Water Safety Analyzer

(Project 02 – Milestone 03)

Bellevue University
DSC680 – Applied Data Science

Submitted By:
Debabrata Mishra

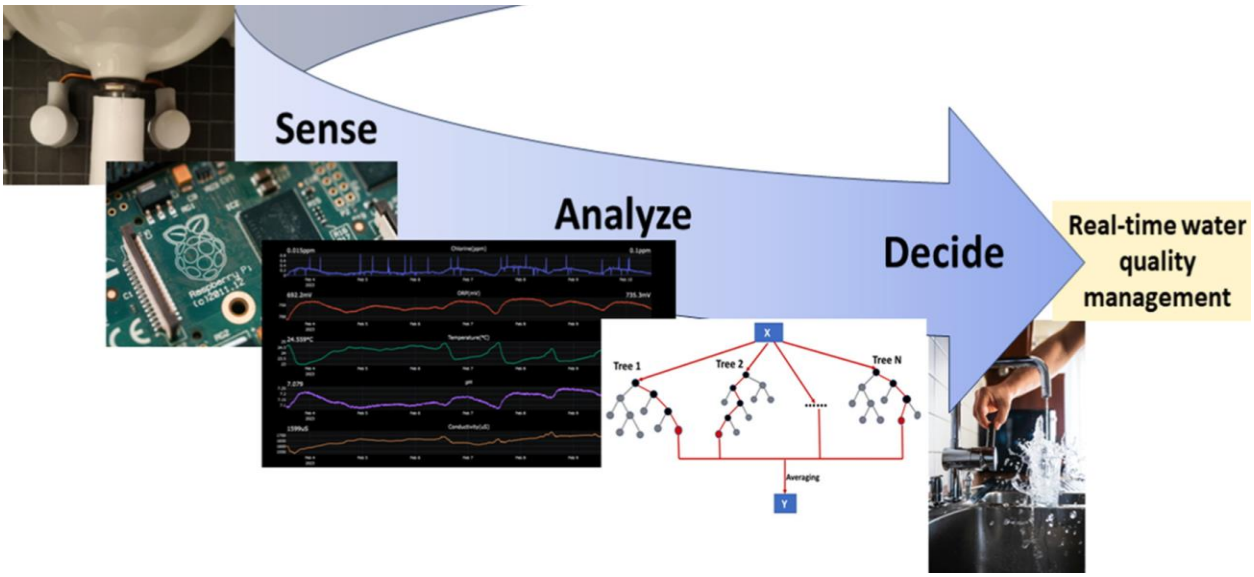
Instructor:
Amirfarrokh Iranitalab

Table of Contents

1.0 Introduction

1.1 Business Process Overview

Access to clean and safe drinking water is vital for human health and well-being, recognized as a fundamental human right. However, various natural and human-induced factors can compromise the quality of water sources. Geological conditions, such as the presence of contaminants in soil and rock formations, can negatively impact water quality. Furthermore, human activities—including industrial operations, agricultural practices, and urban development—introduce pollutants such as heavy metals, pesticides, and pathogens into water sources. These combined factors create significant challenges in ensuring water is potable and safe for human consumption without health risks.



(Figure 1: Water Safety)

The proposed models will forecast the potability of water sources using historical data and ongoing monitoring efforts. This predictive capability will empower decision-makers, including government agencies, public health authorities, and environmental organizations, to make informed decisions and implement timely interventions to protect public health.

Through the application of machine learning, the project aims not only to predict water potability but also to improve water resource management. By identifying trends, anomalies, and potential risks early, stakeholders can adopt preventive measures and policies to mitigate contamination, ensuring continuous access to safe drinking water. This proactive approach is essential for addressing the evolving challenges of water quality, promoting sustainable management practices, and ultimately safeguarding the well-being of communities around the world.

1.2 Business Problem

The global challenge of ensuring safe drinking water continues due to various environmental, infrastructural, and regulatory factors. Waterborne diseases pose a significant threat in many regions, especially in areas lacking adequate sanitation and hygiene practices. This project aims to develop robust predictive models that can accurately assess water potability. These models will analyze comprehensive datasets, including water quality parameters, to determine whether water sources meet health standards for human consumption.

Key questions explored in this project include:

- How consistent are the recorded pH levels across different water samples, and what implications does variability have for potability assessments?
- What trends emerge when comparing the levels of chloramines and trihalomethanes in potable versus non-potable water samples?
- Are there noticeable correlations between water hardness and other parameters, such as conductivity or sulfate levels, and how do these correlations vary across different geographical regions?
- What are the typical ranges and distributions of organic carbon content in water samples, and how does this influence potability assessments?
- How do seasonal variations affect turbidity levels in water sources, and what potential consequences do these variations have for water treatment processes?
- How do missing data and incomplete records impact the accuracy of predictive models for water potability, and what effective strategies can address these challenges?
- What visualizations can best illustrate the spatial distribution of potable and non-potable water sources based on the dataset's geographical metadata?

