Identifying Parameters in Water Potability Analysis Through Machine Learning

Molly Ho

9.22.22

Abstract:

Machine learning has become a rapid and prevalent tool in the growth of environmental science. With the onset of emerging data, data analysts have relied on ML models to reveal hidden trends. In past papers, machine learning has been brought up countless times as an innovative method to rapidly sort through data in big groups. However, developing an artificial neural network model is a sparsely used tool in research due to its lack of human interface, risk factors, and not enough experimentation in real-life applications. In this research, we implemented multiple ANN models to identify factors in a water's chemical levels to identify which has the largest weight or bias in determining whether the water is potable or not. Subsequently, we created a model that uses these parameters to analyze and predict whether a water source will be safe to drink. Water is an essential resource for all beings on Earth, and as the climate changes, the water levels rise, and along comes pollutants to once safe drinking water. By comparing the accuracy between both classifier and regressor ML models, each model was hyper-parameterized to develop a better accuracy, as well as identify the weight each parameter carries towards the potability of the source. The results demonstrate a promising conclusion that the number of solids affects the potability versus other chemical levels. That said, there was not a strong, distinct correlation and every factor plays an important part. The accuracy of these models returned a sub-par 64.77% accuracy with a 0.53 precision, a 0.46 recall score, and a 0.49 f1 score. In conclusion, while the accuracy of the models is ineffective to be used just yet on a bigger scale, they offer a stepping stone in the development of machine learning models in the environmental science fields, resourcing digital infrastructures to analyze and dissect the Earth and its trajectory through climate changes.

1. Introduction

Predicting whether the water is potable or not can be helpful for people who are reliant on bodies of water and redirect them to safer options. It will also be beneficial to apply the algorithm to other places where it is expensive and inefficient to send people out and collect water samples. Over the past couple of decades, researchers have often commented on the lack of funding as a source of error when it comes to data analysis and the accuracy of the research. In the article "Effects of Land Use and Water Quality on Greenhouse Gas Emissions from an Urban River System", greenhouse gasses, air pollution, and raw sewage are among the most significant attributes of dangerous contaminants being leached into bodies of water that are used for drinking and cleaning. Developing countries reliant on these water sources have begun to face disease and drought because climate change has caused the cleanliness to decline. In current statistics, 80% of illnesses in developing countries are linked to water-borne diseases, and roughly 28% of a household has been affected by the illnesses. Identifying major factors that affect the safety of the water will allow better counteractive measures and filtering methods so places can comfortably utilize nearby water sources and reduce the dire clean water issue. The problem faced in the research is an example of supervised learning of mostly regression-based numerical data (except the classification between safe and unsafe to drink). The output of the data is multiple graphs visualizing the accuracy, layout, and trends of the dataset run through multiple ML models. By conducting different baseline ANN models, the research will help create a comprehensible understanding of how to approach this problem in a more extensive scope.

3. Background

Machine Learning models were incorporated into real-life applications in the 1990s, as ML and AI began to become independent fields of study. In Dawood et al. 's research, “Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks”, their work focused on the water quality in urban areas. They created multiple ML models, along with flow charts to examine where the construction of the pipes and urban planning may have caused the variance in water quality. Ultimately, their results demonstrated the materials between lines that have led to the leaching of various metal contaminants in the water. Simultaneously, they discovered the more complex the pipe system is, the more risk for high levels of dangerous contaminants, as the water travels through a longer pip system. Their research offered insight into the risk analysis for urban water systems, but they risked the performance level of their ANN model due to their multi-variable inputs. Furthermore, the dataset used in our research focused on chemical levels, and less on where each sample was collected, so environmental effects are inconclusive in our research. Another article that was used was Rizal et al.’s “Water Quality Predictive Analytics Using an Artificial Neural Network with a Graphical User Interface.” In this research, they conducted a predictive analysis of water potability, centralized in bodies of water in Malaysia. They also implemented a Graphical User Interface (GUI) to assess better the parameters their ANN models used. Their models were highly accurate and helped establish a comprehensible, standard layout of the models that our research used. Within the field of environmental science, among countless other real-world applications, machine learning has become a key role in efficiently and effectively finding trends naked to the human eye, and sorting big amounts of data at a quick, errorless pace. Artificial Intelligence and Machine learning, while not completely autonomical, have made large advances in the past 30 years, while climate change issues are arising quicker as well. Hopefully, with the help of AI and ML models in previous literatures, key, comprehensible factors can be identified and demonstrate solutions that will reduce our carbon footprint and pose new, sustainable methods that starkly decrease our intense use of the dwindling natural resources.

