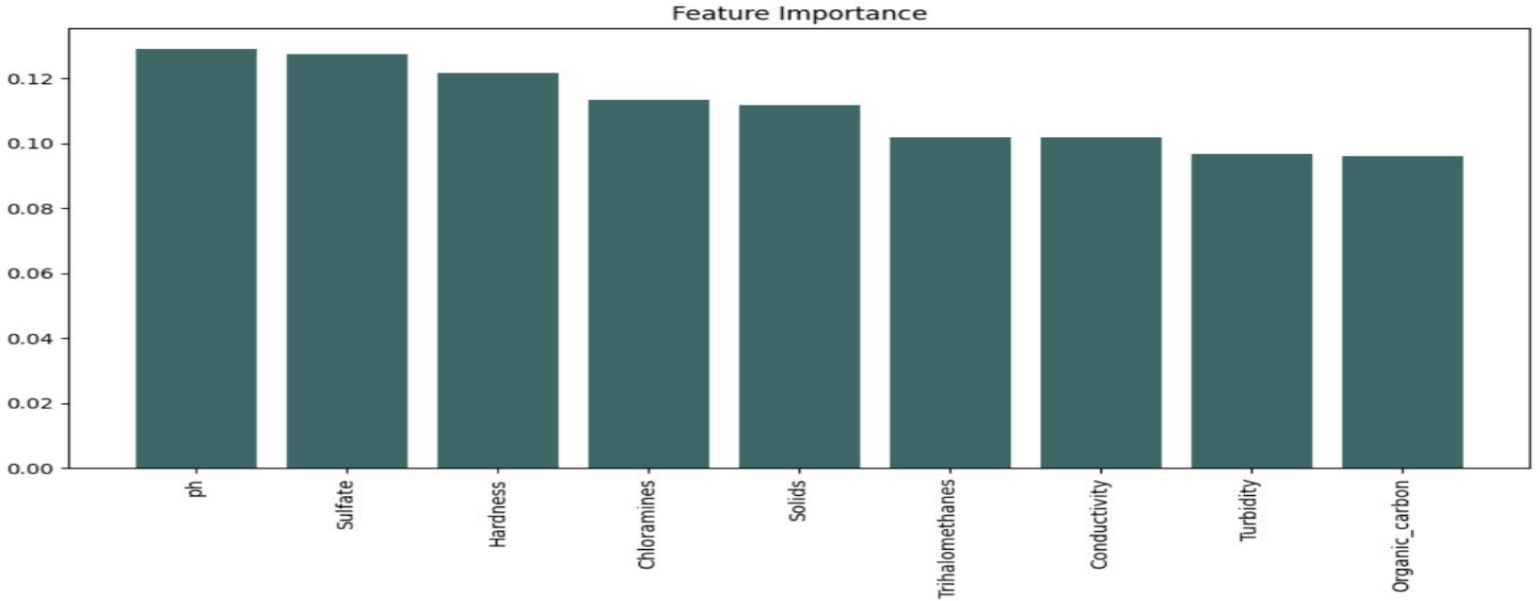
Water Quality Analysis

Analysis & Recommendations about predicting water quality
By Charlotte Sun

Overview

1. What are the key factors that influence water potability?

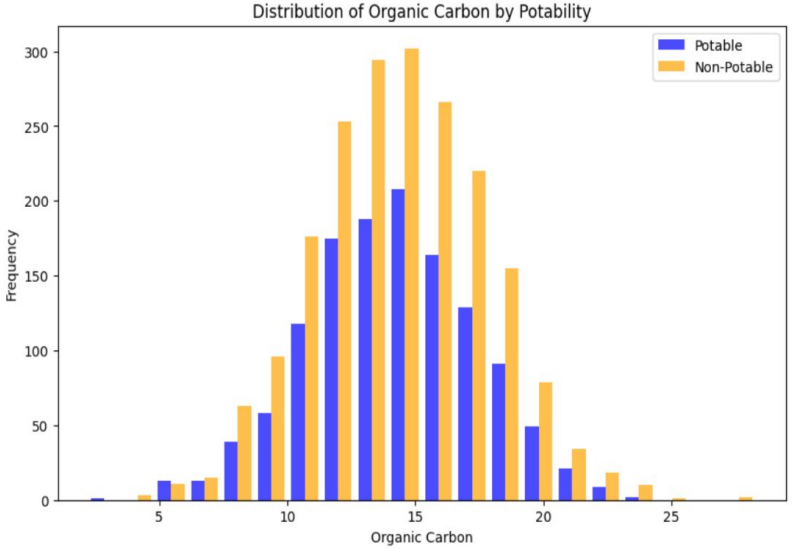
2. Can we accurately predict if water is drinkable based on these factors?



