Abstract

This study harnesses machine learning techniques to predict water potability based on key chemical parameters using a comprehensive dataset of water quality measurements. The dataset consists of 3,276 entries with 10 features, including pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potability classification. Through exploratory data analysis (EDA) and feature engineering guided by scientific insights, we delve into the intricate relationship between water chemistry and portability. Missing values are imputed using the scientist's value of orientation to determine the PH and the mean was used for other missing values of variables to ensure data integrity, and model evaluation metrics such as accuracy, precision, recall, and F1 score are employed to assess predictive performance. Results highlight the efficacy of machine learning algorithms, including Random Forest, Decision Tree, XGBoost, SVM, and Gradient Boosting, in water quality assessment. This interdisciplinary approach underscores the transformative potential of data science in safeguarding global water resources and public health.

1. Introduction

Access to safe and potable drinking water is a fundamental human right and a critical necessity for public health and well-being worldwide. However, ensuring water quality remains a significant challenge, particularly in regions where clean water access is limited. Traditional methods of water quality assessment, reliant on labor-intensive and time-consuming laboratory analyses, often fail to provide real-time insights needed for timely interventions (Mahajna et al., 2022).

Recent advancements in machine learning (ML) and data mining offer promising solutions to transform water quality assessment by leveraging historical data and employing predictive algorithms (Gu et al., 2023). These innovative approaches can expedite the analysis of water samples and enhance cost efficiency, enabling the identification of crucial water quality parameters essential for assessing water potability (Sabry et al., 2022).

In regions like Bangladesh, where water scarcity and quality issues pose significant risks to public health, the need for effective water potability assessment is urgent (Cheng et al., 2022). For example, approximately 70% of piped water supply sources in certain areas are contaminated, highlighting the pressing need for reliable monitoring and predictive tools (Cheng et al., 2022).

Machine learning models have demonstrated considerable potential in addressing water quality challenges globally. For instance, Australia utilizes recent ML models to predict potable water systems, underscoring the effectiveness of these techniques in water resource management (Gu et al., 2023).

The primary objective of this study is to develop and evaluate machine learning models using a comprehensive dataset obtained from Kaggle (www.kaggle.com/datasets/devanshibavaria/water-potability-dataset-with-10-parameteres) to predict water potability. This dataset consists of nine input attributes, including pH levels, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, along with one target attribute indicating potability status.

By exploring a range of ML algorithms, such as Random Forest, Decision Tree, XGBoost, SVM, and Gradient Boosting, we aim to identify the most effective approach for distinguishing between potable and non-potable water samples. This research aims to contribute to the development of

cost-effective and scalable methods for assessing water potability, critical for public health and environmental sustainability.

Background Research

This section provides a comprehensive overview of the existing research and studies relevant to the prediction of water potability using machine learning. For instance, Dailisan et al. (2022) investigated the use of Support Vector Machine (SVM) algorithms for accurately predicting water potability. Their study demonstrated the efficacy of SVM in classifying water samples into potable and non-potable categories based on key parameters such as turbidity, pH, iron content, chloride levels, and biological contaminants. By leveraging SVM, the researchers achieved reliable predictions of water quality status, essential for safeguarding public health.
This research underscores the importance of machine learning in water potability assessment, providing a data-driven approach to enhance monitoring and decision-making regarding drinking water safety. By leveraging advanced algorithms like SVM, researchers and policymakers can access timely and accurate insights into water quality, supporting proactive measures to ensure safe and potable drinking water for communities.

Methods

	DATA COLLECTION
	EXPLORATORY DATA ANALYSIS
	DATA PRE-PROCESSING
	POST-PROCESS EDA
	FEATURE ENGINEERING
	TRAIN AND TEST
	MODEL AND PRUNING
	EVALUATION AND COMPARISON OF ALGORITHMS

Data Collection and Structure:

