phase involved a thorough descriptive analysis of the water quality dataset. This step was crucial to gain an understanding of the fundamental characteristics of the data, such as the distribution of chemical properties and the prevalence of potable versus non-potable samples. It involved techniques like:

- Missing Value Analysis: Identifying and imputing missing data points with a group-specific median approach, ensuring integrity and robustness in the dataset.
- Outlier Detection: Employing the Mahalanobis distance method to ascertain the presence of outliers, which reinforced the data's reliability by confirming the absence of significant anomalies.

3.3 Predictive Modeling

The core of the project revolved around developing predictive models capable of accurately classifying water as potable or non-potable. For the Feature Engineering, we enhance the

dataset with carefully selected transformations and interactions, such as pH binning, logarithmic transformation of skewed features, and polynomial feature creation, especially the significant 'poly_5' (Chloramines*Sulfate interaction).

3.4 Model Selection

We tried to use many models but finally only kept two after evaluation: Random Forest and Gradient Boosting Machines. By selecting random samples from our dataset and creating a diverse training base for each decision tree within the forest, multiple decision trees were then generated, each providing a unique perspective in predicting water potability based on various chemical and physical attributes. The final prediction for each water sample was determined through a majority voting system across all decision trees, thereby harnessing the collective predictive power of the ensemble. Gradient Boosting Machines was identified as another tool in project. GBM falls under the umbrella of ensemble methods but distinguishes itself by sequentially correcting the errors of previous trees, hence boosting the overall predictive power. This iterative refinement makes GBM particularly effective in scenarios where precision is paramount.

3.5 Cross-Validation and Performance Metrics

To ensure the reliability and generalizability of the models, a robust validation approach was adopted:

- Cross-Validation: The models were subjected to 5-fold cross-validation, which helped assess their stability and performance consistency across different subsets of the data.
- Performance Evaluation: A suite of performance metrics, including accuracy, precision, recall, and F1-score, were employed to comprehensively evaluate the models. These metrics provided a multifaceted view of the models' effectiveness, balancing the aspects of error rate and predictive reliability.

3.6 Further Evaluation

We delved deeper into model performance:

- Random Forest Confusion Matrix: The model shows a good balance between Precision and Recall, with both metrics being over 0.80 for each class. The overall accuracy of the model is 0.83, which is relatively high, indicating that the model performs well on the dataset. The macro average and weighted average metrics are also consistent, suggesting that the dataset is relatively balanced between classes and the model is performing equally well for both classes.

```
Random Forest Confusion Matrix:
 [[512 105]
 [100 482]]

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.83      0.83       617
           1       0.82      0.83      0.82       582

    accuracy                           0.83      1199
   macro avg       0.83      0.83      0.83      1199
weighted avg       0.83      0.83      0.83      1199
```

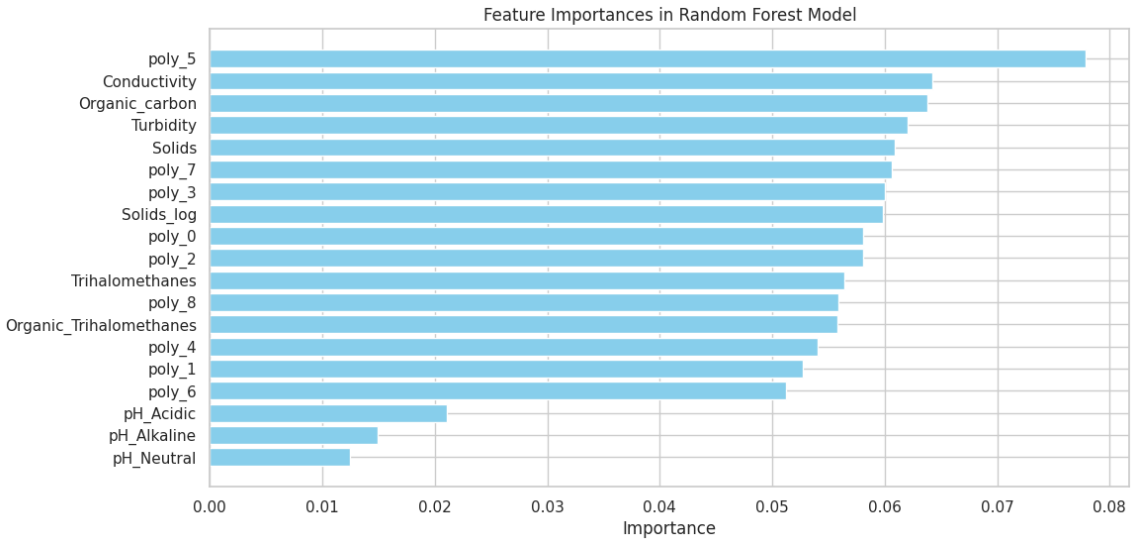
- Gradient Boosting Machine Confusion Matrix: The Gradient Boosting Machine model exhibits a reasonable balance between Precision and Recall, with both metrics being exactly 0.80 for class 1 and slightly varied for class 0, with Precision at 0.82 and Recall at 0.77. The overall accuracy of the model is 0.80, indicating good performance across the dataset. The macro average and weighted average metrics are uniformly 0.80, implying a consistent performance across both classes, despite a modest discrepancy favoring class 1 in terms of recall. This consistency suggests that the dataset is balanced and that the model does not exhibit a significant bias toward either class.

```
Gradient Boosting Machine Confusion Matrix:
 [[478 139]
 [106 476]]

Gradient Boosting Machine Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.77      0.80       617
           1       0.77      0.82      0.80       582

    accuracy                           0.80      1199
   macro avg       0.80      0.80      0.80      1199
weighted avg       0.80      0.80      0.80      1199
```

