Safe Aquifers: Investigating Water Potability Trends for Public Health

Term Project2 – Final Submission

Shanthibooshan Subramanian

Bellevue University

DSC680

Amirfarrokh Iranitablob

Table of Contents

1.0 Introduction 2
1.1 Problem Statement. 3
1.2 Importance/usefulness of solving the problem. 4
2.0 Dataset Overview 4
2.1 Dataset Details 4
2.2 Dataset Dictionary 5
2.3 Data Preprocessing for Analysis and Modeling 6
3.0 Comprehensive Analysis Summary 6
3.1 Data Exploration and Initial Insights 6
3.2 Data Preparation 11
3.3 Model Building and Evaluation. 12
4.0 Conclusion 15
4.1 Outcome of analysis and model building. 15
4.2 Model Deployment and Implementation Decision 15
4.3 Potential challenges and additional opportunities. 16
4.4 Conclusion. 17
4.5 Assumptions and Limitations 17
4.6 Recommendations 18
4.7 Ethical Assessment 18
5.0 References 19

1.0 Introduction

Access to clean and safe drinking water is essential for human health and well-being and a fundamental human right. However, the quality of water sources can be compromised by various natural and anthropogenic factors. Geological conditions, such as contaminants in soil and rock formations, can affect water quality. Additionally, human activities such as industrial operations, agricultural practices, and urbanization contribute pollutants like heavy metals, pesticides, and pathogens to water sources. These factors collectively pose significant challenges in ensuring that water is potable—safe for human consumption without risk of health implications.

These models will predict the potability of water sources based on historical data and ongoing monitoring efforts. This predictive capability empowers decision-makers, including governmental agencies, public health authorities, and environmental organizations, to make informed decisions and implement timely interventions to safeguard public health.

Through the application of machine learning, the project not only aims to predict water potability but also to enhance the management of water resources. By identifying trends, anomalies, and potential risks early on, stakeholders can implement preventive measures and policies to mitigate contamination and ensure continuous access to safe drinking water. This proactive approach is crucial in addressing the dynamic nature of water quality challenges, promoting sustainable water management practices, and ultimately protecting the well-being of communities worldwide.

1.1 Problem Statement.

The global challenge of ensuring safe drinking water persists due to various environmental, infrastructural, and regulatory factors. Waterborne diseases remain a significant threat in many regions, particularly in areas with inadequate sanitation and hygiene practices. The objective of this project is to develop robust predictive models that can accurately assess water potability. These models will analyze comprehensive datasets comprising water quality parameters to predict 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 on 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 like 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 impact potability assessments?
- How do seasonal variations affect turbidity levels in water sources, and what are the potential consequences for water treatment processes?
- How do missing data and incomplete records affect the accuracy of predictive models for water potability, and what strategies are effective for handling these challenges?
- What visualizations 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 goal of predicting water potability, the project aims to offer a thorough insight into the complexities of water quality assessment. This approach allows for a nuanced understanding of the variables affecting potability across diverse water sources. By harnessing advanced machine learning methods, the project endeavors to equip stakeholders with the tools needed to proactively ensure safe drinking water and promote effective resource management strategies.

1.2 Importance/usefulness of solving the problem.

Key stakeholders in this project include governmental agencies responsible for water management, public health authorities, environmental organizations, and communities relying on local water sources. Collaboration among these stakeholders is crucial for data sharing, model validation, and the implementation of predictive insights into operational practices.

2.0 Dataset Overview

The dataset employed for this term project can be accessed on Kaggle and encompasses water quality metrics for 3276 different water samples, each representing various parameters used to assess water potability.

2.1 Dataset Details

The dataset used in this project consists of various water quality parameters collected from different sources. These parameters include measurements such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. Each record in the dataset represents a water sample and indicates whether the water is potable (safe to drink) or not. The dataset is publicly available and has been used extensively for water quality analysis and prediction.

Parameters	**WHO Limits**
pH	**6.5 – 8.5**
Hardness	**200 mg/L**
Solids	**1000 ppm**
Chloramines	**4 ppm**
Sulfate	**1000 mg/L**
Conductivity	**400 µS/cm**
Organic Carbon	**10 ppm**
Trihalomethanes	**80 ppm**
Turbidity	**5 NTU**

These parameters are recommended by WHO as permissible levels for safe drinking water.