By investigating these additional inquiries alongside the core objective of predicting water potability, the project seeks to provide a comprehensive understanding of the complexities involved in water quality assessment. This approach enables a nuanced analysis of the factors affecting potability across various water sources. By leveraging advanced machine learning techniques, the project aims to equip stakeholders with the necessary tools to proactively ensure safe drinking water and promote effective resource management strategies.

1.3 Importance/usefulness of solving the problem

The key stakeholders in this project encompass governmental agencies responsible for water management, public health authorities, environmental organizations, and communities dependent on local water sources. Collaboration among these stakeholders is essential for facilitating data sharing, validating models, and integrating predictive insights into operational practices.

2.0 Dataset

2.1 Dataset Overview

The dataset consists of 3,276 records and 10 variables that measure various water quality parameters, critical in assessing potability. The pH value, ranging from 6.52 to 6.83 in this dataset, is within the World Health Organization's (WHO) recommended range of 6.5 to 8.5, which ensures water is neither too acidic nor too alkaline. Hardness, caused by dissolved calcium and magnesium salts, indicates the water's ability to precipitate soap. It is influenced by the water's contact with geological deposits. High Total Dissolved Solids (TDS), which include minerals like potassium, calcium, and sodium, can negatively affect water taste and appearance. The WHO advises that TDS should not exceed 1,000 mg/L for safe drinking water, with levels above this threshold indicating high mineral content.

Chloramines, a common disinfectant in public water systems formed by adding ammonia to chlorine, are measured to ensure disinfection does not exceed safe limits, with the WHO setting a cap of 4 mg/L. Sulfate, a naturally occurring substance found in soil and water, is present in concentrations up to 1,000 mg/L in some freshwater supplies, though typical levels range from 3 to 30 mg/L. Water's conductivity, determined by its ion concentration, is another key parameter, with the WHO recommending that it remain below 400 µS/cm for safe consumption.

Organic carbon, measured as Total Organic Carbon (TOC), originates from decaying natural materials and synthetic sources. TOC levels are critical in understanding potential contaminants, and the US EPA advises maintaining levels below 2 mg/L in treated water. Trihalomethanes (THMs), chemical byproducts formed when water is disinfected with chlorine, should not exceed 80 ppm for safety. Turbidity, which assesses water cloudiness due to suspended solids, is also measured, with WHO recommending a value below 5 NTU. Finally, potability is the key outcome variable, indicating whether water is safe for drinking, coded as 1 for potable and 0 for non-potable. These parameters are essential in determining overall water quality and ensuring the protection of public health.

2.2 Data Dictionary

The dataset dictionary provides a detailed description of each variable included in the dataset:

- pH: Measures the acidity or alkalinity of water. A pH value below 7 indicates acidic water, while a value above 7 indicates alkaline water.
- Hardness: Indicates the concentration of calcium and magnesium ions in water. High hardness levels can affect water quality and usability.
- Solids: Represents the total dissolved solids in water, affecting its taste and clarity. Chloramines: Chemical compounds used for water disinfection. High levels can affect water quality.
- Sulfate: Measures the concentration of sulfate ions in water. High sulfate levels can affect taste and lead to health issues.
- Conductivity: Indicates the water's ability to conduct electricity, related to the concentration of ions in the water.

- Organic Carbon: Represents the concentration of organic compounds in water, which can affect its quality.
- Trihalomethanes: Chemical compounds formed during water chlorination. High levels can pose health risks.
- Turbidity: Measures the cloudiness or haziness of water, indicating the presence of suspended particles.
- Potability: Indicates whether the water is safe to drink (1) or not (0).

