DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

Note 1: Milestone 3 considerations are integrated in the Introduction section. In addition, there are additional inputs for Milestone 3 in the Preliminary Analysis section. See the segments highlighted in yellow for Milestone 3 inputs.
Note 2: Milestone 4 updates are integrated with Milestone 3. Milestone 4 considerations are highlighted in green for easy identification. Any previous thoughts or concerns that will be removed will have a strikethrough along with an explanation of why the section will be removed.

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 (TSD) 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 help evaluate how well the model performs.~~ The target variable, Potability, will be used to build the model to address the questions of interest posed later in this section.

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

There will be three models evaluated for this project. Logistic Regression, k-Nearest Neighbor, and Decision Tree 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). For my model/evaluation choices, I am going to keep the three already discussed (Logistic Regression, K-Nearest Neighbors, Decision Tree). However, I plan to add some additional models to the analysis such as Random Forest (a common ensemble learning algorithm under "bagging") and Support Vector Machine (SVM) models. In addition, I will explore ensemble methodology for the analysis which was not originally planned. These models will be reviewed and compared later in the analysis. As a result, they will most likely need to be adjusted to some degree. ~~Some of the evaluation metrics that will be considered will be accuracy, precision, recall, F1-Scores, and confusion matrices.~~ 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?
- Which model performs the best to predict water Potability?
- ~~What are the evaluation metrics for the best performing model?~~ Show how the best performing model compares to other models in the analysis.
- Is the model recommended to be deployed?

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

The preliminary analysis for this project involved several steps to better understand the dataset. These steps included an initial overview of the data once loaded into a Pandas data frame, preliminary preparation, univariate analysis, and bivariate analysis~~, and cluster analysis. The cluster analysis was not originally part of the plan from Milestone 2.~~ Removing the clustering analysis since this is a classification problem. The dataset was confirmed to have 3276 records and ten features. The “Potability” was converted to categorical and will be used to consider each record as potable or not. 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%. 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. The imputing method did not significantly change the initial descriptive statistics for each feature and retained the same size of the dataset from the start. Another preliminary data preparation step performed was renaming the columns to all lower case (no spaces) for convenience. For Univariate Analysis, histograms were utilized for each of the numeric variables and a countplot for the target variable. For Bivariate Analysis, a pairplot (assortment of scatterplots and histograms), heatmap, and boxplots were utilized to review variable relationships. ~~A K-Means Model was constructed to understand the dataset better with the potability feature removed. Through utilization of an elbow plot, there appeared to be two main clusters signifying the potable and non-potable data within the dataset.~~ Since this is a classification problem for assessing potable vs. non-potable water sources, this K-Means clustering algorithm does not add much value to the analysis. These initial strides helped understand whether the questions posed for this project could be answered.

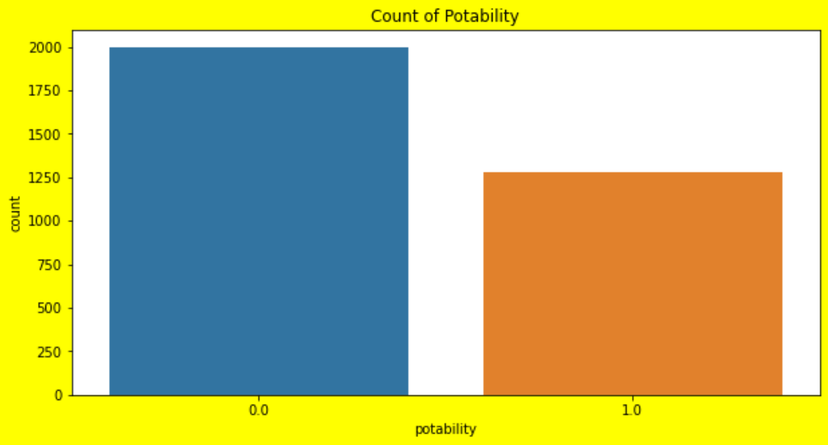
Based on the preliminary steps thus far, I will be able to answer the questions for this project. In fact, half of the questions can be answered after the data understanding phase. The dataset contains 61% of records listed as non-potable and 39% as potable. ~~Interestingly, most of the features within the dataset appear to be normally distributed.~~ Initial statement for the appearance of normal distribution was based the feature distributions illustrated from each histogram. However, after further analysis using Q-Q Plots and Shakiro-Wilks test, only Organic Carbon and Turbidity are normal. With the missing values imputed, the Sulfate, pH, and Trihalomethane features show Leptokurtik signatures due to spikes in median values. The boxplots show similar distributions for each feature regardless of whether the sample was potable or not. There are no features in the dataset that are highly, or even mediumly, correlated with potability. The questions regarding the models will need to be addressed after the analysis progresses further.

The visualizations that are useful for explaining my data are the histograms, countplots, boxplots, heatmap, and scatterplot. Momentarily, I do not need to adjust the driving questions for the project. For the data, I may have to make some adjustments with handling outliers. This will be addressed during the data preparation phase prior to modeling. The Preliminary Analysis Visualization section provides additional insights into these useful plots.

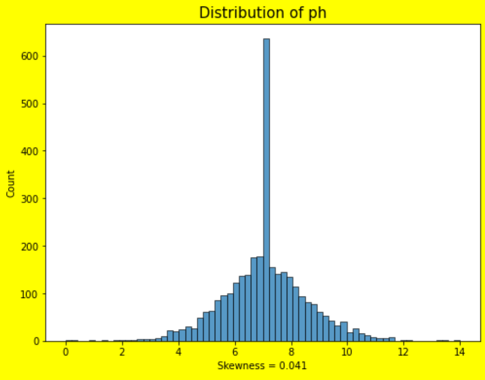
There are a few risks associated with this project. The dataset chosen did not have any details regarding the way the water quality measurements were obtained or consolidated into the CSV file. The measurement method for the values within this dataset are assumed to be valid. In addition, the exploratory 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 three 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 original expectations are still reasonable for this project. The main concern at this point is there does not appear to be any highly correlated variables with the target variable of potability. Although this is a concern, I will continue to move forward with the project and this dataset. I will pursue finding the most accurate model achievable with this dataset through the revised ensemble modeling strategy. The current 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.