2.2 Dataset 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).

2.3 Data Preprocessing for Analysis and Modeling

The data preprocessing phase focused on ensuring the dataset was clean, standardized, and suitable for machine learning model training. Steps included handling missing values through imputation, standardizing numerical features, performing feature selection, addressing class imbalance using oversampling techniques like ADASYN, and ensuring data quality through thorough checks. Finally, the dataset was split into an 80-20 ratio for training and testing to evaluate the model's performance. These efforts were crucial in preparing the dataset to build accurate predictive models for water potability.

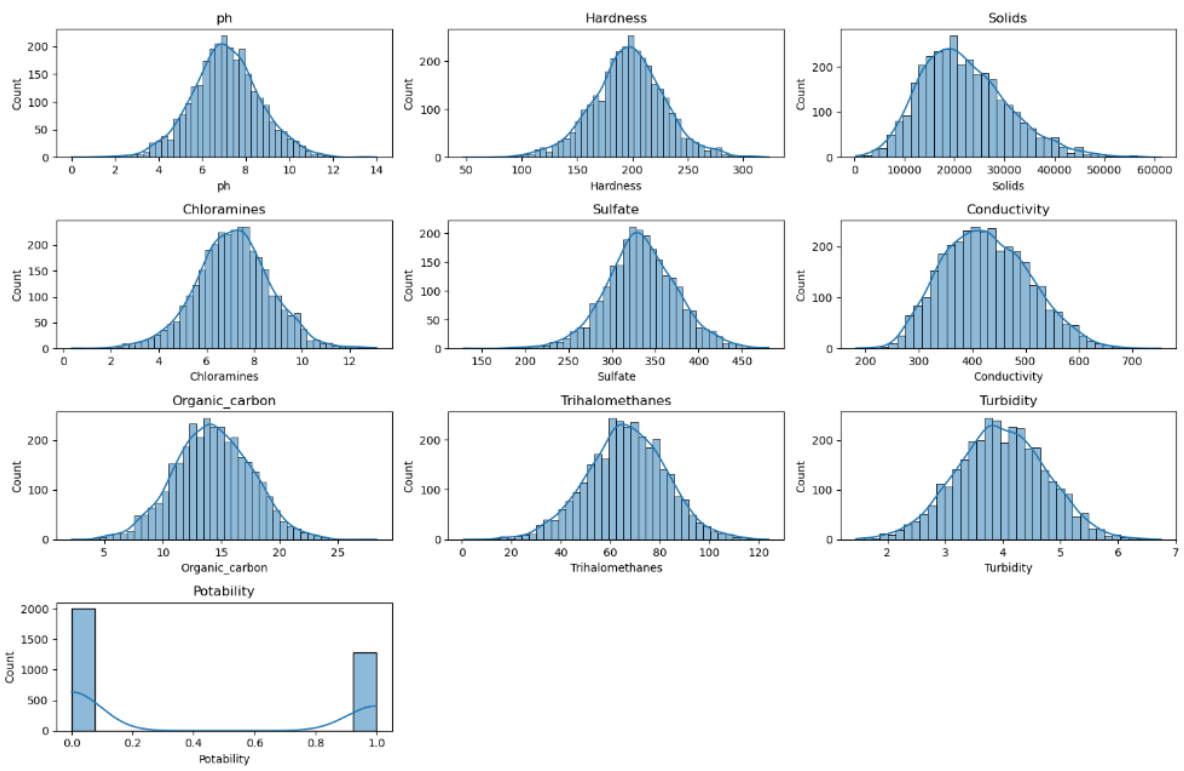
3.0 Comprehensive Analysis Summary

The dataset comprises 3276 records and encompasses 10 columns. Within this dataset, the column labeled "Potability" serves as the target variable for this project, featuring binary values, either "1" or "0," denoting whether the water is potable or not.