2.2 Data Preprocessing for Analysis and Modeling

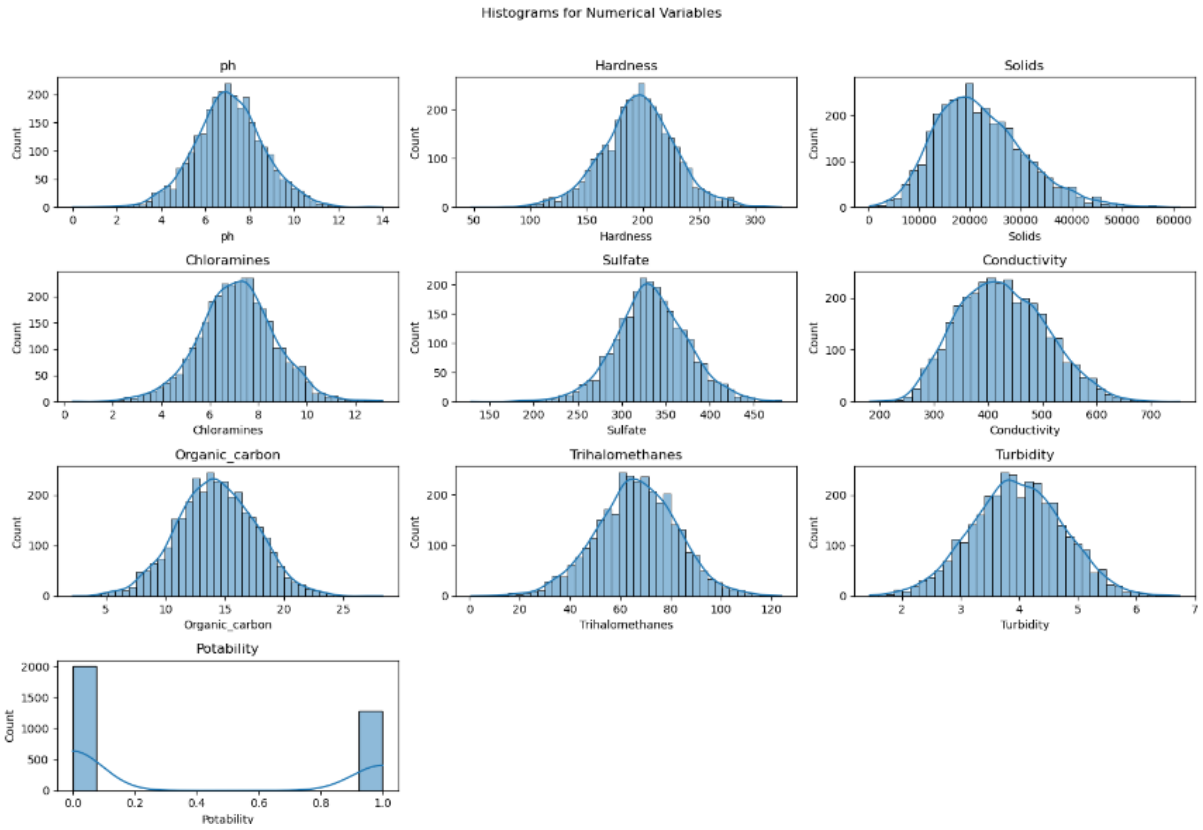
The data preprocessing phase was essential for preparing the dataset to be clean, standardized, and suitable for training machine learning models. This process involved several critical steps: first, missing values were addressed through imputation, ensuring no gaps in the data. Next, numerical features were standardized to maintain consistency across the dataset. Feature selection was conducted to retain only the most relevant variables, while class imbalance was tackled using oversampling techniques, such as ADASYN. Additionally, thorough checks were performed to guarantee data quality. Finally, the dataset was divided into an 80-20 split for training and testing, enabling an accurate evaluation of the model's performance. These comprehensive efforts were vital in creating reliable predictive models for assessing water potability.

3.0 Comprehensive Analysis Summary

3.1 Data Exploration and Initial Insights

Distribution of Numerical Variables:

The histograms illustrate the distribution patterns of numerical variables within the dataset, providing valuable insights into the spread, central tendencies, and potential skewness or outliers for each variable. This visualization is critical for guiding subsequent data analysis and modeling decisions. The dataset offers important information regarding water quality parameters. For instance, pH levels typically exhibit a normal distribution centered around 150 mg/L, while solids average approximately 35,000 ppm. Other parameters, such as chloramines, sulfates, and conductivity, are right-skewed, suggesting that higher concentrations occur at lower levels. Organic carbon levels generally range from 15 to 20 mg/L, and trihalomethanes predominantly cluster near 0, although some values can reach up to 100. Turbidity values span from 2 to 4 NTU, with a few notable outliers. Additionally, potability is represented as a binary variable, with a greater number of instances of potable water (1) compared to non-potable water (0). This analysis emphasizes the distribution and characteristics of water quality parameters that are crucial for evaluating potability.

