CDS6214

Data Science Fundamentals

Project (40%)

Tutorial Section: TT7L

Group Number: G19

ID	Name	Email Address
1211103024	Yap Jack	1211103024@student.mmu.edu.my
1211101186	Tam Li Xuan	1211101186@student.mmu.edu.my
1211101198	Fam Yi Qi	1211101198@student.mmu.edu.my

1221303972	Lee Jia Ying	1221303972@student.mmu.edu.my

YouTube Link	https://youtu.be/lQ-MU1TzX4I

TASK DISTRIBUTION

Lecture Section: TC3L

Tutorial Section: TT7L

Group Number: G19

Domain: Climate and Environment

Task Distribution:

Student Name	Task Done (Be very specific)
Yap Jack	Introduction, Model Construction
Tam Li Xuan	Data Preprocessing, Model Optimization, Result
Fam Yi Qi	Introduction, Hypothesis Testing, Conclusion
Lee Jia Ying	EDA, Dataset Overview

Introduction

Background

Safe drinking water is important to public health. Once human life accidentally drinks non-potable water, they might face severe health problems. This is a critical issue that requires the government and organizations to pay attention to and action about the water quality. The potability of water, or its suitability for human consumption, is determined by a multitude of factors such as hardness, solidity, turbidity, Chloramines, Sulfate, Conductivity, Trihalomethanes, and Organic Carbon. Every indicator provided above plays a significant role in assessing the quality of water. Only when these parameters fall within acceptable ranges, then the water is deemed be potable. Therefore, our project aims to investigate the relationship between these water quality indicators and the potability status and then provide valuable insight that could enhance water safety measures and health outcomes.

Motivation

Nowadays water pollution is escalating at an alarming rate, making potable water become non-potable. There is a common misconception among people today that boiling water can render it safe for consumption. However, it is not accurate. Truly boiling can eliminate certain types of bacteria and viruses, but it does not remove and kill those substances and pollutants. Besides, drinking non-potable water will cause severe health issues such as infections, waterborne disease, gastrointestinal problems, and even death. Therefore, it is important to ensure that the water can be consumed to avoid tragedy.

Problem Proposed

Our primary objective of this study is to conduct a comprehensive analysis of the patterns about the relationship between water quality indicators and potability status. This investigation is geared towards creating a model for identifying whether the water is potable based on various quality indicators and how we can improve water potability.

Impact on the Proposed Problem

The proposed problem brings significant implications for various sectors, including environmental conservation, avoiding illness, and so on. As people understand the relationship between water quality indicators and water potability, they can better monitor the water resources and know which type of water is suitable to consume. This knowledge is essential for individual health and the national public health standards; it also helps the economic activities of the communities; it also ensures sustainable environmental practices across societies. This proves the critical role of water quality for communities, societies, and nations.

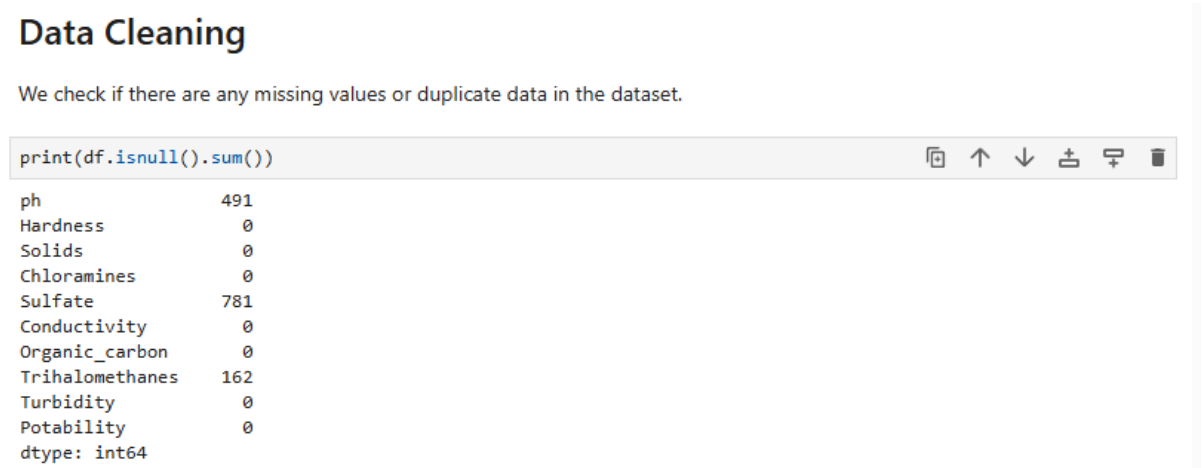
Questions

1. What is the distribution of each water quality feature between potable and non-potable?
2. Does the water quality feature have correlations with water potability?
3. What can be recommended to improve water potability based on the data analysis?
4. Can we predict water potability based on water quality indicators?

Data Preprocessing

1. Checking for Missing Values

Firstly, we examined the dataset for any missing values. Missing values can hinder data analysis and modelling, so it's essential to address them appropriately.



The output shows that there are missing values at column ph, Sulfate, and Trihalomethanes.

2. Checking for Duplicate Rows

Next, we checked the dataset for duplicate rows. Duplicate data can skew analysis results and should be removed to maintain data quality.

```
duplicates = df.duplicated()
duplicate_count = duplicates.sum()
print(f"Number of duplicate rows: {duplicate_count}")
```

```
Number of duplicate rows: 0
```

The output indicated the number of duplicate rows in the dataset. In this case, it was found that there were no duplicate rows.