3.1 Data Exploration and Initial Insights

Distribution of Numerical Variables

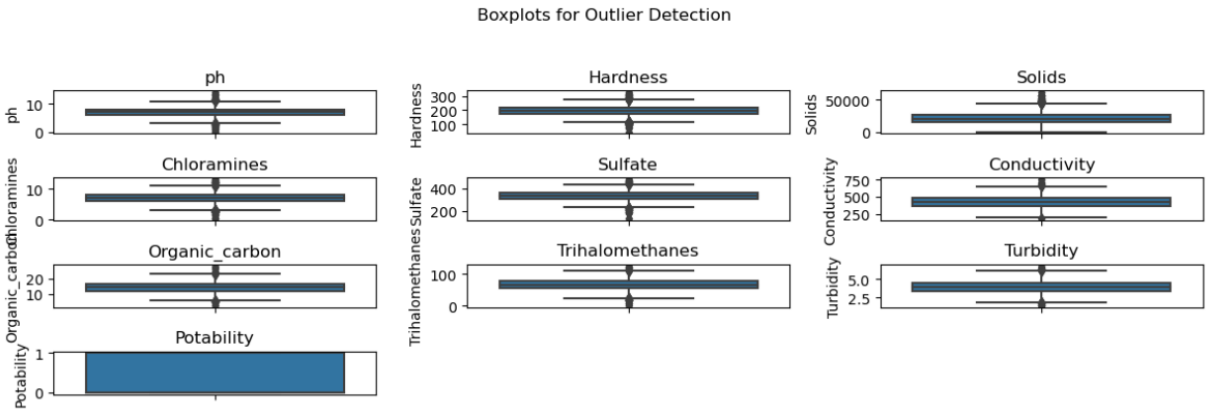
The histograms display the distribution patterns of numerical variables in the dataset. This visualization is crucial for understanding the spread, central tendencies, and potential skewness or outliers in each variable, which informs subsequent data analysis and modeling decisions.



The histograms of water quality parameters reveal that pH levels are normally distributed with a peak around 8-9, though some extreme values are present. Hardness values also follow a normal distribution with a peak at 150. Solids, chloramines, and conductivity are right-skewed, indicating most samples have lower concentrations, with a few showing higher values. Sulfate and organic carbon are normally distributed, peaking at 200-250 and 15-20, respectively. Trihalomethanes and turbidity are right-skewed, like solids and chloramines. The potability bar chart shows an imbalance, with a higher proportion of samples classified as potable. Overall, the histograms suggest that while some parameters are normally distributed, others exhibit skewness, pointing to potential contamination or extreme conditions.

