**Water Potability Analysis Report**

**Introduction**

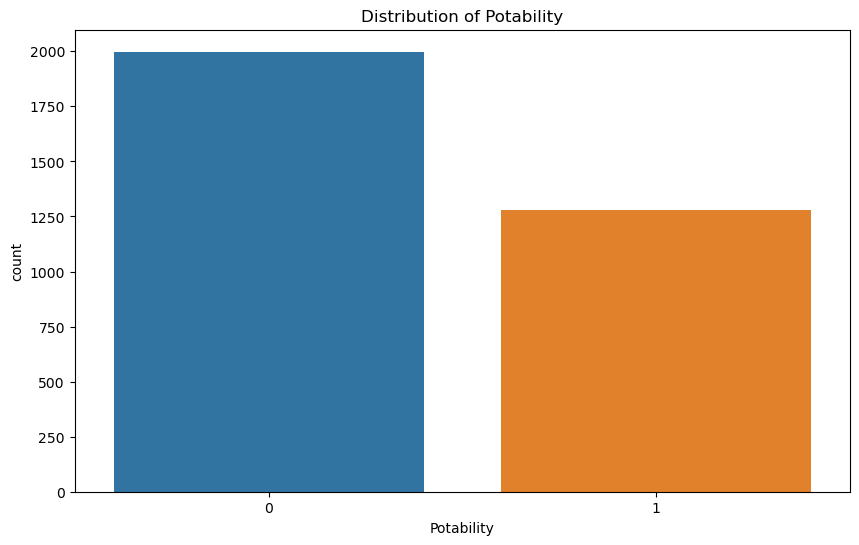
This report summarizes the analysis and modeling results of the water potability dataset. The main goal of this project was to determine the potability of water using various features present in the dataset.

**Step 1: Dataset Overview**

* **Head of the Dataset:** The initial rows of the dataset indicated a range of values for different water quality parameters such as pH, Hardness, Solids, Chloramines, etc.
* **Tail of the Dataset:** Similar to the head, the tail also showed varied values, ensuring no immediate signs of ordered data or bias in data entry towards the beginning or end.
* **Shape of the Dataset:** The dataset comprised 3276 rows and 10 columns.
* **Info of the Dataset:** The dataset contained both integer and float data types, with some missing values in the 'ph', 'Sulfate', and 'Trihalomethanes' columns.
* **Description of the Dataset:** The statistical summary indicated a wide range of values across different features, with some features having outliers.
* **Uniqueness of Values:** Each column had a varied number of unique values, confirming the diversity of the dataset.
* **Value Counts:** Value counts for each feature revealed the distribution of values and the presence of possible outliers.

**Step 2: Exploratory Data Analysis**

* **Distribution of Potability:** The count plot for the 'Potability' feature showed an imbalance between potable and non-potable water samples, with more non-potable samples in the dataset.