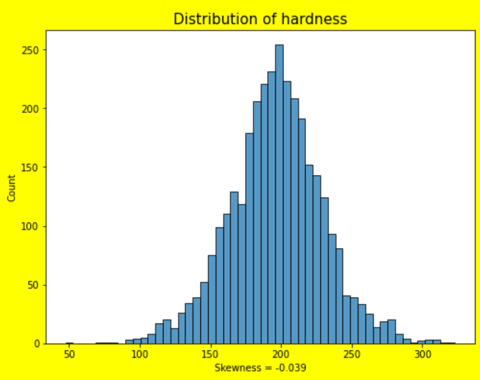
Preliminary Analysis Visualizations

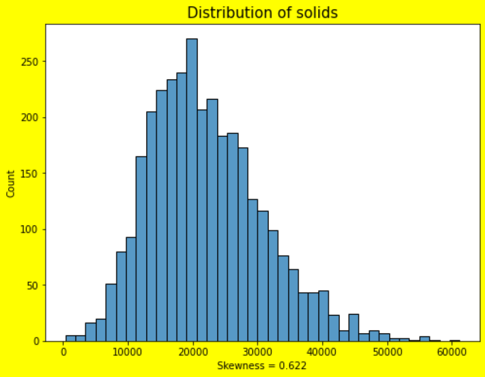
This section provides some insights into the useful Preliminary Analysis Visualizations for the dataset. The first visualization shows the count of potable (1) vs. non-potable (0) records within the dataset.

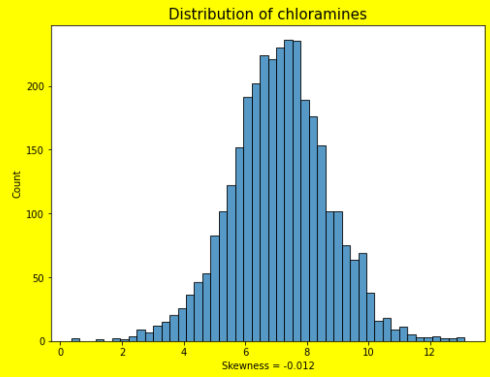


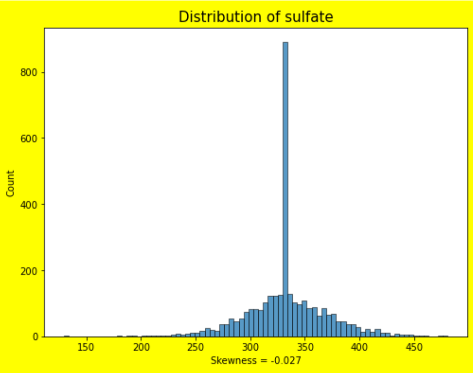
Though there are more non-potable (1998) records compared to potable (1278), there still is a decent sample size for each category. The next visualizations useful for understanding the features within the dataset were histograms. The visualizations below show examples of the data distributions for each feature.