Boxplot Analysis of Water Quality Variables

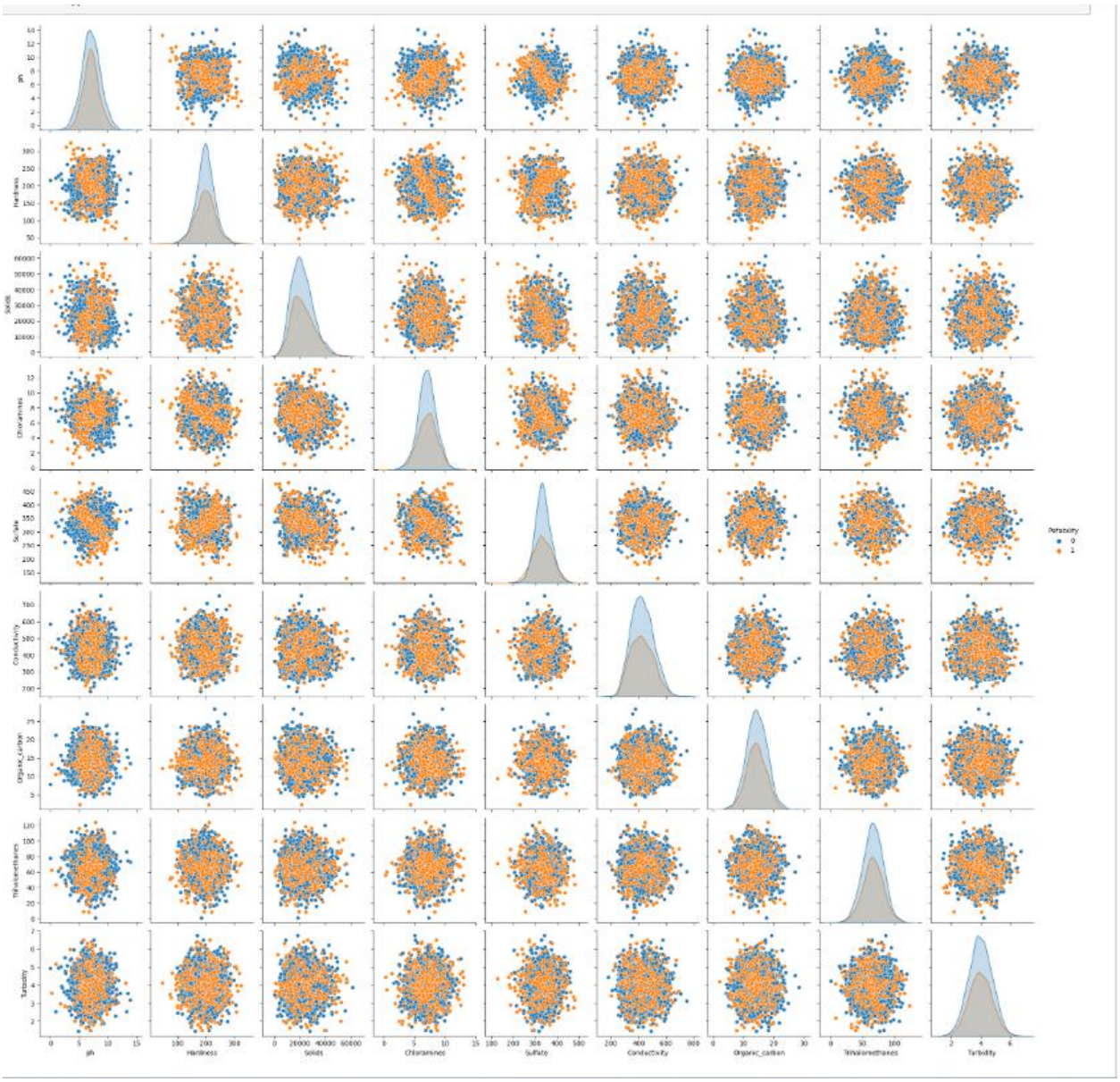
The boxplot visualizes the distribution of various water quality variables. Turbidity, organic carbon, and trihalomethanes are clustered around the median, indicating consistent values. Chloramines, sulfates, and conductivity show a skew towards lower values, suggesting a concentration of data on the lower end. Hardness and solids exhibit a normal distribution centered around the median. Potability, a binary variable, shows more readings for high potability (1) than low potability (0).



This visualization is crucial for the Water Potability project as it provides initial insights into the distribution and characteristics of the water quality parameters. By identifying data skewness, clustering, and potential outliers, this exploratory data analysis aids in understanding the dataset's structure. Such insights are essential for making informed decisions during data preprocessing, feature selection, and model building, ultimately enhancing the accuracy and reliability of predictive models for water potability.

Pairplot to Visualize Relationships

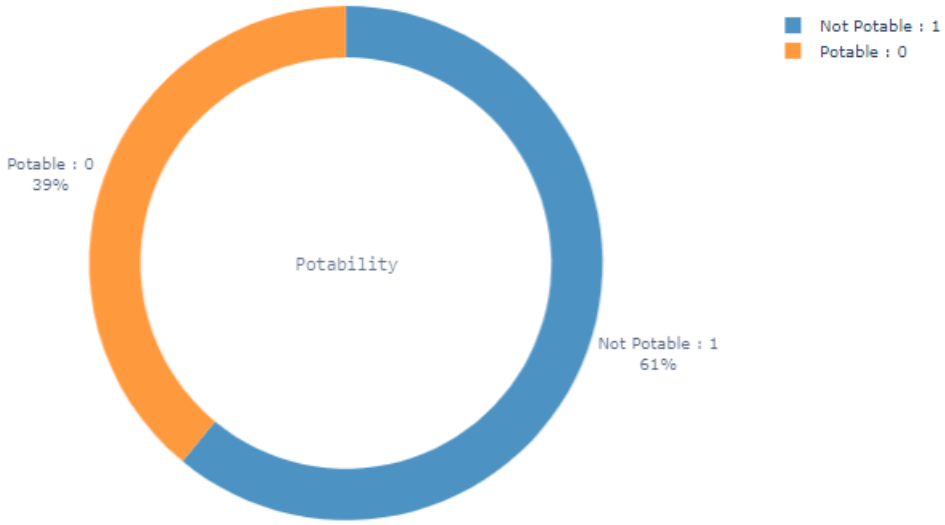
Displaying pairwise relationships between variables, colored by the Potability target, helps in identifying patterns and correlations between water quality parameters. It reveals clusters or groupings associated with potable versus non-potable water, highlighting significant variables and guiding feature selection. This visualization also provides insights into necessary preprocessing steps, enhancing overall data understanding and supporting more effective model building.



