# Introduction

Water is a key factor for human survival. Although water may seem abundant since it makes up roughly 75% of the Earth’s surface, only 0.78% of this resource is accessible for human consumption (Misachi, 2018). According to the article posted by World Atlas (2018), an estimated 2.5% of water on Earth is non-saline, freshwater. Only a fraction of this freshwater, roughly 31%, is accessible. The remaining freshwater is currently frozen in glaciers, ice sheets, or ice caps. Two common sources for freshwater are groundwater (underground aquifers) and surface water such as rivers and lakes (BOSAQ, 2020). Other potential sources for potable water are rain/snow, wastewater, and saltwater processed through desalination systems. The rain/snow fall is essential for helping to replenish the groundwater and surface water, but there are also systems in place to harvest this rainfall/snowfall in a more efficient manner. The wastewater requires extensive filtration to remove contaminants, such as fecal matter, and may be used for irrigation purposes as well. The desalination performed on saltwater is currently high in cost, however this may be a suitable path due to the plethora of saltwater available on Earth. In the United States, a common source of water is considered tap water in which water is pulled from a centralized water supply (BOSAQ, 2020). Unfortunately, about 30% of the world population do not have access to potable water. This is most common in third world or developing countries.

The freshwater availability is currently limited. The costs associated with setting up and managing water filtration and sanitation systems differ depending on the water source. The objective of this project is to construct a model to predict whether water is potable based on water quality measurements. This predictive model is beneficial for multiple reasons. First, a water sanitation company could benefit by testing the potability of the water through certain stages of the sanitation process. By testing the water at different stages and inputting the data into this model, a company may find they are able to eliminate non-value-added steps within their process. Second, an organization such as the Environmental Protection Agency (EPA) may benefit from this model through the testing of raw water sources. Additional bodies of water or water collection systems may be identified and tested to distinguish water sources either approved for drinking or requiring minimal sanitation. Third, individuals will benefit from a potable water prediction model since it will help reduce the consumption of non-potable water. Some example diseases transferred from non-potable water are diarrhea, polio, typhoid, and cholera. According to the World Health Organization (2022), each year there are 829,000 individuals that die from diarrhea transmitted from unsafe drinking water or poor hand hygiene. Many of these individuals, roughly 35.8%, are children under the age of five. A predictive model for drinkable water will help to reduce this unfortunate statistic. Water quality is extremely important and needs to be a major focus worldwide.

The dataset for this model is from Aditya Kadiwal’s post on Kaggle. The file format is Comma Separated Value (CSV) and consists of a consolidation of ten features measured across 3276 unique bodies of water. Unfortunately, there is no information available as to how the measurements were obtained, so the data will need to be explored and understood thoroughly. The ten features captured in this dataset are pH, Hardness, Total Dissolved Solids (TDS), Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potability. The pH test will indicate how acidic (0-6.9), neutral (7), or basic (7.1-14) the sample measures. It serves as a useful measure for water quality since corrosive water can lead to contamination from pipes and appliances. The recommended pH value from the World Health Organization (2022) is between 6.5-8.5, but will vary depending on materials touching the surface of the water. Hardness does not contain any specific guidelines from the World Health Organization. Hardness is a measure, typically in milligrams of calcium per liter, of the volume required to react with soap. There are usually calcium or magnesium cations that contribute to the Hardness in water. Total Dissolved Solids (TDS) represent traces of organic matter and inorganic salts caused from environmental sources. The desired range for TDS is between 500-1000 milligrams per liter (Kadiwal, 2021). Chloramine is generated from chlorine reacting with ammonia and results in odd smells or taste within the water. The desired level of chlorine is below 4 milligrams per liter (Kadiwal, 2021). Sulfate in drinking water may result in stomach issues and contribute to undesired taste. According to Kadiwal, most freshwater sources have sulfate in the 3-30 milligram per liter range. Conductivity will help indicate the amount of dissolved solids in water by measuring how well the water conducts electrical current. The guidelines recommend conductivity below 400 micro-Siemens per centimeter (Kadiwal, 2021). Total Organic Carbon (TOC) results from decomposing natural organic matter and is recommended to be below 2 milligrams per liter for drinking water (Kadiwal, 2021). Trihalomethanes result from chlorine treated water and are recommended to be below 80 parts per million for drinking water (Kadiwal, 2021). Turbidity utilizes light with water to measure waste discharge. The recommended Turbidity for drinking water is below 5 Nephelometric Turbidity Units (NTU). Lastly, the Potability feature indicates whether the water sample is drinkable or not. This data will be useful to solve the problem because it includes nine features based on measurements for water samples across over 3200 bodies of water.The target variable, Potability, will be used to build the model to address the questions of interest posed later in this section.

There will be six models evaluated for this project. Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine (SVM), and Adaboost Models will be constructed, trained, and tested with the selected data. These models were chosen since the framework for the problem includes Supervised Learning with a binary target variable (Potable, Non-Potable). A confusion matrix will be utilized to calculate metrics such as accuracy, precision, recall, and F1-score. The models will be compared based on these different metrics and the best model will be identified. A review of the best model will be discussed and a recommendation will be given whether it is ready for deployment. In terms of what I hope to learn, the high-level questions for this project are shown below:

* How balanced is the dataset for Potable vs. Non-Potable water samples?
* Are there any insights for the feature distributions included in this dataset?
* What are the main features contributing to water Potability? Are there any features highly correlated with water Potability within this dataset?
* Which model performs the best to predict water Potability?
* Show how the best performing model compares to other models in the analysis.
* Is the model recommended to be deployed?

There are a few risks associated with this project. The dataset chosen did not have any details regarding the way the water quality measurements were obtained or consolidated into the CSV file. The measurement method for the values within this dataset are assumed to be valid. In addition, the exploratory analysis performed on the dataset will require a thorough review of missing values, feature distributions, and outlier detection/handling. Another risk associated with this project is the inability to construct a deployable model with the training and test data available. The evaluation metrics will be illustrated to compare the six predictive models. In the event the project does not go as expected, the contingency plan will be to locate an alternate dataset pertaining to water quality. Depending on the dataset(s) available, the scope of the project may need to be modified. For example, the scope may need to be narrowed down to water quality in a particular region/country. The plan is to utilize Python, specifically Jupyter Notebook, to perform the analysis. This project may serve as a template for domain experts in water quality, water treatment, or environmentalists to build from for additional insights.

# Methods/Results

This section will be broken down by Data Understanding, Data Preparation, Model Training (and Testing), and Model Evaluation sections. There may be some overlap discussed between the sections, but this seemed like the most logical approach to breakdown the methods/results.