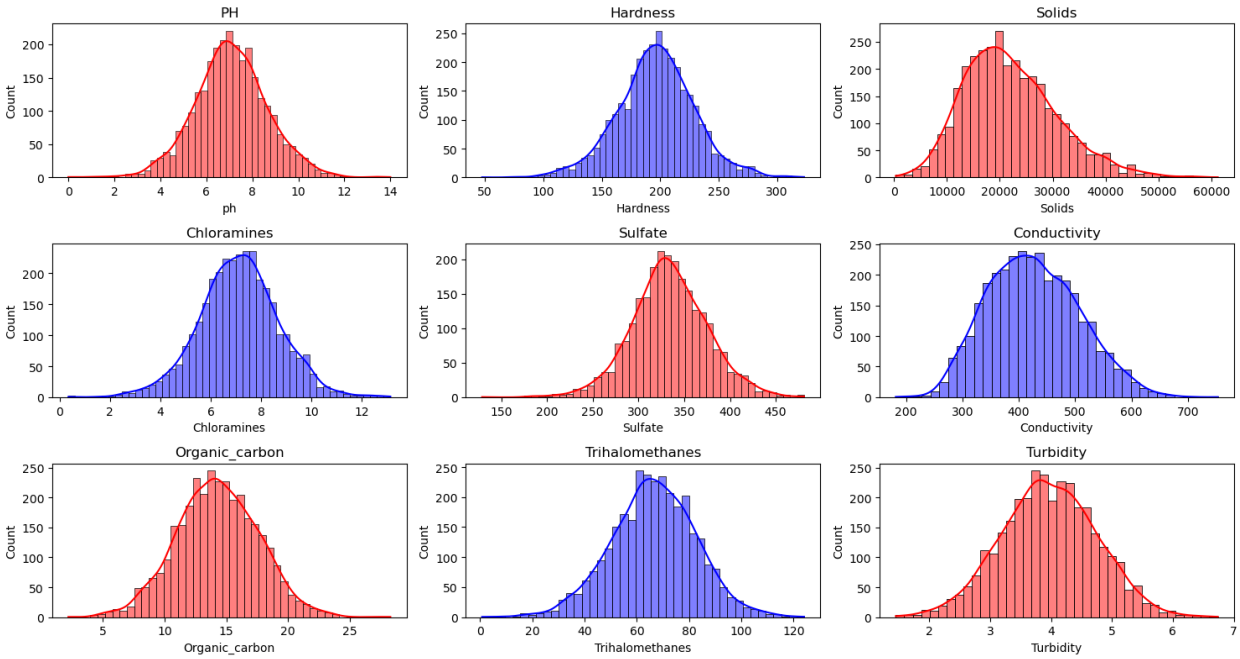
Data for this study was obtained from Kaggle(Water Potability Dataset (kaggle.com). The dataset comprises 3,276 entries, each representing measurements related to water quality, including properties like pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and a binary portability classification (0 or 1). The target variable, potability (1 for safe to drink and 0 for not potable), is crucial for analyzing which factors influence water safety and building predictive models.

```
2  8.099124  224.236259  19909.541732     9.275884         NaN    418.606213
3  8.316766  214.373394  22018.417441     8.059332  356.886136    363.266516
4  9.092223  181.101509  17978.986339     6.546600  310.135738    398.410813

   Organic_carbon  Trihalomethanes  Turbidity  Potability
0       10.379783        86.990970   2.963135           0
1       15.180013        56.329076   4.500656           0
2       16.868637        66.420093   3.055934           0
3       18.436524       100.341674   4.628771           0
4       11.558279        31.997993   4.075075           0
            ph    Hardness        Solids  Chloramines     Sulfate  \
3271  4.668102  193.681735  47580.991603     7.166639  359.948574
3272  7.808856  193.553212  17329.802160     8.061362         NaN
3273  9.419510  175.762646  33155.578218     7.350233         NaN
3274  5.126763  230.603758  11983.869376     6.303357         NaN
3275  7.874671  195.102299  17404.177061     7.509306         NaN

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
3271    526.424171       13.894419        66.687695   4.435821           1
3272    392.449580       19.903225              NaN   2.798243           1
3273    432.044783       11.039070        69.845400   3.298875           1
3274    402.883113       11.168946        77.488213   4.708658           1
3275    327.459760       16.140368        78.698446   2.309149           1
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 256.1 KB
None
```