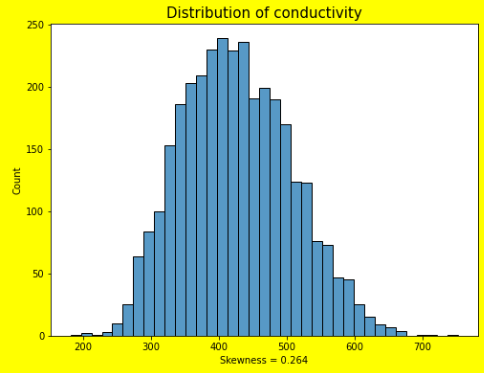


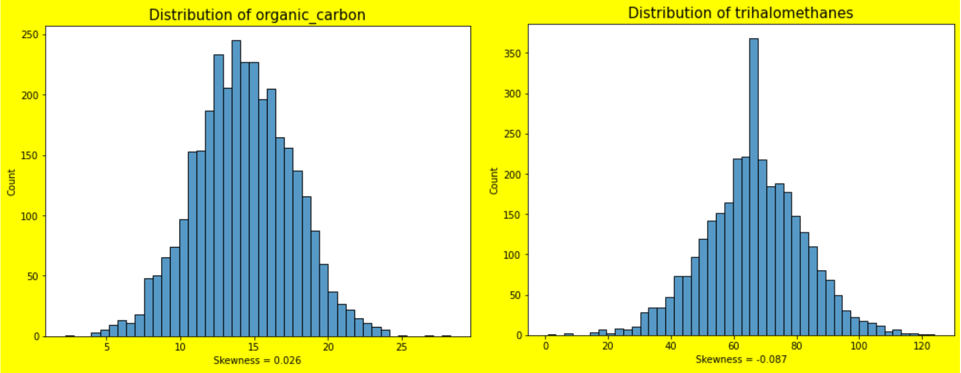


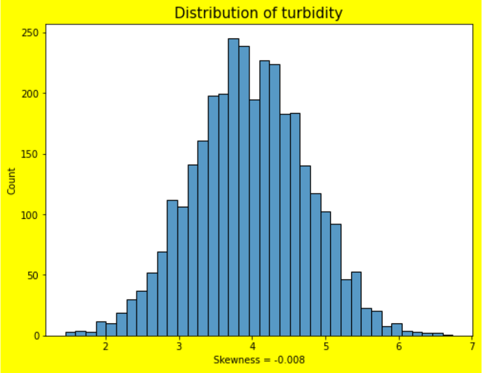




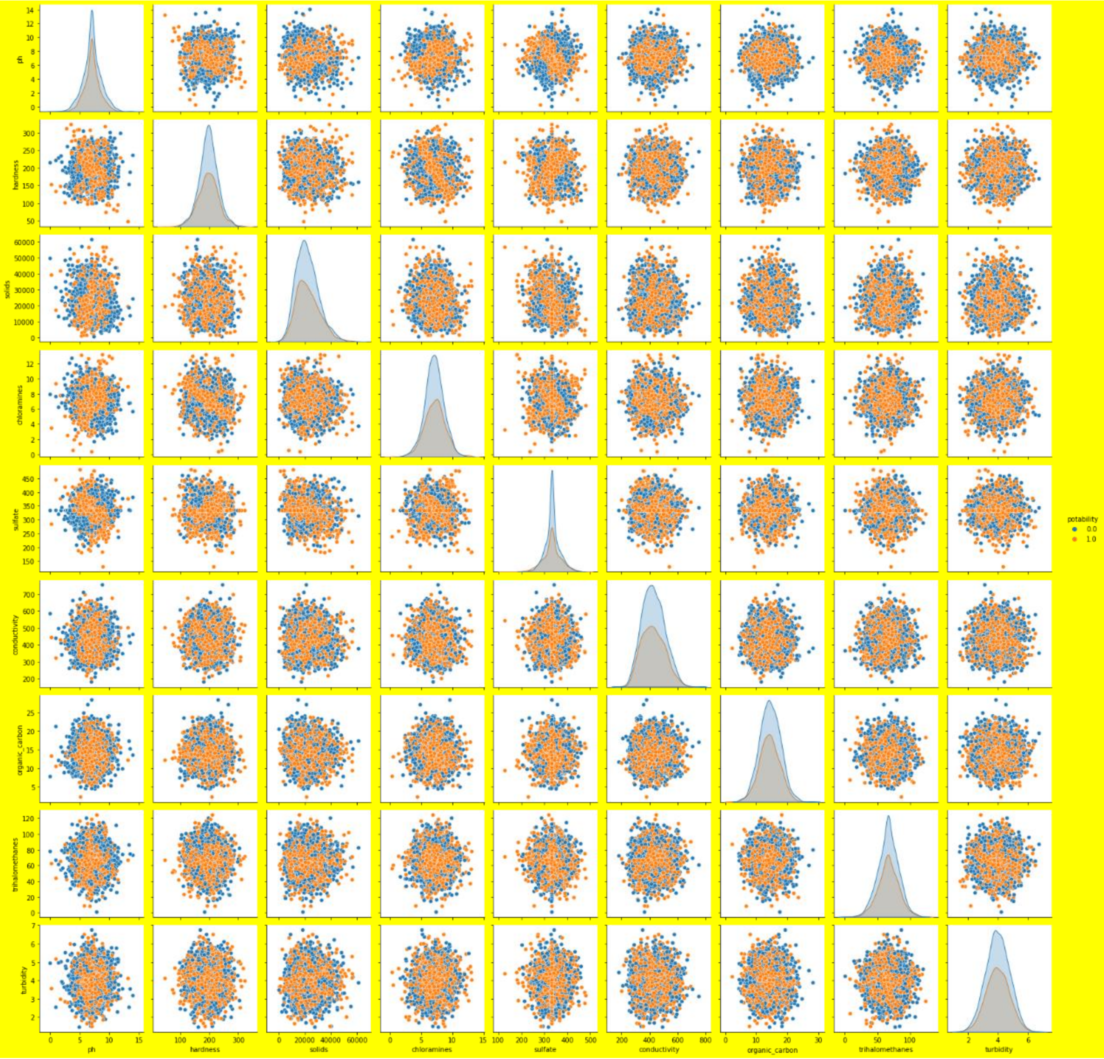




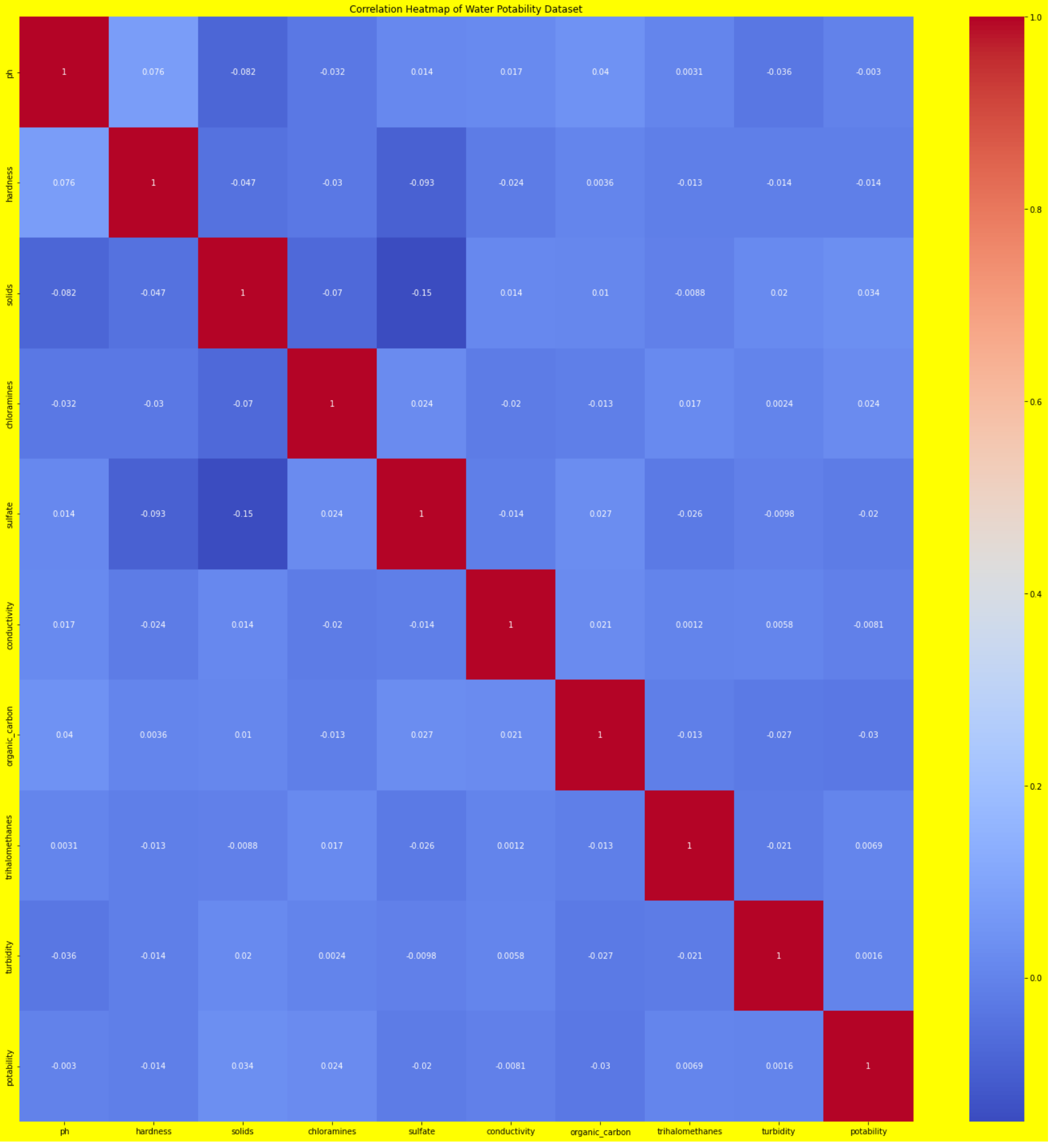




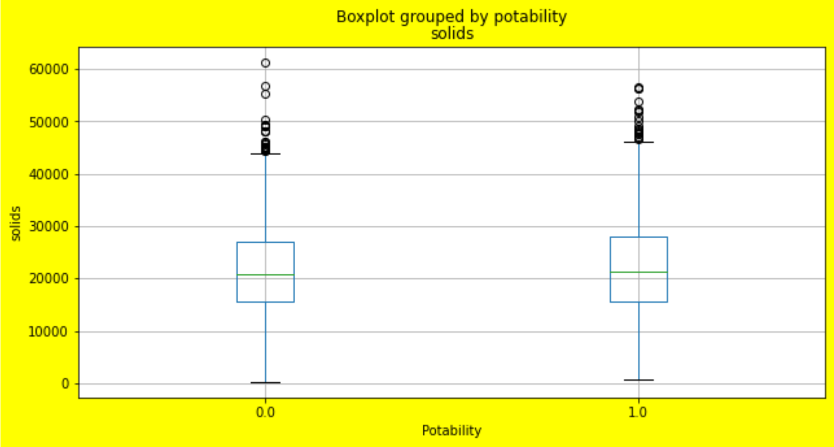
The imputed median values for pH, Sulfate, and Trihalomethanes are illustrated in the histograms for respective visualizations. There are spikes near the median value due to the extra values. Removing the missing values was also explored during the preliminary analysis, however this drastically reduced the number of