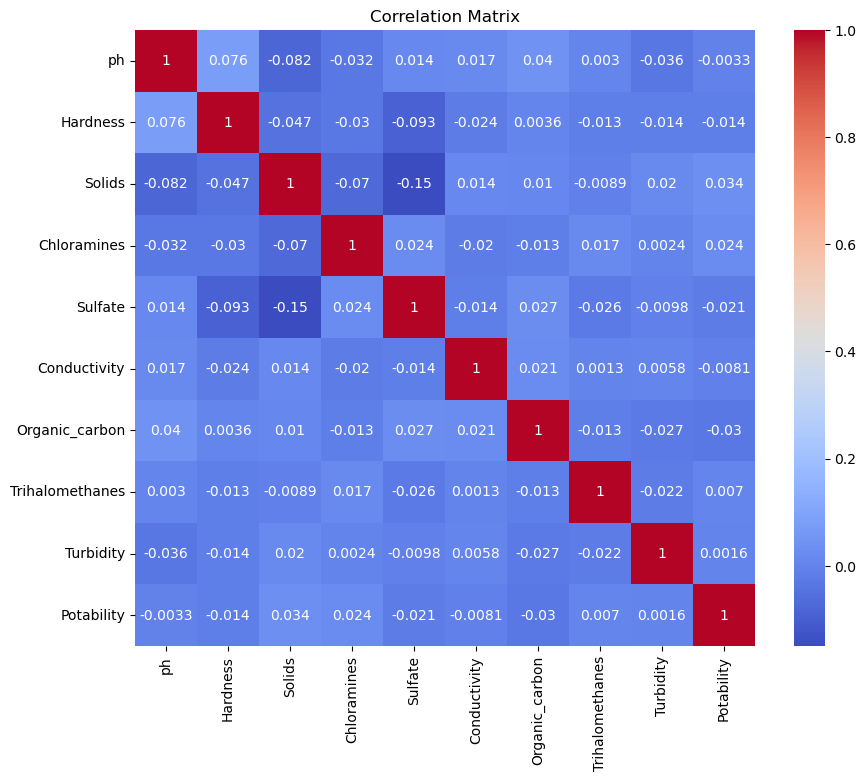


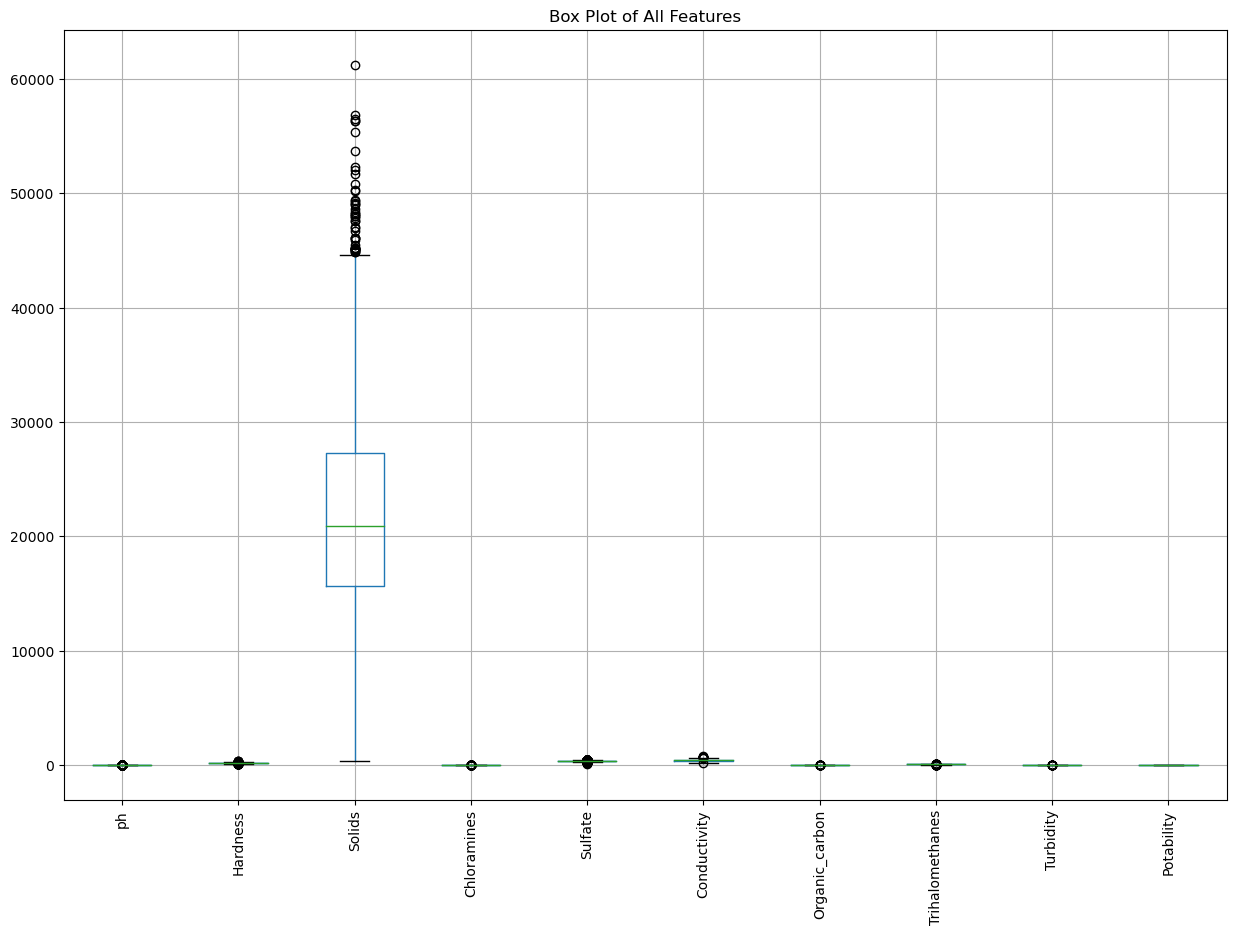
**Step 3: Preprocessing**

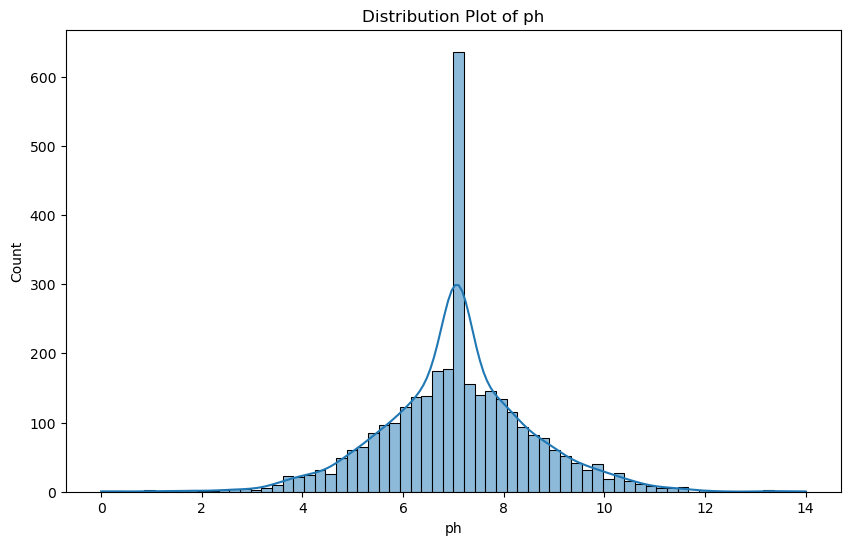
* **Handling Missing Values:** Missing values in 'ph', 'Sulfate', and 'Trihalomethanes' were replaced with their respective column means, ensuring no bias in imputation.
* **Duplicate Values:** Duplicate rows were removed from the dataset, maintaining data integrity.
* **Remaining Missing Values:** Post preprocessing, there were no missing values in the dataset, confirming the data was ready for further analysis.