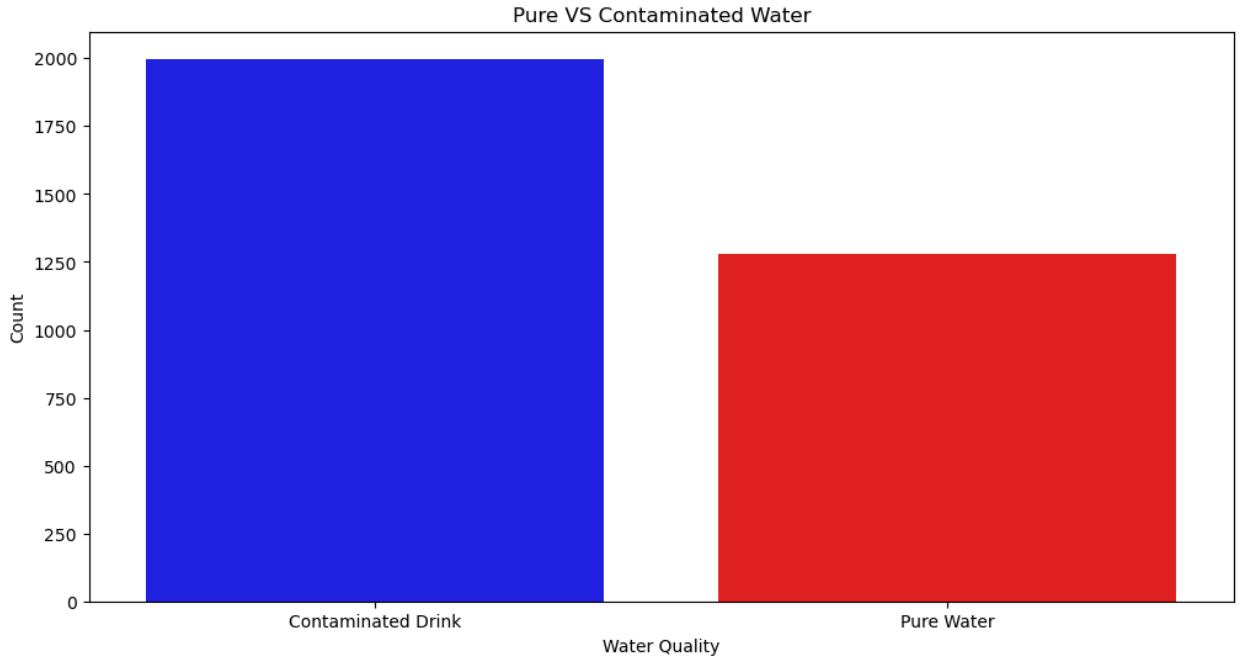
Exploratory Data Analysis

- Box Plots: The distributions of pH, hardness, and other variables exhibit typical characteristics, such as symmetrical shapes and moderate spreads. This suggests consistency in some parameters like pH and organic carbon, while others like solids and sulfate show significant variability.

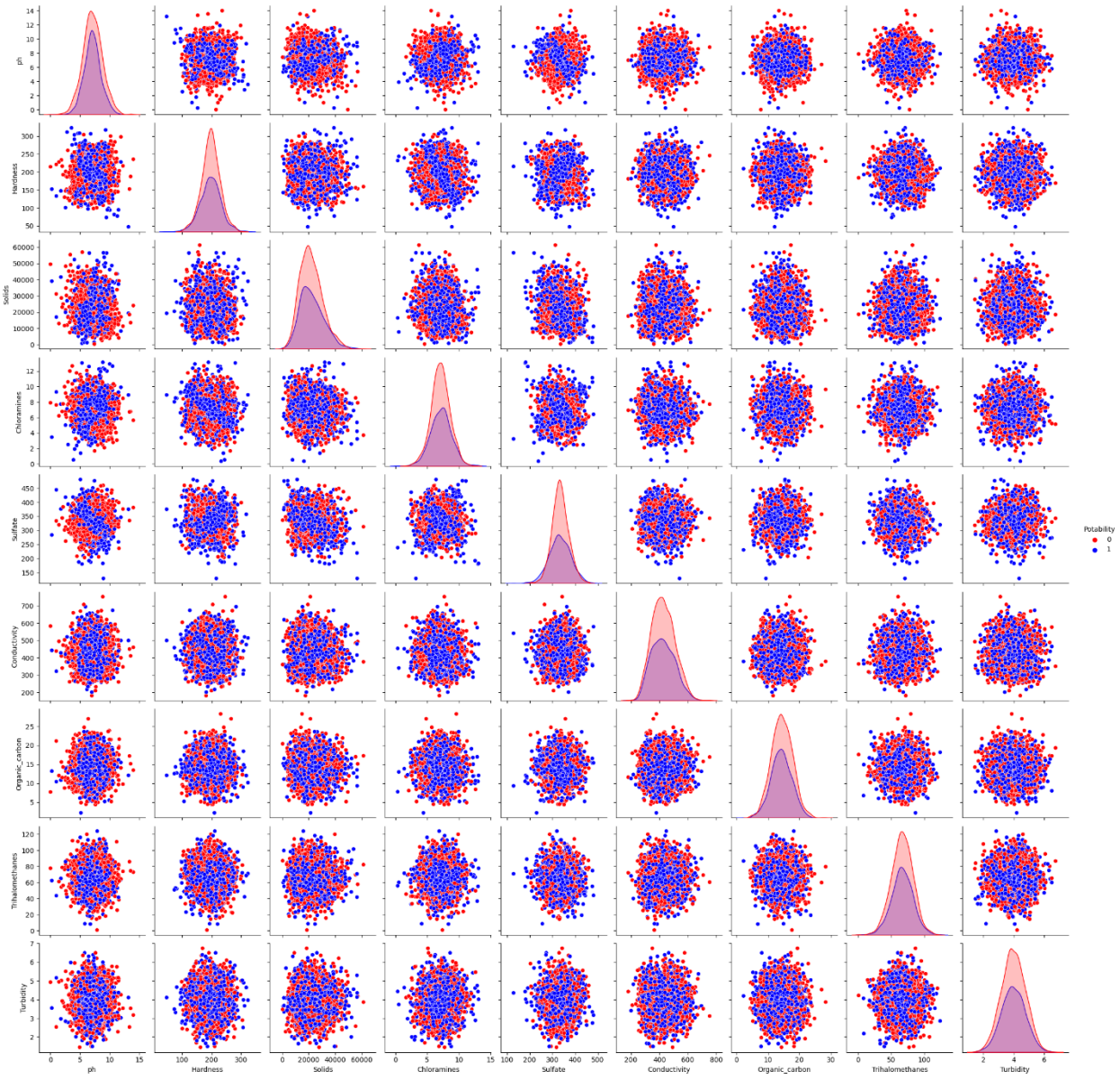
- Histograms with Kernel Density Estimates: Most variables display bell-shaped curves, indicating normal distributions. Variables like solids and conductivity have wider spreads, reflecting greater variability across samples.