The pair plot provides valuable insights into the relationships between water quality parameters and water potability. It reveals that many variables are right-skewed and exhibit outliers. The plot highlights a positive correlation between Hardness and Solids, with visible clustering suggesting possible subgroups within the data. Despite these patterns, significant overlap between potable and non-potable samples suggests that individual variables alone may not be sufficient for strong predictions. This visualization helps in identifying correlations and clusters, which can guide further analysis such as correlation assessment, feature importance evaluation, outlier handling, dimensionality reduction, and model development to enhance the accuracy of potability predictions.

Pie Chart of Potability Distribution

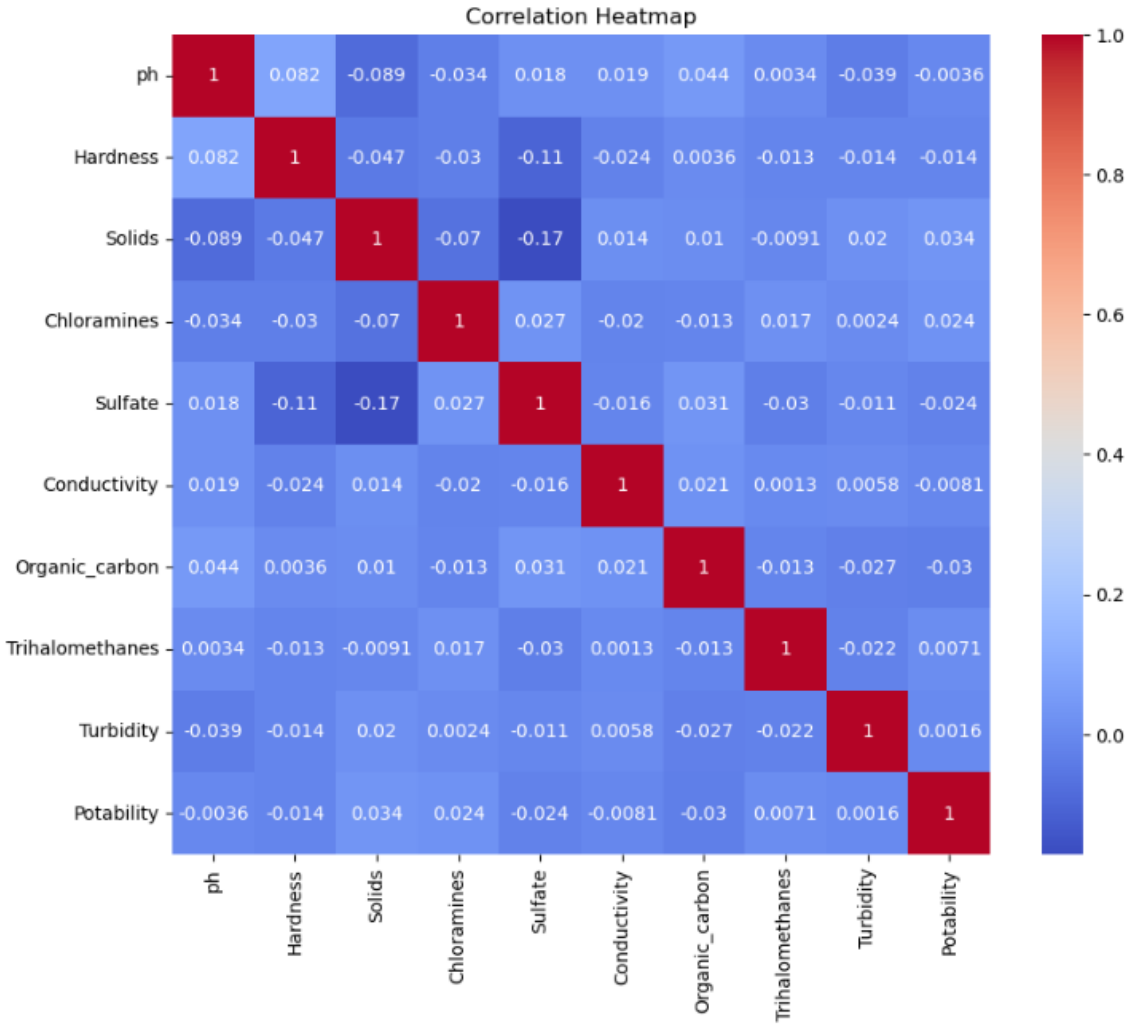
The pie chart visualizes the distribution of water samples based on their potability. It shows the proportion of potable and non-potable samples, with color-coded segments for clear differentiation.