3. Handling Missing Values

Since the dataset had missing values, we decided to fill the missing value in each column. We replace the missing value in the pH column with the mean 'pH' value of their respective group based on the Potability. Similarly, the Sulfate column is replaced with the mean of Sulfate value within each Potability group. Then we do the same for the Trihalomethanes column. After filling all these missing values, to preserve the integrity of the original dataset, the updated data was saved to a new CSV file. This method was chosen because it preserves the overall distribution and statistical properties of the dataset. Sometimes if the missing values are not handled properly, it will cause bias in the analysis.

```
# Replace null values based on the group mean of potable and non-potable
df['ph'] = df['ph'].fillna(df.groupby(['Potability'])['ph'].transform('mean'))
df['Sulfate'] = df['Sulfate'].fillna(df.groupby(['Potability'])['Sulfate'].transform('mean'))
df['Trihalomethanes'] = df['Trihalomethanes'].fillna(df.groupby(['Potability'])['Trihalomethanes'].transform('mean'))

# save cleaned_df to csv
df.to_csv('cleaned_water_quality.csv', index=False)
```

The dataset has been cleaned by filling missing values sand saved to cleaned_water_quality.csv. This ensures the data is ready for further analysis with no missing values and no duplicate entries.

Dataset Overview

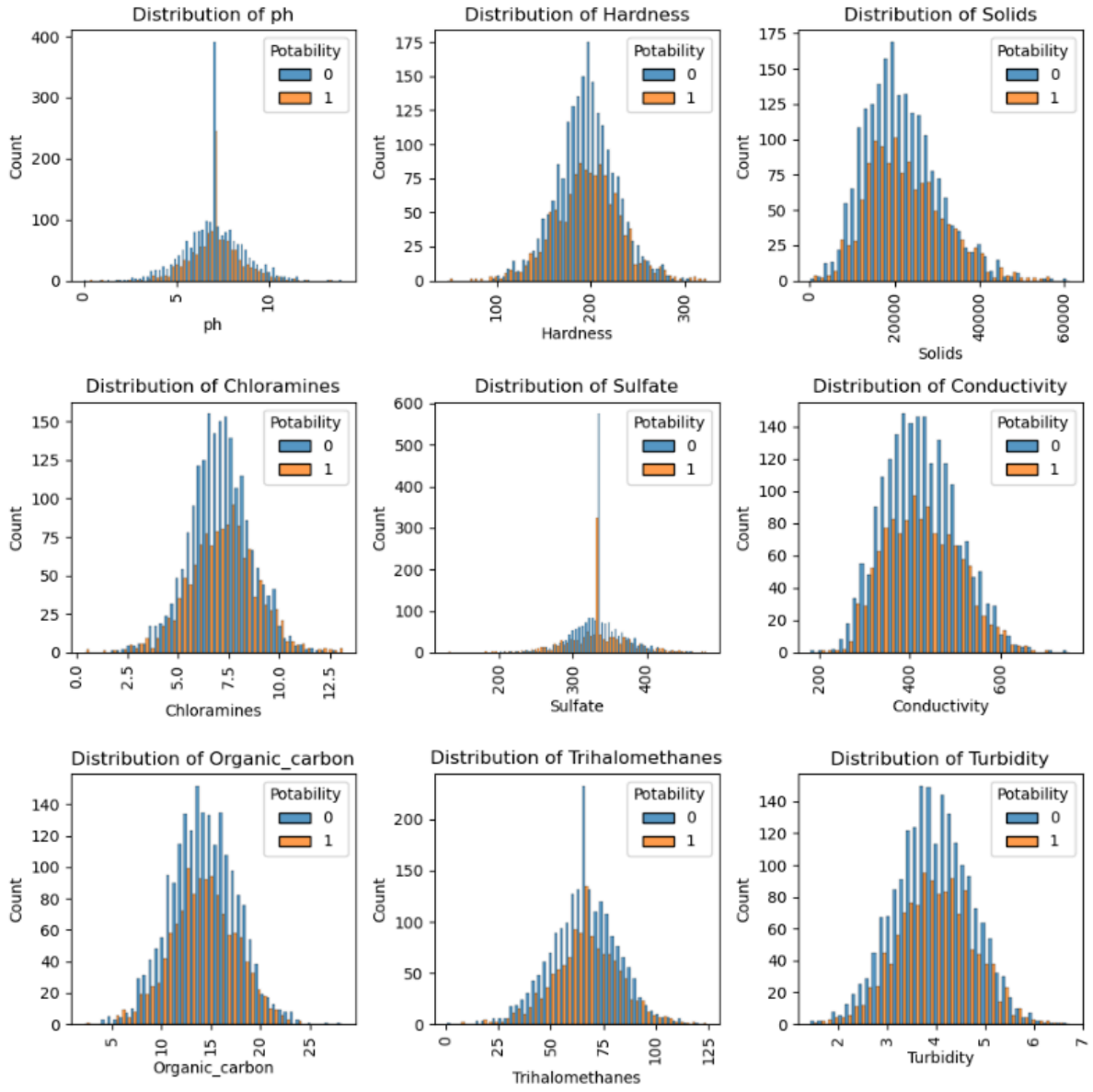
Variable	Data Type	Anticipated Role	Comments
ph	Float	Feature	pH level of water, indicating its acidity or basicity.
Hardness	Float	Feature	Measure of water hardness in mg/L.
Solids	Float	Feature	Total dissolved solids (TDS) in ppm.
Chloramines	Float	Feature	Concentration of

			chloramines in ppm.
Sulfate	Float	Feature	Concentration of sulphate in mg/L.
Conductivity	Float	Feature	Electrical conductivity of water in µS/cm.
Organic_carbon	Float	Feature	Total organic carbon (TOC) content in mg/L.
Trihalomethanes	Float	Feature	Concentration of trihalomethanes in µg/L.
Turbidity	Float	Feature	Measure of water clarity or haziness in NTU.
Potability	Integer	Target	Binary indicator of whether water is safe to drink (1) or not (0).