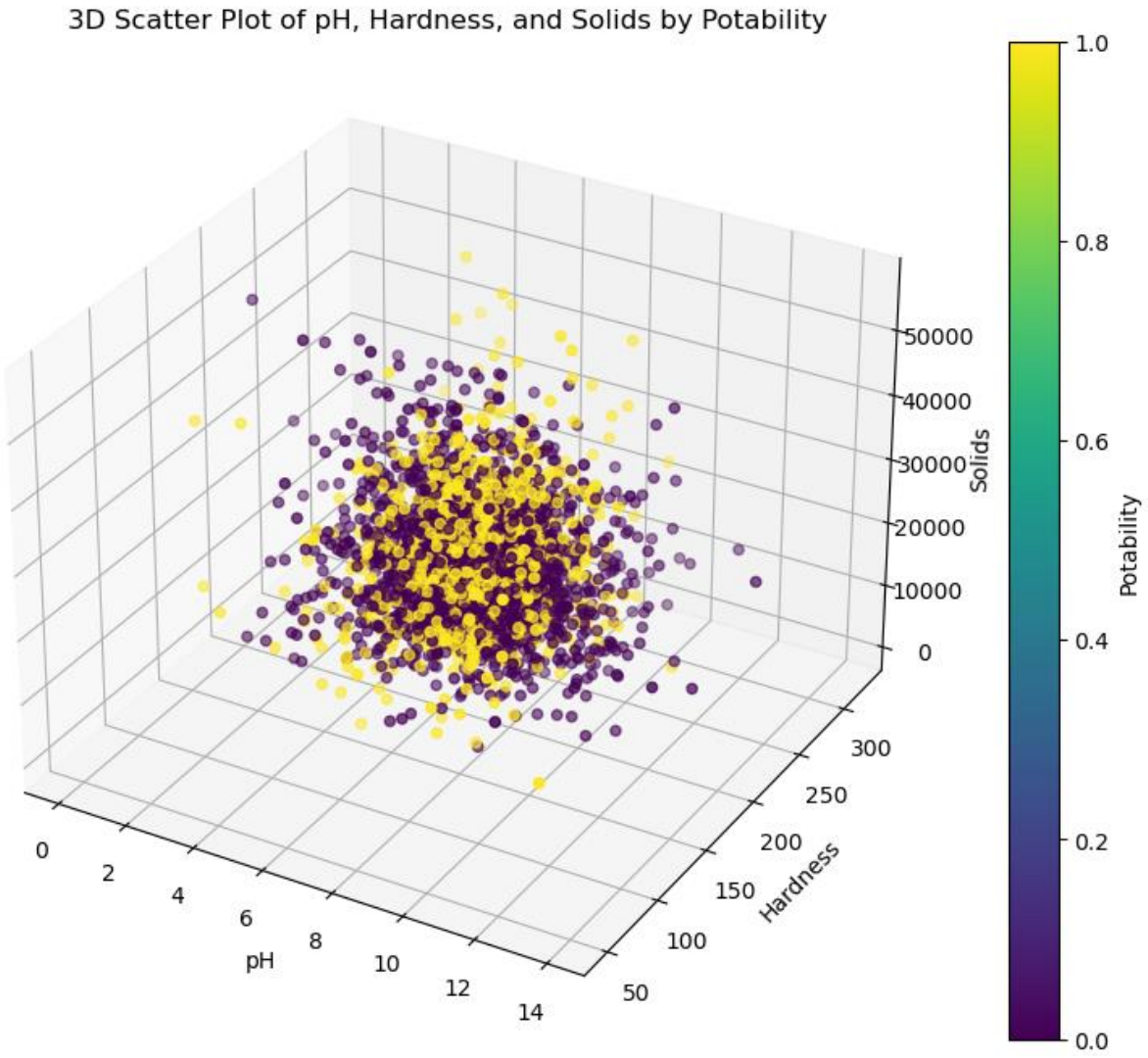
- Bar Chart: The comparison between contaminated and pure water samples highlights a higher prevalence of contamination.