The Data Understanding phase for this project involved several steps to better comprehend the dataset. These steps included an initial overview of the data, preliminary data preparation, univariate analysis, and bivariate analysis. During the initial overview of the data, the dataset was confirmed to have 3276 records and ten features. The “Potability” was treated as a categorical variable considering “0” as non-Potable and “1” as Potable. The remaining nine features were loaded into the data frame as numeric data types. There were 1434 missing values within the Sulfate, pH, and Trihalomethane columns combined. When broken down by percent of missing values, Sulfates had ~24%, pH had ~15%, and Trihalomethane had ~5%. The table below summarizes the missing values found within the dataset.

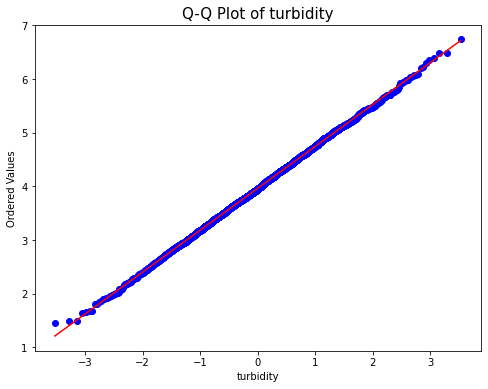
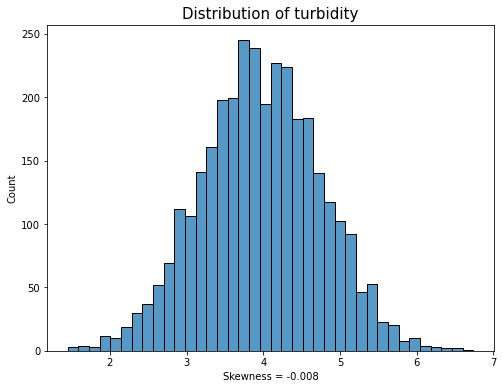
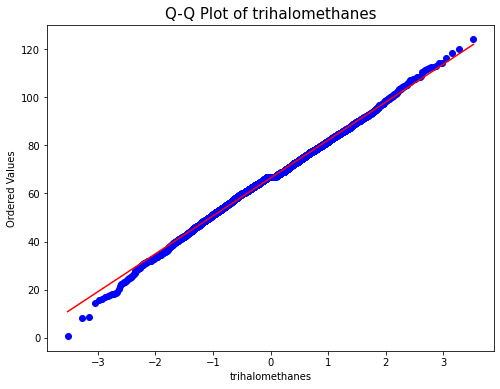
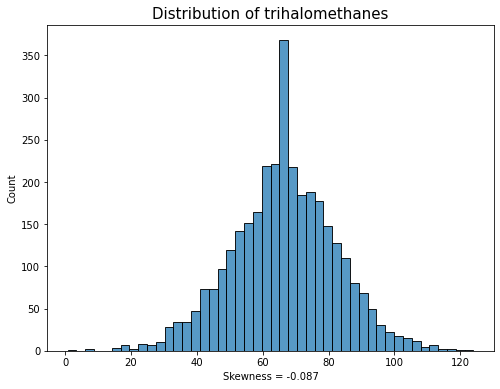
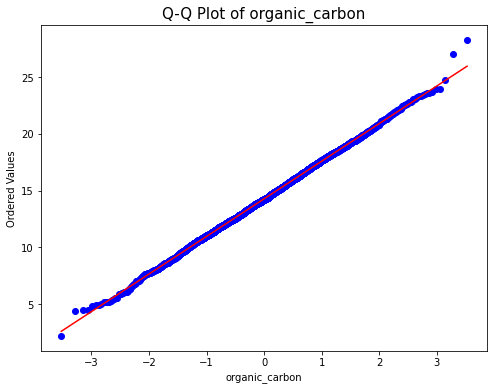
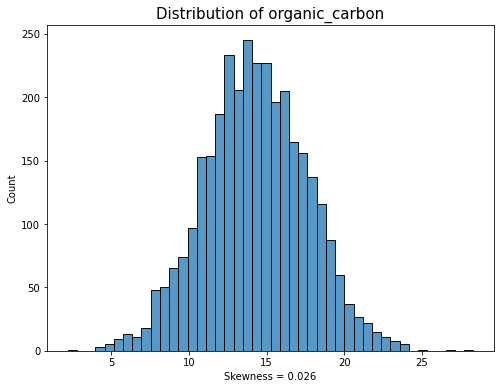
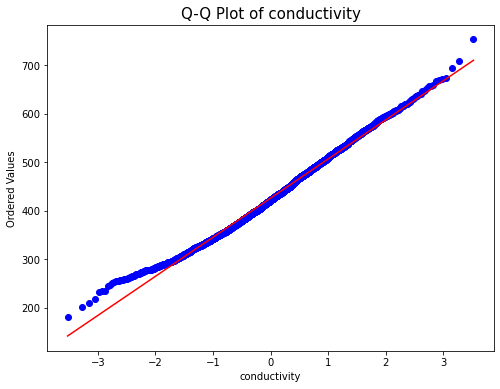
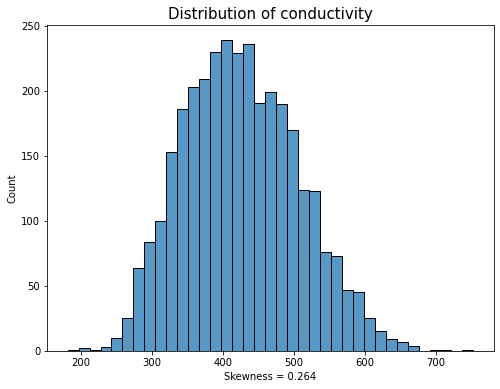
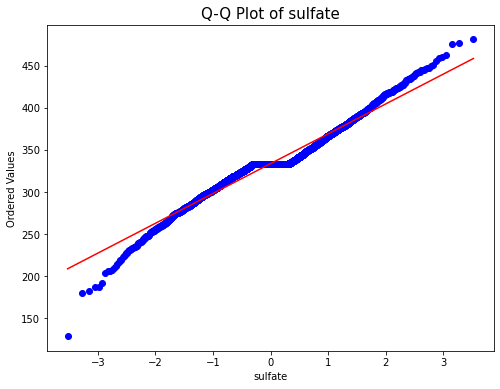
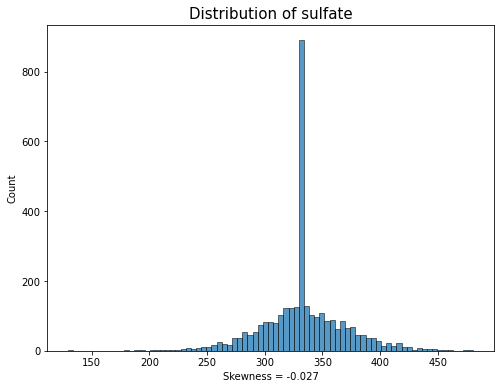
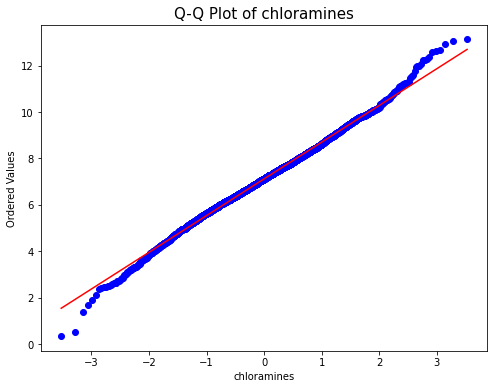
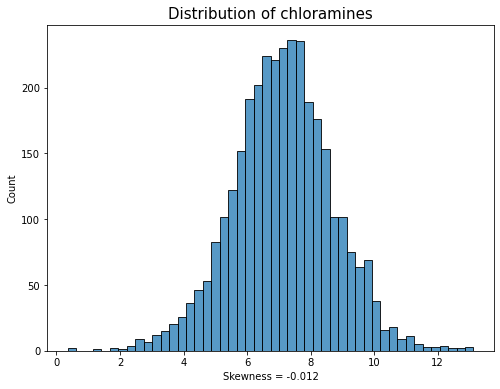
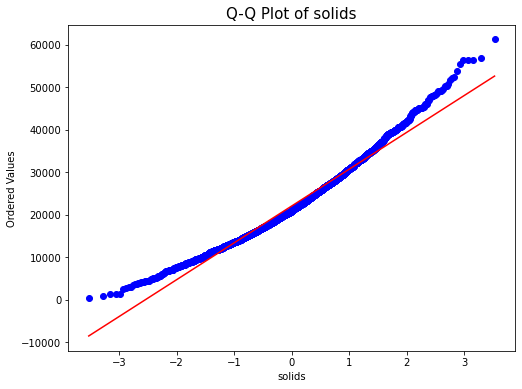
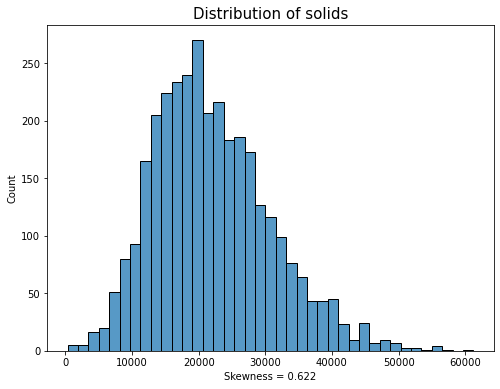
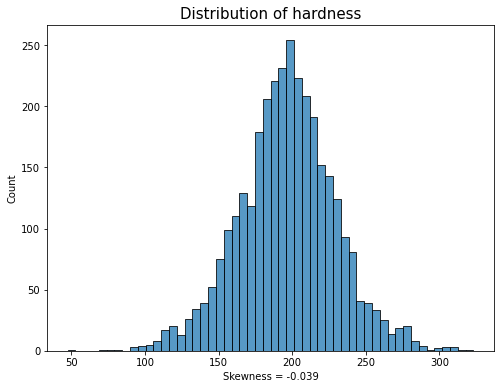
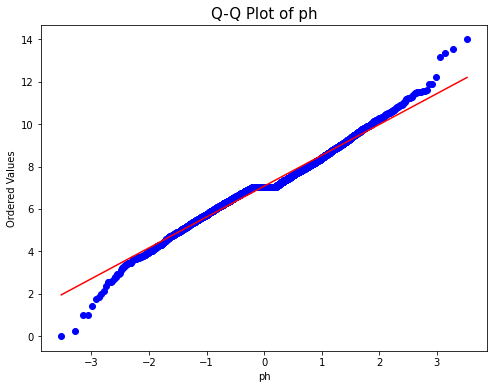
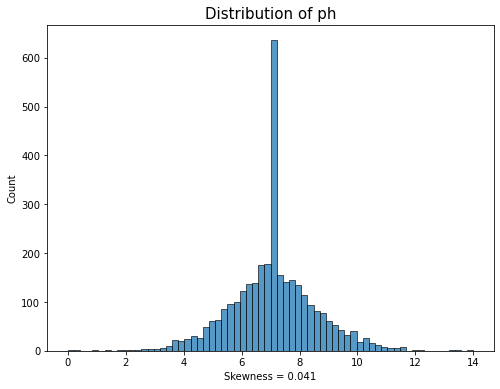
Table 1: Percentage of Missing Values by Feature

