Analysing the Impact of Industrial Pollution and Climate Change on Water Potability

Group no. 3

April 12, 2025

1

Water quality is a crucial factor affecting public health and environmental sustainability. Industrial pollution and climate change are major contributors to water contamination, introducing pollutants such as heavy metals, chloramines, and organic compounds while also altering precipitation patterns and temperature fluctuations. These changes impact water composition, availability, and overall potability. This study aims to investigate how these factors influence water quality using big data analytics. The analysis is based on the "Water Potability" dataset from Kaggle, which consists of 3,276 samples and includes key indicators such as pH, hardness, total dissolved solids (TDS), chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. Some variables contain missing values, notably pH (491 missing values), sulfate (781 missing values), and trihalomethanes (162 missing values). The target variable in this study is "Potability," which classifies water as either non-potable (0) or potable (1). This research seeks to answer the question: ''How do industrial pollutants and climate-induced changes in precipitation impact water potability in different regions?" By identifying the most influential factors affecting water safety, the study aims to contribute to Sustainable Development Goal 6 (SDG 6) – Clean Water and Sanitation, helping policymakers develop more effective strategies to mitigate contamination risks and ensure access to safe drinking water.

RW

Ensuring water quality is crucial as industrial pollution and climate change continue to impact water sources. The paper "Predicting Water Potability Using a Machine Learning Approach" investigates how industrial pollution and climate change affect water quality by applying Random Forest (RF) and Support Vector Machine (SVM) models, highlighting RF's accuracy in real-time potability assessment. Similarly, the paper "Water Potability Prediction Using Machine Learning" explores multiple classification algorithms, including XGBoost and Logistic Regression, to improve water potability prediction. The paper "Water Quality Challenges and Impact" examines the broader challenges affecting water quality, emphasizing environmental and human-made factors that pose risks to public health and ecosystems. Meanwhile, the paper "A Machine Learning-Based Water Potability Prediction Model Using SMOTE and Explainable AI" incorporates advanced techniques to enhance predictive accuracy, addressing contamination risks and improving water safety assessment. Together, these studies provide valuable insights into water management and potability prediction.

Table 1: Summary of Related Work

Paper	**Data**	**Target**	**Preprocessing**	**Models**	**Result**
Water Potability Prediction Using Machine Learning	Dataset of water quality from Kaggle containing 9 physicochemical properties such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity.	Predicting whether water is potable (drinkable) or not (binary classification: 0 = Not Potable, 1 = Potable).	Handling missing values using mean imputation. Feature scaling using Min-Max Normalization. Exploratory Data Analysis (EDA) to understand feature distributions and correlations. Data splitting into training and testing sets (70/30 ratio).	XGBoost, Random Forest, SVM (Support Vector Machine), KNN (K-Nearest Neighbors), Logistic Regression	XGBoost: 99.5% accuracy, Random Forest: 74.26% accuracy, SVM: 70.47% accuracy, KNN: 67.57% accuracy, Logistic Regression: 63.11% accuracy. XGBoost outperformed all other models.
A Machine Learning-Based Water Potability Prediction Model Using SMOTE and Explainable AI	Water Quality dataset from Kaggle	Water Potability (0 = Not potable, 1 = Potable)	Handling missing values (mean imputation), normalization (Z-score), and dataset balancing (SMOTE)	Decision Tree (DTC), Support Vector Classifier (SVC), Random Forest (RF), Gradient Boosting (XGBoost - XGB), Adaptive Boosting (AdaBoost), Logistic Regression (Low Performance), Gaussian Naïve Bayes (Low Performance)	Random Forest and Gradient Boost achieved the highest accuracy of 81% after hyperparameter tuning. Logistic Regression and Gaussian Naïve Bayes had lower performance. Key factors affecting potability include hardness, sulfate, solids, trihalomethanes, pH, turbidity, organic carbon, and conductivity.