Exploratory Data Analysis (EDA)

1. Distribution of Water Quality Features (Histograms)

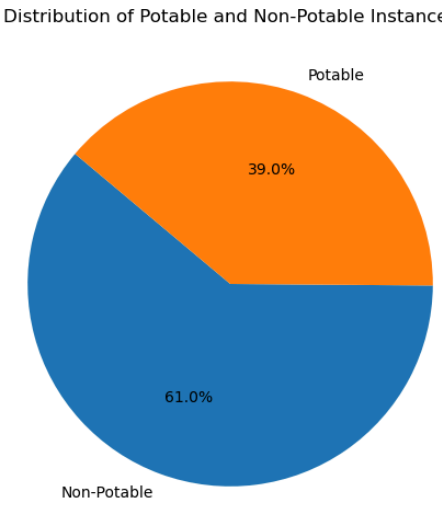
Histograms are used to visualize the distribution of each numerical variable in the dataset.



This visualization provides insights into the distribution and characteristics of each numerical variable. For most of the water quality features of non-potable, the distribution is approximately normally distributed except for solids which are slightly right-skewed. For the water quality features of potable, the distribution is very similar to the distribution of the non-potable features. Since the distribution of potable and non-potable are very similar, it suggest that the features may not have a strong relation with water potability.

2. Visualizing Categorical Variable (Pie Chart)

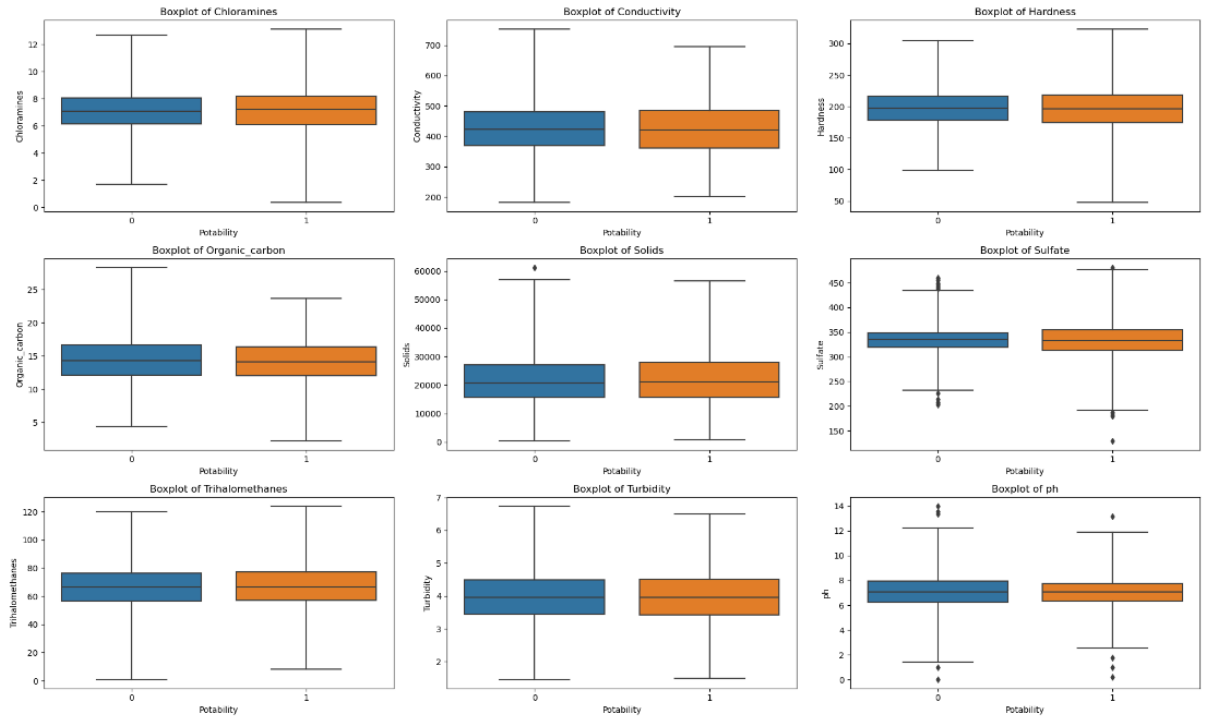
A pie chart is used to visualize the distribution of the categorical variable "Potability" (potable vs. non-potable instances).



This visualization helps understand the balance between potable and non-potable instances in the dataset. The pie chart shows that 39% of water potability is potable and 61% is non-potable.

3. Boxplots for Outlier Detection

Boxplots are employed to identify outliers in numerical variables.