Graphical user interface

Description automatically generated with medium confidence

After exploring the options of removing rows with missing values or imputing values in-place of the missing values, it was decided to use the imputing method with the median of each feature. This option was chosen after evaluating the mean and median values broken down by Potability for each feature with missing values. Another preliminary data preparation step performed was renaming the columns to all lower case (no spaces) for convenience. After performing the preliminary data preparation steps, the next step was to understand the features better through Univariate Analysis. For Univariate Analysis, Histograms and Q-Q Plots were utilized for each of the numeric variables. The figures shown below overview each feature’s Histogram and Q-Q Plot respectively.

Figure 1: Histogram and Q-Q Plots for Univariate Analysis



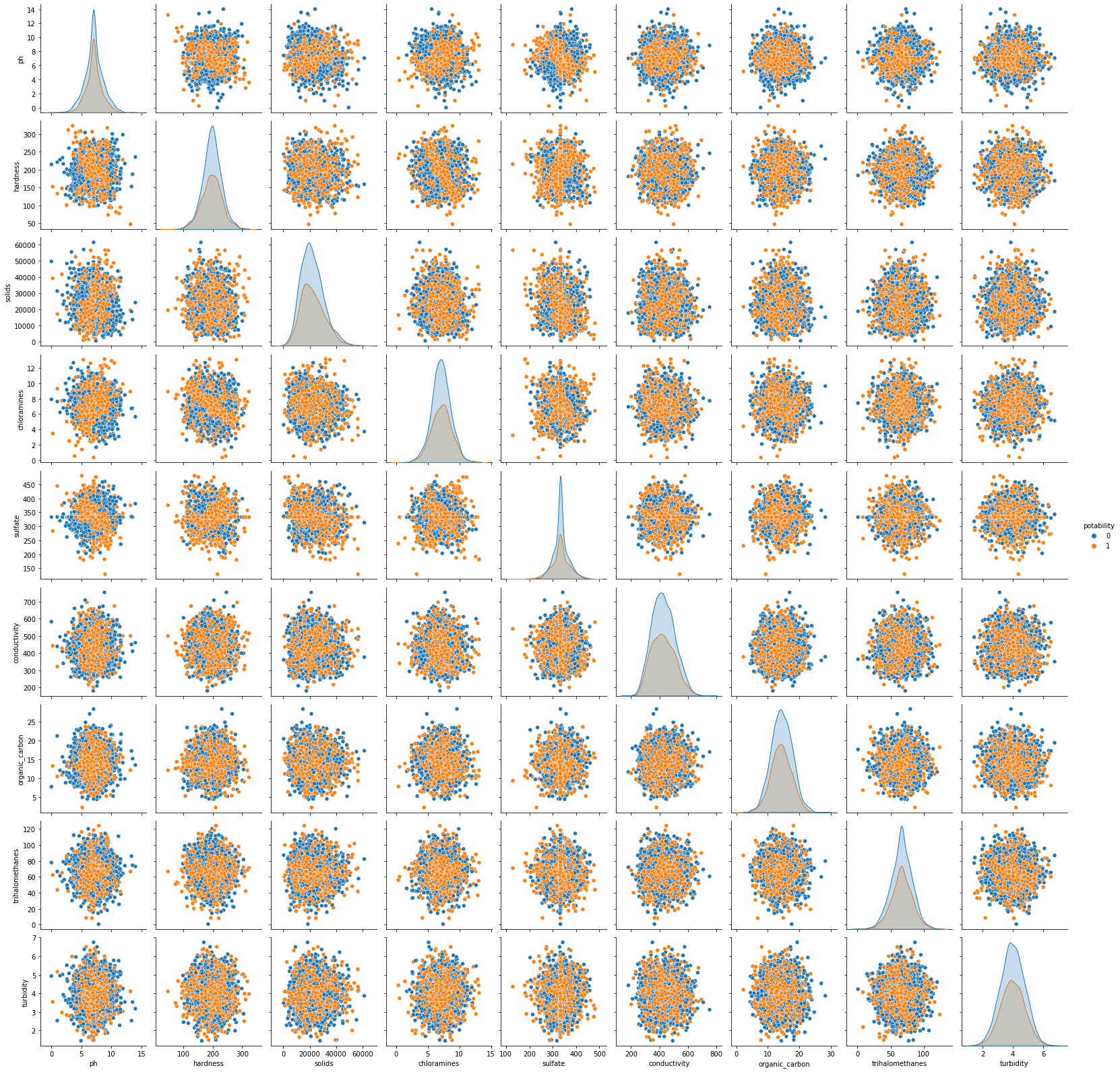
Interestingly, most of the features within the dataset appeared to be normally distributed based on the Histograms. However, after further analysis using Q-Q Plots and Shakiro-Wilks test, only Organic Carbon and Turbidity are normally distibuted. With the missing values imputed, the Sulfate, pH, and Trihalomethane features show Leptokurtik signatures due to spikes in median values. In addition, the option of removing outliers outside of +/- 3 Sigma of the Mean for each feature was explored to see if there was a significant impact on the models. No significant impact was found based on the model metrics, so all the data points were retained for the remainder of the analysis. A countplot was constructed for the target variable to better understand the balance of Potability within the dataset.

Figure 2: Countplot of Potability



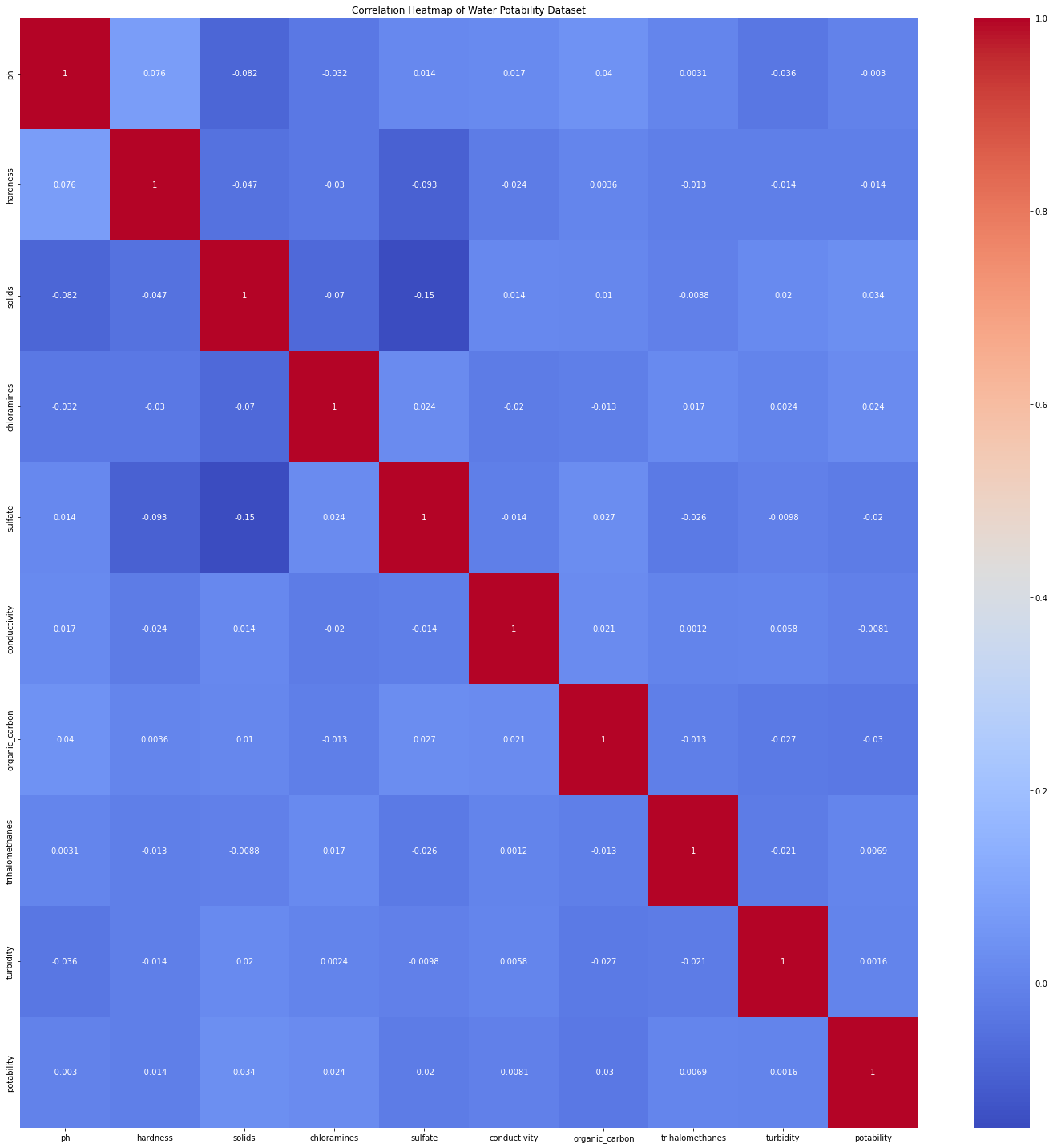
There were 61% of the records classified as non-Potable and 39% considered Potable. Although there is an imbalance in favor of non-Potable data, there are sufficient sample sizes of each category to move forward with the analysis. For Bivariate Analysis, a pairplot (assortment of scatterplots and histograms), correlation heatmap, and boxplots were utilized to review variable relationships. The first visualization generated was the pairplot shown in the figure below.