records within the dataset. A pairplot was useful for understanding the relationships between each variable.



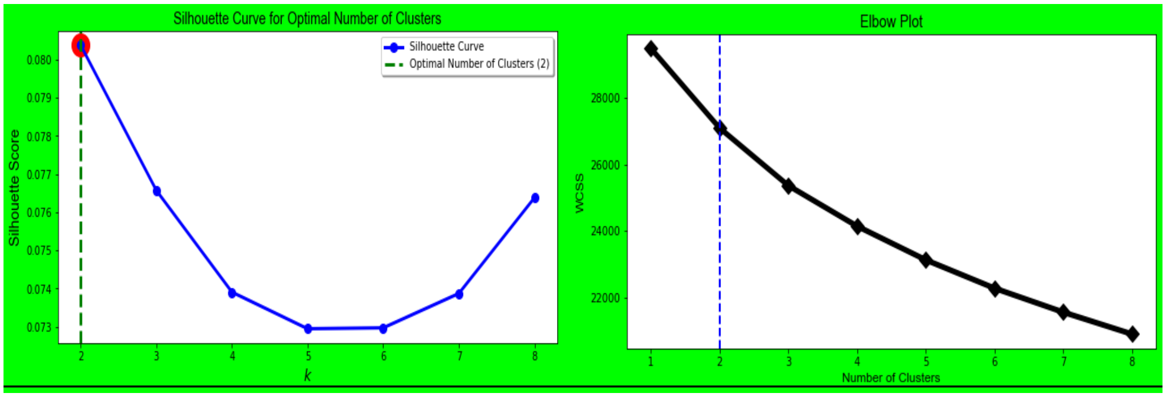
The pairplot shows potable (orange) and non-potable (blue) data points within each feature on a scatterplot or histogram. This chart was useful for trying to understand initial variable relationships. Unfortunately, there were not many useful patterns able to be identified. As a result, I turned to a correlation heatmap to understand which features were most correlated with potability. This is illustrated in the heatmap below:



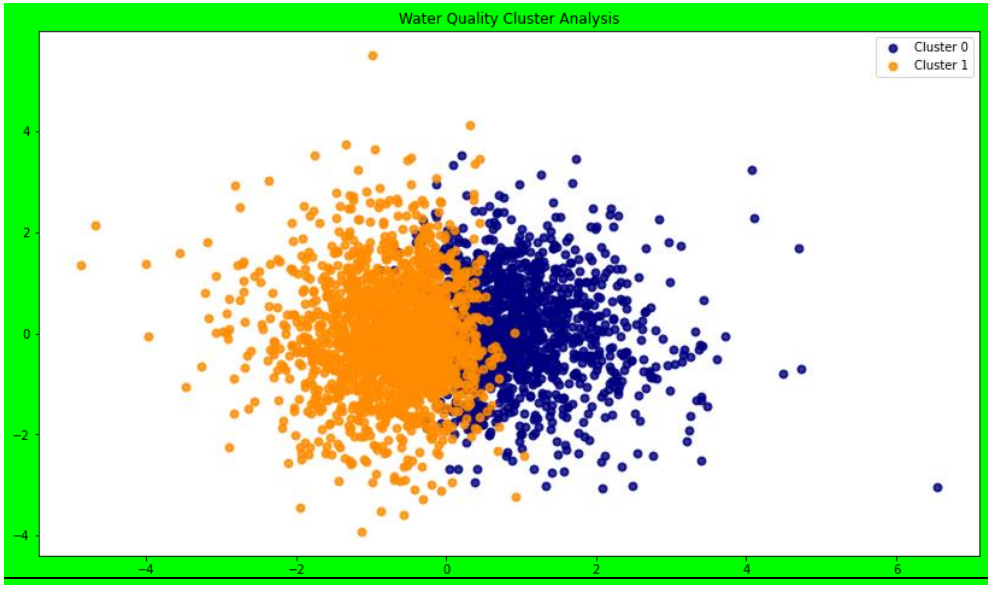
The heatmap helped confirm there are no variables highly correlated with water potability. The variable with the highest correlation was total solids. Boxplots were then utilized as another method of comparing the distributions for the features with respect to potability. The visualization below represents one of the boxplots for the most correlated feature, solids, with potability.



As seen in the boxplot, there is not a major shift for the feature distributions based on water potability. I am leaving the preliminary analysis visualization section in with Milestone 4 to ensure these findings are included in the final project. However, the cluster analysis will be removed since it is not value added as previously mentioned. ~~Lastly, a cluster analysis was performed with K-Means to better understand the dataset. The target feature, potability, was removed from the training data. An Elbow Plot and Silhouette Curve were utilized to confirm two clusters for the dataset. The plots are shown below for reference.~~



~~With 2 clusters identified for the K-Means model, the scatterplot below shows the cluster analysis after transforming the data using PCA. This image helped segregate the data into potable and non-potable water as seen below:~~



~~These visualizations and steps taken in the preliminary analysis were beneficial for understanding whether the project questions could be answered. Thus far, there are a few concerns still pending such as model accuracy and feature correlation with potability. However, it appears the project is still on track and the dataset can be utilized for useful insights to answer the questions previously posed.~~

Milestone 4: Finalizing the Results

Explain your process for prepping the data

The measures taken to prepare the data for modeling were fundamental. The data preparation was broken down into steps for data understanding as well as for predictive modeling groundwork. Upon loading the water potability dataset into a Pandas DataFrame, it was evident that ‘Sulfate’, ‘ph’, and ‘Trihalomethanes’ were missing values. Several options were considered on how to handle these missing values. The mean and median values were compared for each of these features against potable and non-potable categories. The missing values were chosen to be replaced with each feature’s median value after review of how close the median values were for each category. The column names for the DataFrame were also renamed to all lowercase for convenience throughout the analysis. During the Exploratory Data Analysis

(EDA) phase, univariate and bivariate analysis were performed to better understand each feature and the relationships present between the features. For the univariate analysis, the balance within the dataset was noted to be slightly offset in favor of non-potable (61%) compared to potable (49%) records. Each of the numeric features appeared to be normally distributed based on histograms, however it was found that only two (Turbidity and Organic Carbon) of the remaining nine were deemed to be from a normal distribution based on the Shapiro-Wilk test for normality. Q-Q Plots were also utilized to show the appearance of the normal distribution for each numeric feature. The findings from the univariate analysis resulted in no additional data preparation steps. For multivariate analysis, a pairplot was utilized to illustrate the relationships present between the features within the dataset by potability. In addition, a correlation map was generated to see which features were highly correlated with potability. Another method for visualizing the feature relationships was through box plots with respect to the potability category. The box plots showed potential outliers within each feature; however, no outliers were removed to gain an initial understanding of how well the predictive models will perform without data removal. This serves as a potential recommendation to revisit and perform outlier removal with the IQR method for the features that make logical sense. The data preparation steps taken to lay the foundation for the predictive model inputs were focused on next. First, the revised data from the data understanding phase was split into a training (independent variables) and test (dependent variable) datasets. The training set consisted of the nine features other than potability. The test dataset only contained the potability data. Next, the data was split into subsets of training (80%) and testing (20%) to utilize for model construction and assessment. This data was utilized for the Logistic Regression Model. For the remaining models, a standard scalar was utilized on the data in conjunction with a fit and transform for the featured training dataset. The standard scalar was applied to the featured test dataset in addition to a transform (without fit). These were all the steps taken to prepare the data for the predictive models.

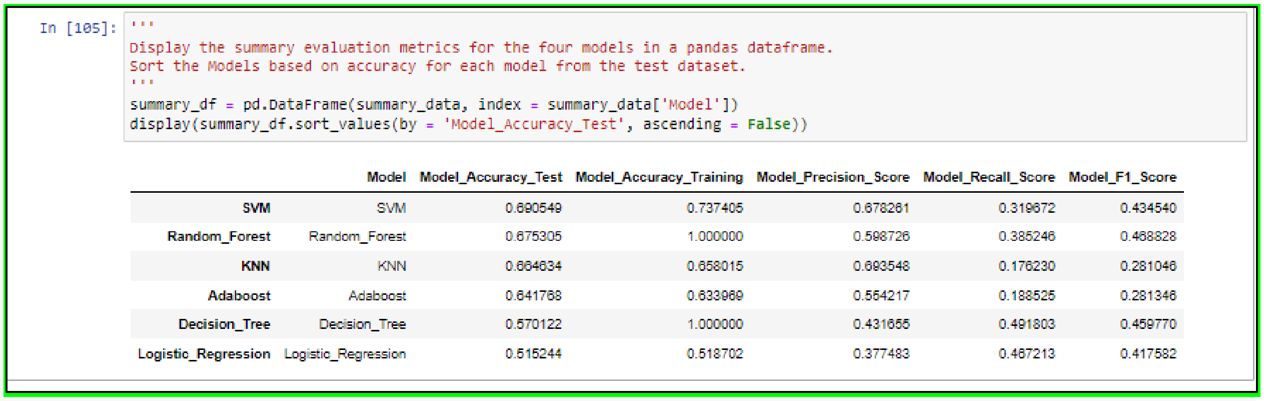
Build and evaluate at least one model

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

Thus far, six predictive models were built and trained. These six models were Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), and AdaBoost. The SVM Model currently had the best performance and the supporting code can be referenced at the end of this document. 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. As previously discussed, the metrics used to evaluate the models were accuracy (test and training sets), precision, recall, and F1 scores. The results are shown in the next section.