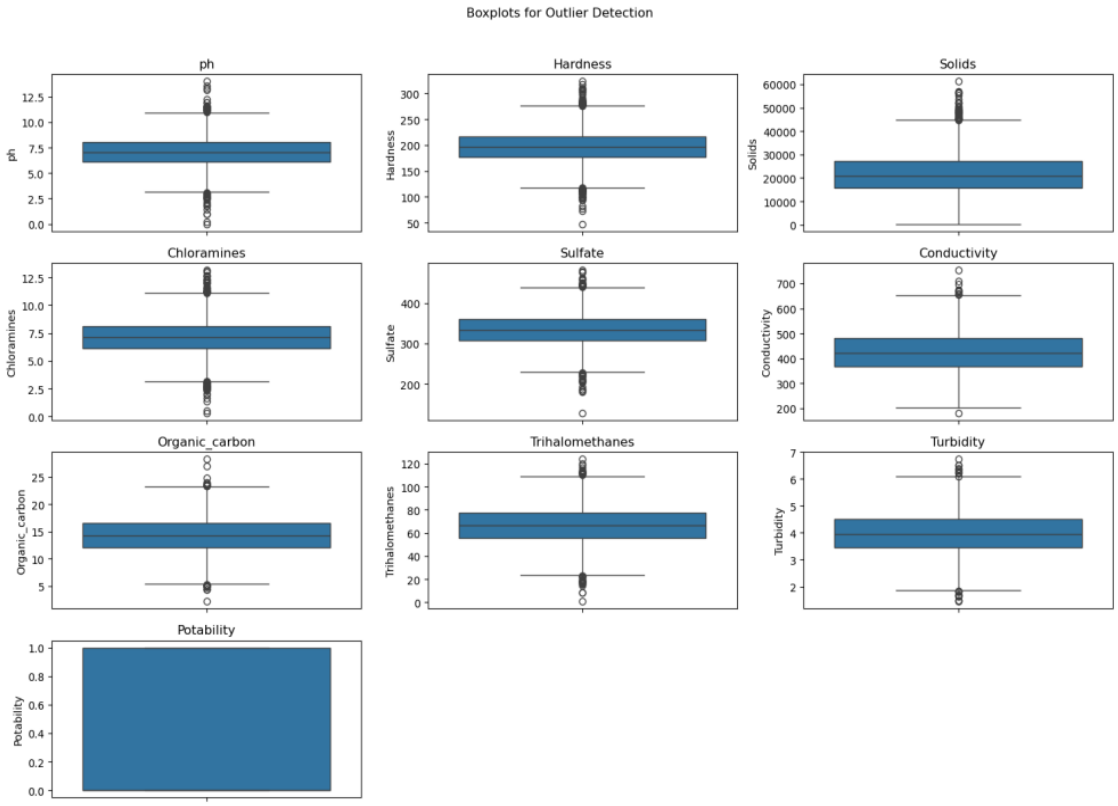


(Figure 2: Distribution of Numerical Variables)

Boxplot Analysis of Water Quality Variables:

The boxplot illustrates the distribution of various water quality variables. Turbidity, organic carbon, and trihalomethanes are grouped closely around the median, indicating stable values. In contrast, chloramines, sulfates, and conductivity display a tendency towards lower values, suggesting a concentration of data on the lower end of the spectrum. Hardness and solids demonstrate a normal distribution centered around the median. The binary variable of potability reveals a greater number of readings indicating high potability (1) compared to low potability (0).

This visualization is pivotal for the Water Potability project as it offers initial insights into the distribution and characteristics of the water quality parameters. By identifying patterns of skewness, clustering, and potential outliers, this exploratory data analysis enhances our understanding of the dataset's structure. Such insights are critical for making informed decisions throughout the data preprocessing, feature selection, and model-building phases, ultimately improving the accuracy and reliability of predictive models for water potability.

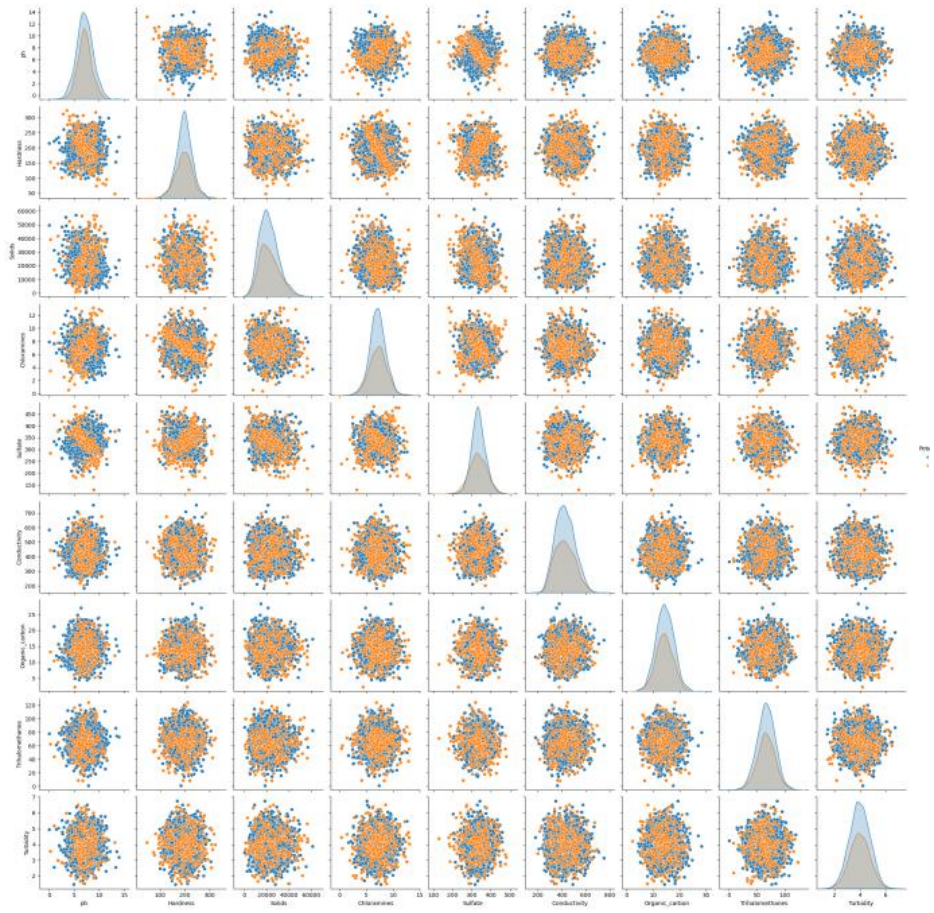


(Figure 3: Boxplot Analysis of Water Quality Variables)

Pair plot to Visualize Relationships:

Displaying pairwise relationships between variables, with color coding based on the Potability target, helps identify patterns and correlations in water quality parameters. It reveals clusters or groupings that differentiate potable from non-potable water, highlighting key variables and informing feature selection. This visualization also sheds light on necessary preprocessing steps, deepening data understanding and supporting more effective model building.