Figure 3: Pairplot of Water Quality Data



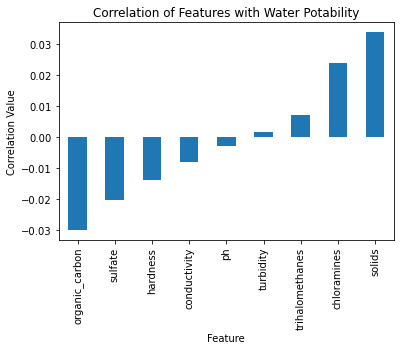
The pairplot helps illustrate the lack of strong relationships for each variable with respect to Potability. A correlation heatmap was constructed to help confirm this conclusion. The correlation heatmap is shown in the figure below.

Figure 4: Correlation Heatmap of Water Quality Data



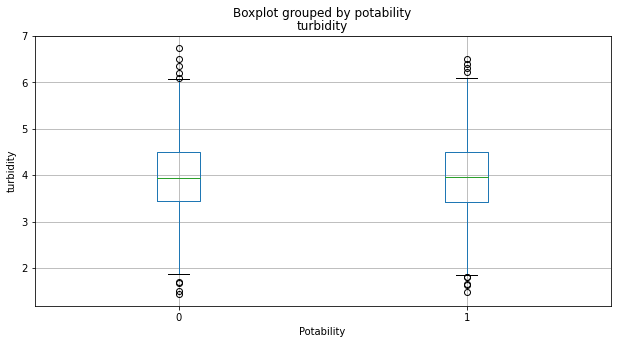
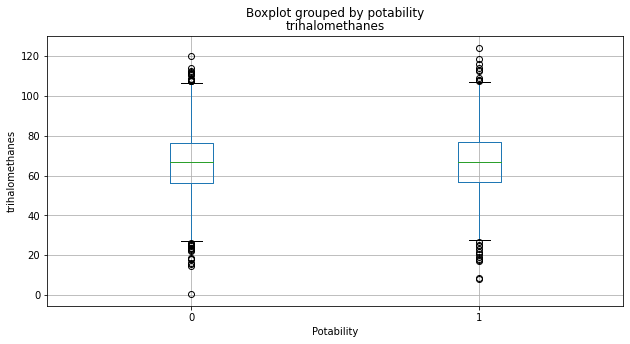
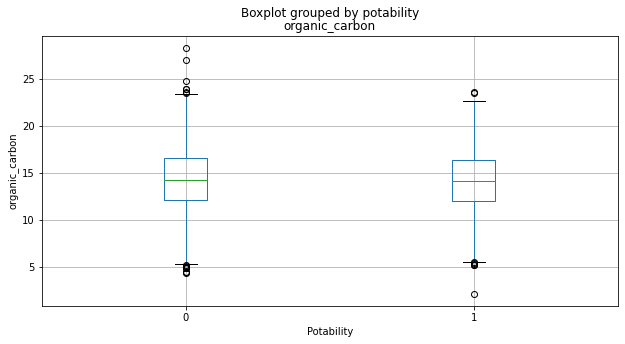
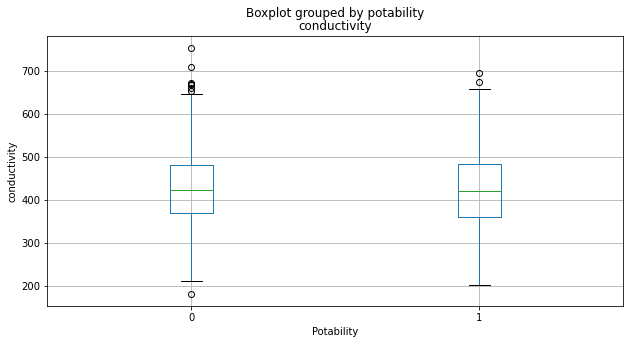
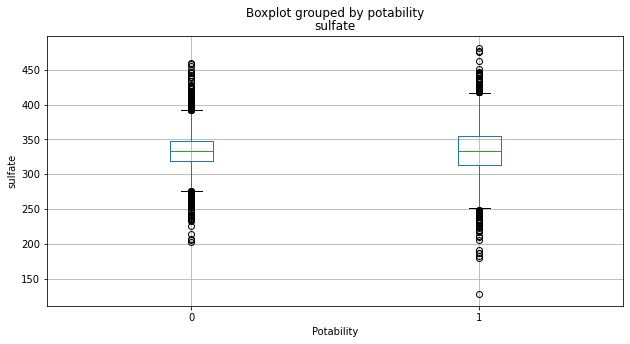
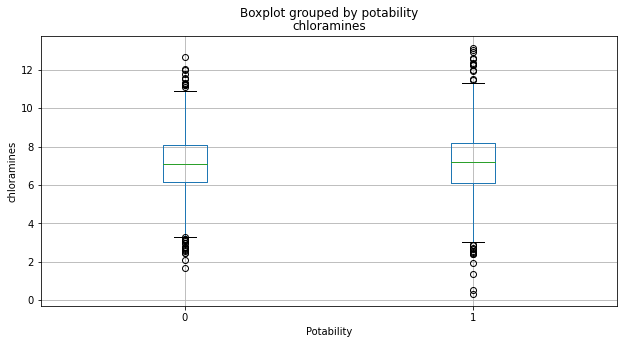
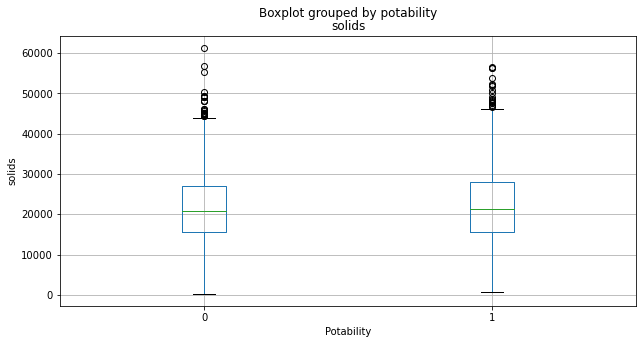
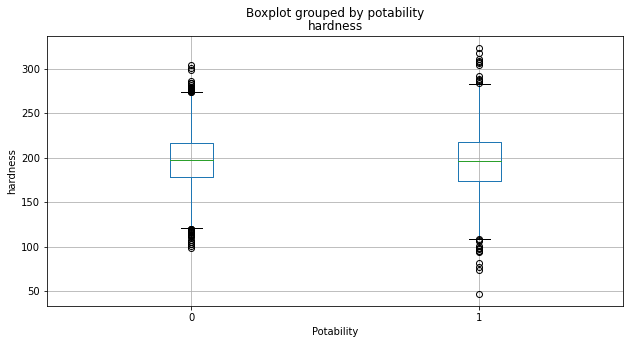
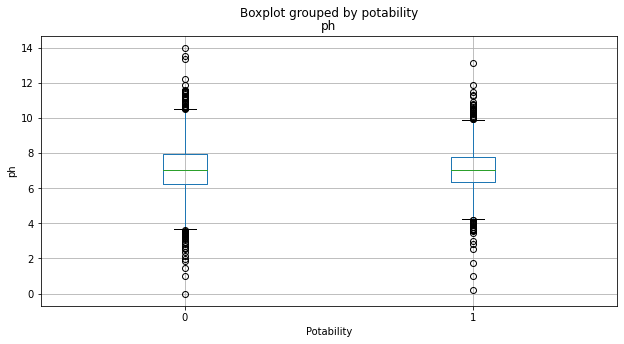
The correlation heatmap shows no strong relationship between any of the features within the dataset, including Potability. Based on this finding, additional datasets and other potential features were pursued. However, it is recommended to consult domain experts for water quality to understand alternative feature options. The analysis continued forward to determine the best model that could be achieved utilizing the current dataset. Prior to the Modeling Training section, two additional plots were constructed to better understand the relationships with Potability. The Bar Chart below illustrates which features in the dataset have the strongest correlation, positive or negative, with Potability.