4. Dataset

The dataset used for the project contained different chemical levels of the bodies of water, and whether it was safe to drink or not. The data was entirely numerical, with one classification column for potability: 0 = unsafe and 1 = safe. The dataset is linked here:

https://www.kaggle.com/datasets/adityakadiwal/water-potability

To split the dataset into training and testing data, we split it 60% training versus 40% testing. The training set is bigger because, for the models, we wanted each one to find as many possible trends as possible in the data to build off of. These models are all cases of supervised learning, where the models are trained with data from the same pool the testing set is, thus being able to detail its accuracy based on the classifications of potability it predicted and the true values. Additionally, the data was already fairly preprocessed, but any data point with null values in any of the columns was removed to clean it further. Additionally when completing a principal component analysis of the data, even if the pH was above or below the safe drinking levels, in comparison to the other chemicals, some were still potable. To be safe and legitimize the model, an additional restraint on the pH data was created, and removed any data points above or below 6.5 - 8.5 on the scale. The final amount of data points after the cleaning totaled 2786 entries. While cleaning the data, we did a principal component analysis (PCA) to see if there were any

Below are 4 scatterplots of the data compared between two features in the dataset, with the different colors corresponding to the possibilities of each data point. The line through the middle shows the estimated slope of the respective levels where the potability is 0. While there is a small correlation, the slope and the deviation of the data are a big difference. Furthermore, with the code programmed to visualize this data, there was an error that resulted in being unable to find the exact equation of the line. With more research and time, we could have solved this issue to better display

A linear regression model that predicts potability (0, 1) from all other features.
Regression lines that are predicting:

- Solids level from PH for data points where the potability = 0
- Solids from pH levels for data points where the potability = 0

- Solids from sulfate level for data points where the potability = 0
- Solids from Trihalomethane levels for data points where the potability = 0
- Solids from Organic Carbon levels for data points where the potability = 0