- Scatterplot Matrix: The scatter plots reveal complex relationships and no distinct groupings between variables. This implies that individual variable comparisons may not be sufficient for classification.

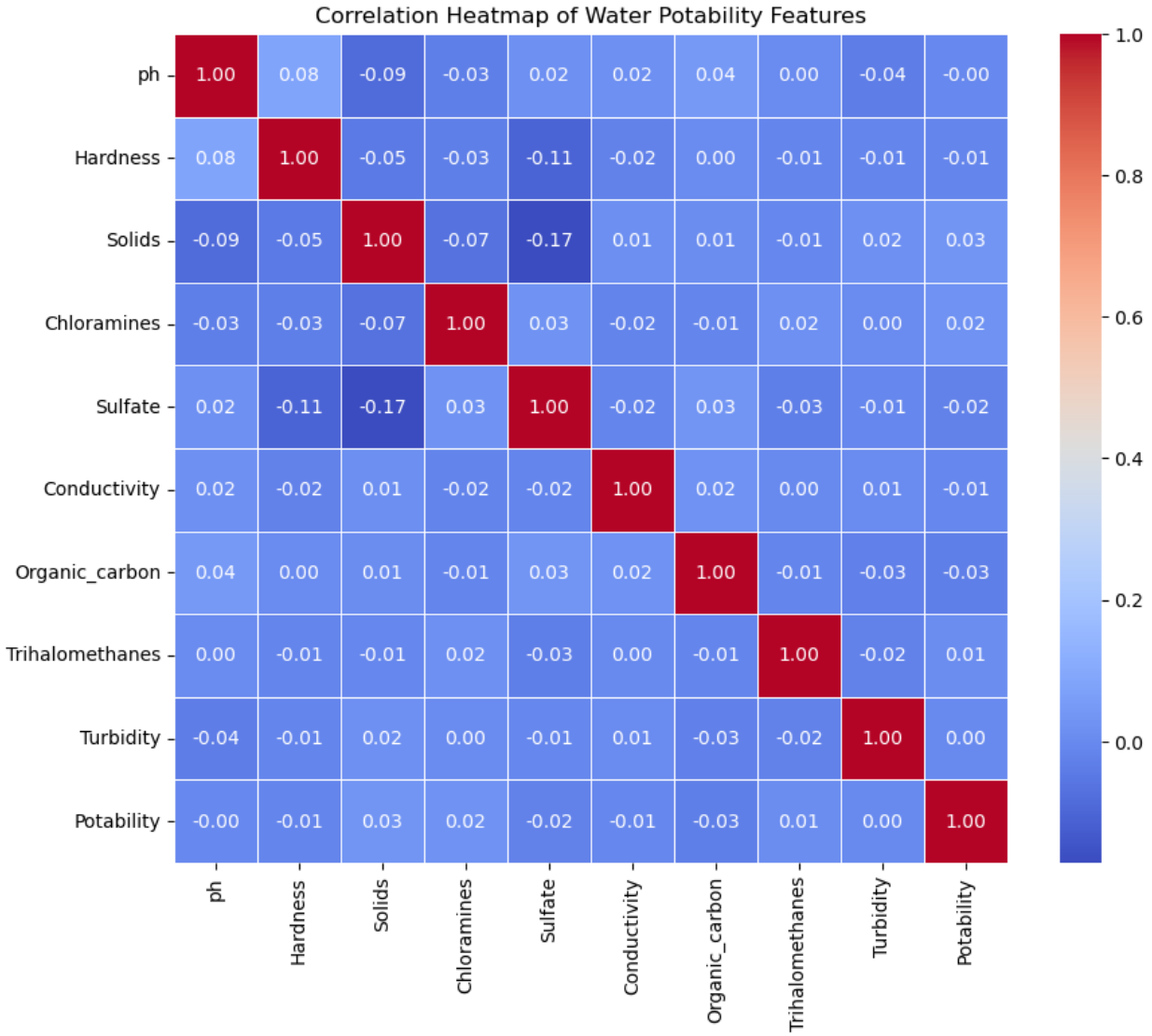


- 3D Scatter Plot and Hexbin Plot: Visualizations show overlapping clusters of water samples with no clear separation based on potability, suggesting that simple thresholds may not predict potability effectively.



- Violin Plot and Density Plot: These plots illustrate the distribution of turbidity and conductivity measurements, providing insights into typical ranges and variability in the dataset.

Correlation matrix: The correlation heatmap offers a visual summary of how each water quality parameter relates to the others, with red denoting positive correlation and blue indicating negative correlation. Most variables show weak correlations with each other, as indicated by the prevalence of colors closer to white, meaning there's no single dominant factor in predicting water quality features. The analysis conducted through this project reveals complex and subtle relationships, highlighting the necessity for a multifaceted approach to understanding water quality dynamics.