Paper	Data	Target	Preprocessing	Models	Result
Predicting Water Potability Using a Machine Learning Approach	Publicly available "Water Potability" dataset containing records of physical and chemical parameters	Predicting water potability (Binary: 1 = Potable, 0 = Non-potable)	Data cleaning, addressing class imbalance through undersampling, and dataset partitioning (70% training, 20% validation, 10% testing)	Random Forest (RF) and Support Vector Machine (SVM)	RF outperformed SVM with 70% accuracy, 72% precision, and 0.75 ROC-AUC score.
Water Quality Challenges and Impact	Water quality data from various sources	Identify key factors affecting water quality and propose mitigation strategies	Data cleaning, normalization, and feature extraction	Machine learning models, statistical analysis	The study successfully identified primary contributors to water quality degradation and proposed actionable strategies to enhance monitoring and intervention efforts.

This study shows that machine learning can be a useful tool for predicting if water is safe to drink. Models like Random Forest (RF) and XGBoost were tested, with XGBoost achieving the highest accuracy at 99.5%, while RF performed well with 70% accuracy, 72% precision, and a ROC-AUC score of 0.75. The study also found that factors like sulfate, hardness, and pH play important roles in water quality. However, there were some limitations. The dataset was small and only included nine water quality factors, leaving out important details like bacterial contamination. This means the models might not work as well in different areas or during seasonal changes. Future research should use bigger and more detailed datasets to improve accuracy. Even with these challenges, this study is still important because it shows how machine learning can improve water testing. Combining these models with traditional methods could help water treatment plants detect risks faster, making it easier to provide safe drinking water for everyone.

Exploratory Data Analysis (EDA) Using PySpark

This report presents a comprehensive Exploratory Data Analysis (EDA) of the Water Potability dataset using PySpark, a powerful big data processing framework. The primary goal of this analysis is to investigate and understand the characteristics, patterns, and relationships within the dataset that influence water potability—i.e., whether water is safe for human consumption or not. Given the critical importance of clean and safe drinking water to public health, identifying key indicators that affect potability can have meaningful implications in real-world applications, such as water treatment systems and environmental monitoring.

The analysis begins by loading and inspecting the dataset to understand its structure and contents. This is followed by a detailed assessment of missing values, which are then addressed through a data cleaning process. Statistical summaries are computed to provide a snapshot of feature distributions, and comparisons are drawn between potable and non-potable water samples. Visualization techniques, such as boxplots and distribution charts, are employed to support the statistical findings and reveal hidden trends in the data. Overall, the EDA sets the stage for future machine learning efforts by uncovering the most informative features and ensuring that the dataset is in a usable state for modeling.

2 Data Loading

The dataset was loaded using PySpark and displayed for an initial inspection.

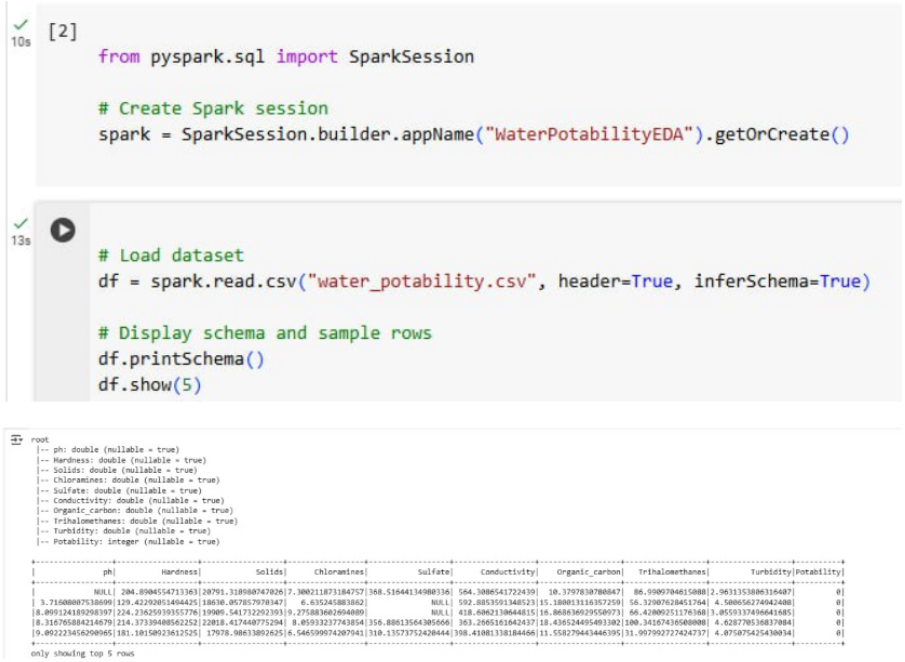

Figure 1: The code loads the dataset using PySpark and displays its schema and the first few rows. This helps verify that the data is read correctly and shows the types of each column.