Inferantial Outcomes

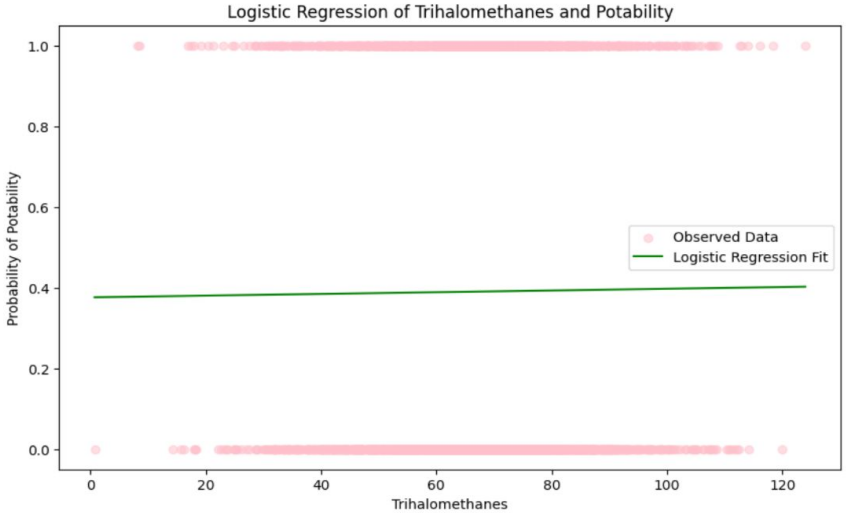
Does organic carbon significantly affect water potability?:

T-TEST: Fail to reject null hypothesis. Organic Carbon doesn't affect water potability siginificantly.