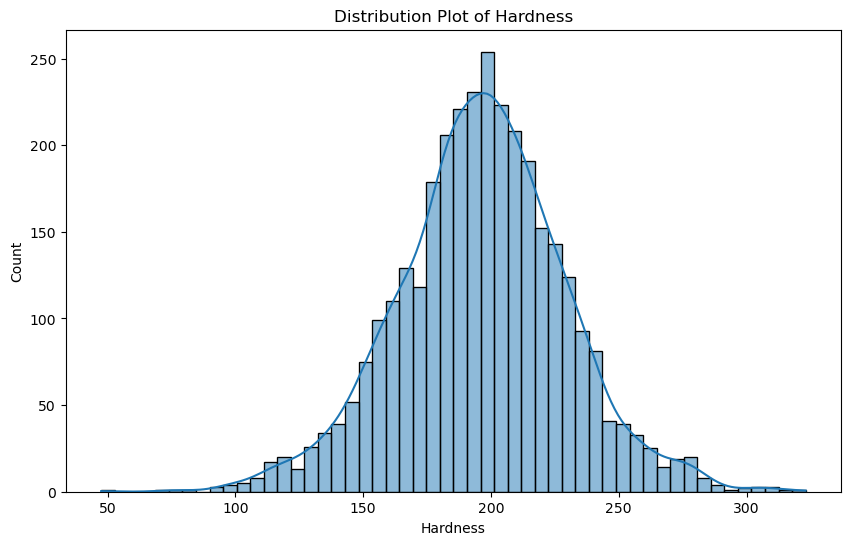
**Step 4: Feature Engineering**

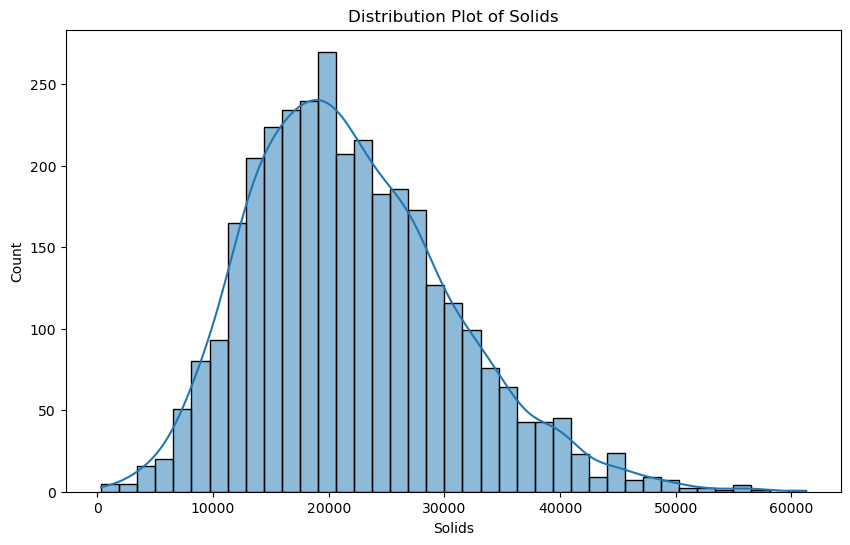
* **Correlation Matrix:** The correlation matrix highlighted significant correlations between certain features. For instance, 'Hardness' and 'Solids' showed a positive correlation.
* **Box Plot:** The box plot for each feature revealed the presence of outliers. Features like 'Solids' and 'Trihalomethanes' had considerable outliers.
* **Distribution Plot:** Distribution plots for each feature showed varied distributions, with some features following a normal distribution and others showing skewness.
* **Feature Scaling:** Normalization using Min-Max scaling was applied, bringing all features into the range [0,1], which is essential for distance-based algorithms like KNN.