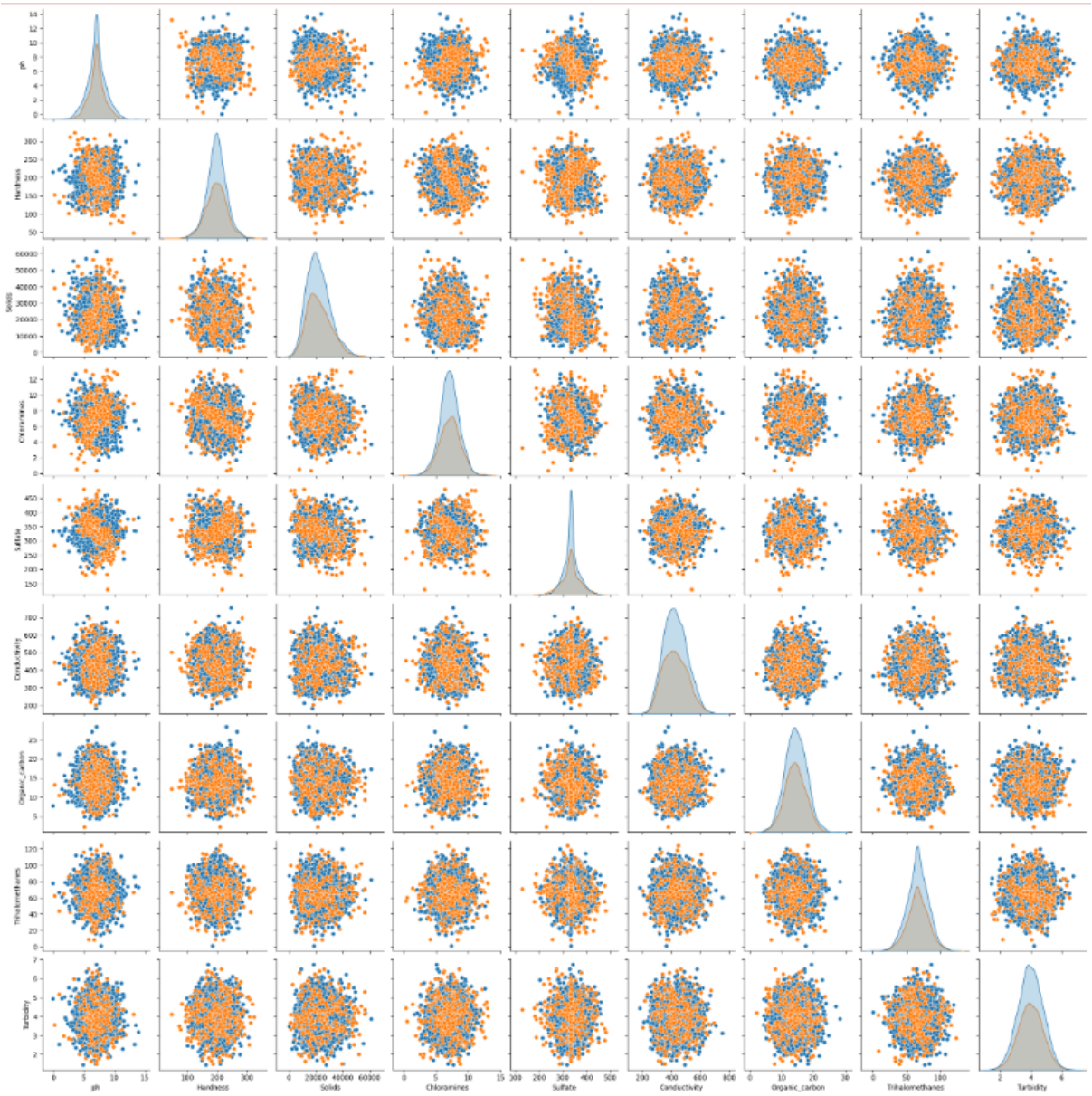


The boxplots provide a visual representation of the spread and skewness of data and help identify potential outliers; in addition, we extended the whiskers of all boxplots from 1.5 to 3 times the interquartile range (IQR) because it provides a clearer view of the distribution.

Looks like there are outliers at the non-potable boxplot of solids. Moreover, there are outliers in the boxplot of sulfate and ph. We decided to count outliers as actual data. Furthermore, the boxplots also show that the distribution of potable and non-potable water features are very similar; therefore, it supports that there is little to no relationship with water potability.