Interpret your results

The results from the six models are shown in the table below:



```
In [105]: '''
          Display the summary evaluation metrics for the four models in a pandas dataframe.
          Sort the Models based on accuracy for each model from the test dataset.
          '''
          summary_df = pd.DataFrame(summary_data, index = summary_data['Model'])
          display(summary_df.sort_values(by = 'Model_Accuracy_Test', ascending = False))
```

	Model	Model_Accuracy_Test	Model_Accuracy_Training	Model_Precision_Score	Model_Recall_Score	Model_F1_Score
SVM	SVM	0.690549	0.737405	0.678261	0.319672	0.434540
Random_Forest	Random_Forest	0.675305	1.000000	0.598726	0.385246	0.468828
KNN	KNN	0.664634	0.658015	0.693548	0.176230	0.281046
Adaboost	Adaboost	0.641768	0.633969	0.554217	0.188525	0.281346
Decision_Tree	Decision_Tree	0.570122	1.000000	0.431655	0.491803	0.459770
Logistic_Regression	Logistic_Regression	0.515244	0.518702	0.377483	0.467213	0.417582

As previously learned about in this program, 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 my 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. The conclusion and recommendations are formulated in the next section.

Begin to formulate a conclusion/recommendations

There are a few recommendations to improve the accuracy and learn which models are most repeatable. The first recommendation is to try and remove outliers using the IQR method for features that make logical sense. For example, it does not make sense to perform the outlier removal on ph since this data is constrained on a scale from 0-14. Second, explore how the models perform against cross-validation. This may provide useful insights into how repeatable a model is in terms of accuracy (or one of the other metrics considered in this analysis). Third, it would be beneficial to investigate hyperparameter tuning to improve the performance of the model(s). The questions posed for this analysis are still capable of being answered. Based on the Exploratory Data Analysis (EDA), this dataset contains 61% non-potable and 39% as potable. The features appear to be normally distributed based on the histogram visualizations, however only two of the nine features are. None of the features within this dataset are highly (or even mediumly) correlated to water potability. The features most correlated with potability are Total Solids (positive) and Organic Carbon (negative). The model with the highest accuracy is currently SVM, but this could change based on the recommendations previously listed in this section. The metrics for the best performing model are outlined in the previous section. The models constructed and trained so far are not ready for deployment. I'll work on the recommendations to improve the models and then revisit whether the model(s) are ready to be deployed.

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

References

Kadiwal, Aditya. (2021, April 25). Water Quality: Drinking water potability. *Kaggle*. Water Quality | Kaggle

BOSAQ. (2020, March 18). Everything You Need To Know About Water Resources. *BOSAQ Blog.* Everything you need to know about water resources | Bosaq

Misachi, John. (2018, February). What Percentage of Earth’s Water Is Drinkable? *WorldAtlas*. What Percentage of the Earth's Water Is Drinkable? - WorldAtlas

World Health Organization. (2022, March 21). Drinking-Water. *World Health Organization Newsroom.* Drinking-water (who.int)

World Health Organization. (2022, March 21). Drinking-water quality guidelines. *Water Sanitation and Health*. Water Sanitation and Health (who.int)

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

Milestone 2 Requirements:

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?

Data selection and your project proposal are due this week. While you might decide to add additional data sources as the project progresses, you should have a good idea of your initial dataset by this milestone.

Milestone 2 should include the information outlined in the introduction above. Additional items to address are the following:

- What types of model or models do you plan to use and why?
- How do you plan to evaluate your results?
- What do you hope to learn?
- Assess any risks with your proposal.
- Identify a contingency plan if your original project plan does not work out.
- Include anything else you believe is important.

The proposal should be a minimum of three pages, double-spaced. You should treat this proposal as the start of your final project paper submission. But also remember this is only the initial proposal. Your findings might take you in a different direction for the final submission.

Please submit Milestone 2 in Blackboard under the group submission link.

This should be submitted through the group assignment submission regardless if it is an independent project or multi-person group.

Also, post your Milestone 2 in your Teams project folder for peer reviews.

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

Milestone 3 Requirements

Milestone 3 should include all the information from Milestone 2, updated if necessary. Additionally, you should start to explore your data. Items to address include the following:

- Will I be able to answer the questions I want to answer with the data I have?
- What visualizations are especially useful for explaining my data?
- Do I need to adjust the data and/or driving questions?
- Do I need to adjust my model/evaluation choices?
- Are my original expectations still reasonable?

Please submit Milestone 3 in Blackboard under the group submission link.
This should be submitted through the group assignment submission regardless if it is an independent project or multi-person group.
Also, post your Milestone 3 in your Teams project folder for peer reviews.

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

Milestone 4 Requirements

In Milestone 4, most of the technical work for the project should be done. You should include the information from Milestone 3 and address the following additional items:

- Explain your process for prepping the data
- Build and evaluate at least one model
- Interpret your results
- Begin to formulate a conclusion/recommendations

Please submit Milestone 4 in Blackboard under the group submission link.
This should be submitted through the group assignment submission regardless if it is an independent project or multi-person group.
Also, post your Milestone 4 in your Teams project folder for peer reviews.