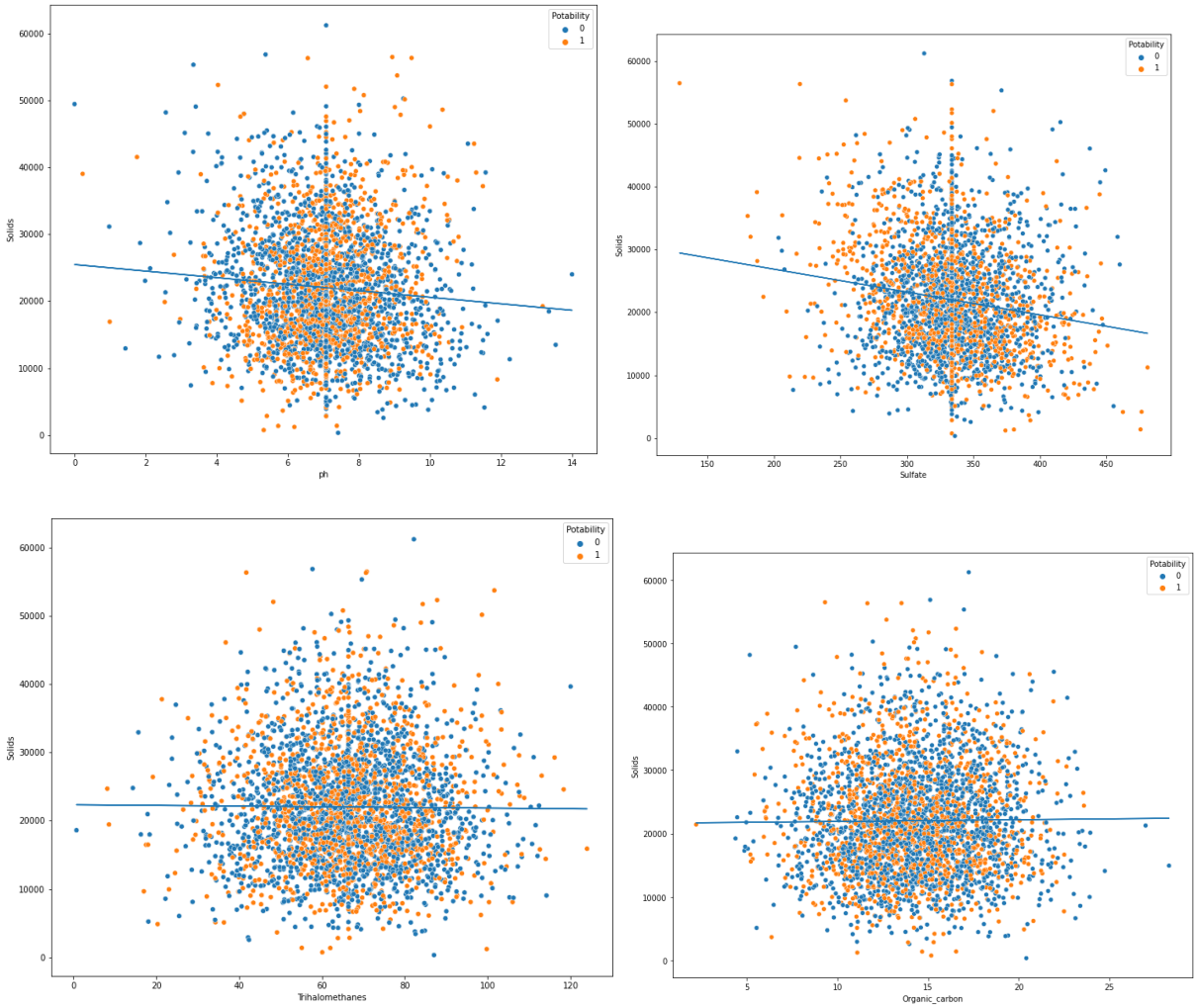


Figure 1) a, b, c, d: different graphs that compare 2 features in the chemical levels of the water data contrasted by their potability

While at the moment, the potable vs. not potable points are evenly interspersed, the model may catch onto trends in the dataset.

5. Methodology/Models

The dataset had 3216 entries, with a couple of null values in columns, so in order to clean the data and make sure the model will face fewer errors, any incomplete data was removed. Additionally, because the pH levels for safe drinking water should be between 6.5 and 8.5 on the pH scale, any data points above or below were taken out as well. Next, we fitted the data to four sklearn models: Linear Regression, DecisionTree Classifier, DecisionTree Regressor, and SVM. For every model, the data was split into 60% training data and 40% testing data. We imported the Pandas, Numpy, Seaborn, and Sklearn libraries into our code.

Next, we followed a similar code to create a Decision Tree Classifier model. The DecisionTree branches off in each new layer of the model, creating a complex multi-layered model.
To try to create better accuracy, we conducted a Gridsearch for the criterion of the Decision Tree Classifier and used hyperparameter tuning to tighten up the models. We hyperparameterized the max_depth levels, as well as the criteria in the DTC model, and subsequently ran it through code to find the best values for each. The *gini* criterion and *entropy* were run through a function to analyze which of the max depths–layers of nodes in a decision tree– would return the best accuracy. Our *Random Forest Regressor* model was put through a Random Search in order to tune it to improve its accuracy. It returned the best parameters and the values for each parameter to have it run at its peak performance. For Linear Regression, we reviewed multiple graphs of different combinations to see which had trends and which were not so closely related.

6. Results and Discussion

The results are a promising start in applying Machine Learning methods to environmental data, however, the results from the multiple models displayed subpar accuracies. After hyper parameterizing the Random Forest Regressor Model, it had the best accuracy of all the models, still at a mediocre but improving 64.77% accuracy.

These were the parameters for the Decision Tree Classifier Model:

class `sklearn.tree.`**`DecisionTreeClassifier`**(**, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0*) [source]

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import precision_score, recall_score, f1_score

dt_class = DecisionTreeClassifier(random_state = 0)
dt_class.fit(X_train, y_train)

predictions = dt_class.predict(X_test)
dtc_score = accuracy_score(y_test, predictions)

dtc_precision = precision_score(y_test, predictions)
dtc_recall = recall_score(y_test, predictions)
dtc_f1 = f1_score(y_test, predictions)


print('Accuracy: {:.2}'.format(dtc_score))
print(dtc_precision, dtc_recall, dtc_f1)
```