Figure 5: Bar Chart showing feature correlations with Water Potability.



The top two features with a positive correlation of water Potability are Total Dissolved Solids and Chloramines. The top two features with a negative correlation for water Potability are Organic Carbons and Sulfates. Lastly, Boxplots were utilized to show feature distributions with respect to water Potability in an alternate manner.

Figure 6: Bivariate Boxplots for each feature with respect to Potability.

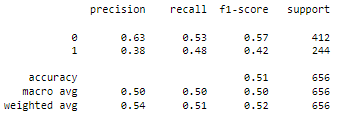
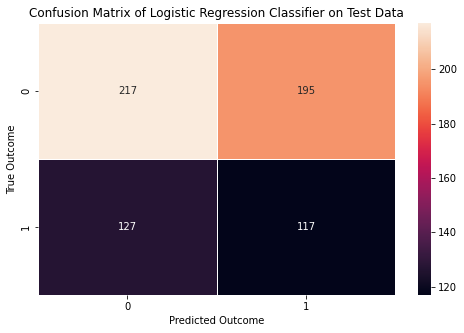
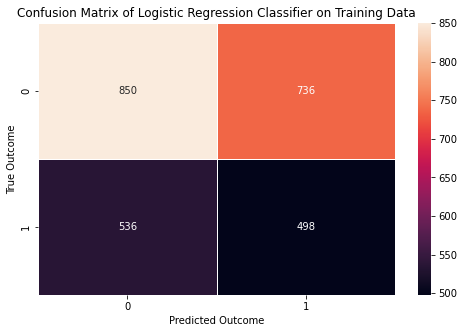


These plots also helped show the distributions are not much different for each respective feature when broken down by Potable and non-Potable. The next step was to prepare the data for modeling.

The Data Preparation phase was relatively straight forward. The preliminary data preparation performed in the Data Understanding phase consisted of renaming the columns, addressing the missing values, and portraying the descriptive statistics for the numeric features. This section of Data Preparation was intended to prepare the data for model training and testing. The data was split into random train (80%) and test (20%) subsets of the overall dataset. A standard scalar was applied to the featuring training and test datasets and used for all model training except for the Logistic Regression Model. The standard scalar was fit and transformed to the training subset, and then transformed for the test subset. The Logistic Regression Model was trained without the standard scaler applied.

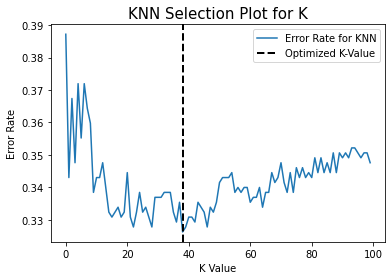
The next phase entailed model training and testing using the subset data. As previously mentioned, the models considered for this analysis were Logistic Regression, K-NN (K-Nearest Neighbor), Decision Tree, Random Forest, Support Vector Machine (SVM), and Adaboost. These models were chosen since the problem being addressed is a classification problem. The evaluation metrics considered for this analysis were overall model accuracy, precision, recall, and F1-Score. A confusion matrix and classification report were output for each model during this phase. All models were created using the sklearn package from Python. The first model trained was the Logistic Regression Model. The Logistic Regression Model had the solver set to ‘liblinear’, multi\_class = ‘ovr’ (binary classification problem), and class\_weight set to ‘balanced’. This model was fit to the non-scalar training data and predictions were generated for both the training and test data. The confusion matrix for each respective scenario and classification report are outlined below.

Figure 7: Logistic Regression Confusion Matrices (Training and Test) and Classification Report.



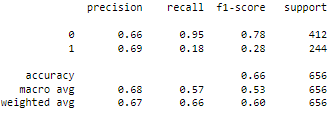
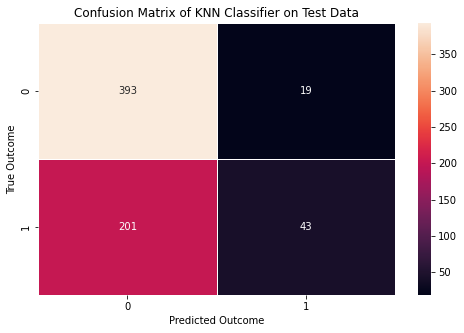
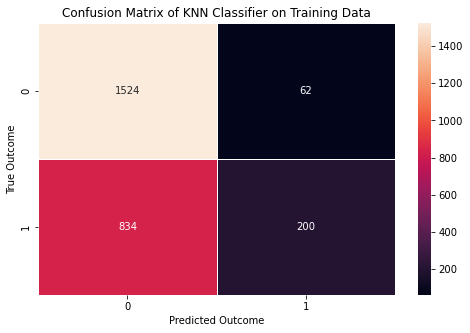
The remaining models all utilized the training data with the standard scalar applied as previously specified. The next model was K-NN. The optimum K-value was determined for this model by plotting the error rate for the model against a range of K-values from 0 to 100. As seen in the plot below, the optimum K-value was found to be 38.

Figure 8: Optimum K-Value for K-NN Model.