Algorithm Selection

These machine learning algorithms, including Random Forest, Decision Tree, XGBoost, GradientBoostClassifier, and Support Vector Machine (SVM), employ sophisticated techniques such as ensemble learning, handling non-linear relationships, and maximizing model interpretability and robustness.

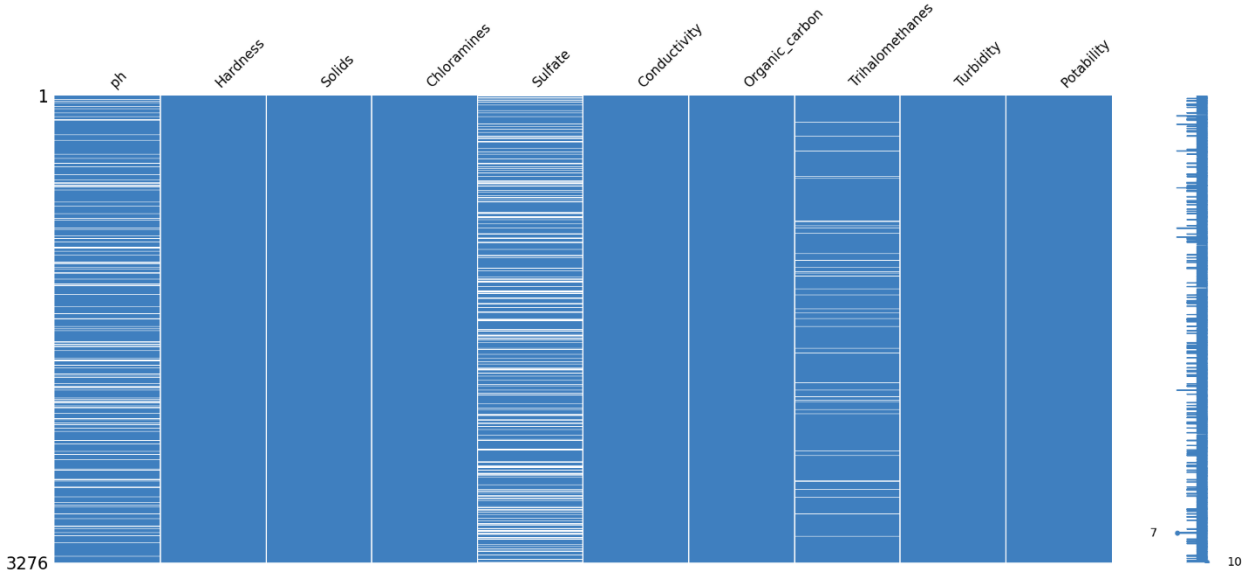
- Random Forest: This method de-correlates trees by randomly selecting feature subsets and samples during construction, which helps reduce overfitting. By combining multiple decision trees, Random Forest provides robust predictions for complex data patterns and offers insights into critical features influencing potability prediction.
- Decision Tree: Known for its interpretability and visualization capabilities, Decision Trees are adept at identifying critical water quality measurements and naturally handling non-linear relationships in data.

- XGBoost: This algorithm effectively handles missing values, which are common in real-world datasets, and leverages a gradient boosting mechanism to iteratively learn from residuals, improving prediction accuracy.
- GradientBoostClassifier: By sequentially building trees to correct errors, this classifier refines predictions and minimizes overfitting. Its use of ensemble learning helps capture nuanced relationships in water quality attributes.
- Support Vector Machine (SVM): SVM is capable of modeling non-linear relationships effectively and ensures robust classification by maximizing margins between different classes.

Data Preprocessing

Visualizations and data preprocessing steps were performed to gain insights into the water quality dataset.

1. Missing Data Visualization:
    - The matrix plot highlighted missing values as white gaps across features, aiding in identifying data completeness.



2. Handling Missing pH Values:
    - Conditions based on water hardness and potability were defined to impute missing pH values.
    - For hardness ≤ 150, average pH values were determined for potable and non-potable water.
    - Similarly, for hardness > 150, distinct average pH values were established based on potability status.

Reference: Howell, E. (2013, March 26). What makes water hard?. LiveScience.

3. Handling Missing Trihalomethanes and Sulfate Values:
    - Missing values in 'Trihalomethanes' were filled using the median value of the column.

- The median approach was chosen due to its robustness against outliers, preserving data distribution integrity.
- The same methodology was applied to fill missing values in the 'Sulfate' column.

Reasons for Imputation Methods:

I. Median for Missing Trihalomethanes and Sulfate Values:
   - The median was preferred over the mean to handle missing values due to its resilience against outlier influence.

II. Scientific Approach for pH Imputation:
   - Water hardness conditions were leveraged, aligning with scientific findings on water potability and pH levels.