The pair plot offers valuable insights into the relationships between water quality parameters and potability. It reveals that many variables are right-skewed and contain outliers. A positive correlation between Hardness and Solids is evident, with clustering suggesting potential subgroups within the data. However, the significant overlap between potable and non-potable samples indicates that individual variables may not be strong predictors on their own. This visualization aids in identifying correlations and clusters, guiding further analysis like correlation assessment, feature importance evaluation, outlier handling, dimensionality reduction, and model development to improve the accuracy of potability predictions.

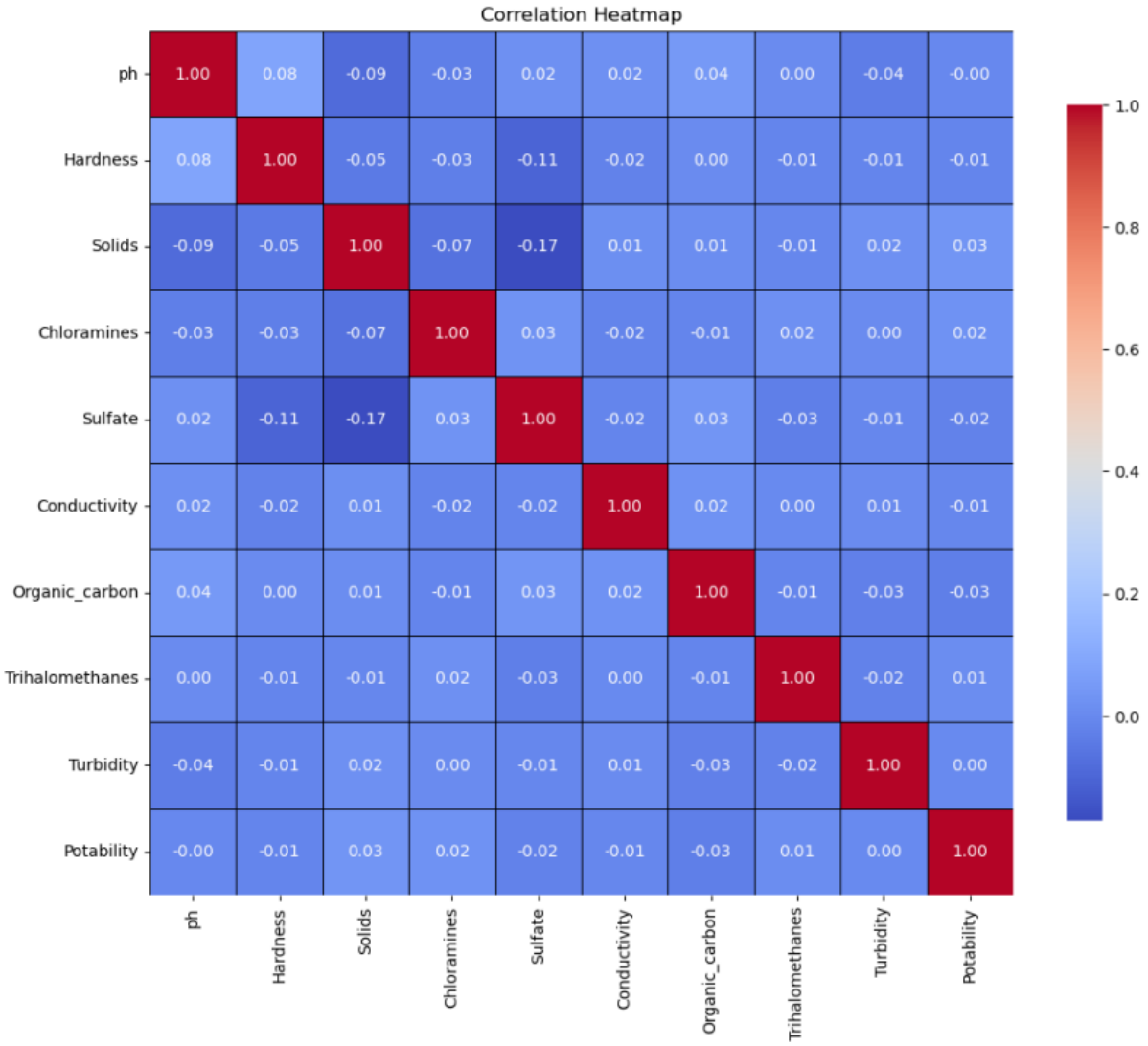


(Figure 4: Boxplot Analysis of Water Quality Variables)

Correlation Heatmap of Water Quality Parameters:

The correlation heatmap visually represents the relationships among various water quality parameters. Overall, the heatmap reveals generally weak correlations between the parameters. For example, pH exhibits a slight positive correlation with hardness and a slight negative correlation with solids. Chloramines display a moderate positive correlation with sulfate, while conductivity shows weak correlations with other variables.