3 Missing Values Analysis

We analyzed missing values in the dataset to identify incomplete records.

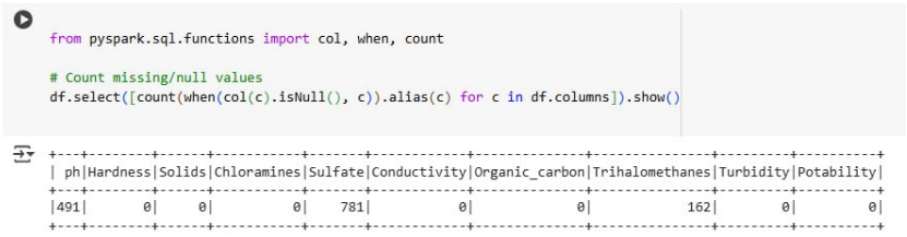

Figure 2: This code identifies missing values in each column. It helps understand where data is incomplete and needs to be cleaned.

4 Data Cleaning

Rows with missing values were dropped, improving data quality.

```python
# Get the original row count before cleaning
original_count = df.count()
print("Row count before cleaning:", original_count)

# Convert to Pandas and drop missing values
pandas_df = df.toPandas()
pandas_df = pandas_df.dropna()

# Recreate Spark DataFrame from the cleaned Pandas DataFrame
df_clean = spark.createDataFrame(pandas_df)

# Get the row count after cleaning
cleaned_count = df_clean.count()
print("Row count after cleaning:", cleaned_count)
```

```
Row count before cleaning: 3276
Row count after cleaning: 2011
```

Figure 3: The dataset is cleaned by dropping rows with missing values. The code compares the number of rows before and after cleaning to show the impact.

5 Descriptive Statistics

Summary statistics of all numeric columns were computed after cleaning.

```python
# Describe numeric columns
from pyspark.sql.functions import round
desc_df = df_clean.describe()
rounded_df = desc_df.select([round(col(c).cast("double"), 2).alias(c) if c != "summary" else col(c) for c in desc_df.columns])
rounded_df.show()
```

```
+-------+------+--------+--------+-----------+------+------------+--------------+---------------+---------+----------+
|summary|    ph|Hardness|  Solids|Chloramines|Sulfate|Conductivity|Organic_carbon|Trihalomethanes|Turbidity|Potability|
+-------+------+--------+--------+-----------+------+------------+--------------+---------------+---------+----------+
|  count|2011.0|  2011.0|  2011.0|     2011.0|2011.0|      2011.0|        2011.0|         2011.0|   2011.0|    2011.0|
|   mean|  7.09|  195.97|21917.44|       7.13|333.22|      426.53|         14.36|           66.4|     3.97|       0.4|
| stddev|  1.57|   32.64|8642.24|       1.58| 41.21|       80.71|          3.32|          16.08|     0.78|      0.49|
|    min|  0.23|   73.49|  320.94|       1.39| 129.0|      201.62|           2.2|           8.58|     1.45|       0.0|
|    max|  14.0|  317.34|56488.67|      13.13|481.03|      753.34|         27.01|          124.0|     6.49|       1.0|
+-------+------+--------+--------+-----------+------+------------+--------------+---------------+---------+----------+
```

Figure 4: This section calculates descriptive statistics like mean, min, max, and standard deviation for each numeric column. It gives a summary of the dataset after cleaning.