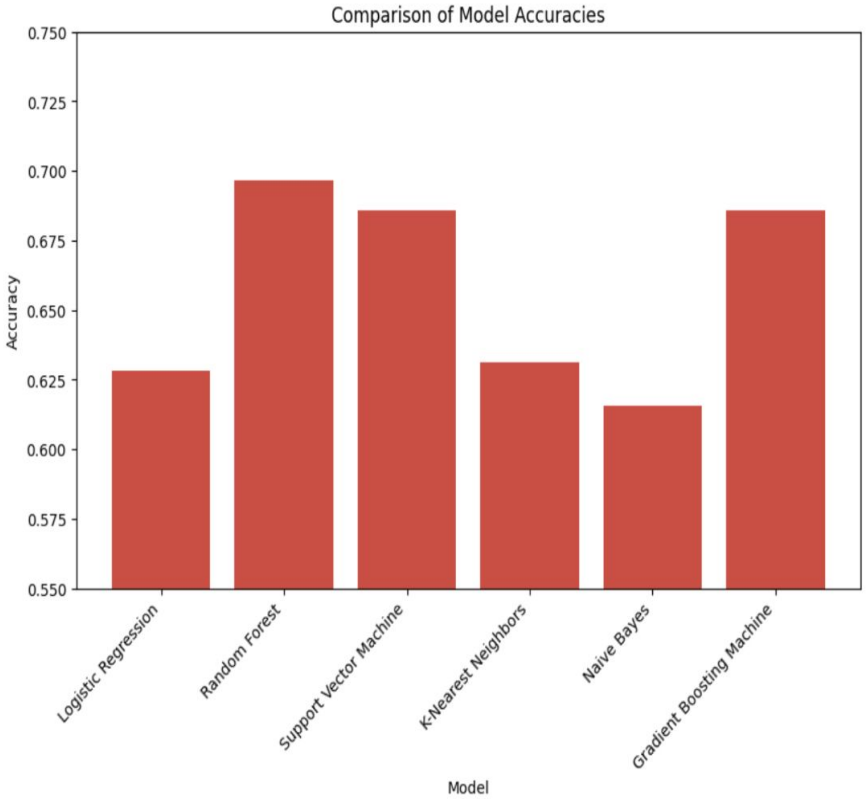


Is there a significant relationship between Trihalomethanes(THMs) and potability?:

Logistic Regression: Fail to reject the null hypothesis, the p-value is greater than 0.05, there is no siginificant relationship between Trihalomethanes(THMs) and potability.



What models are trained?

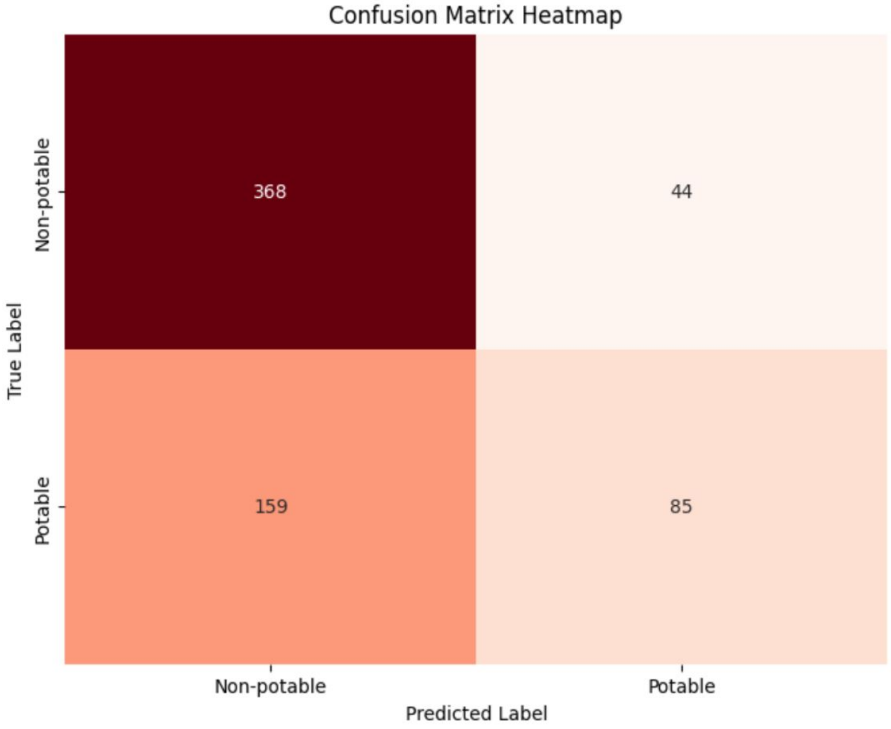


Model Accuracy

- Logistic Regression 0.628049
- Random Forest 0.696646
- Support Vector Machine 0.685976
- K-Nearest Neighbors 0.631098
- Naive Bayes 0.615854
- Gradient Boosting Machine 0.685976

Random Forest, Support Vector Machine and **Gradient Boosting Machine** perform relatively better!

The Best Result



Stacking was applied using: **Random Forest, Gradient Boosting, Support Vector Machine, and KNN** as base models, with **Logistic Regression** as the meta-model**.** The original training set resulted in the same best accuracy of 0.69 as before, but with a 0.01 improvement in recall and f1-score for Class 1(Potable). Consequently, this model was selected as the final result.

Conclusion and Improvement Suggestions:

The final model achieved an accuracy of 0.69.

Future improvements could focus on exploring non-linear feature combinations, advanced data augmentation, and further model ensembling to enhance performance.

While the model can be used for preliminary water potability screening, it's recommended to combine predictions with external data and expert judgment in business applications.