4. Pairplot

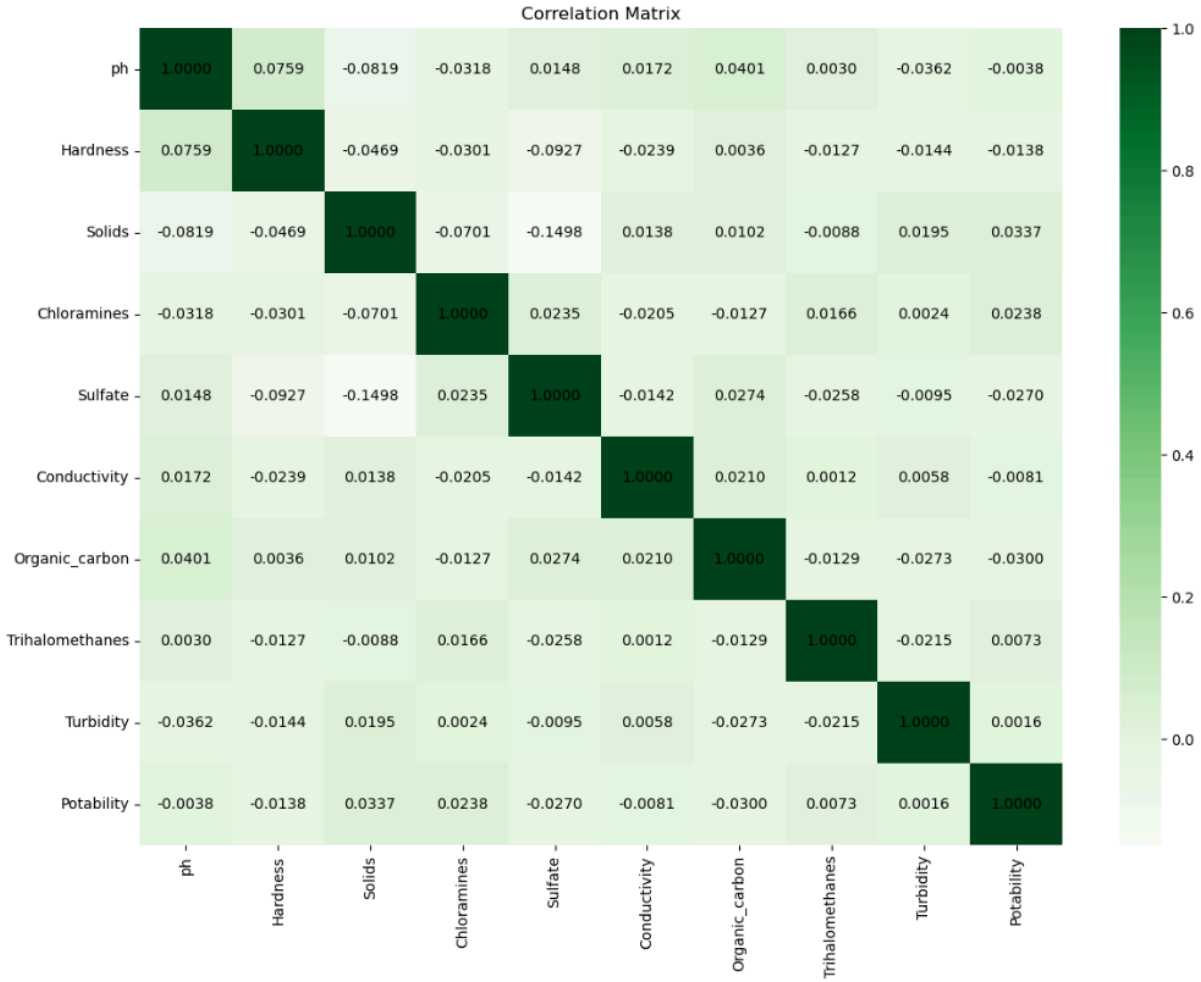
A pairplot overlays smooth density curves on the diagonal plots instead of histograms, providing a more nuanced view of variable distributions.



The scatter plots in the pairplot show a cluster in the middle without clear separation between potable and non-potable water samples, it suggests that the features might not have strong discriminatory power for predicting water potability. Therefore, no linear relation between features.

5. Correlation Analysis

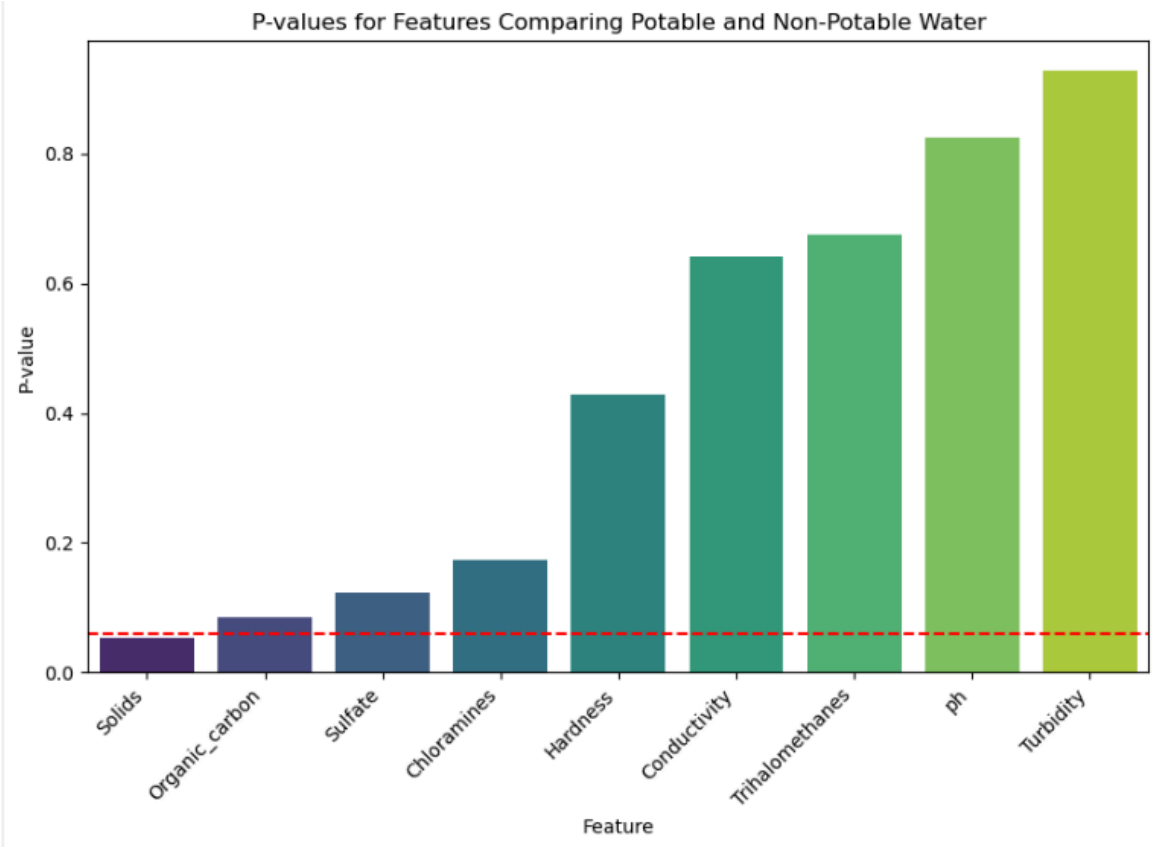
To further explore the hypothesis, a correlation matrix is generated to explore relationships between numerical variables.



