Dirt water report

1. Introduction

Drinking water is a basic necessity for all humans. We will evaluate the physical properties of samples from water bodies and create a model to classify them as potable or non-potable. Water quality is relevant for people living close to those water bodies to evaluate their availability of water or companies and the government responsible for providing drinkable water to the population. This information can also be used as the water quality indicator to health researchers trying to analyze the effects of the water quality on health issues and the well-being of the patients.

The objective is to check the data quality of our dataset and test different classifiers, sampling techniques, and scaling methods to find the most accurate classifier for this data. This data is available on the public data repository Kaggle at https://www.kaggle.com/adityakadiwal/water-potability/activity.

2. Exploratory Data Analysis

The original data comprises the target variable: potable and non-potable, and water's physical properties are presented in nine (9) numeric features such as pH, hardness, chloramines, etc. Thus, we started with a total of 3276 entries representing different water bodies.

Looking for missing values, we had Sulfate with 781 null entries (23.8%), pH 491 (15.0%), and Trihalomethanes 162 (4.95%). Fig. 1 shows all parameters distributions following normal distributions. Potable and non-potable distributions are overlapping and have the same means and interquartile distances. Outliers are present but not in a significant number so that we won't worry about them.

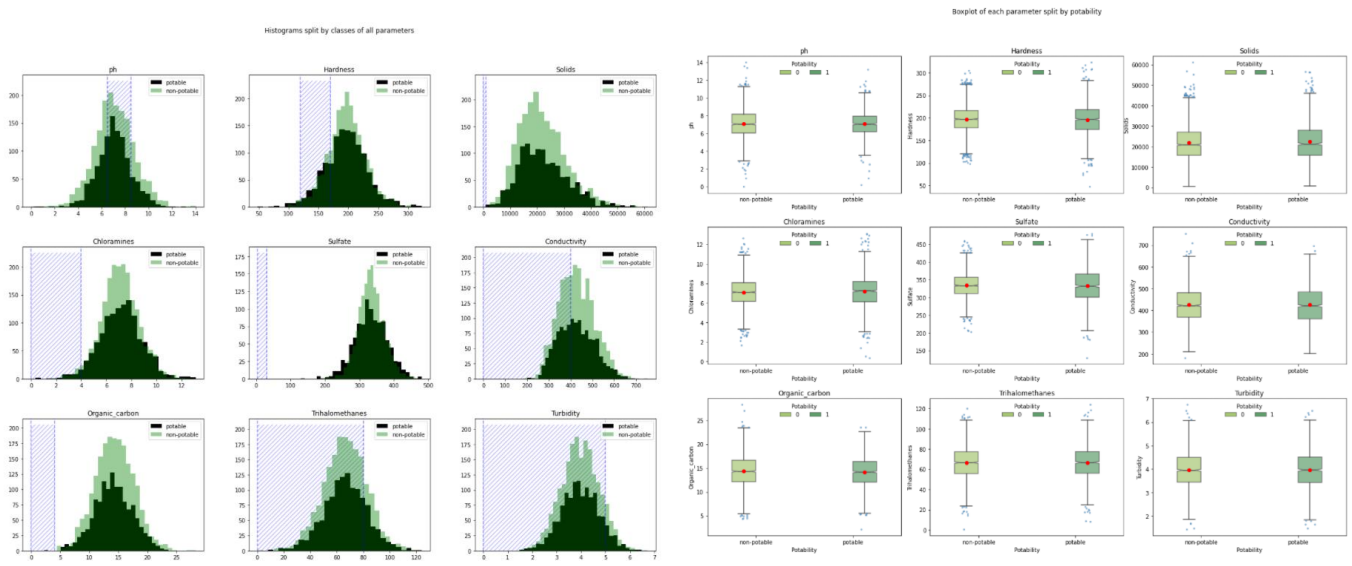

Figure 1 – Left: histogram comparing the frequency of potable (dark green) and non-potable (light green) for each feature and the range considered potable hatched in blue. Right: boxplot of the parameters notice the presence of few outliers overall.

Analyzing the correlation among the parameters (see Fig. 2), we didn't find any relevant correlation even for features that should be correlated, and none of them had a remarkable correlation to our target, the potability.

The similar distribution for both classes and the misalignment with the expected values raised concerns about the classification. So we created a comparison of the expected class versus the classification shown on Fig. 2. It would be expected the majority of the entries matching the expected values, predominance of the dark green in the columns. Still, we got a lot of entries classified as non-expected for both classes.