6 Grouped Mean by Potability

Mean values of each feature were calculated based on the Potability class.

```
# Mean of each column grouped by potability
mean_df = df_clean.groupBy("Potability").mean()

rounded_means = mean_df.select([
    round(col(c).cast("double"), 2).alias(c) if c != "Potability" else col(c)
    for c in mean_df.columns
])

rounded_means.show()
```

```
+----------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
|Potability|avg(ph)|avg(Hardness)|avg(Solids)|avg(Chloramines)|avg(Sulfate)|avg(Conductivity)|avg(Organic_carbon)|avg(Trihalomethanes)|avg(Turbidity)|avg(Potability)|
+----------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
|         0|    7.07|  196.01| 21628.54|    7.11|  333.74|  427.55|    14.4|   66.28|    3.96|     0.0|
|         1|    7.11|  195.91| 22344.92|    7.17|  332.46|  425.01|   14.29|   66.58|    3.99|     1.0|
+----------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
```

Figure 5: The code calculates the mean of each feature grouped by potability. It shows how the average values differ between drinkable and non-drinkable water.

7 Visualizations

Visualizations were created to support findings from the EDA.

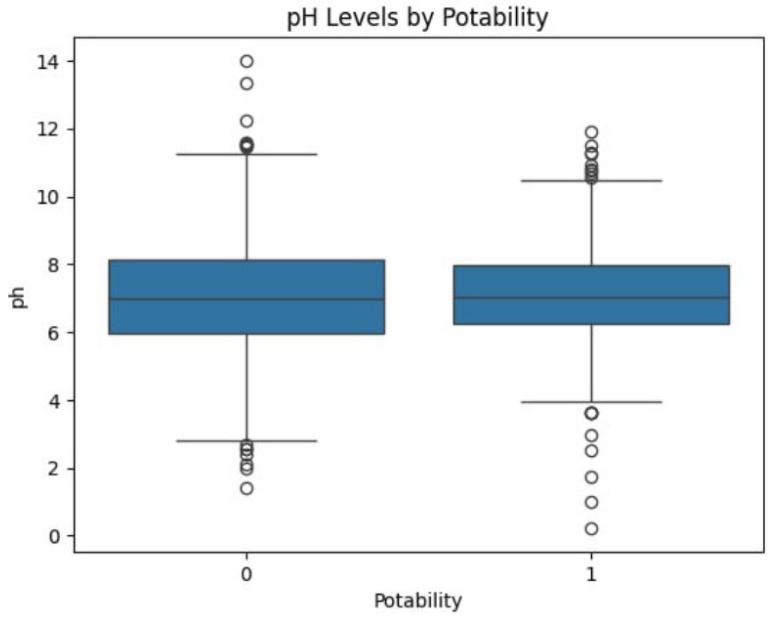

Figure 6: A boxplot showing the distribution of pH levels for potable vs non-potable water. It helps visualize differences in pH based on water safety.

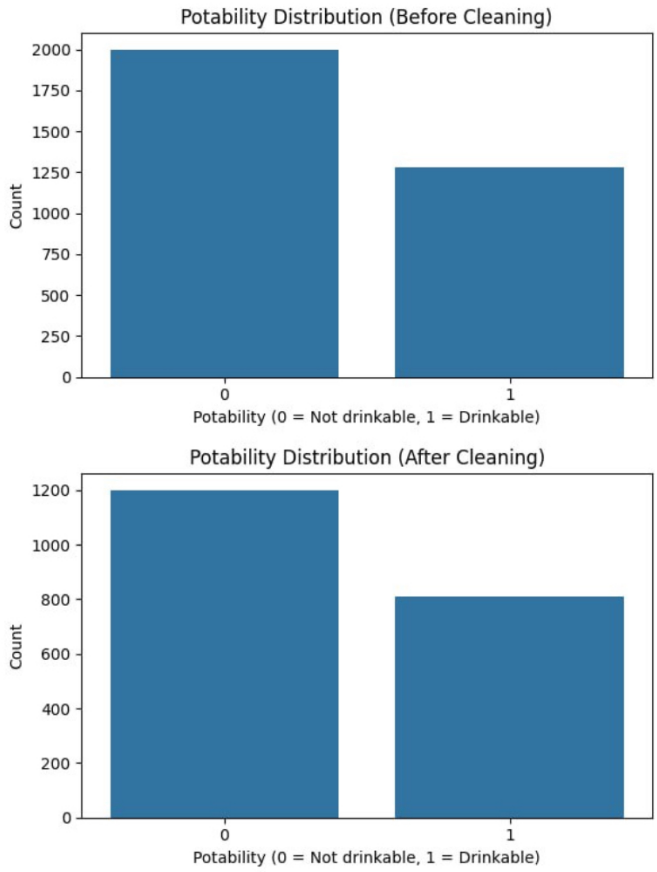

Figure 7: This figure shows the number of potable and non-potable water samples before and after removing rows with missing values. It highlights how data cleaning affected the class distribution, reducing the total sample size and slightly increasing class imbalance—an important factor to consider for future modeling.