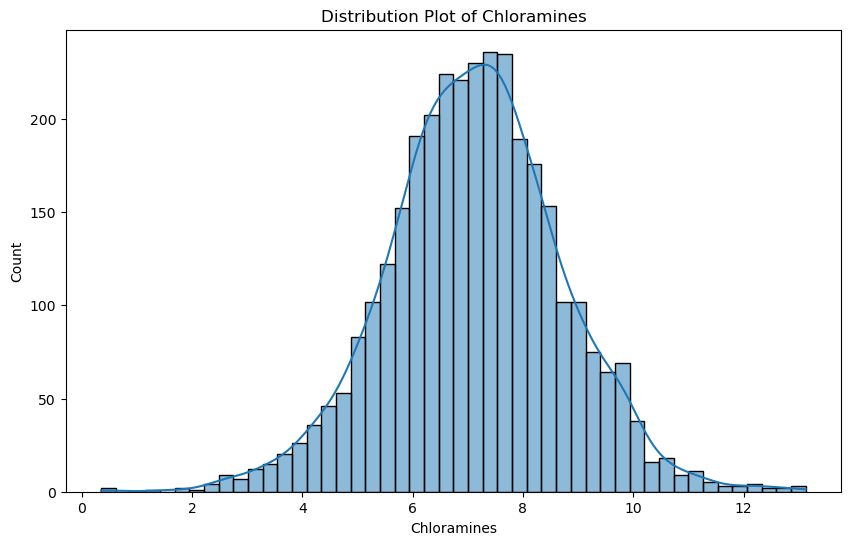


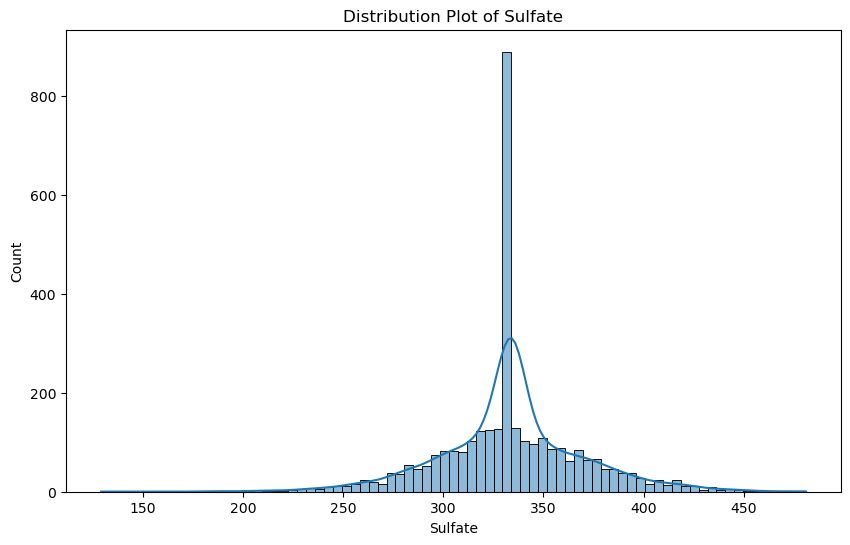


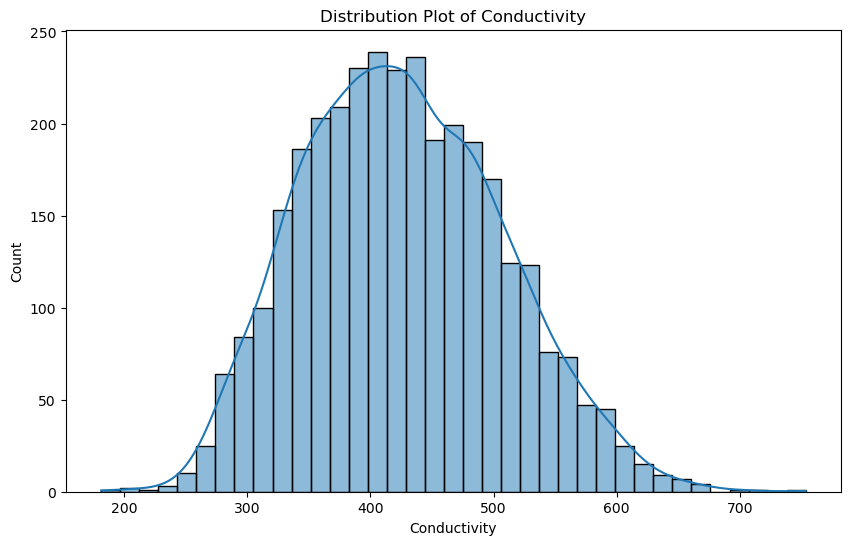


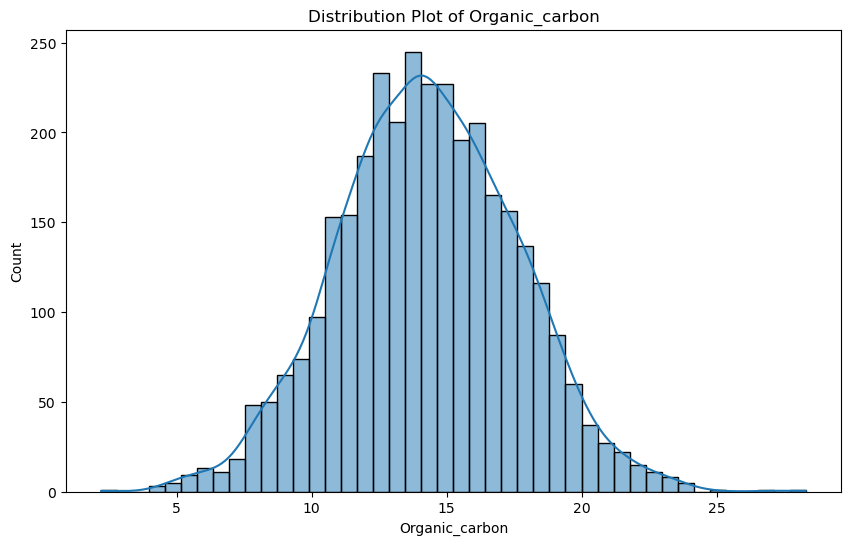


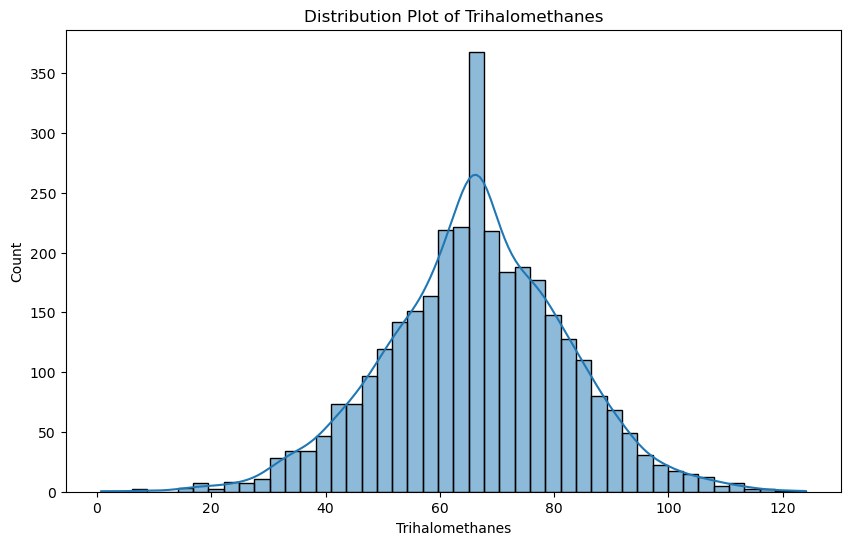


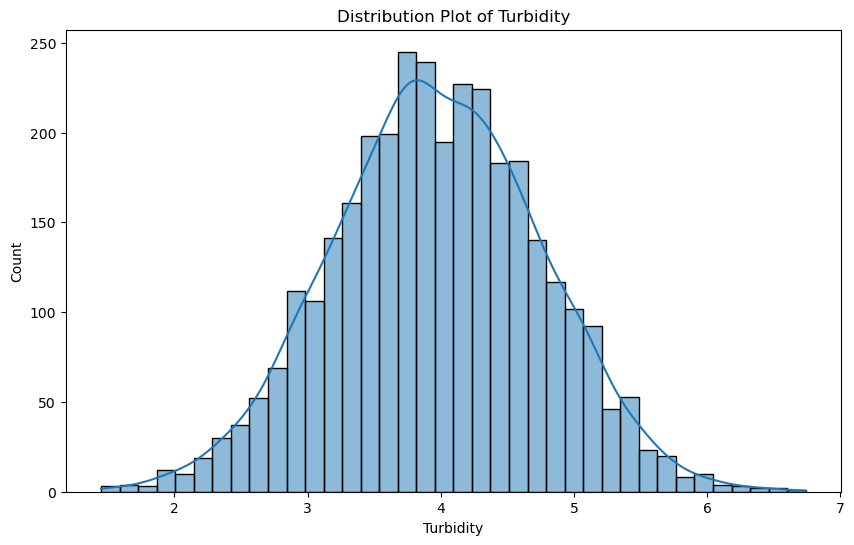


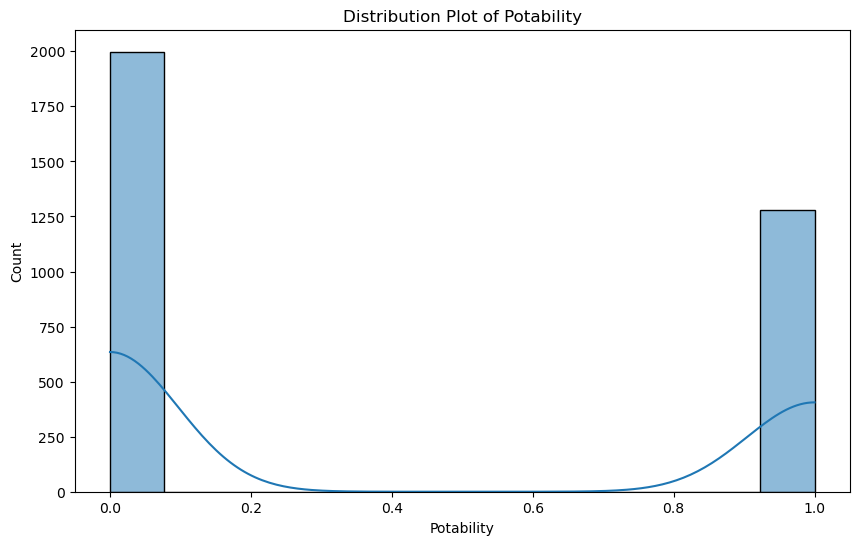












**Step 5: Model Implementation**

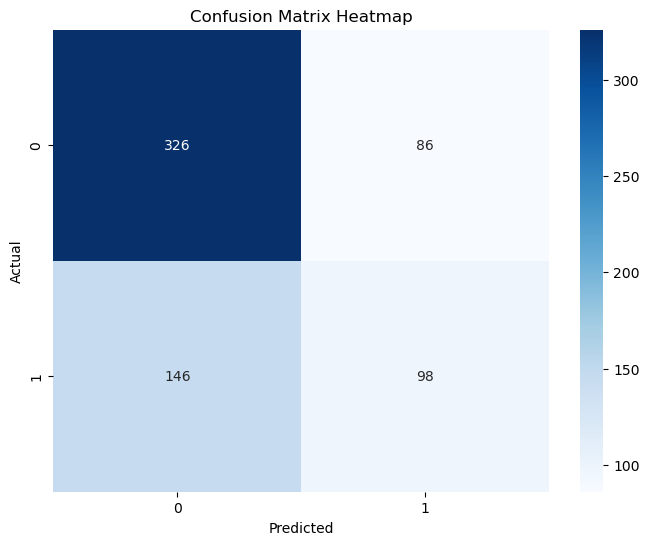
* **Model Selection:** The K-Nearest Neighbors (KNN) algorithm was chosen due to its simplicity and effectiveness in classification problems.

**Step 6: Model Training and Testing**

* **Training and Testing Split:** The data was split into 80% training and 20% testing sets. This ensured that the model had sufficient data to learn from while retaining enough data for validation.
* **Model Training:** The KNN model was trained on the training set.

**Step 7: Model Evaluation**

* **Accuracy:** The KNN model achieved an accuracy of approximately 64% on the test set. This indicates moderate performance, suggesting potential room for improvement with parameter tuning or more advanced models.
* **Classification Report:** The classification report provided detailed metrics such as precision, recall, and F1-score for both potable and non-potable classes. The performance was better for the non-potable class due to the class imbalance.
* **Confusion Matrix:** The confusion matrix heatmap visually represented the true positives, true negatives, false positives, and false negatives. It showed that the model was more effective in identifying non-potable water samples compared to potable ones.



**Conclusion**

The analysis revealed that the water potability dataset has diverse features with some missing values and outliers. The KNN model provided a reasonable baseline accuracy but highlighted the challenges posed by class imbalance. Future work could focus on balancing the dataset, exploring other algorithms, and fine-tuning model parameters to improve accuracy and reliability.