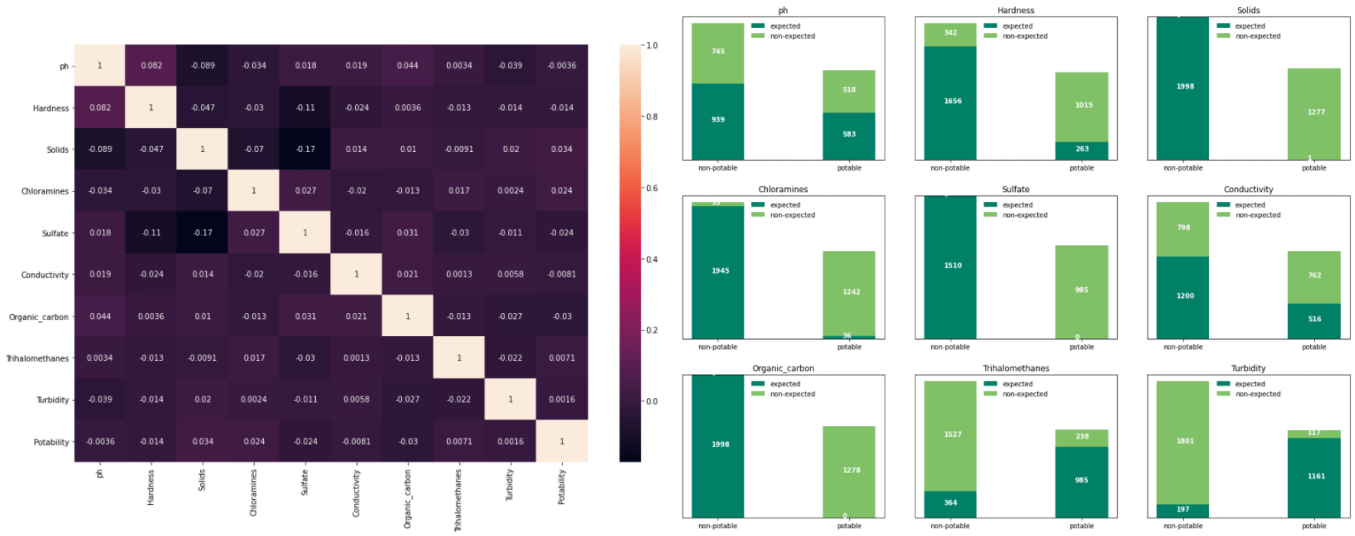

Figure 2 – Left: heatmap of correlation; no remarkable correlation. Right: counting the expected result (dark green) and non-expected (light green) for each classification in the data.

We performed a Principal Component Analysis (PCA) to check if reducing the number of parameters by finding some combination of parameters that could explain the variance.

3. Modeling

We treated the data by removing all entries with missing data: 1265 rows excluded and 2011 for the model data. Fig. 3 shows that excluding the rows with any null feature did not change the potability distribution.

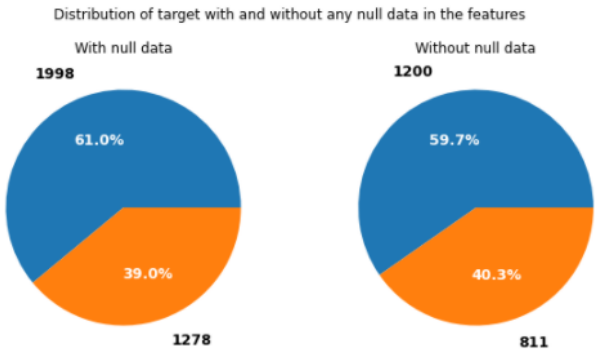

Figure 3 – Pie chart of the potability for data with any null data and after their exclusion.

Data scaling used three (3) methods: StandardScaler, MinMaxScaler, and RobustScaler. For the sampling, were considered the Random and Stratified sampling. The classification models tested are:

1. **Logistic Regression (Logit model)**: models the probability of well-defined classes or events (categorical) that can be binary or linear. Classification is the main application of this model.
2. **K Nearest Neighbours (KNN)**: uses the distance between the values to group them, assuming that the data is similar when close.
3. **Decision Tree**: creates a set of rules to make the decision. These rules are evaluated and based on the decision, and you move to the following node till you reach the final classification (leaf).
4. **Random Forest**: consists of a large number of individual decision trees that operate as an ensemble.
5. **Ada-boost** or **Adaptive Boosting** is an ensemble boosting classifier proposed by Yoav Freund and Robert Schapire in 1996. It combines multiple classifiers to increase the accuracy of classifiers.
6. The **XGBoost** stands for eXtreme Gradient Boosting, a boosting algorithm based on gradient boosted decision trees algorithm.