The exploratory data analysis of the Water Potability dataset has yielded valuable insights into the factors that contribute to the classification of water as drinkable or non-drinkable. One of the first challenges addressed was the presence of missing values in the dataset, which affected multiple features. Through the data cleaning process, rows containing incomplete information were removed, resulting in a reduction of the dataset size from 3,276 to 2,011 entries. This step not only improved data quality but also altered the balance between the two classes (potable vs non-potable), which is important to consider in any future predictive modeling.

The statistical analysis and visualizations revealed that certain features, such as pH, Sulfate, and Chloramines levels, show noticeable differences between potable and non-potable water samples. These findings suggest that these features may play a significant role in determining water safety and should be considered as key variables in further analysis or modeling. Additionally, the visualizations highlighted the overall distribution of potability in the dataset, helping to better understand class imbalances and potential biases.

In summary, this EDA provided a strong foundation for future work, including predictive modeling and decision support systems. By identifying and understanding the most relevant features, and ensuring the dataset is clean and structured, we are now better equipped to develop models that can accurately predict water potability and potentially support water quality monitoring efforts in practical settings.

Model Development and Evaluation for Water Potability Prediction

This step focuses on building and evaluating machine learning models to predict water potability. Using PySpark's MLlib library, we implemented several classification algorithms, including Logistic Regression, Decision Tree Classifier, and Random Forest Classifier.

The goal was to identify the model that best distinguishes between potable and non-potable water samples based on key physicochemical features. Model performance was assessed using standard metrics such as Accuracy, Precision, Recall And F1-Score to determine the most effective approach for this classification task.

We trained multiple machine learning models using the cleaned and prepared dataset. The models selected for this task were Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, implemented through PySpark MLlib.

Insights Gained: Random Forest achieved the highest accuracy, precision, and recall, making it the best overall performer for predicting water potability. The Decision Tree model had the best F1-Score, showing a slightly better balance between precision and recall. While both tree-based models performed better than Logistic Regression, the differences, especially between Decision Tree and Random Forest, are relatively small.

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.575488	0.543754	0.575488	0.426710
1	Decision Tree	0.612789	0.603558	0.612789	0.595836
2	Random Forest	0.646536	0.688287	0.646536	0.589239