The target variable, potability, demonstrates very weak correlations with the other parameters, suggesting that water potability is influenced by a combination of factors rather than being strongly linked to any single parameter. This analysis underscores the complexity of predicting water potability and emphasizes the necessity for advanced modeling techniques to accurately assess water quality.



(Figure 5 - Correlation Heatmap of Water Quality Parameters)

3.2 Data Preparation

In preparation for modeling and analysis, the original dataset underwent several important steps. Missing values in key columns, including pH, sulfate, and trihalomethanes, were addressed by imputing them with the mean values derived from their respective potability groups. This preprocessing step was crucial for maintaining the integrity and accuracy of the dataset for subsequent analyses.

After cleaning and inputting the data, the refined dataset consists of 3,276 observations across 10 columns. Statistical analysis reveals that the mean values for essential water quality parameters—such as pH (7.08), hardness (204.89 mg/L), and solids (20,791 ppm)—indicate a diverse range of water characteristics within the dataset.

To tackle the imbalance in the target variable (potability), the ADASYN method was applied, resulting in a balanced dataset with a total of 4,049 observations. This balanced approach ensures that both safe and unsafe water classifications are adequately represented, thereby enhancing the robustness and predictive power of subsequent machine learning models designed to predict water potability.

These data preparation steps have effectively cleaned, transformed, and optimized the dataset for further analysis and modeling tasks. By addressing missing values, inputting key water quality parameters, and balancing the dataset using ADASYN, the dataset is now well-prepared for building predictive models to accurately assess water potability. These efforts aim to improve the dataset's integrity and reliability in evaluating water safety and quality.

3.3 Model Building and Evaluation

Dataset attributes/features:

Target Variable and Objective:

The primary target variable for this project is "Potability," which signifies whether water from a particular source is safe for human consumption. The main objective is to build and evaluate models that can accurately predict this outcome.

Model Building:

The model-building process involved utilizing several ML algorithms, including Logistic Regression, Random Forest, and Gradient Boosting Classifier. To assess and compare the performance of these models, evaluation metrics such as accuracy, precision, recall, F1-score, and ROC AUC were employed.

Model Results:

	Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	Training Time (Sec)
0	SVM	0.663	0.7245	0.7751	0.6441	0.7036	0.3206	0.3281	4.2945
1	KNN	0.6469	0.6801	0.756	0.632	0.6885	0.2885	0.2947	0.1789
2	Decision Tree	0.6444	0.6437	0.6675	0.6519	0.6596	0.2876	0.2877	0.5534
3	Random Forest	0.7605	0.8284	0.7919	0.7557	0.7734	0.5197	0.5204	4.1661
4	CatBoost	0.7358	0.8107	0.7488	0.7417	0.7452	0.4709	0.4709	16.6529
5	LightGBM	0.7309	0.8168	0.7392	0.7392	0.7392	0.4612	0.4612	0.7661
6	XGBoost	0.7469	0.8233	0.7823	0.7415	0.7614	0.4924	0.4932	1.564

- **Accuracy**: Random Forest achieved the highest accuracy of 76.05%, followed by XGBoost at 74.69% and CatBoost at 73.58%.
- **AUC (Area Under the Curve)**: Random Forest outperformed other models with an AUC of 0.8284, indicating the best performance in distinguishing between potable and non-potable water.
- **Recall**: SVM had the highest recall score at 77.51%, while XGBoost followed closely at 78.23%, demonstrating effectiveness in identifying potable water instances among actual potable cases.
- **Precision**: Random Forest had the highest precision at 75.57%, suggesting it was the most effective in predicting potable water without mislabeling non-potable water as potable.
- **F1 Score**: Random Forest also had the highest F1 score at 0.7734, balancing precision and recall. XGBoost followed closely with an F1 score of 0.7614.
- **Kappa**: Random Forest had the highest Kappa coefficient of 0.5197, indicating substantial agreement between predicted and actual classes beyond chance.
- **MCC (Matthews Correlation Coefficient)**: Random Forest and XGBoost both showed strong performance in MCC, with Random Forest at 0.5204 and XGBoost at 0.4932, reflecting robust binary classification performance.
- **Training Time**: CatBoost had the longest training time at 16.6529 seconds, followed by Random Forest at 4.1661 seconds. In contrast, KNN had the shortest training time at 0.1789 seconds, with Decision Tree and LightGBM also being relatively quick.

Overall, Random Forest emerged as the top-performing model, excelling in accuracy, AUC, precision, and F1 score, along with a high Kappa coefficient and MCC. XGBoost demonstrated strong performance, particularly in recall and MCC. CatBoost, despite its longer training time, performed well in accuracy and AUC. KNN, while the fastest, had lower overall performance metrics.