The donut chart effectively communicates the distribution of water samples between potable and non-potable categories. By clearly showing that 61% of samples are non-potable and 39% are potable, this visualization helps quickly grasp the balance in the dataset. It highlights the imbalance in water quality, making it easier to understand the prevalence of non-potable samples, which is crucial for targeting areas needing improvement.

Correlation Heatmap of Water Quality Parameters

The correlation heatmap provides a visual representation of the relationships between various water quality parameters. The heatmap indicates that there are generally weak correlations among the parameters. For instance, pH shows a slight positive correlation with Hardness and a slight negative correlation with Solids. Chloramines have a moderate positive correlation with Sulfate, while Conductivity shows weak correlations with other variables



Potability, the target variable, exhibits very weak correlations with other parameters, indicating that water potability is influenced by a combination of factors rather than being strongly associated with any single parameter. This analysis highlights the complexity of predicting water potability and underscores the need for sophisticated modeling techniques to accurately assess water quality.

3.2 Data Preparation

In preparation for modeling and analysis, the original dataset underwent the following steps:

After identifying missing values in key columns such as pH, Sulfate, and Trihalomethanes, they were handled by imputing them with the mean values derived from their respective

'Potability' groups. This preprocessing step ensured that the dataset maintained integrity and accuracy for subsequent analysis.

Following data cleaning and imputation, the refined dataset comprises 3276 observations and 10 columns. Statistical insights reveal that the mean values for important water quality parameters, such as pH (7.08), Hardness (204.89 mg/L), and Solids (20791 ppm), are indicative of a diverse range of water characteristics across the dataset.

To address the imbalance in the target variable (Potability), the ADASYN method was employed, resulting in a balanced dataset with 4049 total observations. This balanced approach ensures that both safe and unsafe water classifications are adequately represented, enhancing the robustness and predictive power of subsequent machine learning models used to predict water potability.

These data preparation steps ensure that the dataset is cleaned, transformed, and optimized for subsequent analysis and modeling tasks. By addressing missing values, imputing key water quality parameters, and balancing the dataset using ADASYN, the dataset is now prepared for building predictive models to accurately predict water potability. These efforts aim to enhance the dataset's integrity and reliability in evaluating water safety and quality.

3.3 Model Building and Evaluation.

Target Variable and Objective:

The target variable for this project is 'Potability,' indicating whether water from a source is safe for human consumption. The focus is on building and evaluating models to accurately predict this outcome.

Model Building:

The model building involved employing various machine learning algorithms such as Logistic Regression, Random Forest, and Gradient Boosting Classifier. Evaluation metrics including accuracy, precision, recall, F1-score, and ROC AUC were utilized to assess and compare model performance.

Model Results:

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	Training(Sec)
SVM	0.66	0.72	0.77	0.64	0.70	0.32	0.32	8.9sec
KNN	0.65	0.68	0.75	0.63	0.68	0.29	0.29	0.06sec
Decision Tree	0.64	0.64	0.63	0.65	0.65	0.28	0.28	2.55sec
Random Forest	0.76	0.83	0.78	0.75	0.76	0.52	0.52	13sec
CatBoost	0.73	0.81	0.74	0.74	0.74	0.47	0.47	32sec
LightBGM	0.73	0.81	0.73	0.73	0.73	0.46	0.46	1.18sec
XGBoost	0.74	0.82	0.78	0.74	0.76	0.49	0.49	2.82sec

- **Accuracy:** Random Forest achieved the highest accuracy of 75.80%, followed by XGBoost at 74.69% and CatBoost at 73.58%. (Note: The initial accuracy reported as 76.30% may have been a mistake or a typo in your initial summary.)