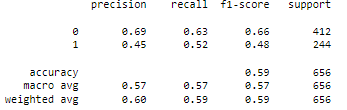
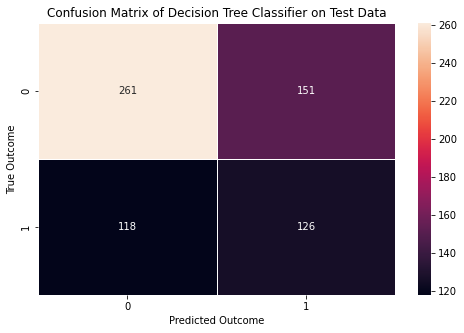
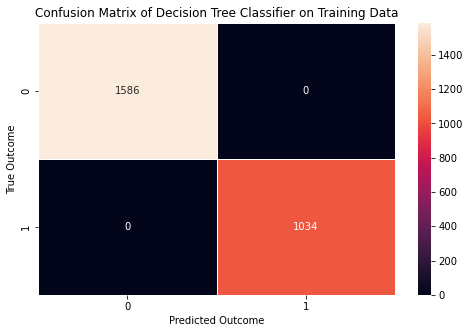
The K-NN model was created with a K-value set to 38. The model was fit to the standardized training subsets and predictions were generated on both training and test datasets. The confusion matrices and classification report are shown below.

Figure 9: K-NN Confusion Matrices (Training and Test) and Classification Report



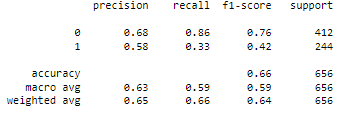
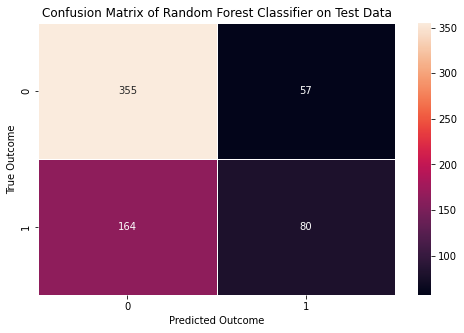
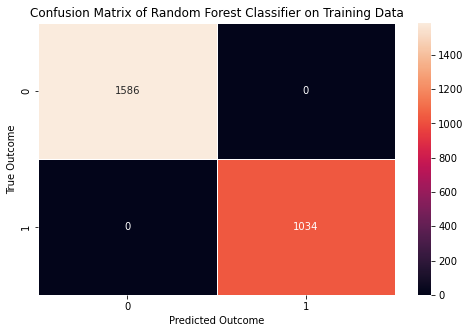
A Decision Tree Classifier was created and trained next. No model parameters were tuned. The model was fit to the standardized scaler training data and predictions were generated for both training and test datasets. The confusion matrices and classification report are shown below.

Figure 10: Decision Tree Confusion Matrices (Training and Test) and Classification Report



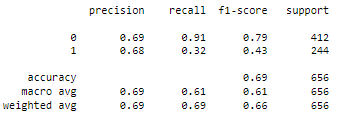
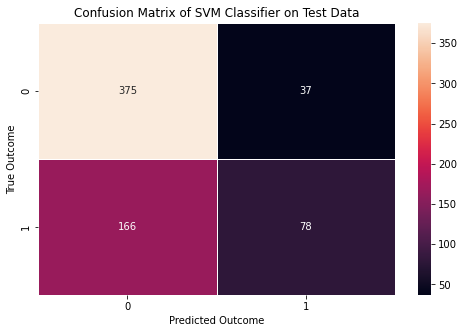
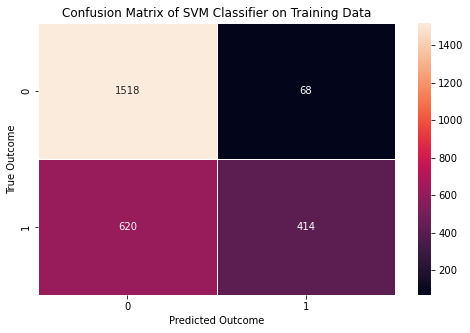
After the Decision Tree Classifier, a Random Forest Model was generated. The number of estimators was set to 78 based on a plot of Model Error Rate against number of Estimators. The optimum number of estimators would change each time the code was run. The model was fit to the standardized scaler training data and predictions were generated for both training and test datasets. The confusion matrices and classification report are shown below.

Figure 11: Random Forest Confusion Matrices (Training and Test) and Classification Report



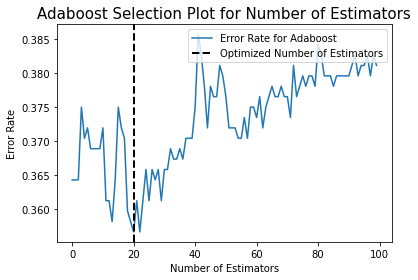
A Support Vector Machine (SVM) was created and trained next. No model parameters were tuned. The model was fit to the standardized scaler training data and predictions were generated for both training and test datasets. The confusion matrices and classification report are shown below.

Figure 12: SVM Confusion Matrices (Training and Test) and Classification Report

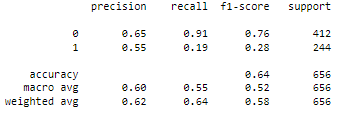
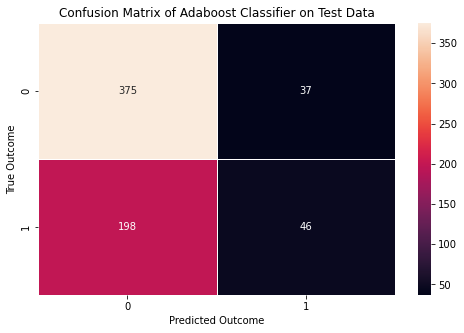
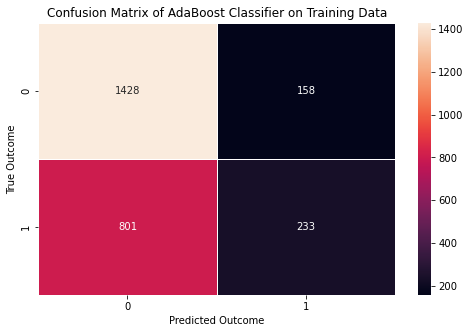


The last model created was Adaboost. The number of estimators was set to 20 based on a plot of Error Rate versus the Number of Estimators from 0 to 100.

Figure 13: Optimum Number of Estimators for Adaboost Model

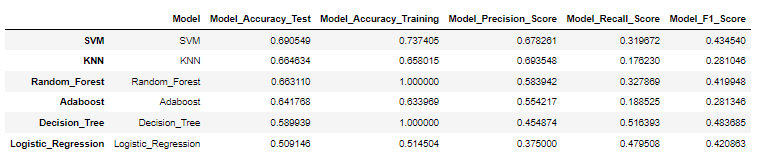


The Adaboost model was created with the Number of Estimators set to 20. The model was fit to the standardized scaler training data and predictions were generated for both training and test datasets. The confusion matrices and classification report are shown below.