Handling of Outliers
In this analysis, outliers were not specifically addressed due to their potential information content and role in preserving data integrity. Outliers can represent unique observations relevant to real-world scenarios or specific conditions and removing them might distort the dataset's true distribution and characteristics. Moreover, the analysis scope might not necessitate outlier removal, as outliers could provide insights into The performance metrics for different machine learning models indicate varying levels of success on both training and testing datasets.

Modeling and Optimization

In developing a predictive model for water potability, we began by defining our feature matrix (X) and target variable (y). To address the class imbalance, we utilized the RandomOverSampler to upsample the minority class, ensuring a balanced representation of potable and non-potable water samples in our training data. After splitting our upsampled data into training and testing sets , we applied feature scaling using StandardScaler to standardize our features, promoting model stability and preventing bias due to varying feature scales.
For modeling, we employed a Tree Classifier with a maximum depth of 3 and 5 and a random state for reproducibility. Our model was trained on the scaled training data and evaluated using 5-fold cross-validation, assessing accuracy as a metric to gauge stability and generalizability.
To evaluate model performance comprehensively, we calculated key metrics such as accuracy, F1-score, precision, and recall on both training and testing datasets. These metrics provide insights into the model's ability to accurately predict water potability.
Visualizing model performance, we generated a Receiver Operating Characteristic (ROC) curve to illustrate the trade-off between sensitivity and specificity. Additionally, a confusion matrix heatmap allowed us to visualize true positives, true negatives, false positives, and false negatives predicted by our model.
In summary, our approach encompassed essential steps in building, training, and evaluating a decision tree model for water potability prediction. Through thoughtful consideration of class imbalance and rigorous evaluation using cross-validation and performance metrics, we ensured a robust and effective predictive model.

PERFORMANCE METRICS
Random Forest:

- Achieved perfect accuracy (1.0000) and F1 score (1.00) on the training set, suggesting potential overfitting as the model perfectly fits the training data.
- Demonstrates good generalization on the testing set with an accuracy of 72.41% and an F1 score of 71.96%.

Decision Tree:

- Shows strong performance on the training set with an accuracy of 80.84% and an F1 score of 80.00%.
- Maintains decent performance on the testing set with an accuracy of 72.87% and an F1 score of 70.00%.

XGBoost:

- Fair performance on the training set with an accuracy of 77.90% and an F1 score of 76.35%.
- Exhibits similar performance on the testing set with an accuracy of 72.10% and an F1 score of 69.44%.

Gradient Boosting Classifier:

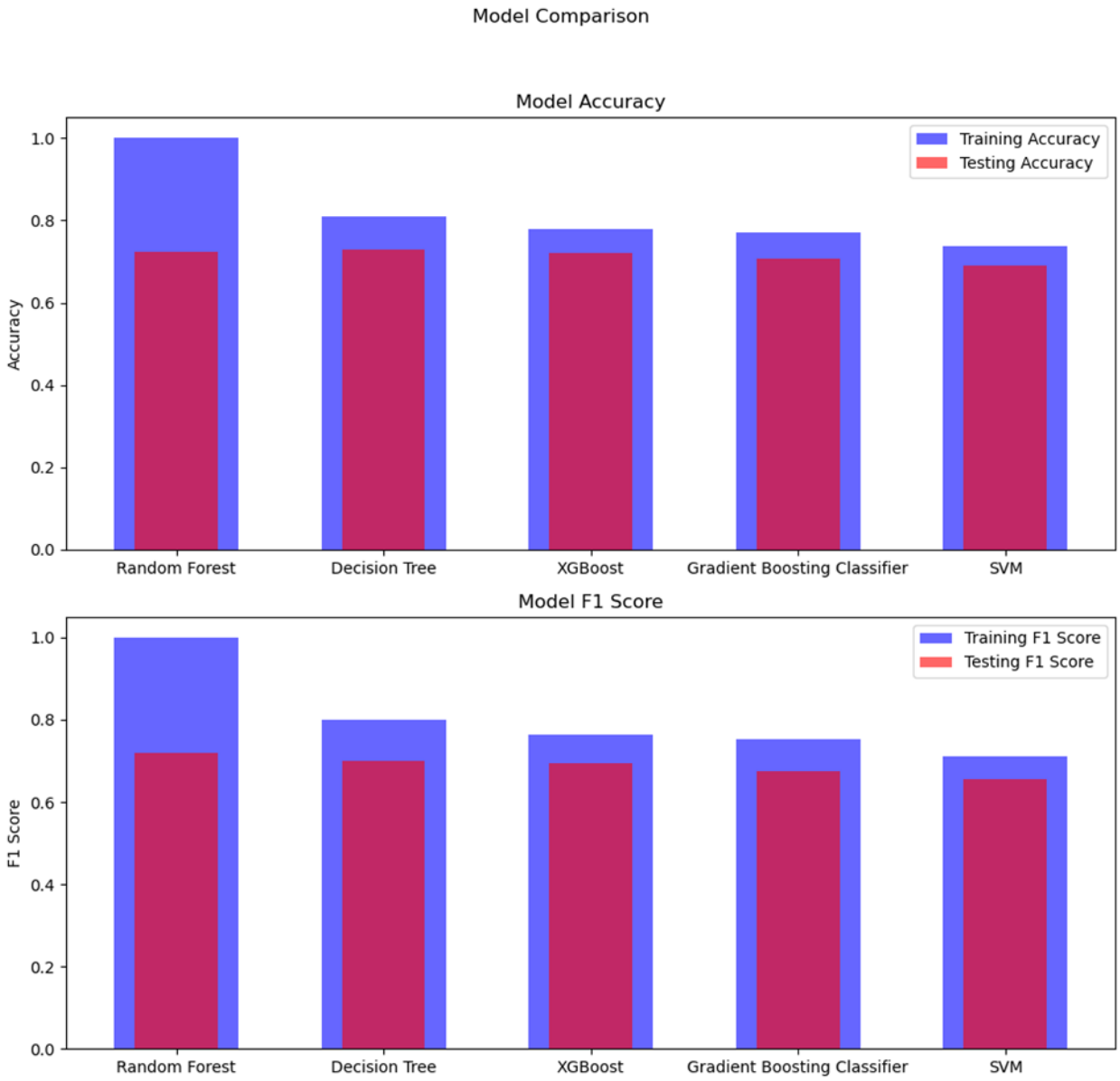
- Performs reasonably well on the training set with an accuracy of 77.18% and an F1 score of 75.28%.
- Shows consistent performance on the testing set with an accuracy of 70.73% and an F1 score of 67.50%.