```
Accuracy: 0.59
0.5346938775510204 0.4628975265017668 0.49621212121212127
```

Below are the correlating results:

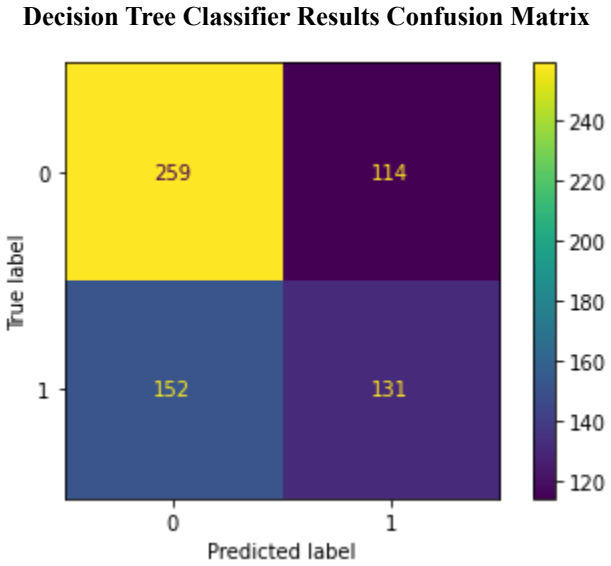


Figure 2: Confusion Matrix of DTC model

On the left is the confusion matrix for the Decision Tree Classifier model that visualizes the errors in a classification model outputted against the data correctly attributed. In the top left corner are the true negative data points that state the amount of the testing set that the model deduced to be unsafe (0), which actually were 0. On the bottom right are the true positive points, which were guessed to be safe (1) and were correct. To the right, there is a relative frequency table of the data to further display the As noted above, the model was only ~60% accurate in correctly identifying the testing set. The results, while the majority correct, are poor for a real-life application of the model. After tuning the model's parameters, unfortunately, the accuracy did not increase and decreased as a result of overfitting the model.

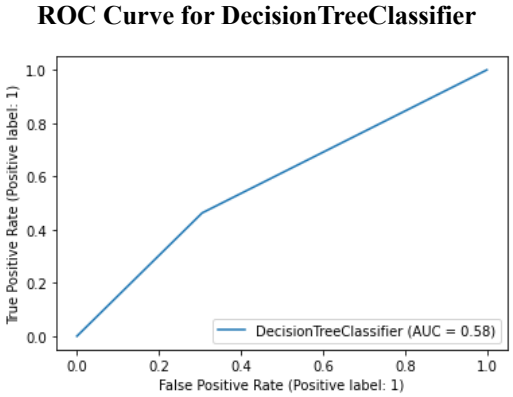
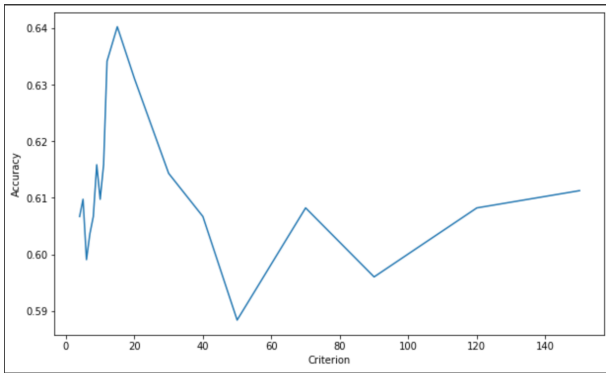


Figure 3: ROC curve of previous model

The ROC Curve (Receiving Operating Characteristic) of our DTC model further demonstrates how weak the classifier model is for the data. Since the curve is close to linear, and the area under the curve (AUC) is 0.58–the lowest possible AUC is 0.5–the DTC model created needs better hyper parametrizing and tuning.

Accuracy vs. max_depth *Gini*



Accuracy vs. max_depth *Entropy*

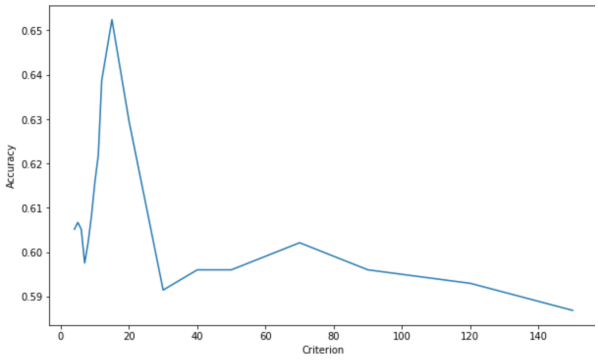


Figure 4) a, b

As seen above, the accuracy peaks around 64.77% for both “gini" and “entropy” near 15 layers. While it is an improvement, a ~65% accuracy is not sufficient for further applications because having a model only able to guess ⅔ of the data correctly is not useful for real-world applications, and could lead to serious consequences if it’s incorrect. In our case, it could falsely predict a certain body of water is safe to drink and then lead to illnesses as a result of the water not actually being safe, but contaminated with harmful substances. Another opportunity to further improve the model is to tune more parameters, other than the 2 main features we hyper parametrized. Another prospective idea is to create ensemble models with ones like Decision Tree Classifier, to enhance its performance and overall accuracy by utilizing the best of each model, and tuning accordingly.

7. Conclusion

The ultimate goal for this project was to identify and create multiple ML models that would help classify water sources to identify those that are safe to drink and consume quickly. In the end, the research was a semi-success, because we developed working models, albeit with subpar performance accuracies. One of the reasons that the model had a poor performance was the dataset used. While researching past projects using this dataset, all the algorithms and ANNs built had a similar accuracy score. Additionally, each water source is unidentified, so the environmental standings and climate in which the water was sourced could have aided in increasing the model's accuracy. It would have also been ideal to track and subsequently group the water sources by either location and range, or climate/similar ecosystems because then new trends and patterns might have emerged. Another possible reason for the poor model performance could have been our code itself. Despite everything working and visualizing the data how we wanted, our models are still fairly baseline, with minimal hyperparameter tunings.

In the future, it would be ideal to create ensemble models of the Decision Tree Classifier and Random Forest Regressor models, because those were our highest performing models. Another method to work by is applying the algorithms to a different dataset to find any other holes or inconsistencies that were harder to spot given our data. Classifier Models and Regression models were comparably similar in their performance output, but working on other ANNs such as Lasso Regression, Bayesian Linear Regression, or K-Nearest Neighbor might prove to be better.