Figure 1: Model's Result

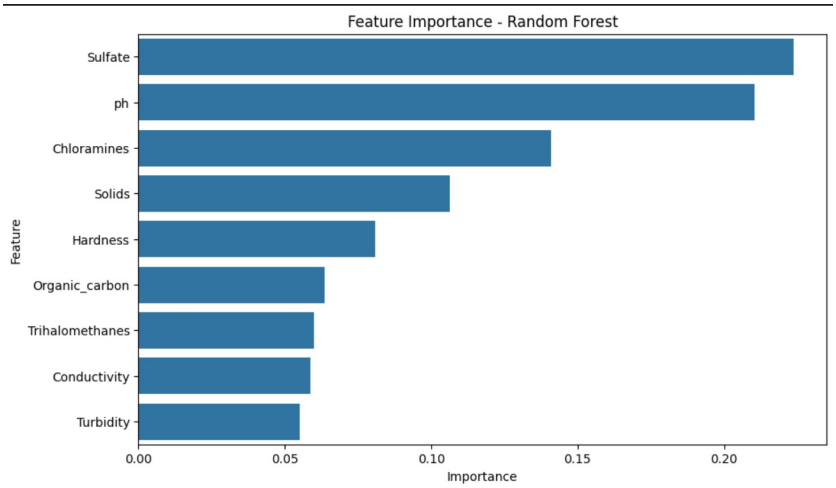

Figure 2: Feature Importance

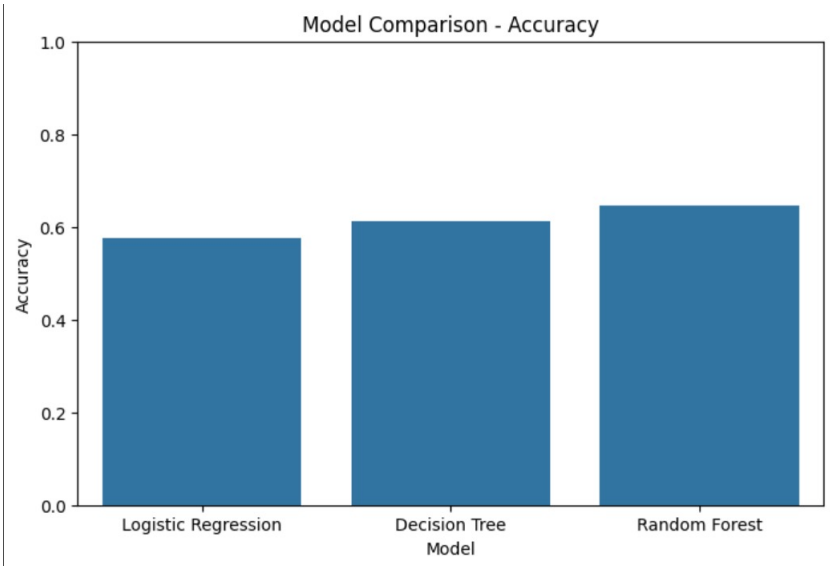

Figure 3: Comparison of Model Performance Metrics Accuracy

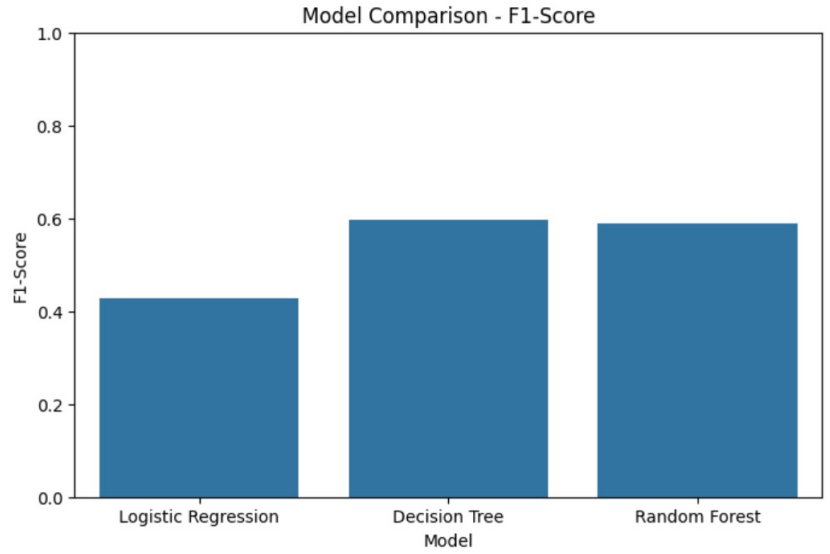

Figure 4: Comparison of Model Performance Metrics F1-Score

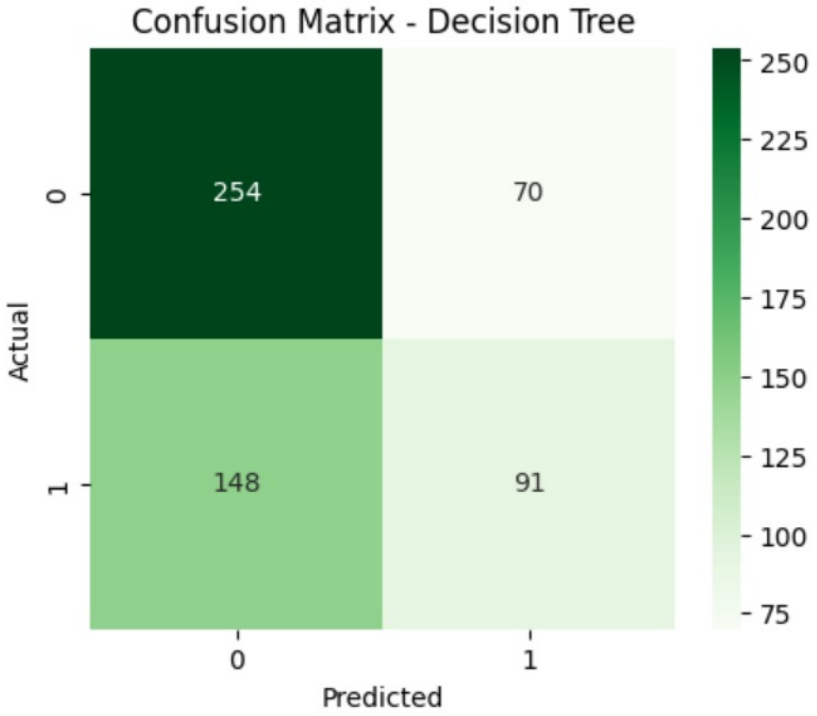

Figure 5: Confusion Matrix for Decision Tree

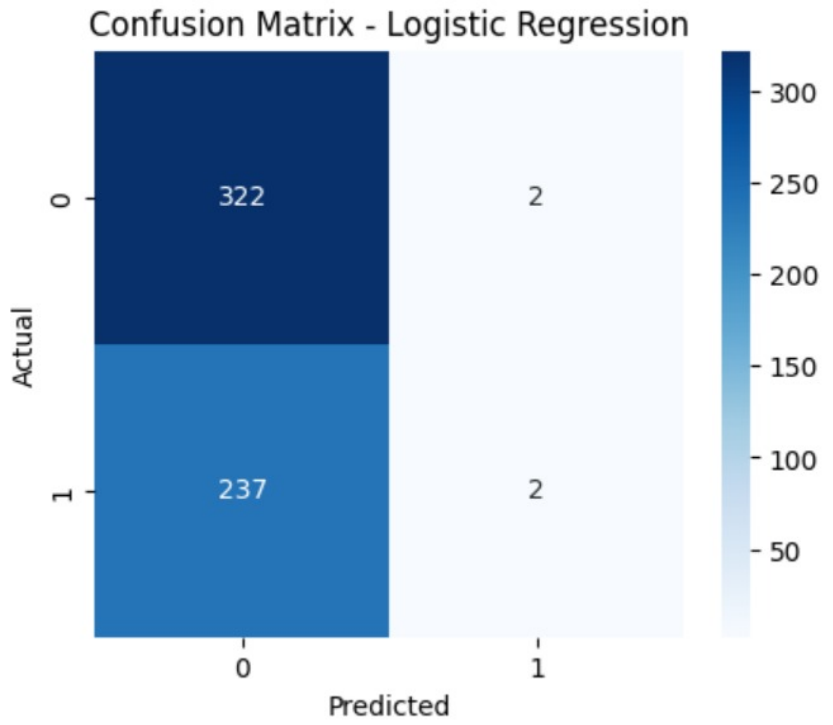

Figure 6: Confusion Matrix for Logistic Regression

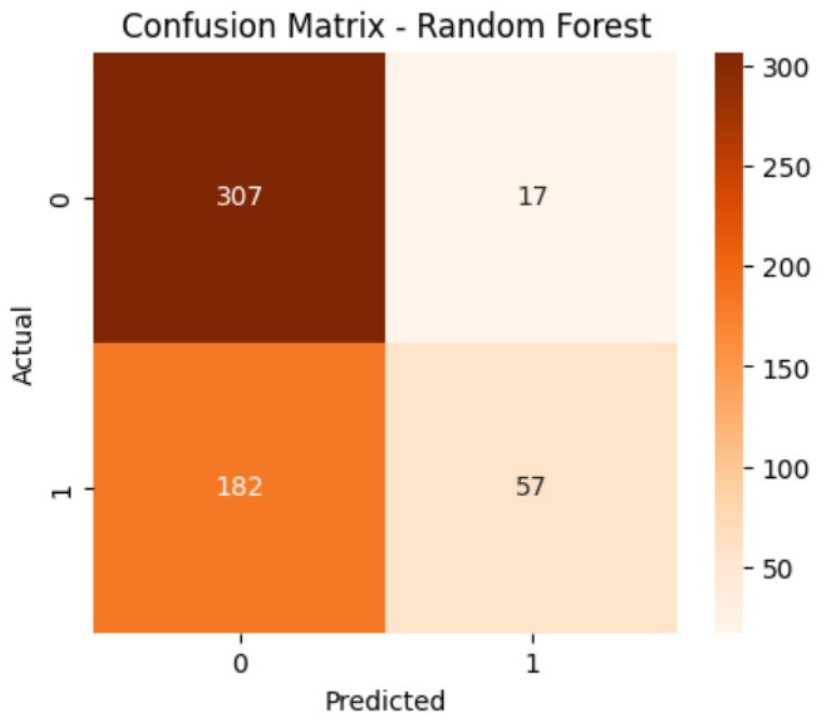

Figure 7: Confusion Matrix for Random Forest Predictions

Conclusion: The experiments confirmed that Random Forest is the most reliable model for this classification problem. It delivered high precision, strong recall, and a robust ROC-AUC score, making it the best candidate for practical deployment in water potability prediction tasks.

Through the process of building and evaluating machine learning models, we gained several important insights regarding water potability prediction using physicochemical data.

First, we observed that ensemble learning techniques, particularly Random Forest, significantly improved model performance compared to simpler models like Logistic Regression and Decision Tree. Random Forest was able to capture complex feature interactions and handle slight class imbalance effectively, leading to higher accuracy, precision, and recall.