The correlation matrix helps identify highly correlated variables, providing insights into potential multicollinearity issues. Overall, The correlation matrix analysis revealed that all water quality features have near zero correlation to other features and the water potability. This finding indicates a lack of linear relationship between parameters such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, and water potability. Therefore, the water quality features have no relationship with the features and water potability. This shows that one feature alone cannot determine the water potability, suggesting that all water quality features are unique; therefore, all water quality features are essential and should be included in the model.

6. Hypothesis Testing

The hypothesis testing using t-tests evaluates the significance of differences in each water quality feature between potable and non-potable water samples.



The hypothesis testing results provide insights into which water quality features play a significant role in determining water potability. Features with lower p-values indicate stronger associations with potability status, whereas non-significant features suggest minimal impact on distinguishing between potable and non-potable water samples. We decided to change the alpha level from 0.05 to 0.06 because all features do not show a significant relationship at 0.05 level. Therefore, this adjustment is a tolerable approach since it gives us a slightly higher acceptance threshold; it also does not compromise the integrity of our findings.

Build a Model

1. Model Construction and Comparison

For model construction and comparison, we have selected four regression models: Random Forest Classifier, Decision Tree Classifier, Logistic regression, and K-nearest Neigbors. We will train each model, compare their results, and choose the best-performing model to proceed with further analysis.

	Training Set Accuracy	Testing Set Accuracy
Random Forest Classifier	1.0	0.8247

Decision Tree Classifier	1.0	0.7470
Logistic Regression	0.6057	0.6280
K-nearest Neigbors Classifier	0.6672	0.5381

From the table, it is evident that the Random Forest Classifier outperforms other models with a training set accuracy of 1.0 and a testing set accuracy of 0.8247. Therefore, the Random Forest Classifier is considered the best model and will be selected for further model training and analysis.

2. Model Selection and Optimization

After identifying the Random Forest Classifier as the best-performing model, we focused on optimizing its performance through hyperparameter tuning. Despite the Random Forest Classifier showing the best accuracy compared to other models, the accuracy indicates that this model is overfitting, as evidenced by the higher accuracy in the training set compared to the test set. The goal was to improve the model's accuracy and generalizability by systematically adjusting key hyperparameters using Grid Search Cross-Validation (GridSearchCV).

Definition:

Overfitting: An overfitting model preforms very well on the training data but poorly on the test data. This happens when the model learns not only the underlying patterns but also the noise and specific details unique to the training dataset, which do not generalize to new, unseen data.
Good fit: A good fit model performs well on both the training and test data. This indicates that the model has successfully learned the underlying patterns in the data without overfitting to noise or specific details of the training set, allowing it to make accurate predictions on new, unseen data.

Hyperparameter Tuning: We explored the following hyperparameters:

```
# Define the parameter grid
param_grid = {
    'n_estimators': [50, 100, 200], # Number of trees in the forest
    'max_depth': [10, 15, 20], # Maximum depth of the trees
    'min_samples_split': [2, 5, 10], # Minimum number of samples required to split an internal node
    'min_samples_leaf': [1, 3, 5] # Minimum number of samples required to be at a leaf node
}
```

Optimization Process: Using GridSearchCV with 5-fold cross-validation, we identified the following optimal hyperparameters:

- `n_estimators`: 200
- `max_depth`: 15
- `min_samples_split`: 2
- `min_samples_leaf`: 5

These hyperparameters were selected based on their performance in cross-validation, balancing model complexity and generalizability.

Results

	Original Model	Optimized Model
Training Set Accuracy	1.0	0.9489
Testing Set Accuracy	0.8247	0.8064

The optimized Random Forest model achieved an accuracy of 94.89% on the training set and 80.64% on the testing set. However, the testing set accuracy has decreased. Therefore, we have decided to use the original model since the optimized model did not show promising results. The detailed evaluation metrics of the original model, including precision, recall, F1-score, and ROC-AUC show the model's robust performance in predicting water potability.