Our results posed a necessary addition to the application of Machine Learning algorithms in the area of Environmental Science. While still functional, our research demonstrates the efficiency at which data can be processed, synthesized, and returned at a better caliber than humans. Hopefully, our research stands as a baseline model to stem forth the development of clean water solutions, collect different sources, and quickly determine its metrics for its best possible use. Our hope is that we will be able to enhance our models, and later incorporate computer vision to examine the climate, shape, and body from satellite images and determine certain metrics that can identify its portability from just an image.

Acknowledgments

Thank you so much to Sharon Chen, who guided and mentored me in how to write a research paper, develop the models, and then gave a quick yet wonderful crash course in data science. I will carry on all your advice with me! Thank you to Inspirit AI for giving me the opportunity to work with them as well, and offer countless resources for my advancement of AI, ML, and presentation skills.

References

Asadnia, M., & Beheshti, A. (Eds.). (n.d.). *Artificial Intelligence and Data Science in Environmental Sensing*. https://doi.org/10.1016/C2020-0-03497-3

Dawood et al., T. (n.d.). Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks. *Elsevier*. https://doi.org/10.1016/j.jclepro.2020.125266

Goliber, S., Black, T., Catania, G., Lea, J. M., Olsen, H., Cheng, D., Bevan, S., Bjørk, A., Bunce, C., Brough, S., Carr, J. R., Cowton, T., Gardner, A., Fahrner, D., Hill, E., Joughin, I., Korsgaard, N. J., Luckman, A., Moon, T., ...Zhang, E. (2022). TermPicks: a century of Greenland glacier terminus data for use in scientific and machine learning applications. *The Cryosphere*, *16*(8), 3215. https://link.gale.com/apps/doc/A713733059/EAIM?u=hill25409&sid=bookmark-EAIM&xid=64a16229

Ho et al, L. (n.d.). Effects of land use and water quality on greenhouse gas emissions from an urban river system. *European Geosciences Union: Biogeosciences*. https://doi.org/10.5194/bg-2020-311

Lu, H., & Ma, X. (n.d.). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Elsevier*. https://doi.org/10.1016/j.chemosphere.2020.126169

Rizal et al, N. (n.d.). Water Quality Predictive Analytics Using an Artificial Neural Network with a Graphical User Interface. *MDPI*. https://doi.org/10.3390/w14081221

Schley, T. (2021, May 21). Artificial intelligence predicts river water quality with weather data. *Penn State Research*. https://www.psu.edu/news/research/story/artificial-intelligence-predicts-river-water-quality-weather-data/

Zhong et al., S. (n.d.). Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *American Chemical Society*. https://pubs.acs.org/doi/full/10.1021/acs.est.1c01339

9. Appendix

Features in the Dataset (taken from the description in the dataset):

pH: The indicator of acidic or alkaline condition of water status.

Hardness: Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water.

Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

Sulfates: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

Solids: Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced an unwanted taste and diluted color in the appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized.

Chloramines: Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

Conductivity: Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceed 400 μS/cm.

Organic Carbon: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to the US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is used for treatment.

Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

Relative Frequency Table for DTC Model:

		Model's Predictions		
		0 (Not Potable)	1 (Potable)	
Testing Set	0 (Not Potable)	**39.48%**	17.38%	56.86%
	1 (Potable)	23.17%	**19.97%**	43.14%
		62.65%	37.35%	100.00%
	Accuracy of DTC:		**59.45%**	