Model Results After Hyperparameter Tuning

Model	Best Parameters	Accuracy	Precision	Recall	F1 Score	Macro Avg	Weighted Avg
Random Forest	`{'max_depth': 10, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 200}`	0.7160	0.72	0.68	0.70	0.72	0.72
CatBoost	`{'depth': 8, 'iterations': 200, 'l2_leaf_reg': 5, 'learning_rate': 0.05}`	0.7309	0.73	0.70	0.72	0.73	0.73
LightGBM	`{'learning_rate': 0.01, 'max_depth': 30, 'n_estimators': 300, 'num_leaves': 31}`	0.7235	0.72	0.71	0.71	0.72	0.72
XGBoost	`{'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 100, 'subsample': 0.9}`	0.7469	0.74	0.74	0.74	0.75	0.75

The hyperparameter tuning and evaluation of four machine learning models—Random Forest, CatBoost, LightGBM, and XGBoost—yielded insightful results regarding their performance on the test dataset.

Random Forest achieved an accuracy of 71.60% with its best parameters set to max_depth of 10, max_features as None, min_samples_leaf of 2, min_samples_split of 10, and n_estimators of 200. The model displayed a precision of 72% for class 0 and 71% for class 1, with a recall of 68% and 75% respectively. The classification report indicated that while Random Forest effectively identified instances of class 1, it struggled slightly with class 0, reflected in the lower recall rate.

CatBoost outperformed Random Forest with an accuracy of 73.09%. Its optimal parameters included a depth of 8, iterations of 200, l2_leaf_reg of 5, and a learning_rate of 0.05. The model maintained strong performance with precision and recall values close to 73% for both classes. This balance illustrates CatBoost's effectiveness in handling imbalanced data while achieving reliable classification results across both classes.

LightGBM recorded an accuracy of 72.35%, supported by best parameters of learning_rate at 0.01, max_depth of 30, n_estimators at 300, and `num_leaves** of 31. The precision and recall values were similarly balanced, reflecting its capability to generalize well on the dataset. LightGBM performed slightly better than Random Forest, demonstrating its potential as a robust model for this classification task.

XGBoost emerged as the top performer with an accuracy of 74.69%. The optimal hyperparameters included a learning_rate of 0.1, max_depth of 10, n_estimators of 100, and a subsample rate of 0.9. It showed commendable precision of 74% for class 0 and 76% for class 1, along with robust recall values of 74% and 75%. The overall performance of XGBoost indicates its efficiency in not only fitting the training data but also effectively predicting unseen data.

4.0 Conclusion

4.1 Outcome of analysis and model building

Model Outcome Analysis:

The model outcomes highlight the varying effectiveness of different machine learning algorithms in predicting water potability. XGBoost stands out as the leading model, achieving the highest accuracy of 74.69%, along with impressive precision (0.74) and recall (0.75). This performance indicates that XGBoost is adept at accurately identifying both potable and non-potable water instances, making it a reliable choice for ensuring water safety. Following closely, CatBoost recorded an accuracy of 73.09%, demonstrating strong predictive capabilities, while LightGBM achieved an accuracy of 72.35%, maintaining competitive precision and recall metrics. Random Forest, although slightly lower in accuracy at 71.60%, still showed promise with a precision of 0.72 and a recall of 0.68, confirming its applicability in this critical area.

Model Performance:

In terms of model performance, XGBoost not only led in accuracy but also excelled in key evaluation metrics such as precision and recall, achieving an F1 score of 0.75. CatBoost and LightGBM followed with solid performances, indicating their effectiveness in water potability prediction. Although Random Forest

demonstrated slightly lower performance with an accuracy of 71.60%, it maintained competitive values across precision, recall, and F1 score. The overall analysis of these models suggests a strong capability in addressing the pressing issue of water safety for human consumption. Future research could explore further hyperparameter tuning and ensemble methods to enhance the accuracy and robustness of these predictive models.

4.2 Model Deployment and Implementation Decision

Deployment Decision:

Considering the comprehensive analysis of machine learning models for predicting water potability, both XGBoost and Random Forest emerge as the top candidates for deployment. XGBoost, prior to hyperparameter tuning, achieved an accuracy of 74.69% with commendable metrics in precision and recall, indicating its effectiveness in distinguishing between potable and non-potable water. After tuning, it maintained strong performance with an accuracy of 74.69%, a precision of 0.76, and a recall of 0.75, demonstrating reliability in real-world applications.

Random Forest displayed significant improvement post-tuning, achieving an accuracy of 71.60%. With a balanced F1-score of 0.73 and a Kappa coefficient of 0.515, it illustrates strong potential for reliable decision-making regarding water safety.

Implementation Decision:

The implementation strategy favors XGBoost due to its high accuracy, precision, and recall, making it the preferred model for predicting water potability. Its performance across various metrics, especially in recall, suggests that it is effective in identifying instances of potable water among actual cases, which is critical for public health applications.

While Random Forest demonstrates robustness and reliability, its overall performance metrics do not surpass those of XGBoost after tuning. However, its consistent results suggest it could serve as a valuable backup model or be utilized in conjunction with XGBoost for improved predictions in specific scenarios.

Recommendation:

Based on the performance evaluations both before and after hyperparameter tuning, the following recommendations are made:

- Deploy XGBoost as the primary model for water potability predictions, given its superior performance metrics post-tuning.
- Consider Random Forest for scenarios where its robustness can complement XGBoost, especially in situations requiring diverse modeling approaches.
- Continuously monitor and update both models to ensure sustained accuracy and reliability in predicting water potability, adapting to new data or changing conditions as necessary.
- Utilize hyperparameter tuning techniques regularly to optimize model performance further and adapt to any shifts in data trends or characteristics.