Classification Report:

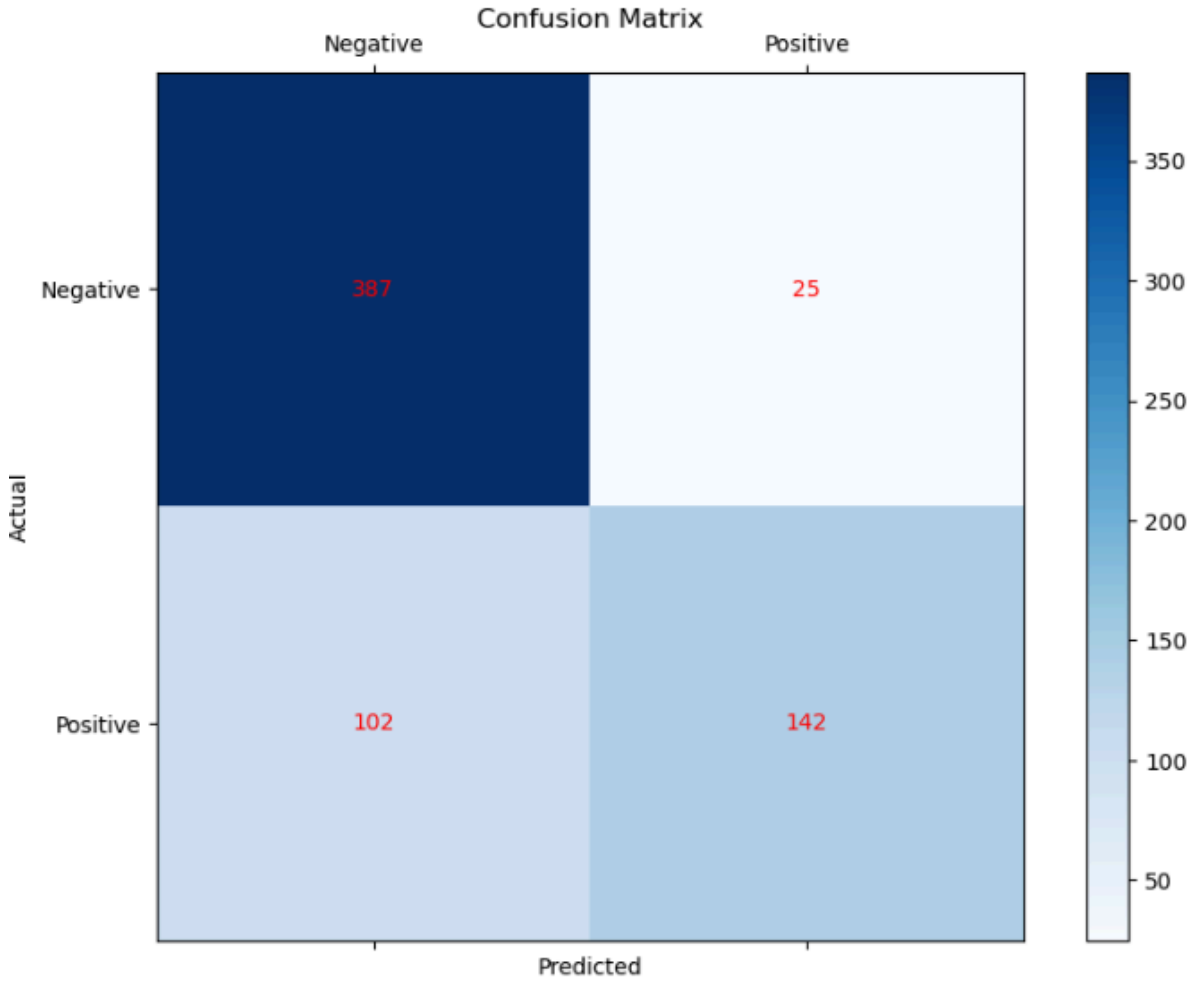
```
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.94      0.86       412
           1       0.85      0.58      0.69       244

    accuracy                           0.81       656
   macro avg       0.82      0.76      0.78       656
weighted avg       0.81      0.81      0.80       656
```