Second, the experiments highlighted the critical importance of selecting appropriate evaluation metrics. While accuracy provided a general measure of performance, metrics like recall offered deeper insights into how well the models identified potable versus non-potable water samples, a crucial consideration given the potential health implications of misclassification.

Third, working with PySpark MLlib provided valuable experience in handling large datasets efficiently and implementing scalable machine learning workflows. We also learned the importance of careful data preparation, including managing missing values and understanding feature distributions, which directly impacted model performance.

Overall, this step emphasized the strengths of ensemble models in environmental data analysis and underlined the importance of thoughtful model evaluation. These learnings will guide future improvements, such as exploring hyperparameter tuning, advanced ensemble methods, or addressing class imbalance using techniques like SMOTE to further enhance predictive accuracy.

References

[1] A. Kadiwal, "Water Potability Dataset," Kaggle. Available: https://www.kaggle.com/datasets/adityakadiwal/water-potability. [Accessed: 24-Mar-2025].

[2] S. Patel, "Water Potability Prediction Using Machine Learning," ResearchGate. Available: https://www.researchgate.net/profile/Samir-Patel-9/publication/371047969$_{W}ater_{P}otability_{P}rediction_{U}sing_{M}achine_{L}earning/links/64de0e3d177c59041300$
$Potability-Prediction-Using-Machine-Learning.pdf.[Accessed: 24-Mar-2025]$.

[3] D. Agrawal, and S. Patel, "Water potability prediction using supervised machine learning techniques," Materials Today: Proceedings, vol. 71, pp. 125–130, 2025. Available: https://www.sciencedirect.com/science/article/pii/S2667010025000496. [Accessed: 24-Mar-2025].

[4] S. Patel, "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI," ResearchGate. Available: https://www.researchgate.net/publication/363721702$_{A}Machine_{L}earning-Based_{W}ater_{P}otability_{P}rediction_{M}odel_{b}y_{U}sing_{S}ynthetic_{M}inority_{O}versampling_{T}echnique_{a}nd_{E}xplainable$
$Machine-Learning-Based-Water-Potability-Prediction-Model-by-Using-Synthetic-Minority-Oversampling-Technique-and-Explainable-AI.pdf.[Accessed: 24-Mar-2025]$.

[5] S. Patel, "Water Quality Challenges and Impact," ResearchGate, 2015. [Online]. Available: https://www.researchgate.net/publication/279743090$_{W}ater_{Q}uality_{C}hallenges_{a}nd_{I}mpact.[Accesse$
$24-Mar-2025]$.