- **AUC (Area Under the Curve):** Random Forest outperformed other models with an AUC of 0.8370, indicating the best performance in distinguishing between potable and non-potable water.

- **Recall:** SVM and XGBoost had the highest recall scores, with SVM at 77.51% and XGBoost at 78.23%, showing their effectiveness in identifying potable water instances among actual potable cases.

- **Precision:** Random Forest had the highest precision at 75.58%, 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 0.770, balancing precision and recall. XGBoost followed closely with an F1 score of 0.761.

- **Kappa:** Random Forest had the highest Kappa coefficient of 0.515, 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 0.515 and XGBoost at 0.493, reflecting robust binary classification performance.

- **Training Time:** CatBoost had the longest training time at 32.55 seconds, followed by Random Forest at 13.08 seconds. In contrast, KNN had the shortest training time at 0.07 seconds, with Decision Tree and LightGBM also being relatively quick.

Random Forest emerged as the top-performing model overall, excelling in accuracy, AUC, precision, and F1 score, and having a high Kappa coefficient and MCC. XGBoost also 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

Best Model: Random Forest emerged as the most effective model overall. It initially achieved an accuracy of 75.80% and an AUC of 0.8370, demonstrating superior performance in distinguishing between classes. After hyperparameter tuning, its accuracy was 75.43% with an F1-score of 0.75, reflecting its consistent effectiveness.

Strong Performers: LightGBM excelled in the hyperparameter tuning phase, achieving the highest accuracy of 75.68% among the tuned models. Initially, it had an accuracy of 73.09% and an AUC of 0.8168. XGBoost initially showed the highest accuracy at 74.69%, but its accuracy dropped to 72.47% after tuning, although it maintained strong recall and precision.

Other Models: CatBoost provided solid performance with an accuracy of 75.31% after tuning, up from 73.58% initially. SVM achieved an accuracy of 66.30%, notable for its high recall but lower precision. KNN had an accuracy of 64.69%, with very fast training but lower accuracy. Decision Tree had the lowest accuracy at 64.32% and performed poorly compared to the other models.

4.0 Conclusion

4.1 Outcome of analysis and model building.

The analysis aimed to predict water potability using machine learning models, evaluating their performance based on key metrics including accuracy, AUC (Area Under the Curve), recall, precision, F1-score, kappa, MCC (Matthews correlation coefficient), and training time.

Random Forest emerged as the most effective model overall, achieving the highest accuracy of 75.80% and an AUC of 0.8370 in its initial performance. Even after hyperparameter tuning, it maintained strong results with an accuracy of 75.43% and an F1-score of 0.75, demonstrating its consistent effectiveness across key metrics.

LightGBM performed best in the tuning phase, with an accuracy of 75.68%, surpassing other models after adjustment. Initially, it had a solid performance but improved further post-tuning. XGBoost, despite showing the highest initial accuracy of 74.69%, experienced a drop to 72.47% after tuning, though it retained strong recall and precision.

CatBoost also showed notable improvement, achieving an accuracy of 75.31% after tuning. SVM and KNN had lower overall performance metrics, with Decision Tree being the least effective.

In summary, Random Forest, LightGBM, and XGBoost are the most promising models for predicting water potability. Random Forest stands out as the most consistent performer across all evaluation metrics, providing the highest accuracy and a strong AUC. This makes it particularly valuable for decision-makers aiming to ensure safe drinking water and manage resources effectively.

4.2 Model Deployment and Implementation Decision

The analysis of various machine learning models for predicting water potability has highlighted Random Forest and XGBoost as the top performers. These models demonstrated superior accuracy and robustness in distinguishing between potable and non-potable water sources, with strong metrics such as F1-score and MCC

Deployment Strategy:

- **Random Forest** is recommended for deployment due to its highest accuracy of 75.80%, an F1-score of 0.770, and an MCC of 0.515. It consistently performed well across multiple metrics, making it a reliable choice for practical implementation.