The final phase for this section consists of Model Evaluation. A summary of the evaluation metrics was consolidated into a table. The table below shows the six models ordered from highest to lowest accuracy on the test data. The training accuracy, precision, recall, and F1 scores (on the test data) are also shown in the table for reference.

Table 2: Model Evaluation Metrics



The SVM Model had the best performance. At this point in time, formal hyperparameter tuning has not been performed on all the models. This is an area that can be explored to improve the model(s) performance. The accuracy of a model is the ratio of the total number of correct predictions over the total number of predictions. Precision focuses on the total number of true positive values divided by the sum of true and false positives. This metric is important to consider when it is important to keep the false positive error low. In this case, this is an extremely important metric since we do not want to misclassify a water record as potable when it is not. This could be dangerous from a health perspective. Recall is an important metric when considering false negative errors. This would be an example of misclassifying water as non-potable when the water is potable. This type of error could be very costly for an organization responsible for sanitization or further cleaning/processing for water that is okay. Lastly, the F1 score is calculated by multiplying two times the ratio of precision multiplied by recall over the sum of precision and recall. This metric is useful for understanding models with a good balance between precision and recall. The model that had the best overall accuracy on the test data was SVM. The Random Forest and Decision Tree Classifiers both had 100% accuracy on the training data, but much lower accuracy on the test data. This indicates these models are overfit in their current state. The KNN Model had the highest Precision Score, but the lowest Recall Score. The Decision Tree Classifier had the highest Recall Score, but one of the lowest accuracies with the test data. The Random Forest Classifier had the highest F1 Score and was the second-best model in terms of accuracy with the test data. Logistic Regression Model had the lowest overall accuracy on both training and test datasets. AdaBoost was consistent in terms of accuracy across the training and test data, however it also had one of the lowest Recall Scores. There are trade-offs with each of these models and none of them are in a current state to be deployed. Cross-Validation was utilized to understand how well the models would perform over multiple iterations.

Figure 14: Model Accuracy with Cross-Validation (20 Iterations)



The top three models are Support Vector Machine, Random Forest, and K-Nearest Neighbor. The boxplot above shows how each model scored with cross-validation over 20 iterations. Although the SVM Model performs the best overall, this model is not recommended to be deployed in the current state. It is recommended to check with domain experts for additional features to include in the dataset that may be higher correlated to water potability. This analysis served useful for understanding the water quality dataset and setting up a standard method for predictive modeling, however the SVM model is not recommended to be deployed now.

# Conclusion

The high-level questions posed for this project were all addressed during the analysis. The dataset was imbalanced towards non-Potable records by ~22%. This could explain why the models were generally better at predicting non-Potable water compared to Potable water. The Univariate Analysis did not lead to any additional data preparation steps but helped provide a solid understanding of each feature. The Bivariate Analysis helped identify the lack of correlation between Potability and the features included in this dataset. Although there are no features highly correlated with water Potability, the features that had the highest positive and negative correlations were recognized. The top two features with a positive correlation of water Potability were Total Dissolved Solids and Chloramines. The top two features with a negative correlation for water Potability were Organic Carbons and Sulfates. The top performing model was SVM. The SVM model generally performs better than the other models as seen after cross-validation with 20 iterations. At this point in time, the SVM model is not recommended to be deployed. The methods used in this analysis are solid foundations to build from and can be applied with other water quality datasets. However, there are a few recommendations for individuals interested in building from this analysis to improve model accuracy. First, obtain the insights from Subject Matter Experts (SME) for water quality to see what additional features need to be included in this dataset. Another option could be a calculated feature based on the features that are already within the dataset. These features need to be higher in correlation with Potability. Second, recommend performing hyperparameter tuning on all the models. One can use GridSearch or another method to try to optimize the model parameters. There are currently a few recommendations for future actions to approach the problem from an alternative perspective. One recommendation is to focus data collection on a sanitation process with key input variables and output responses monitored for a set period. A final recommendation is to collect data from known potable sources and non-potable sources with key metrics (identified by domain experts) captured for analysis.

# References

Abbott, D. (2014). Applied Predictive Analytics. (1st ed., pg. 23). John Wiley & Sons, Inc.

BOSAQ. (2020, March 18). Everything You Need To Know About Water Resources. *BOSAQ Blog.* [Everything you need to know about water resources | Bosaq](https://bosaq.com/water-resources/#:~:text=%20Everything%20you%20need%20to%20know%20about%20water,from%20beneath%20the%20earth%27s%20surface%2C%20located...%20More%20)

Kadiwal, Aditya. (2021, April 25). Water Quality: Drinking water potability. *Kaggle*. [Water Quality | Kaggle](https://www.kaggle.com/datasets/adityakadiwal/water-potability)

Misachi, John. (2018, February). What Percentage of Earth’s Water Is Drinkable? *WorldAtlas*. [What Percentage of the Earth's Water Is Drinkable? - WorldAtlas](https://www.worldatlas.com/articles/what-percentage-of-the-earth-s-water-is-drinkable.html)

World Health Organization. (2022, March 21). Drinking-water quality guidelines. *Water Sanitation and Health*. [Water Sanitation and Health (who.int)](https://www.who.int/teams/environment-climate-change-and-health/water-sanitation-and-health/water-safety-and-quality/drinking-water-quality-guidelines)

World Health Organization. (2022, March 21). Drinking-Water. *World Health Organization Newsroom.* [Drinking-water (who.int)](https://www.who.int/news-room/fact-sheets/detail/drinking-water)

# Milestone 5 Requirements:

Note that you need to submit two items for your final project submission. The paper should

include an introduction, a summary of your methods and results, and a conclusion as outlined

above. In addition, submit an audio/video presentation with slides summarizing your project. A

good goal for the length of your presentation is 10-15 minutes. Think about this as a high-level

presentation you would give to your CEO.

At minimum, your final project paper should include the following.

Introduction

* Problem statement
* Explain why the problem is important/interesting
* Who would be interested in solving this problem, i.e., who would you be trying to sell
* this project to?
* Where did you get your data?
* Why is this data useful to solve the problem?

Methods/Results

* What did you find out by exploring the data?
* Are there any visualizations that help tell a story with your data?
* What steps did you perform to prepare the data?
* What type of types of modeling are you using on your data?
* What metric(s) are you using to measure your results?
* Why did you choose the metric(s) you chose?

Conclusion

* What did you learn?
* What recommendations would you make based off your analysis?
* Is your model ready for deployment?
* What work still needs to be done?

References

* Include at least three properly cited references at the end of your paper
* Also include in-text citations

# Water Quality Dataset

CSV Attachment - 