Each model tested a set of parameters using the different scaling and sampling adopted and selected the best parameters combination. Fig. 4 showcases the scatterplot of the mean accuracy versus the mean run time and the ratio accuracy over time. Considering the accuracy, Random Forest is the winner, but KNN and Logistic Regression are technically tied in the lead when we compare the ratio.

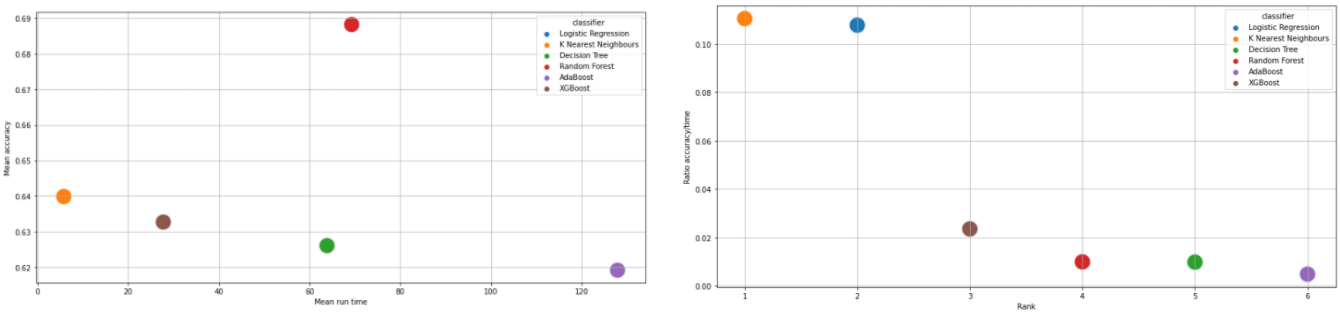

Figure 4 – Left: scatterplot of the mean run time versus the mean accuracy. Right: ratio of accuracy over time.

The evaluation of the sampling technique shows the Random sampling with slightly higher accuracy than the Stratified. The scaling method had no significant impact on the performance.

4. Conclusion

The goal was to classify samples from water bodies as potable or not. We tested six different classifiers exploring a large set of parameters and evaluating three scaling methods and two sampling techniques. Data was split between training and test set (70/30%).

Modeling

The best model considering the accuracy was the **Random Forest** adopting the parameters' min_samples_leaf': 2 and 'n_estimators': 500 achieving an accuracy of 70%. It used the random sampling technique and MinMaxScaler data scaling and took approximately 70 seconds to run.

When we added the processing time to the analysis, **KNN** had the best accuracy over time ratio. However, the extra time to run the Random Forest is small given the gain in accuracy obtained. So we can call the Random Forest our best option over the classifiers tested here. Table 1 presents the top 5 classifiers by accuracy.

Rank	Scaling	Sampling	Classifier	Accuracy	Run time training (s)	Run time test (s)	Best parameters
1	MinMaxScaler	random	Random Forest	0.70	70.94	0.09	{'min_samples_leaf': 2, 'n_estimators': 500}
2	StandardScaler	random	Random Forest	0.69	68.71	0.07	{'min_samples_leaf': 2, 'n_estimators': 350}
3	MinMaxScaler	stratified	Random Forest	0.69	68.44	0.04	{'min_samples_leaf': 2, 'n_estimators': 200}
4	RobustScaler	random	Random Forest	0.69	70.48	0.07	{'min_samples_leaf': 2, 'n_estimators': 350}
5	RobustScaler	stratified	Random Forest	0.68	68.11	0.06	{'min_samples_leaf': 2, 'n_estimators': 350}

Table 1 – Top 5 models based on accuracy.

Data

Besides reaching 70% of accuracy, the data quality raises some concerns and could be classified as unreliable data with poor quality:

- Classification measured is discrepant with the expected classification.

- All parameters have a heavy overlapping distribution for potable and non-potable.
- The absence of correlation between parameters knowingly correlated, for example, the known correlation between pH and potability.

Final thoughts

The physical properties presented are accurate predictors of portability, as can be found in the academic literature. However, the analysis of this data showed us a substantial discrepancy between the expected and the measured. Unless we deal with data from areas without alternatives, those samples should never be classified as potable. So the only reasonable explanation is that the classification field is wrong.

Conclusion

Our model can not be used as a predictor for water potability based on the physical properties of the water due to the issues of the data quality.

Alternative solution

Solving this problem would demand redoing all the sampling and measurements or fixing the classification of the database.