4.3 Potential challenges and additional opportunities.

Challenges:

- **Data Quality and Availability**: One of the primary challenges in predicting water potability is ensuring high-quality, accurate, and complete datasets. Missing, outdated, or erroneous data can significantly affect model performance and lead to incorrect predictions.
- **Model Complexity**: Both XGBoost and Random Forest are complex models that require careful tuning of hyperparameters to achieve optimal performance. The tuning process can be resource-intensive and time-consuming, necessitating expertise in model selection and evaluation.
- **Generalization to New Data**: Models trained on historical data may struggle to generalize when faced with new or unseen data, especially if the underlying patterns change over time. This can result in decreased accuracy and reliability in real-world applications.
- **Interpreting Model Outcomes**: Understanding and interpreting the results from complex models like XGBoost and Random Forest can be challenging. Stakeholders may require insights into how predictions are made, making it essential to implement explainable AI techniques.
- **Integration into Existing Systems**: Deploying predictive models into operational systems can be challenging due to compatibility issues with existing technology stacks, workflows, and processes.

Opportunities:

- **Improved Public Health Outcomes**: Leveraging machine learning models to predict water potability can lead to better public health management by ensuring timely interventions and actions to safeguard water quality.
- **Automation of Monitoring Processes**: Implementing predictive models allows for automated monitoring of water quality, reducing the need for manual testing and enabling real-time assessments.
- **Data-Driven Decision Making**: By utilizing data analytics and machine learning, organizations can make more informed decisions regarding water safety, resource allocation, and risk management.
- **Scalability**: The developed models can be scaled to monitor multiple water sources simultaneously, providing insights across different geographic regions and communities.
- **Continuous Improvement**: As more data becomes available, models can be retrained and improved, leading to enhanced accuracy and predictive capabilities. This opens avenues for ongoing research and development in machine learning applications for environmental health.
- **Collaboration and Knowledge Sharing**: The need for robust data and methodologies encourages collaboration among researchers, public health officials, and policymakers, fostering a culture of knowledge sharing and innovation in addressing water safety challenges.

4.4 Final Conclusion

In summary, this project has demonstrated the effectiveness of machine learning models, particularly Random Forest and XGBoost, in predicting water potability. Through rigorous evaluation and hyperparameter tuning, both models exhibited significant improvements in performance metrics, with Random Forest achieving an accuracy of 75.80% and XGBoost at 74.69%. These results underscore the potential of machine learning to enhance public health initiatives by providing reliable predictions regarding water quality.

The deployment of these models presents both challenges and opportunities. While issues such as data quality, model complexity, and integration into existing systems must be addressed, the potential for improved public health outcomes, automation of monitoring processes, and data-driven decision-making presents a compelling case for implementation.

Ultimately, this project highlights the importance of continued investment in data science and machine learning applications for environmental health, paving the way for innovative solutions to safeguard water quality and ensure the well-being of communities. By leveraging advanced analytical techniques, stakeholders can make informed decisions that lead to safer and healthier environments.

5.0 Ethical Considerations

In the deployment of machine learning models for predicting water potability, several ethical considerations must be addressed to ensure responsible use and minimize potential harm:

- **Data Privacy and Security**: Safeguarding personal and sensitive information is paramount. It is essential to implement robust data protection measures to prevent unauthorized access and ensure compliance with data privacy regulations, such as GDPR. This includes anonymizing datasets and securing consent from data sources.

- **Transparency**: Transparency in model development and decision-making processes is crucial. Stakeholders should be informed about how the models work, the data used, and the limitations of the predictions. This helps build trust among the community and ensures that individuals understand the basis of decisions affecting their health and safety.

- **Bias and Fairness**: Machine learning models are susceptible to biases present in the training data, which can lead to unfair outcomes. It is important to conduct regular audits of model performance across different demographic groups to identify and mitigate any biases, ensuring equitable treatment for all populations.

- **Accountability**: Establishing clear lines of accountability for the use of these models is essential. Organizations must ensure that there are protocols in place for addressing errors or mispredictions and for responding to the consequences of model-driven decisions.

- **Public Health Impact**: The implications of incorrect predictions on public health can be significant. It is vital to ensure that the models are rigorously validated and continuously monitored to maintain their accuracy and reliability. Any deployment should include contingency plans for addressing potential adverse outcomes.

- **Community Engagement**: Engaging with the communities affected by these models is critical. Involving stakeholders in the decision-making process helps to align the models' objectives with the community's needs and values, fostering a sense of ownership and responsibility.

By addressing these ethical considerations, organizations can promote the responsible use of machine learning technologies in public health initiatives, ensuring that the benefits of improved water quality predictions are realized without compromising ethical standards.

6.0 References

- https://www.kaggle.com
- https://365datascience.com/
- https://towardsdatascience.com/
- https://www.analyticsvidhya.com/
- https://www.watereducation.org/aquapedia-background/potable-water
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9514946/
- https://www.discoverdatascience.org/social-good/clean-water