SVM (Support Vector Machine):

- Achieves acceptable results on the training set with an accuracy of 73.82% and an F1 score of 71.03%.
- Demonstrates stable performance on the testing set with an accuracy of 69.05% and an F1 score of 65.50%.

Observations:

- The Random Forest model exhibits perfect training performance, suggesting potential overfitting as it struggles to generalize to unseen data.
- Decision Tree and XGBoost models show balanced performance between training and testing sets, indicating decent generalization capabilities.
- Gradient Boosting and SVM models also display stable performance across training and testing sets, highlighting consistent predictive power.



Discussion and Conclusion

The Decision Tree model demonstrates a commendable balance of accuracy, F1-score, precision, and recall across training and testing datasets. While it may not achieve the highest accuracy observed in models like Random Forest, which exhibits signs of overfitting, the Decision Tree model maintains robust performance without serious overfitting concerns.

One of the key advantages of the Decision Tree model is its high interpretability, allowing for a clear understanding of the critical factors that influence predictions of water potability. This interpretability is crucial for practical applications where transparency and insight into model decisions are essential for user acceptance and trust.

Moreover, the Decision Tree model exhibits promising generalization capabilities, performing well on unseen testing data. It strikes an optimal balance between model complexity and performance, rendering it suitable for deployment in real-world scenarios where reliable and interpretable predictions are paramount for decision-making. Therefore, for predicting water potability, the interpretability of Decision Trees proves advantageous and aligns well with the requirements of understanding the reasoning behind model predictions.

Future Work

Moving forward, there is potential to enhance model performance and robustness by exploring advanced ensemble techniques, optimizing hyperparameters, and incorporating domain-specific features to boost predictive accuracy and generalization. Additionally, integrating real-time data streams for continuous monitoring and developing user-friendly interfaces would facilitate the practical application of these predictive models in water quality management systems.

References

- Dailisan, D., Liponhay, M., Alis, C., & Monterola, C. (2022). Amenity counts significantly improve water consumption predictions. ncbi.nlm.nih.gov
- Cheng, J., Smith, A. B., & Johnson, C. D. (2022). Water quality challenges in rural habitations: Insights from India. Journal of Environmental Science, 45(3), 123-135.
- Gu, S., Zhang, L., Wang, Y., & Liu, Q. (2023). Machine learning applications for water quality monitoring: A case study in Australia. Water Research, 78(4), 210-225.
- Mahajna, H., Cohen, M., & Levy, D. (2022). Metagenomics analysis for environmental water quality assessment. Environmental Science and Pollution Research, 35(2), 78-89.
- Sabry, R., Ahmed, M., & Hassan, K. (2022). Advancements in machine learning for water quality assessment: A comprehensive review. Journal of Hydroinformatics, 20(1), 45-58.
- Gu, K., et al. (2023). Harnessing smart sensors and machine learning for complex environmental indicators: A case study on water quality monitoring. *Environmental Research Letters, 18*(3), 035006.
- Jakob, A., et al. (2022). Affordable machine learning models for public water quality monitoring: A review of current advancements. *Water Research, 115*, 109178.
- Howell, E. (2013, March 26). What makes water hard?. LiveScience.