3.7 Prescriptive Insights

Based on the predictive models' results, the project offers prescriptive insights:

- Strategic Recommendations: The feature importance analysis, especially the prominence of 'poly_5', guides toward more focused chemical testing and monitoring strategies in water quality assessment.



- Policy Implications: The high accuracy and robustness of the models present an opportunity for policymakers and public health officials to adopt more data-centric approaches in water quality regulations and public health initiatives.

3.8 Conclusion

In essence, the modeling approach in this project was not merely a series of computational steps but a strategic blend of various analytical methods, each contributing to a deeper understanding and more accurate prediction of water potability. This methodical approach ensures that the project's findings are not just statistically sound but also practically relevant and actionable in the real world, thus bridging the gap between data analysis and practical application.

4. Result and insights

4.1 Analytical Work Overview

The analytical journey commenced with a thorough data preprocessing routine, where missing values were judiciously imputed using group-specific medians. This approach preserved the intrinsic characteristics of each subgroup, ensuring a robust foundation for further analysis. Subsequent outlier detection using the Mahalanobis distance revealed a commendable level of data cleanliness, as no significant outliers were identified at both 95% and 99% confidence levels.

In the realm of feature engineering, strategic steps were taken to enhance the dataset's predictive capability. pH values were categorized into discrete groups and encoded to capture their non-linear impact on water potability. The transformation of the 'Solids' feature through logarithmic scaling addressed its right-skewed distribution, thereby normalizing its effect. Polynomial features were introduced to unravel complex, non-linear relationships, particularly highlighting 'poly_5' (Chloramines*Sulfate interaction) as a significant predictor.

The model development phase witnessed the deployment of various algorithms, with the Random Forest and Gradient Boosting Machines emerging as frontrunners. The Random

Forest model, after hyperparameter optimization, achieved an accuracy of 82.90%, signifying its robustness. The Gradient Boosting Machine, following optimization, demonstrated a respectable accuracy of 79.57%.

```
Optimized Random Forest Test Accuracy: 0.8290241868223519

Optimized Gradient Boosting Machine Test Accuracy: 0.7956630525437864
```

4.2 Fulfilling Project Objectives

- Precision in Predictive Modelling:

The attained accuracy levels, notably the 82.90% by the Random Forest model, underscore a high degree of predictive precision. This aligns perfectly with the project's goal of accurately determining water potability.

- Further Evaluation:

The random forest Machine model's confusion matrix and classification report also indicated a strong performance, especially in terms of balancing the false positives and false negatives, thus validating the robustness of the feature engineering process.

- Insightful Feature Utilization:

The feature importance analysis, particularly the emphasis on 'poly_5', offers profound insights into the interplay of chemical properties in water. Understanding these dynamics is pivotal for accurately assessing water quality and directs attention to crucial factors in potability assessments.

- Model Robustness and Reliability:

The absence of significant outliers and the model's performance across various metrics bolster confidence in the robustness and reliability of the analytical approach. This is instrumental in ensuring that the models can be reliably applied to diverse and real-world data scenarios.

- Informing Further Research and Policy:

The results lay a substantial groundwork for informing future research directions. Specifically, the nuanced understanding of feature interactions and their impact on potability can guide more targeted data collection and policy formulation in water quality management.

In essence, the analytical work has not only met the project's objectives with high accuracy and insightful feature analysis but also set a strong precedent for future research and practical applications in the domain of water quality assessment.