Predictive Model Code from Jupyter Notebook (SVM)

Support Vector Machine (SVM) Model

In [80]:

```
'''
Create the SVM Classifier.
'''
svm = SVC()
```

In [81]:

```
'''
Fit the SVM Classifier on the training dataset.
'''
svm_classifier = svm.fit(X_train, y_train )
svm_classifier
```

Out[81]:

```
SVC()
```

In [82]:

```
'''
Obtain the y_predictions for the SVM Classifier.
'''
y_predictions_svm = svm.predict(X_test)
```

In [83]:

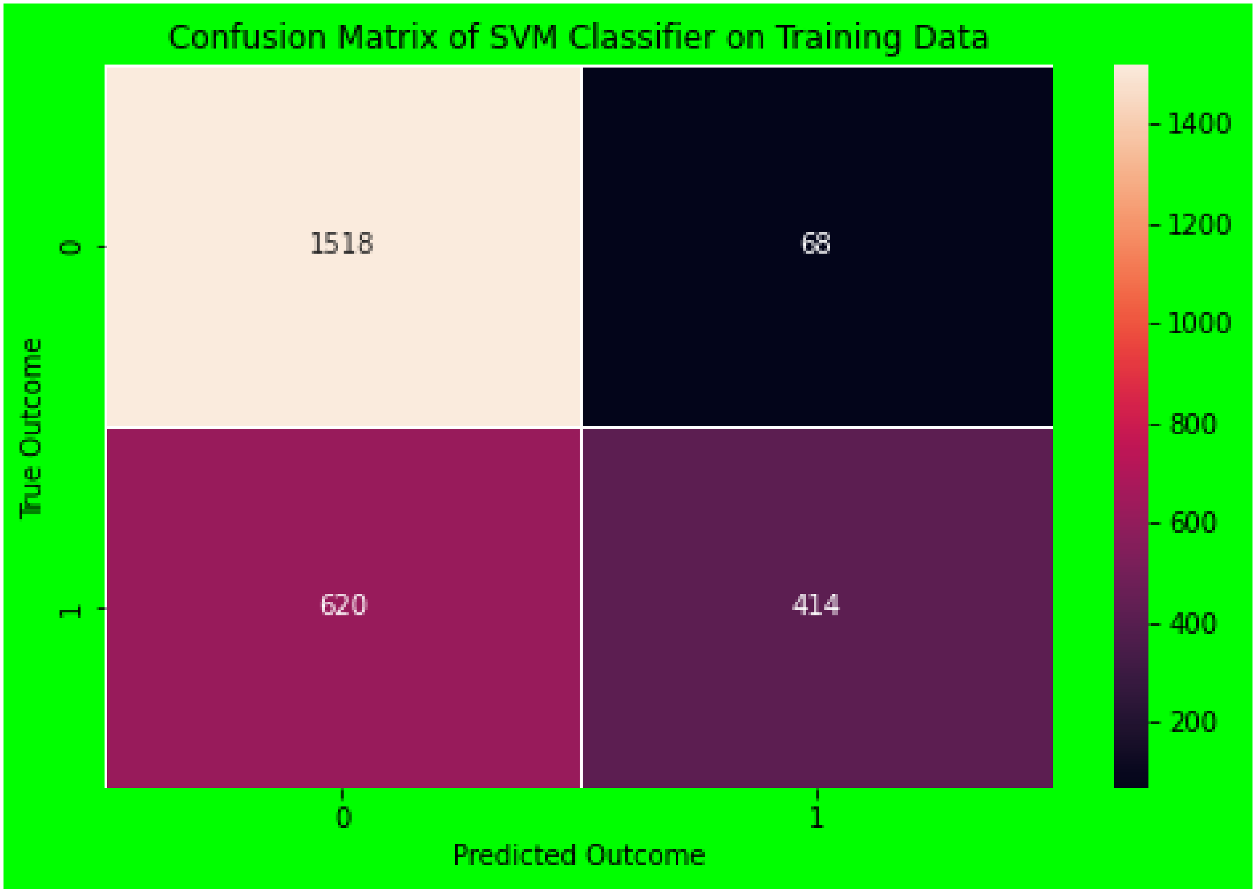
```
'''
Obtain the y_prediction probabilities for each record in the training dataset.
'''
y_predictions_svm_train = svm.predict(X_train)
```

In [84]:

```
'''
Generate a Confusion Matrix for the Support Vector Machine Model based on the training dataset.
'''
cm_svm_train = confusion_matrix(y_train, y_predictions_svm_train)
```

In [85]:

```
'''
Plot the confusion matrix so it is clearly labelled and illustrated.
'''
f, ax = plt.subplots(figsize = (8,5))
sns.heatmap(cm_svm_train, annot = True, linewidths = 0.5, fmt = ".0f", ax = ax)
plt.xlabel('Predicted Outcome')
plt.ylabel('True Outcome')
plt.title('Confusion Matrix of SVM Classifier on Training Data')
plt.show()
```



In [86]:

```
'''
Obtain the y_predictions for the SVM classifier.
'''
y_predictions_svm_test = svm.predict(X_test)
```

In [87]:

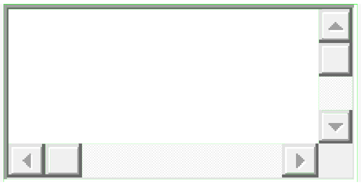
```
'''
Generate a confusion matrix based on the test data set.
'''
```

```
cm_svm = confusion_matrix(y_test, y_predictions_svm_test)
cm_svm
```