- **XGBoost,** with an accuracy of 74.69%, a recall of 78.23%, and a precision of 74.15%, is also a strong candidate for deployment. Its performance in recall and precision makes it suitable for cases where these metrics are prioritized.

Implementation Decision:

- **Random Forest** offers a balanced approach with high accuracy, robustness, and reliability in predicting water potability, making it the preferred choice for practical implementation.

- **XGBoost** provides competitive performance and could be considered based on specific operational requirements or preferences, particularly where high recall is essential.

4.3 Potential challenges and additional opportunities.

A potential challenge was handling imbalanced data, as the number of potable water samples was much lower than non-potable ones. This imbalance could lead to a biased model that favors the majority class. Techniques such as oversampling with ADASYN were employed to mitigate this issue. Another challenge was ensuring the interpretability of the model. It is crucial that stakeholders, particularly those in water treatment facilities and public health authorities, trust and understand the model's predictions. Efforts were made to maintain a balance between model complexity and interpretability.

The predictive model developed in this project has several potential future uses and applications. Beyond predicting water potability, the model could be adapted to predict other water quality issues, such as contamination with specific chemicals or microbial pathogens. Expanding the dataset to include samples from more diverse regions would improve the model's robustness and generalizability. Furthermore, collaboration with environmental agencies could enhance data collection efforts, leading to more comprehensive analyses and better-informed public health decisions.

4.4 Conclusion.

In conclusion, the analysis and model-building phase highlighted **Random Forest** as the most effective classifier for predicting water potability, achieving the highest accuracy of 75.80% and strong recall rates for both potable and non-potable water. Random Forest also excelled with an F1-score of 0.770 and an MCC of 0.515, demonstrating its superior predictive power and consistency.

CatBoost and XGBoost also performed well, with XGBoost showing high recall (78.23%) and competitive accuracy (74.69%), and CatBoost achieving an accuracy of 73.58% and a solid F1-score. However, Random Forest's overall performance, particularly in accurately identifying potable water sources, set it apart as the most effective model.

This phase underscores the importance of selecting models that not only achieve high overall accuracy but also balance targeted performance metrics tailored to specific objectives, such as ensuring safe drinking water and mitigating health risks associated with contaminated water sources.

4.5 Assumptions and Limitations

Throughout the analysis, several assumptions were made. Firstly, it is assumed that the dataset is representative of the general water quality conditions across various regions. The data collected is presumed to reflect a wide range of water sources and contamination scenarios. Secondly, it is assumed that the water treatment processes remained consistent throughout the data collection period. This consistency is crucial to ensure that the model's predictions are reliable and not influenced by changes in water treatment methodologies. Lastly, the project assumes that the machine learning models used will generalize well to new, unseen data, maintaining their predictive accuracy outside the initial dataset.

Despite comprehensive analysis, several limitations should be acknowledged to ensure accurate interpretation and application of findings. The dataset size is relatively small, which may affect the model's ability to generalize to broader populations or different geographic areas. A larger dataset would provide more robust training and validation, potentially improving model accuracy. Additionally, there may be inherent biases in the data collection methods, such as the specific times and locations where samples were taken, which could impact the model's performance. These biases need to be carefully managed to ensure the model's predictions remain accurate and reliable.

4.6 Recommendations

Based on the findings and outcomes of this project, several recommendations are proposed. We recommend deploying the Random Forest model to predict water potability due to its superior performance. Continuous monitoring of water quality using the identified key parameters is essential. Integrating the predictive model into water treatment facilities' workflows will enable real-time monitoring and early detection of potential contamination issues. It is also recommended to refine the model continuously with new data and advanced techniques to improve its accuracy and reliability. Additionally, using the model as a supplementary tool to existing water quality monitoring systems will enhance overall water safety efforts

4.7 Ethical Assessment

An ethical assessment is a critical component of this project. Ensuring data privacy and security throughout the project lifecycle is paramount. Measures must be taken to protect sensitive information and comply with relevant data protection regulations. Additionally, addressing biases in the model is essential to prevent misclassification and ensure fairness. The model should be designed and tested to minimize any potential biases that could disproportionately affect certain populations or regions. Ethical considerations must be embedded in every stage of the project to ensure the outcomes are just and equitable for all stakeholders.

5.0 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/