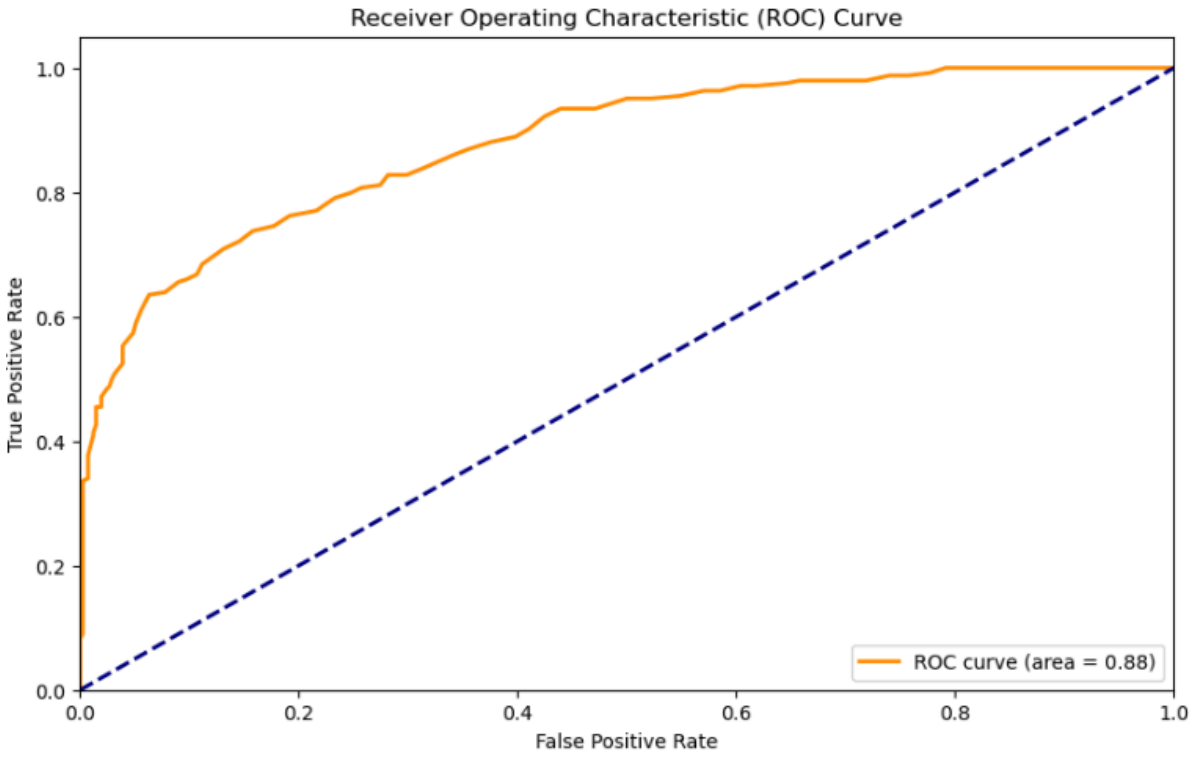
The classification report explains each class's accuracy, recollection, F1-score, and support. For class 0 (non-potable water), the model has a precision of 0.79, showing that it correctly detects 79% of non-potable forecasts. It has a high recall of 0.94, meaning that it properly identifies 94% of all genuine non-potable instances. For class 0, the F1-score, which calculates accuracy and recall, is 0.86. The accuracy for class 1 (potable water) is 0.85, showing that 85% of the predicted potable instances are correct; however, the recall is 0.58, showing that the model only determines 58% of the actual potable situations. Class 1's F1 score is 0.69. The weighted average precision, recall, and F1-score of the model are around 0.81, 0.81, and 0.80, respectively, yielding an overall accuracy of 81%.

Confusion Matrix:



The model accurately identified 142 potable samples (TP) and 387 non-potable samples (TN). However, it also wrongly classified 25 non-potable samples as potable (FP) and 102 potable samples as non-potable (FN). This matrix shows that, while the model is good at identifying non-potable water, it improves at properly identifying potable water.

ROC-AUC Curve:



The ROC-AUC curve shows the model's ability to identify differences across classes. The model performs well in terms of balancing sensitivity and specificity, shown by its AUC of 0.88, which shows that it can identify between potable and non-potable water.

Conclusion

In conclusion, this project aims to conduct a comprehensive analysis on the relationship between water quality features and potability status, with the final goal of building a model to predict water potability. After a thorough exploration and analysis of the dataset, several key findings have appeared:

1. **What is the distribution of each water quality feature between potable and non-potable?**
   Through the distribution of water quality features, we found out that all features have approximately similar distribution for potable and non-potable water. The distribution of ph, hardness, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity is approximately normally distributed for both potabilities. However, the distribution of solids is slightly right-skewed for both potabilities.

2. **Does the water quality feature have correlations with water potability?**

The correlation analysis reveals that all water quality features have near zero correlation to water potability. Therefore, we have concluded that the individual water quality features do not have a significant correlation with water potability.

3. **What can be recommended to improve water potability based on the data analysis?**
   Based on the hypothesis test, we have concluded that solid has a high significant difference between potable and non-potable water. Therefore, we should focus on monitoring and regulating solids to improve water potability. This means reducing the levels of solids in water could improve the water potability.

4. **Can we predict water potability based on water quality indicators?**
   Yes, in our Random Forest Classifier, we achieved a training accuracy of 100% and testing accuracy of 82.47%. Therefore, this indicates that water potability can be reasonably predicted based on the water quality features.

Challenges encountered and limitations of current work

Building a predictive model for water potability involves navigating through various challenges. Therefore, several limitations occur and could have a huge impact to the overall project. Understanding these limitations is very significant for future improvements of predictive models for water potability.

1. **Dataset Searching**: One challenge encountered was in searching for a dataset large enough for statistical significance because some datasets were either to small or lacked the variables needed to address the topic.

2. **Near Zero Correlation**: One challenge was dealing with variables that exhibited correlations that were near zero with the target variable (potability). With the correlation, it brings a big challenge in identifying which features were relevant that will affect the accurate predictions.

3. **Overfitting**: Our random forest classifier displayed a 1.0 training data accuracy but only a 0.8247 on testing data accuracy. Therefore, the model indicated overfitting so we applied the regularisation techniques and cross-validation to enhance our model. Although we tried our best to optimize our model, it still did not achieve a good result which indicates that further improvements are needed for this model.

4. **Model Complexity**: It is quite challenging for us to find the right balance between model complexity and performance because there are many complex models, that carry the risk of overfitting, and have the potential to capture intricate patterns. However, simpler models may not encompass all pertinent information.

5. **Model Accuracy:** There is still an ability to improve since the accuracy of the model on the testing set is 82.47%, Enhancements in feature engineering, thoughtful model selection, and fine-tuning hyperparameters are necessary. Additionally, the low correlation of features with the target variable suggests that more sophisticated feature extraction techniques may be needed to reveal hidden patterns in the data.

References

1) *Water quality*. (2021, April 25). Kaggle . - https://www.kaggle.com/datasets/adityakadiwal/water-potability