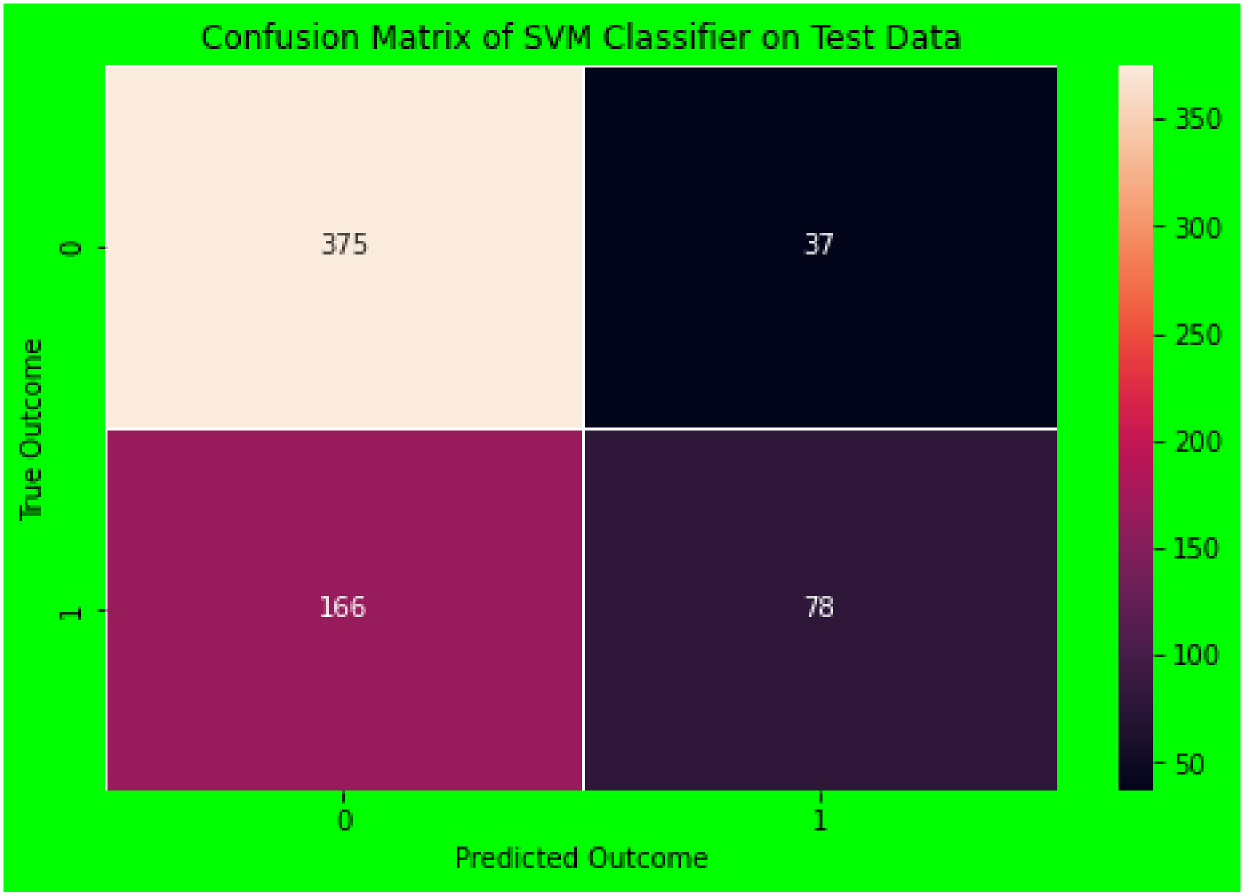
Out[87]:

```
array([[375,  37],
       [166,  78]], dtype=int64)
```

In [88]:



```
'''
Plot the confusion matrix so it is clearly labelled and illustrated.
'''
f, ax = plt.subplots(figsize = (8,5))
sns.heatmap(cm_svm, annot = True, linewidths = 0.5, fmt = ".0f", ax = ax)
plt.xlabel('Predicted Outcome')
plt.ylabel('True Outcome')
plt.title('Confusion Matrix of SVM Classifier on Test Data')
plt.show()
```



In [89]:

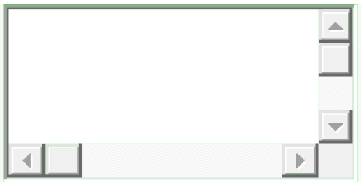
```
'''
Show the classification report for the test data.
'''
print(classification_report(y_test ,y_predictions_svm_test))
```

```
              precision    recall  f1-score   support

           0       0.69      0.91      0.79       412
           1       0.68      0.32      0.43       244

    accuracy                           0.69       656
   macro avg       0.69      0.61      0.61       656
weighted avg       0.69      0.69      0.66       656
```

In [90]:



```
'''
Calculate the accuracy for the model based on training and test data. Also, report the
Precision, Recall, and F1 score for the model predications against the test data.
'''
accuracy_svm_train = accuracy_score(y_train, y_predictions_svm_train)
accuracy_svm_test = accuracy_score(y_test, y_predictions_svm_test)
precision_svm = precision_score(y_test, y_predictions_svm_test)
recall_svm = recall_score(y_test, y_predictions_svm_test)
f1_svm = f1_score(y_test, y_predictions_svm_test)
print("Accuracy of SVM Model on training data is:{}".format(accuracy_svm_train))
print("Accuracy of SVM Model on testing data is:{}".format(accuracy_svm_test))
print("Precision of SVM Model on testing data is:{}".format(precision_svm))
print("Recall of SVM Model on testing data is:{}".format(recall_svm))
print("F1 Score of SVM Model on testing data is:{}".format(f1_svm))
```

```
Accuracy of SVM Model on training data is:0.7374045801526717
Accuracy of SVM Model on testing data is:0.6905487804878049
Precision of SVM Model on testing data is:0.6782608695652174
Recall of SVM Model on testing data is:0.319672131147541
F1 Score of SVM Model on testing data is:0.43454038997214484
```

DCS 630 Predictive Analytics (DSC630-T302 2231-1)
Bellevue University
Course Project: Prediction of Water Quality
Milestone 4: Finalizing the Results
Author: Jake Meyer
Date: 10/30/2022

Water Quality Dataset

water_potability.csv

CSV Attachment -