5. Conclusions

5.1 Project Context and Motivation

The project embarked on a crucial mission to analyze water quality data with the goal of accurately determining potability, a vital concern in public health and environmental management. The motivation was rooted in leveraging advanced data analytics to provide a more nuanced and data-driven approach to water quality assessment, transcending traditional methods that may lack the granularity offered by machine learning techniques.

5.2 Key Findings and Their Implications

- Model Efficacy and Reliability:

The Random Forest model, post-optimization, demonstrated a high accuracy rate of 82.90%. This is not merely a statistic but a testament to the model's ability to effectively discern between potable and non-potable water samples with substantial reliability. Such precision in prediction is crucial for applications in public health where accurate water quality assessment can directly influence community well-being.

- Further Evaluation:

The detailed evaluation of the models using various performance metrics, including F1-scores and accuracy, highlights their potential for real-world application, providing a tool that can significantly impact decision-making in water resource management. The balanced accuracy in identifying potable and non-potable water samples enhances the models' credibility, suggesting their applicability in policy formulation and public health directives.

- Feature Importance and Insight Generation:

The identification of 'poly_5' (the interaction term between Chloramines and Sulfate) as a significant feature brings forth an insightful revelation about the chemical dynamics influencing water quality. This finding underscores the importance of considering not just individual chemical measures, but also their interactions. It paves the way for more sophisticated water testing protocols that can capture these complex relationships.

- Business and Societal Value:

The analytical work transcends technical accomplishment and taps into significant business and societal value. For water treatment facilities and public health authorities, the ability to accurately predict water potability based on complex chemical interactions leads to more informed decision-making. It equips these entities with a predictive tool that enhances their operational efficiency, risk management, and public health response.

- Strategic Decision Making and Policy Influence:

The results obtained – particularly the high accuracy of the predictive models – have the potential to inform strategic decision-making processes in environmental management and public health policies. The insights gleaned from feature importance can guide policy development, focusing on key chemical indicators and their interactions for regular monitoring and regulation.

- Direction for Future Research:

The project's findings open avenues for further research, particularly in exploring the causative links between identified chemical interactions and water potability. It also encourages the exploration of additional predictive factors that may further refine the assessment models. Future research may be directed towards sorting out non-potable water, distinguishing between those that are easy to purify and those that are difficult to purify.

5.3 Conclusion

In summary, the project successfully harnesses the power of machine learning to offer a sophisticated, reliable, and insightful approach to water potability assessment. The best model is Random Forest model, post-optimization, demonstrated a high accuracy rate of 82.90%. This is not merely a statistic but a testament to the model's ability to effectively discern between potable and non-potable water samples with substantial reliability. The problem of drinking water is a big problem for human beings. I hope our project will make even a small contribution

to the happiness of human beings.

6. reference

[1] WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene, (2019). Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines. World Health Organization (WHO) and the United Nations Children's Fund (UNICEF): https://www.who.int/publications/i/item/9789241512893

[2] https://www.kaggle.com/datasets/balavashan/drinking-water-dataset

[3] "Drinking-water." World Health Organization (WHO), 21 March 2022, https://www.who.int/newsroom/fact-sheets/detail/drinking-water

[4] Cabral, João PS. "Water Microbiology. Bacterial Pathogens and Water – PMC." NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996186/.

[5] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," OpenAI, 2023. [Online]. Available: https://chat.openai.com/.

[6] Patel, J., Amipara, C., Ahanger, T.A., Ladhva, K., Gupta, R.K., Alsaab, H.O., Althobaiti, Y.S. and Ratna, R., "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI", Computational Intelligence and Neuroscience: CIN, 2022.

[7] Pal, O.K., "The Quality of Drinkable Water using Machine Learning Techniques", Int. J. Adv. Eng. Res. Sci., 8, p.5. 2021.

[8] Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I., "Performance analysis of the water quality index model for predicting water state using machine learning techniques", Process Safety and Environmental Protection, 169, pp.808-828, 2023.

[9] Aldhyani, T.H., Al-Yaari, M., Alkahtani, H. and Maashi, M., "Water quality prediction using artificial intelligence algorithms", Applied Bionics and Biomechanics, 2020

[10] Addisie, M.B., "Evaluating Drinking Water Quality Using Water Quality Parameters and Esthetic Attributes", Air, Soil and Water Research, 15, p.11786221221075005, 2022.

[11] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R. and García-Nieto, J., "Efficient water quality prediction using supervised machine learning", Water, 11(11), p.2210, 2019.